ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN CUỐI KÌ
Môn học: Kho dữ liệu và OLAP

Đề tài: Xây dựng kho dữ liệu FurnitureStore

Giảng viên hướng dẫn: Đỗ Thị Minh Phụng

Nhóm sinh viên thực hiện:

Võ Ngọc Lệ Xuân - 21521692

Bùi Hữu Bằng - 21520151

Thứ Tư, Ngày **22** Tháng **05** Năm **2024**

MỤC LỤC

NHẬN XÉT CỦA GIÁO VIÊN 10

CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU 11

1. Lý do chọn đề tài 11

2. Giới thiệu về dataset 11

2.1. Giới thiệu nguồn dữ liệu 11

2.2. Nguồn tải dataset 12

2.3. Hướng chủ đề 12

2.4. Mô tả chi tiết thuộc tính dataset 12

3. Mô hình bông tuyết 15

3.1. Mô tả bảng Dim_Shipmode: 15

3.2. Mô tả bảng Dim_Product 16

3.3. Mô tả bảng Dim_Category 16

3.4. Mô tả bảng Dim_Customer 17

3.5. Mô tả bảng Dim_Location 17

3.6. Mô tả bảng Dim_Date 18

3.7. Mô tả bảng Fact 18

4. Nội dung 15 câu truy vấn 19

CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS) 21

1. Chuẩn bị công cụ và Data Warehouse: 21

1.1. Tải và cài đặt công cụ SQL Server Data Tools vào SQL Server: 21

1.2. Tải công cụ SSDT cho phiên bản Visual Studio Community 2022 22

1.3. Chuẩn bị kho dữ liệu: 27

2. Thực hiện quá trình SSIS 28

2.1. Tạo mới project SSIS 28

2.2. Tạo bảng các bảng Dim và bảng Fact 30
2.2.1. Tạo kết nối đến Server 30
2.2.2. Tạo các bảng Dim và Fact 32
3. Tiến hành chạy dữ liệu 107
3.1. Kết quả chạy dữ liệu cho dự án 107
3.2. Tiền xử lý dữ liệu, tạo các bảng Dim và Fact 108
3.3. Tạo bảng Dim_Product 108
3.4. Merge Dims và Fact: 109
4. Kết quả cơ sở dữ liệu: 109
CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS) 110
1. TẠO PROJECT SSAS TRONG VISUAL STUDIO 2022 110
1.1. Tạo mới Project SSAS trong Visual Studio 110
1.2. Tạo DataSource 112
1.3. Tạo Data Source Views 117
1.4. Xác định các chiều và độ đo cho kho dữ liệu 123
1.5. Thêm thuộc tính vào Dimensions 129
1.5.1. Dim Customer 129
1.5.2. Dim Ship Mode 130
1.5.3. Dim Location 131
1.5.4. Dim Date 133
1.5.5. Dim Product 134
1.6. Tạo Hierachy 135
1.6.1. Dim Location 135
1.6.2. Dim Date 144
2. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ SSAS TRÊN CÁC KHỐI CUBE, NGÔN NGỮ MDX VÀ PIVOT TABLE 160

2.1. Cho biết khách hàng (mã khách hàng, tên khách hàng) có số tiền mua hàng cao nhất trong năm 2017? 160
2.1.1. Phân tích dữ liệu trên Cube 160
2.1.2. Truy vấn bằng ngôn ngữ MDX: 161
2.1.3. Sử dụng Pivot Table 162
2.1.4. Kết luận: 162
2.2. Đưa ra số lượng sản phẩm bán ra của từng loại nội thất? 163
2.2.1. Phân tích dữ liệu trên Cube 163
2.2.2. Truy vấn bằng ngôn ngữ MDX 163
2.2.3. Sử dụng Pivot Table 165
2.2.4. Kết luận: 165
2.3. Cho biết thông tin khách hàng (Tên khách hàng) đặt hàng nhiều nhất trong bang Minnesota? 165
2.3.1. Phân tích dữ liệu trên Cube 165
2.3.2. Truy vấn bằng ngôn ngữ MDX 167
2.3.3. Sử dụng Pivot Table 167
2.3.4. Kết luận: 168
2.4. Quý 4 của năm nào có doanh thu bán hàng cao nhất? 169
2.4.1. Phân tích dữ liệu trên Cube 169
2.4.2. Truy vấn bằng ngôn ngữ MDX 169
2.4.3. Sử dụng Pivot Table 171
2.4.4. Kết luận: 171
2.5. Theo từng tháng, quý, năm hãy liệt kê số đơn hàng đã được cửa hàng bán ra và tổng giá trị của các đơn hàng đó? 171
2.5.1. Phân tích dữ liệu trên Cube 171
2.5.2. Truy vấn bằng ngôn ngữ MDX 173

2.5.3. Sử dụng Pivot Table 173
2.5.4. Kết luận: 173
2.6. Liệt kê những thành phố có phần trăm lợi nhuận của đơn hàng trên 30%? 173
2.6.1. Phân tích dữ liệu trên Cube 173
2.6.2. Truy vấn bằng ngôn ngữ MDX 176
2.6.3. Sử dụng Pivot Table 176
2.6.4. Kết luận: 176
2.7. Hãy đưa ra tỉ lệ lợi nhuận thu được của top 3 sản phẩm có số lượng mua cao nhất của từng Bang? 177
2.7.1. Phân tích dữ liệu trên Cube 177
2.7.2. Truy vấn bằng ngôn ngữ MDX 178
2.7.3. Sử dụng Pivot Table 179
2.7.4. Kết luận: 181
2.8. Cho biết tổng số tiền cửa hàng đã chiết khấu cho khách hàng khi mua loại sản phẩm 'Bookcases' theo từng khu vực của nước Mỹ? 182
2.8.1. Phân tích dữ liệu trên Cube 182
2.8.2. Truy vấn bằng ngôn ngữ MDX 184
2.8.3. Sử dụng Pivot Table 184
2.8.4. Kết luận: 185
2.9. Với từng loại khách hàng, hãy đưa ra tổng số lượng các đơn hàng được bán (quantity) theo từng loại hình vận chuyển? 186
2.9.1. Phân tích dữ liệu trên Cube 186
2.9.2. Truy vấn bằng ngôn ngữ MDX 186
2.9.3. Sử dụng Pivot Table 187
2.9.4. Kết luận: 187

2.10. Hãy liệt kê thông tin khách hàng và loại khách hàng chi tiền mua sản phẩm thuộc phân loại tables hoặc chairs nhiều nhất? 188
2.10.1. Phân tích dữ liệu trên Cube 188
2.10.2. Sử dụng ngôn ngữ MDX 189
2.10.3. Sử dụng Pivot Table 190
2.10.4. Kết luận: 191
2.11. Cho biết tổng doanh số của từng khu vực bán hàng? 192
2.11.1. Phân tích dữ liệu trên Cube 192
2.11.2. Sử dụng ngôn ngữ MDX 193
2.11.3. Sử dụng Pivot Table 193
2.11.4. Kết luận: 194
2.12. Liệt kê các sản phẩm có số lượng mua từ 50 đến 100? 194
2.12.1. Phân tích dữ liệu trên Cube 194
2.12.2. Sử dụng ngôn ngữ MDX 195
2.12.3. Sử dụng Pivot Table 195
2.12.4. Kết luận: 195
2.13. Đưa ra 20% phần trăm sản phẩm có doanh thu cao nhất, trong đó chỉ ra doanh thu theo từng loại hình vận chuyển được sử dụng (liệt kê nếu có)? 196
2.13.1. Phân tích dữ liệu trên Cube 196
2.13.2. Sử dụng ngôn ngữ MDX 196
2.13.3. Sử dụng Pivot Table 197
2.13.4. Kết luận: 197
2.14. Thống kê lợi nhuận, phần trăm lợi nhuận từng loại khách hàng, sắp xếp theo phần trăm lợi nhuận giảm dần? 197
2.14.1. Phân tích dữ liệu trên Cube 197
2.14.2. Sử dụng ngôn ngữ MDX 198

2.14.3. Sử dụng Pivot Table 198
2.14.4. Kết luận: 198
2.15. Liệt kê Top 10 khách hàng có số đơn hàng ít nhất chọn sử dụng dịch vụ giao hàng First Class và danh sách tương tự nhưng sử dụng dịch vụ giao hàng Standard Class? 199
2.15.1. Phân tích dữ liệu trên Cube 199
2.15.2. Sử dụng ngôn ngữ MDX 201
2.15.3. Sử dụng Pivot Table 201
2.15.4. Kết luận: 201
CHƯƠNG 4. QUÁ TRÌNH LẬP BÁO BIỂU(SSRS) 203
1. QUÁ TRÌNH SSRS VỚI VISUAL STUDIO REPORTING SERVICE VÀ REPORT BUILDER 203
1.1. Cài đặt các công cụ cần thiết: 203
1.1.1. Cài đặt SQL Server Reporting Services: 203
1.2. Quá trình tạo lập báo biểu bằng công cụ Microsoft SSRS(Report Builder) 218
1.2.1. Tạo project SSRS 218
1.2.2. Tạo kết nối với kho dữ liệu: 221
1.3. Thực hiện lập báo biểu trên Report Builder: 225
1.3.1. Báo biểu 1 (Report Grouping): Đưa doanh số, lợi nhuận, tổng số lượng mua của từng loại sản phẩm 225
1.3.2. Báo biểu 2 (Report Grouping + Total + Matrix): Báo cáo về doanh thu của các thành phố thuộc nước Mỹ từ trong khoảng thời gian 2014 - 2017 theo các khoảng thời gian tháng, quí, năm 230
1.3.3. Báo biểu 3(Report Chart): Với từng loại khách hàng(Segment) hãy đưa ra doanh thu từ năm 2014 đến năm 2017. Kết quả của báo biểu là một biểu đồ dạng Column Chart. 244
2. QUÁ TRÌNH SSRS VỚI POWER BI 257

2.1. Cài đặt các công cụ cần thiết 257
2.1.1. Cài đặt Power BI 257
2.2. Quá trình tạo lập báo biểu bằng công cụ Power BI 258
2.2.1. Tạo project SSRS trên Power BI 258
2.2.2. Tạo kết nối với kho dữ liệu 259
2.3. Thực hiện lập báo biểu trên Power BI 262
2.3.1. Báo biểu 4: Thống kê số sản phẩm được bán ra theo từng Bang (có thể lọc theo từng loại sản phẩm) 262
2.3.2. Báo biểu 5: Thống kê số đơn hang, số sản phẩm đặt mua theo từng sản phẩm, giá trị đơn hang và địa chỉ giao hàng (thành phố) 265
2.3.3. Báo biểu 6: Thống kê và so sánh số tiền chiết khấu và đơn hàng mua của các khu vực khảo sát 268
CHƯƠNG 5. DATA MINING 272
1. Bộ dữ liệu dùng cho quá trình Data mining 272
2. Thiết lập các công cụ cần thiết để thực hiện quá trình Data Mining 272
2.1. Cài đặt Visual Studio 2019 272
2.2. Cài đặt SSDT 2019 273
3. Thực hiện quá trình Data Mining 274
3.1. Quá trình đổ dữ liệu vào cơ sở dữ liệu 274
3.2. Quá trình Data Mining bằng công cụ SSAS của Visual Studio 2019 và thuật toán Decision Trees(Cây quyết định): 281
3.2.2. Xác định tập luật: 308
3.3. Khai thác dữ liệu bằng thuật toán Clustering(gom cụm) 309
3.3.1. Khai báo mô hình 309
3.3.2. Phân tích mô hình: 312
3.3.3. Xác định tập luật: 317

3.4. Thực hiện khai thác dữ liệu với thuật toán Naïve Bayes 318
3.4.1. Khai báo mô hình 318
3.4.2. Phân tích mô hình 320
3.4.3. Xác định tập luật: 324
3.5. So sánh mức độ độ hiệu quả của các thuật toán: 324
3.5.1. Sử dụng Lift Chart 324
3.6. Kiểm tra chức năng Filtered Model 326
3.7. Dự đoán với các giá trị tương lai 339

NHẬN XÉT CỦA GIÁO VIÊN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU

1. Lý do chọn đề tài

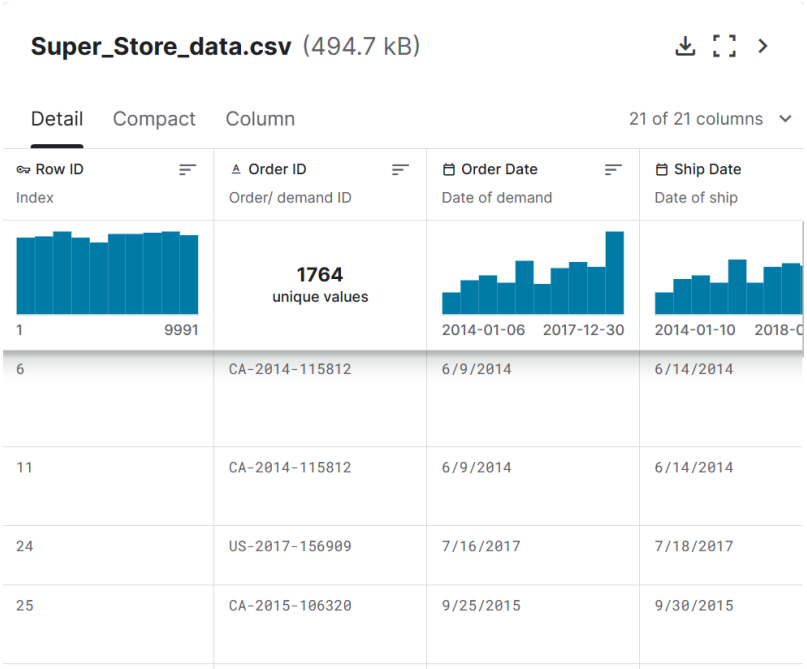
Với sự phát triển kinh tế và gia tăng dân số, nhu cầu về nội thất luôn tăng, đặc biệt khi người tiêu dùng sẵn sàng đầu tư vào việc cải thiện không gian sống. Ngành này cung cấp một loạt sản phẩm đa dạng, từ ghế, bàn, tủ, giường đến các phụ kiện trang trí, giúp doanh nghiệp dễ dàng đáp ứng nhu cầu đa dạng của khách hàng.

Ngoài ra, sự phát triển của công nghệ và thương mại điện tử mở ra nhiều cơ hội mới, cho phép các cửa hàng mở rộng thị trường qua các nền tảng trực tuyến. Kinh doanh nội thất cũng liên quan đến xu hướng bền vững, sử dụng vật liệu thân thiện với môi trường, thu hút khách hàng quan tâm đến yếu tố này. Đam mê về thiết kế nội thất và mong muốn đóng góp vào việc xây dựng không gian sống tiện nghi và đẹp mắt cũng là lý do chúng em chọn đề tài này.

2. Giới thiệu về dataset

2.1. Giới thiệu nguồn dữ liệu

- Tên dataset: Sales for Furniture Store.



Hình 1.1 Sales for Furniture Store dataset (Kaggle)

- Năm cập nhật gần nhất: 2024.
- Tác giả của bộ dữ liệu này là Zahra' Alaa Tag el-Dein, là một AI developer.
- Nguồn gốc của bộ dữ liệu thuộc về chương trình thực tập sinh khoa học máy tính của Mentorness(Mentorness Machine Learning Internship). Mentorness là một nền tảng tuyển dụng dựa trên bằng chứng công việc. Đơn vị này đặt mục tiêu hoạt động đó chính là đóng vai cầu nối giữa những ứng cử viên đang tìm kiếm công việc phù hợp với kỹ năng của họ và những công ty gặp khó khăn trong việc tìm kiếm những ứng cử viên thích hợp.
- Bộ dữ liệu thu thập các thông tin chi tiết về việc kinh doanh như thông tin chung của đơn hàng như: Thông tin về khách hàng, loại hình giao hàng, địa chỉ giao hàng,… và thông tin chi tiết khác như: Chi tiết đơn hàng, doanh số, dữ liệu khuyến mãi của một cửa hàng bán nội thất trong vòng 10 năm trở lại đây.
- Bộ dữ liệu gồm 2,121 dòng dữ liệu và 21 cột thuộc tính.

2.2. Nguồn tải dataset

- Link dataset: https://www.kaggle.com/datasets/zahraaalaatageldein/sales-for-furniture-store

2.3. Hướng chủ đề

- Kho dữ liệu theo hướng chủ đề: Kinh doanh.

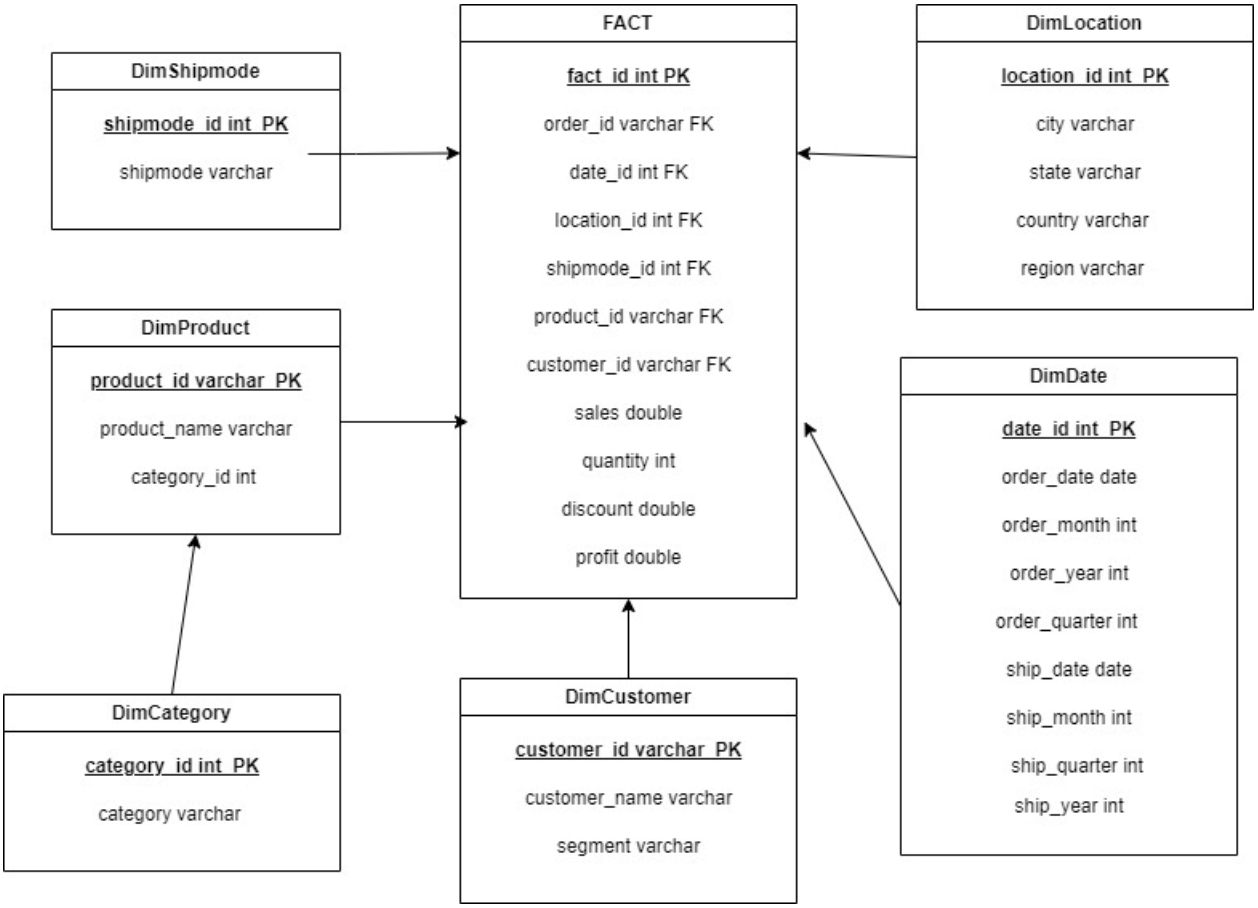
2.4. Mô tả chi tiết thuộc tính dataset

STT	Tên thuộc tính	Tên đầy đủ	Kiểu dữ liệu	Ý nghĩa
1	Row ID	Row Identification	Integer	Số nhận dạng duy nhất của dòng dữ liệu
1	Order ID	Order Identification	Varchar	Đây là mã nhận dạng duy nhất của một đơn hàng. Nếu như có 2 dòng liên tiếp có dùng một Order

				ID thì có nghĩa là đó là 2 sản phẩm trong cùng một đơn hàng
2	Order Date	Order Date	Date	Hiển thị ngày đặt hàng của đơn hàng. Định dạng ngày: MM/dd/yyyy
3	Ship Date	Ship Date	Date	Hiển thị ngày đơn hàng được giao tới cho bên vận chuyển để vận chuyển đến cho khách hàng. Định dạng ngày: MM/dd/yyyy
4	Ship Mode	Ship Mode	Varchar	Phương thức vận chuyển được chọn cho đơn hàng . Phương thức vận chuyển có thể nhận 1 trong 4 giá trị sau: - Standard Class. - Second Class. - First Class. - Same Day. Phương thức có
5	Customer ID	Customer Identification	Varchar	
6	Customer Name	Customer Name	Varchar	Tên khách hàng

7	Segment	Segment	Varchar	Loại khách hàng
8	Country	Country	Varchar	Quốc gia
9	State	State	Varchar	Bang
10	Region	Region	Varchar	Vùng
11	Product ID	Product Identification	Varchar	Mã sản phẩm
12	Product Name	Product Name	Varchar	Tên sản phẩm
13	Category	Category	Varchar	Loại sản phẩm
14	Sales	Sales	Float	Đơn giá
15	Quantity	Quantity	Integer	Số lượng
16	Discount	Discount	Float	Phần trăm khuyến mãi
17	Profit	Profit	Float	Lợi nhuận

3. Mô hình bông tuyết



Hình 1.2 Sơ đồ bông tuyết cho cơ sở dữ liệu chủ đề Sales for Furniture Store

3.1. Mô tả bảng Dim_Shipmode:

Gồm có 2 thuộc tính, 4 dòng dữ liệu thể hiện thông tin về phương thức vận chuyển của đơn hàng.

Khoá chính	Tên thuộc tính	NOT NULL	Ý nghĩa	Kiểu dữ liệu
X	SHIPMODE_ID		Mã phương thức vận chuyển	int
	SHIPMODE		Tên phương thức vận chuyển	varchar

3.2. Mô tả bảng Dim_Product

Gồm có 4 thuộc tính, dòng dữ liệu chứa thông tin của các sản phẩm được mua trong các đơn hàng.

Khoá chính	Tên thuộc tính	NOT NULL	Ý nghĩa	Kiểu dữ liệu
X	PRODUCT_ID		Mã sản phẩm tự động tăng	Int
	PRODUCT_SYSTEM_ID		Mã sản phẩm trong dữ liệu góc	
	PRODUCT_NAME		Tên sản phẩm	varchar
	CATEGORY_ID		Mã loại sản phẩm	int

3.3. Mô tả bảng Dim_Category

Gồm có 2 thuộc tính, 4 dòng dữ liệu chứa thông tin các loại sản phẩm.

Khoá chính	Tên thuộc tính	NOT NULL	Ý nghĩa	Kiểu dữ liệu
X	CATEGORY_ID		Mã loại sản phẩm	int
	CATEGORY		Tên loại sản phẩm Nhận các giá trị: - Bookcases: tủ sách - Chairs: Ghế - Furnishing; Đồ trang trí nội thất - Tables: Bàn	varchar

3.4. Mô tả bảng Dim_Customer

Gồm có 3 thuộc tính, dòng dữ liệu. Trong đó, mỗi dòng dữ liệu chứa thông tin về mỗi khách hàng có trong bộ dữ liệu gốc.

Khoá chính	Tên thuộc tính	NOT NULL	Ý nghĩa	Kiểu dữ liệu
X	CUSTOMER_ID		Mã khách hàng	varchar
	CUSTOMER_NAME		Tên khách hàng	varchar
	SEGMENT		Loại khách hàng Nhận các giá trị: - Consumer: Khách hàng tiêu dùng - Home Office: Văn phòng tại gia - Corporate: Công ty	varchar

3.5. Mô tả bảng Dim_Location

Gồm có 5 thuộc tính, dòng dữ liệu. Trong đó, mỗi dòng chứa thông tin của địa điểm mua hàng có trong bộ dữ liệu ban đầu.

Khoá chính	Tên thuộc tính	NOT NULL	Ý nghĩa	Kiểu dữ liệu
X	LOCATION_ID		Mã địa điểm	int
	CITY		Tên thành phố	varchar
	STATE		Tên tiểu bang	varchar
	COUNTRY		Tên quốc gia	varchar
	REGION		Tên vùng lãnh thổ	varchar

3.6. Mô tả bảng Dim_Date

Gồm có 9 thuộc tính, dòng dữ liệu. Trong đó, mỗi dòng chứa cặp thông tin ngày đặt hàng và ngày giao hàng của đơn hàng trong bộ dữ liệu gốc.

Khoá chính	Tên thuộc tính	NOT NULL	Ý nghĩa	Kiểu dữ liệu
X	DATE_ID		Mã địa điểm	int
	ORDER_DATE		Ngày đặt hàng	date
	ORDER_MONTH		Tháng đặt hàng	int
	ORDER_YEAR		Năm đặt hàng	int
	ORDER_QUARTER		Quý đặt hàng	int
	SHIP_DATE		Ngày giao hàng	date
	SHIP_MONTH		Tháng giao hàng	int
	SHIP_YEAR		Năm giao hàng	int
	SHIP_QUARTER		Quý giao hàng	int

3.7. Mô tả bảng Fact

Gồm 11 thuộc tính, dòng dữ liệu. Là bảng sự kiện của mô hình, có chứa mô

Khoá chính	Tên thuộc tính	NOT NULL	Ý nghĩa	Kiểu dữ liệu
X	FACT_ID		Mã bảng sự kiện	int
	ORDER_ID		Mã đơn hàng	varchar
Khoá ngoại	DATE_ID		Mã ngày	int

Khoá ngoại	LOCATION_ID		Mã địa điểm	int
Khoá ngoại	SHIPMODE_ID		Mã phương thức vận chuyển	int
Khoá ngoại	PRODUCT_ID		Mã sản phẩm	varchar
Khoá ngoại	CUSTOMER_ID		Mã khách hàng	varchar
	SALES		Đơn giá	double
	QUANTITY		Số lượng hàng	int
	DISCOUNT		Phần trăm khuyến mãi	double
	PROFIT		Lợi nhuận	double

4. Nội dung 15 câu truy vấn

1. Cho biết khách hàng (mã khách hàng, tên khách hàng) có số tiền mua hàng cao nhất trong năm 2017?
2. Đưa ra số lượng sản phẩm bán ra của từng loại nội thất?
3. Cho biết thông tin khách hàng (Tên khách hàng) đặt hàng nhiều nhất trong bang Minnesota?
4. Quý 4 của năm nào có doanh thu bán hàng cao nhất?
5. Theo từng tháng, quý, năm hãy liệt kê số đơn hàng đã được cửa hàng bán ra và tổng giá trị của các đơn hàng đó?
6. Liệt kê những thành phố có phần trăm lợi nhuận của đơn hàng trên 30%?
7. Hãy đưa ra tỉ lệ lợi nhuận thu được của top 3 sản phẩm có số lượng mua cao nhất của từng Bang?
8. Cho biết tổng số tiền cửa hàng đã chiết khấu cho khách hàng khi mua loại sản phẩm ‘Bookcases’ theo từng khu vực của nước Mỹ?
9. Với từng loại khách hàng, hãy đưa ra tổng số lượng các đơn hàng được bán (quantity) theo từng loại hình vận chuyển?
10. Hãy liệt kê thông tin khách hàng và loại khách hàng chi tiền mua sản phẩm thuộc phân loại tables hoặc chairs nhiều nhất?
11. Cho biết tổng doanh số của từng khu vực bán hàng?

12. Liệt kê các sản phẩm có số lượng mua từ 50 đến 100?
13. Đưa ra 20% phần trăm sản phẩm có doanh thu cao nhất, trong đó chỉ ra doanh thu theo từng loại hình vận chuyển được sử dụng (liệt kê nếu có)?
14. Thống kê lợi nhuận, phần trăm lợi nhuận từng loại khách hàng, sắp xếp theo phần trăm lợi nhuận giảm dần?
15. Liệt kê Top 10 khách hàng có số đơn hàng ít nhất chọn sử dụng dịch vụ giao hàng First Class và danh sách tương tự nhưng sử dụng dịch vụ giao hàng Standard Class?

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

1. Chuẩn bị công cụ và Data Warehouse:

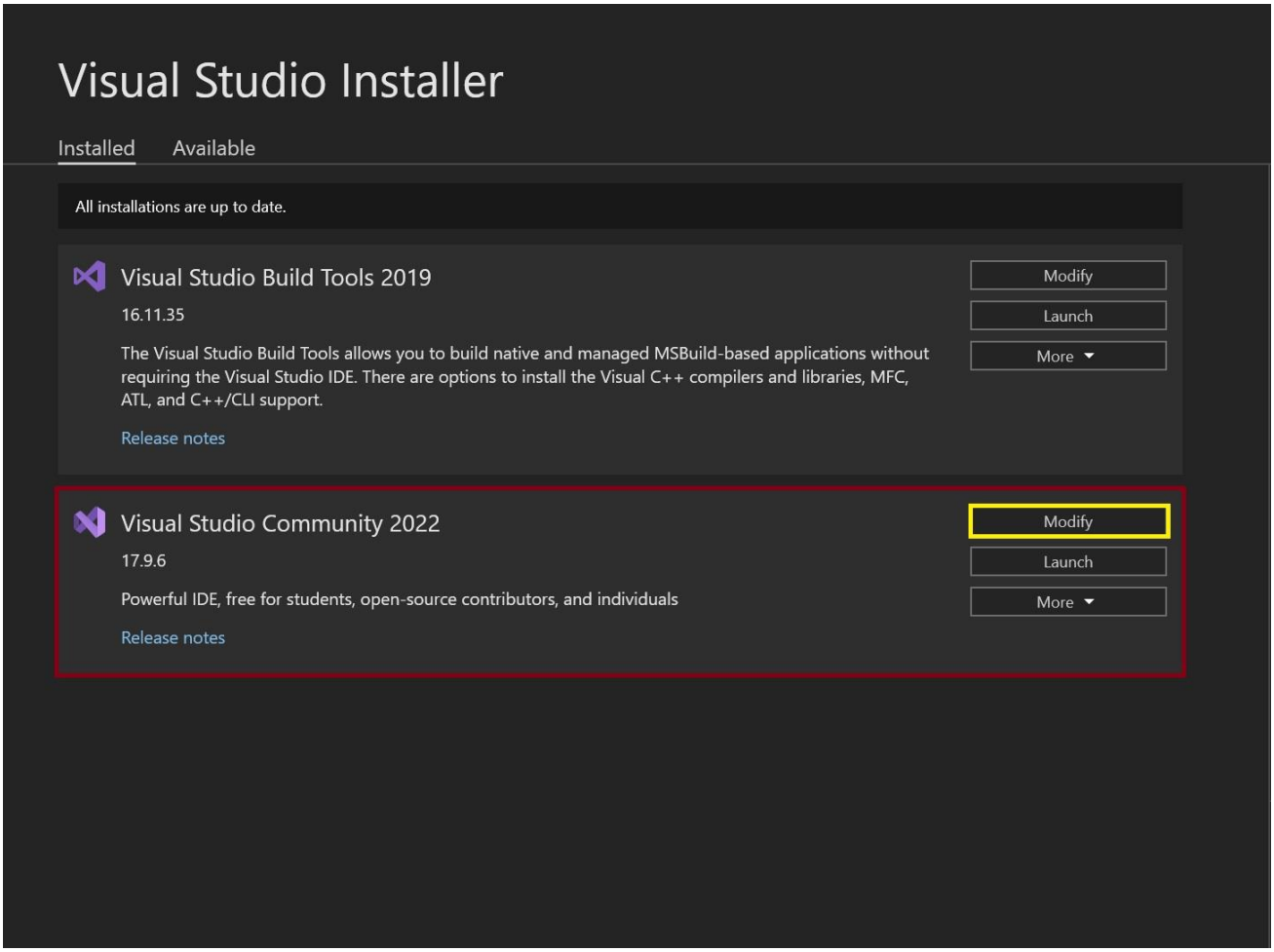
Để thực hiện quá trình SSIS thì ta cần chuẩn bị và cài đặt các công cụ sau:

- Visual Studio Community 2022
- SQL Server Integration Services Project

1.1. Tải và cài đặt công cụ SQL Server Data Tools vào SQL Server:

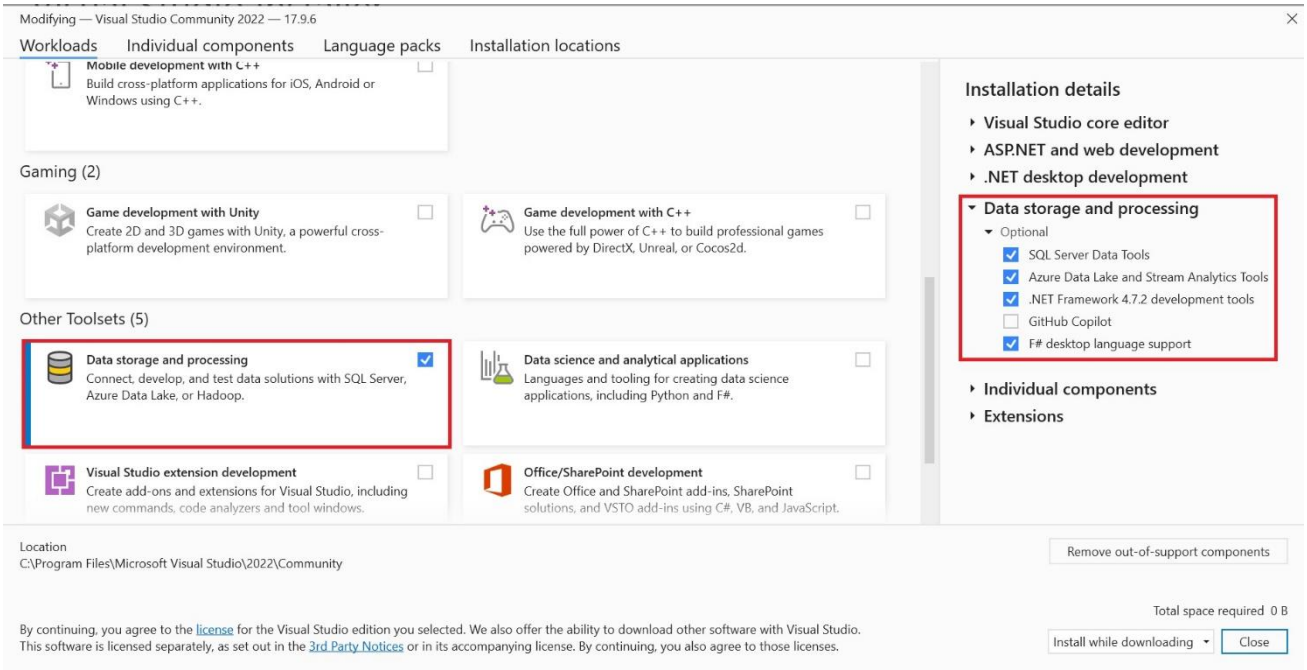
Bước 1: Tải Visual Studio Community 2022 và tiến hành cài đặt trên máy theo chỉ dẫn của trình cài đặt. Nếu như máy đã có sẵn Visual Studio Community 2022 thì ta có thể đến bước 2.

Bước 2: Mở Visual Studio Installer và chọn vào chức năng Modify của phần mềm Visual Studio Community 2022



Hình 2.1 Quy trình tải công cụ Server Data Tools (1)

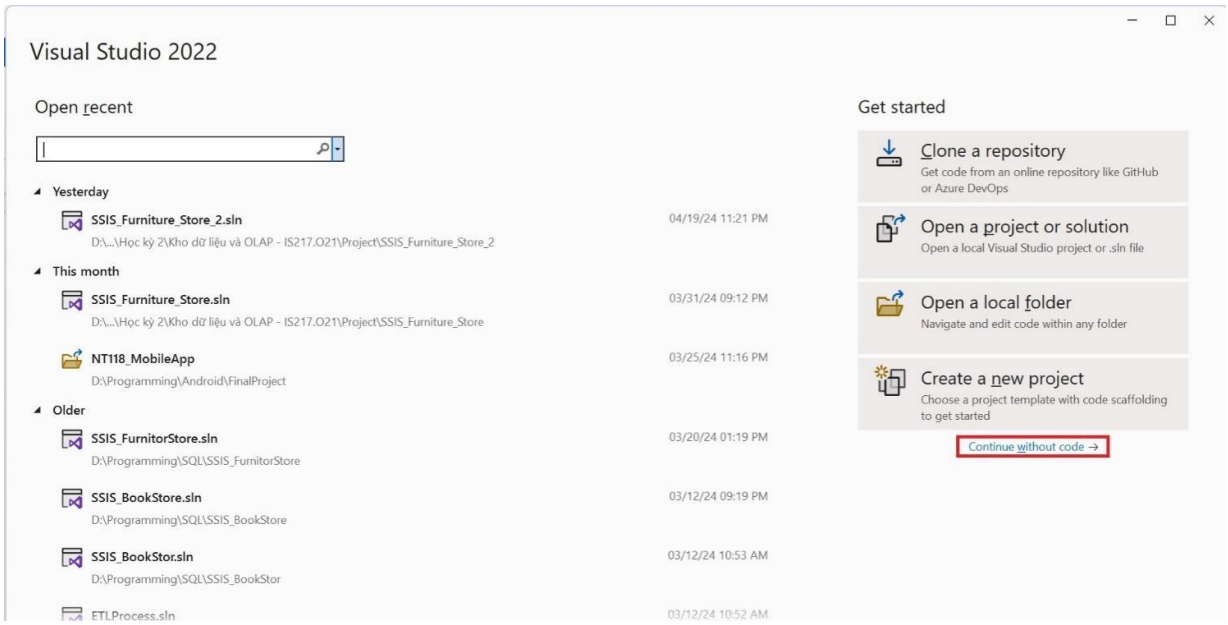
Bước 3: Chọn **SQL Server Data Tools** và các cài đặt khác dưới mục **Data storage and processing** trong danh sách workload.



Hình 2.2 Quy trình tải công cụ Server Data Tools (2)

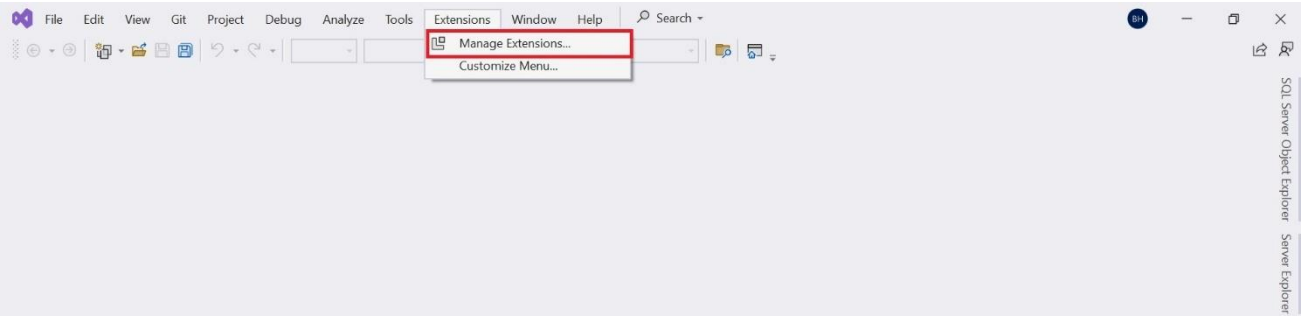
1.2. Tải công cụ SSDT cho phiên bản Visual Studio Community 2022

Bước 1: Mở Visual Studio Community 2022 và chọn **Continue without code**



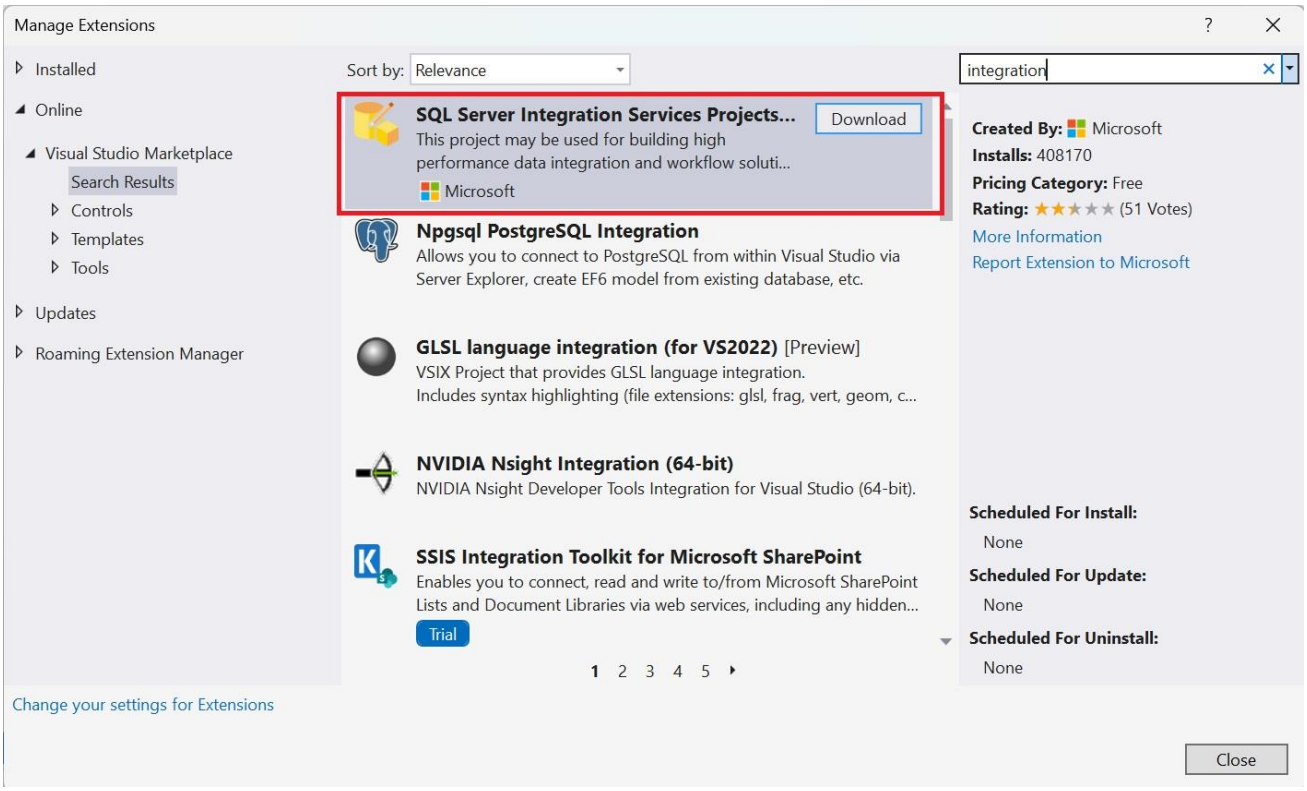
Hình 2.3 Quy trình tải công cụ SSDT(1)

Bước 2: Trong giao diện chính trên thanh công cụ chọn **Extension → Manage Extensions…**



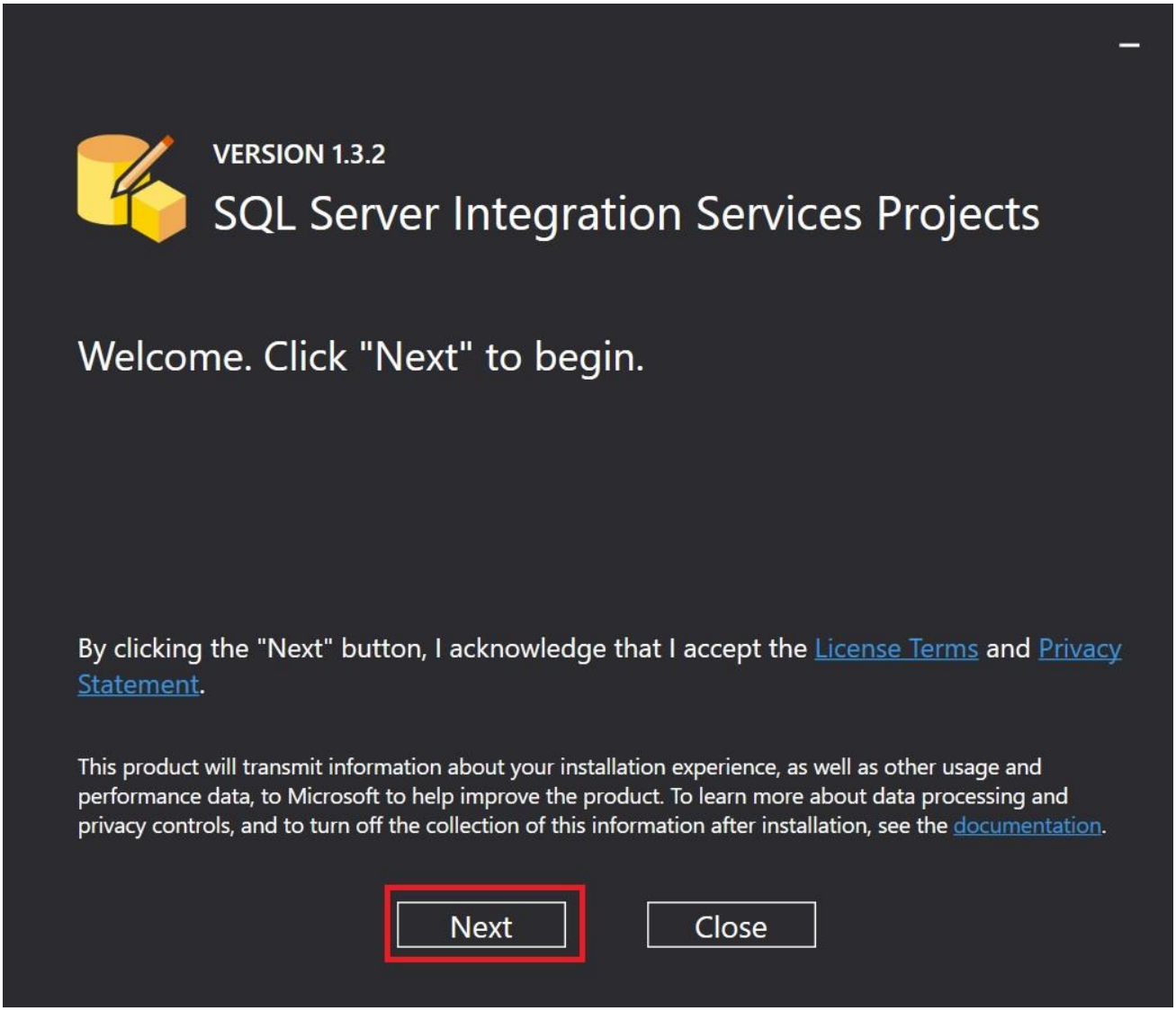
Hình 2.4 Quy trình tải công cụ SSDT(2)

Bước 3: Tìm kiếm và tải về công cụ **SQL Server Integration Services Projects.**



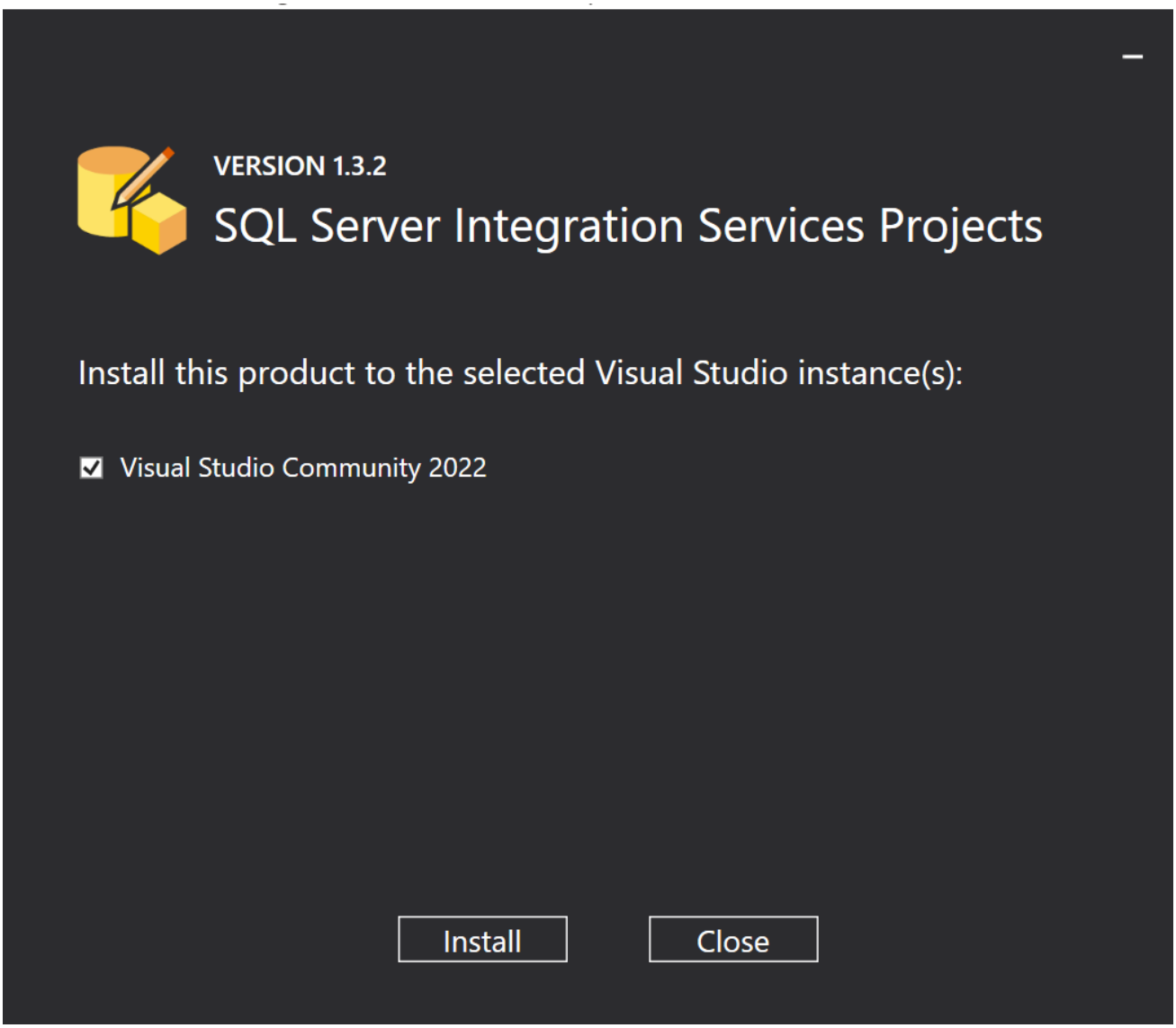
Hình 2.5 Quy trình tải công cụ SSDT(3)

Bước 4: Khi công cụ đã được tải về thì mở file .exe công cụ vừa tải xuống và nhấn **Continue**.



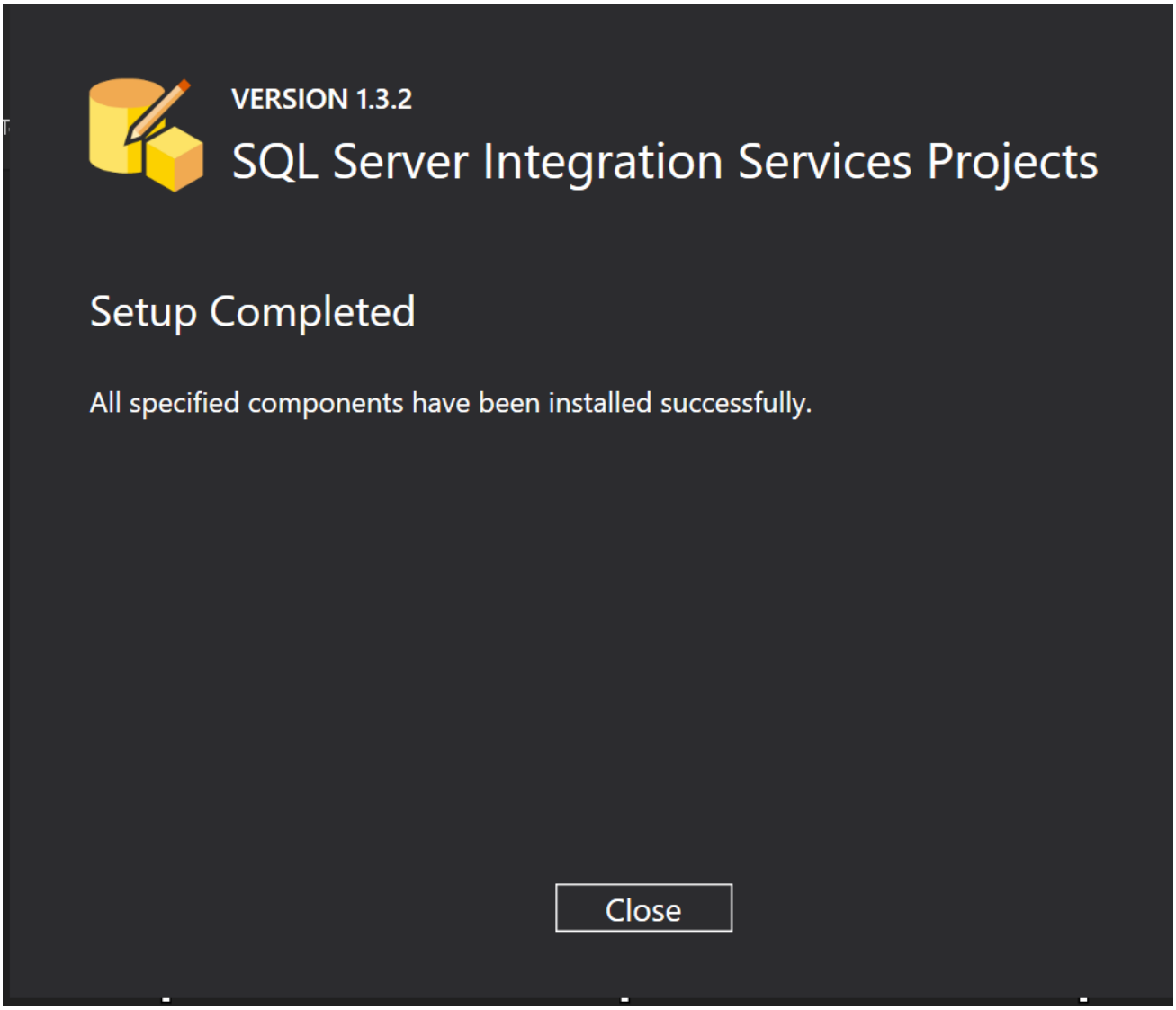
Hình 2.6 Quy trình tải công cụ SSDT(4)

Bước 5: Tick chọn vào ô Visual Community 2022 và chọn Install để bắt đầu quá trình cài đặt.



Hình 2.7 Quy trình tải công cụ SSDT(5)

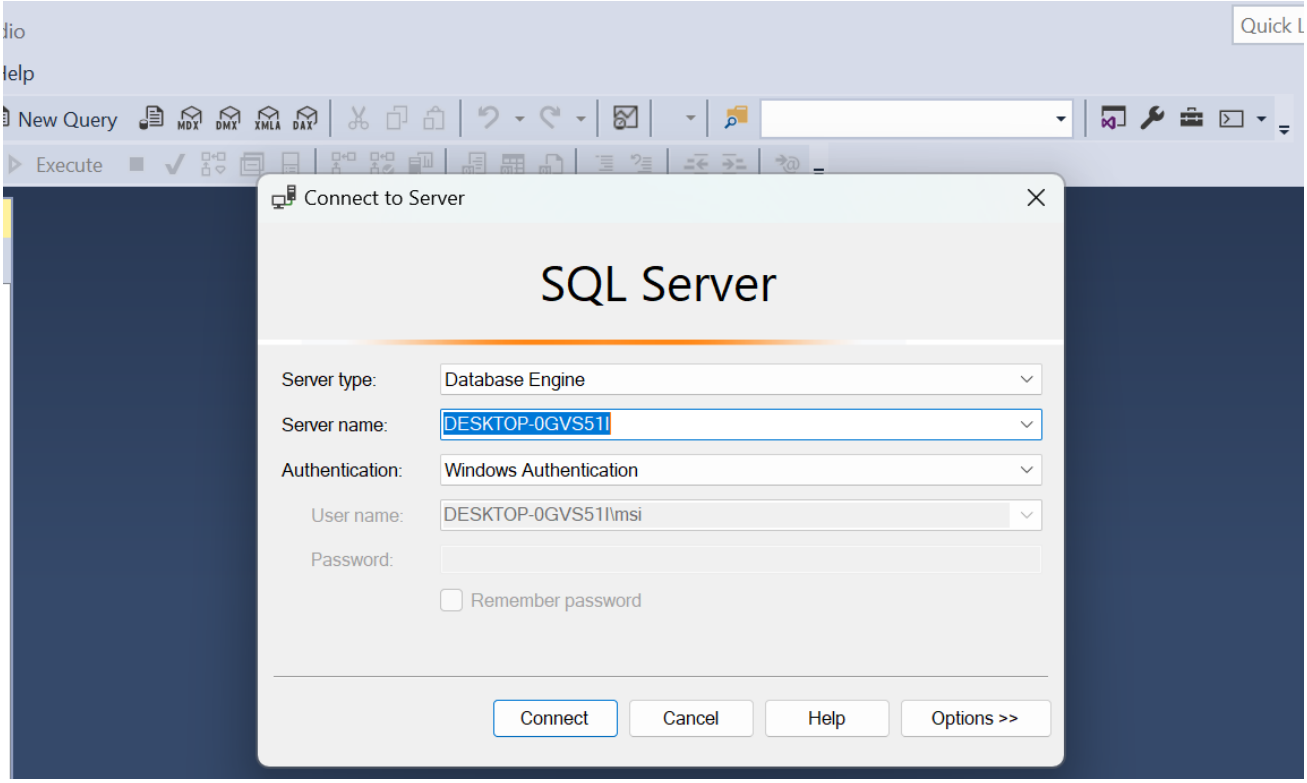
Bước 6: Sau khi cài đặt xong thì sẽ nhận được thông báo:



Hình 2.8 Quy trình tải công cụ SSDT(6)

1.3. Chuẩn bị kho dữ liệu:

Bước 1: Mở SQL Server Management Studio 20 và kết nối tới server bằng tài khoản user của window (Windows Authentication).

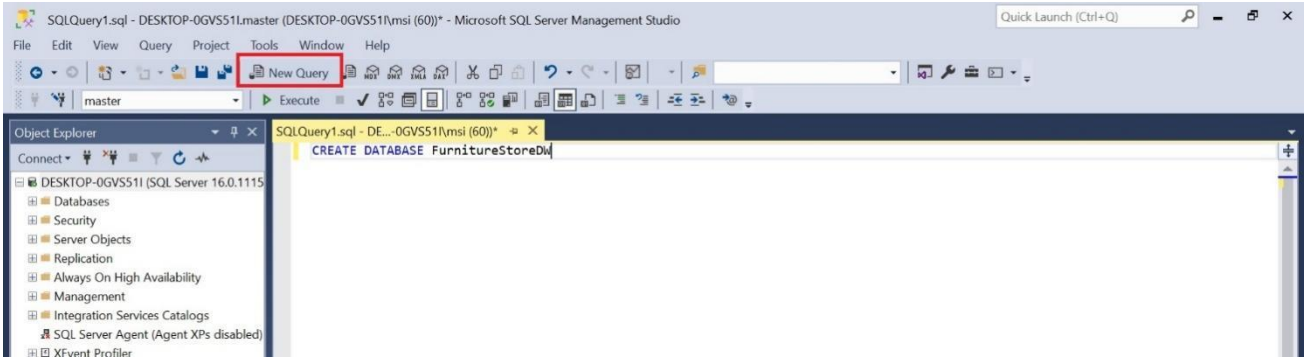


Hình 2.9 Mở SQL Server Management Studio và kết nối với server

Bước 2: Trên thanh công cụ chọn **New Query** và khởi tạo một cơ sở dữ liệu có tên là FurnitureStoreDW, đây chính là cơ sở dữ liệu nơi chứa các bảng Dim và bảng Fact cùng với các dữ liệu của các bảng đó.

Câu lệnh:

```
CREATE DATABASE FurnitureStoreDW
```

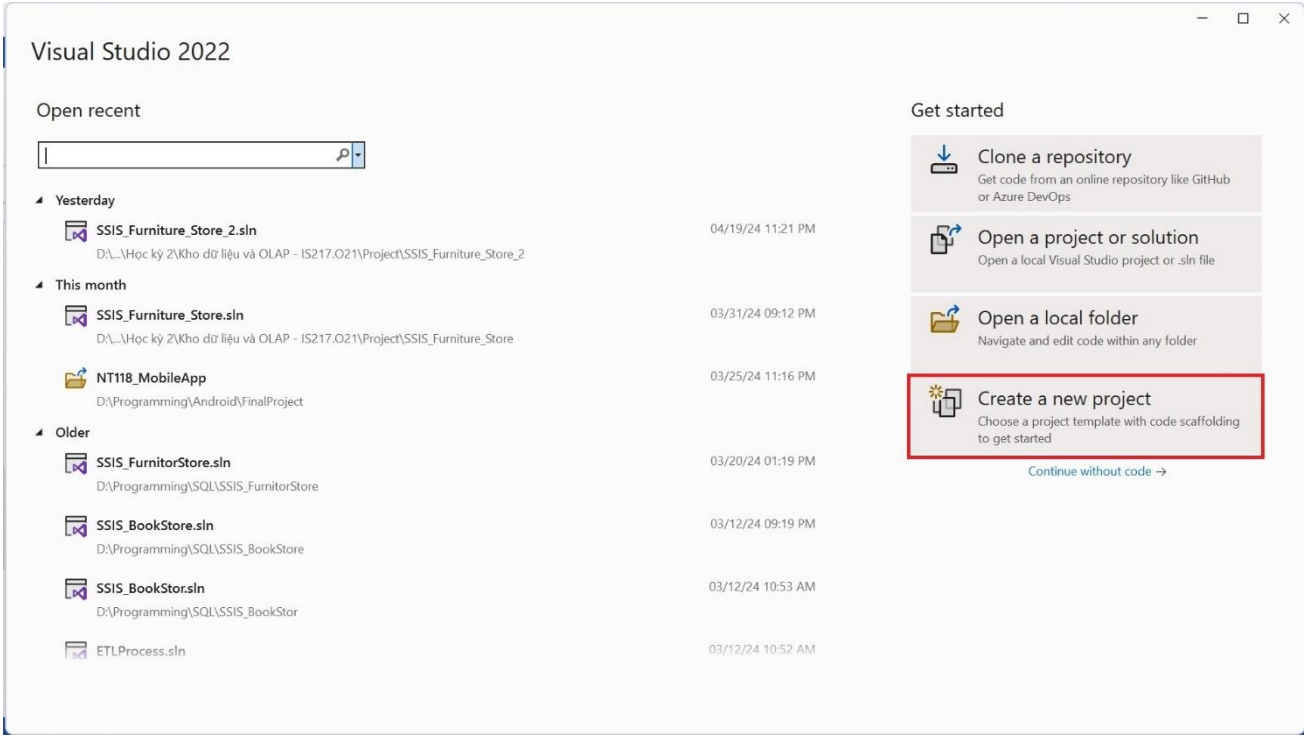


Hình 2.10 Tạo kho dữ liệu dùng để chứa các bảng Dim và bảng Fact

2. Thực hiện quá trình SSIS

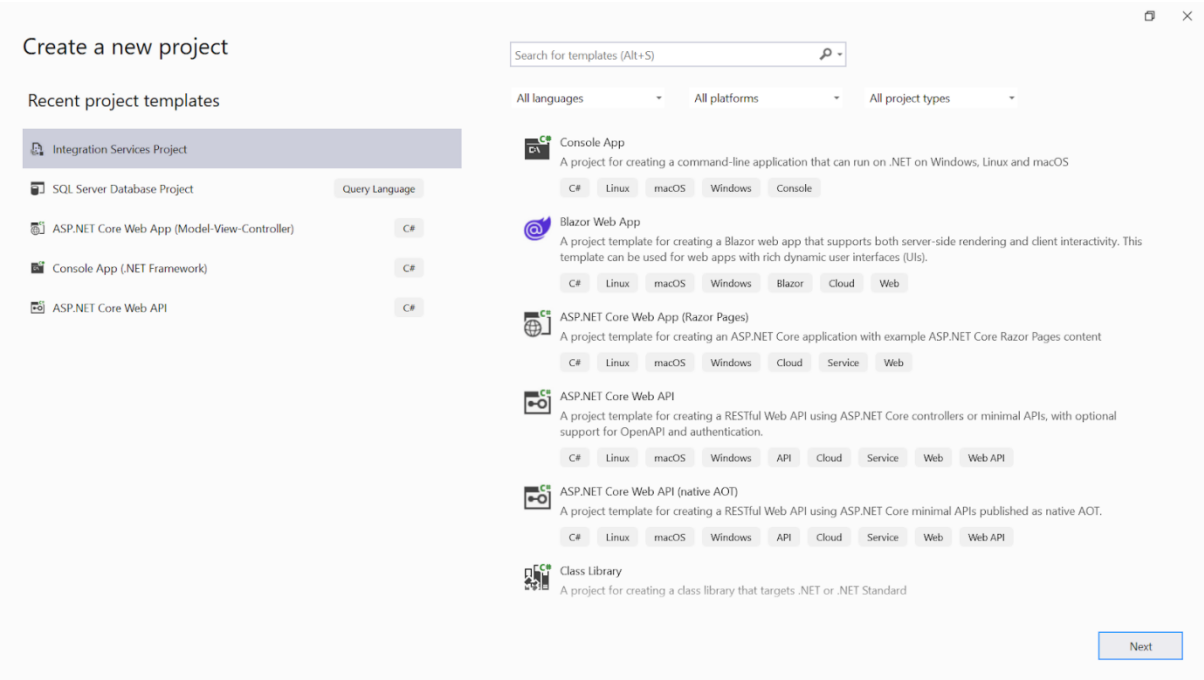
2.1. Tạo mới project SSIS

Bước 1: Mở Visual Studio Community 2022 và chọn Create a new project



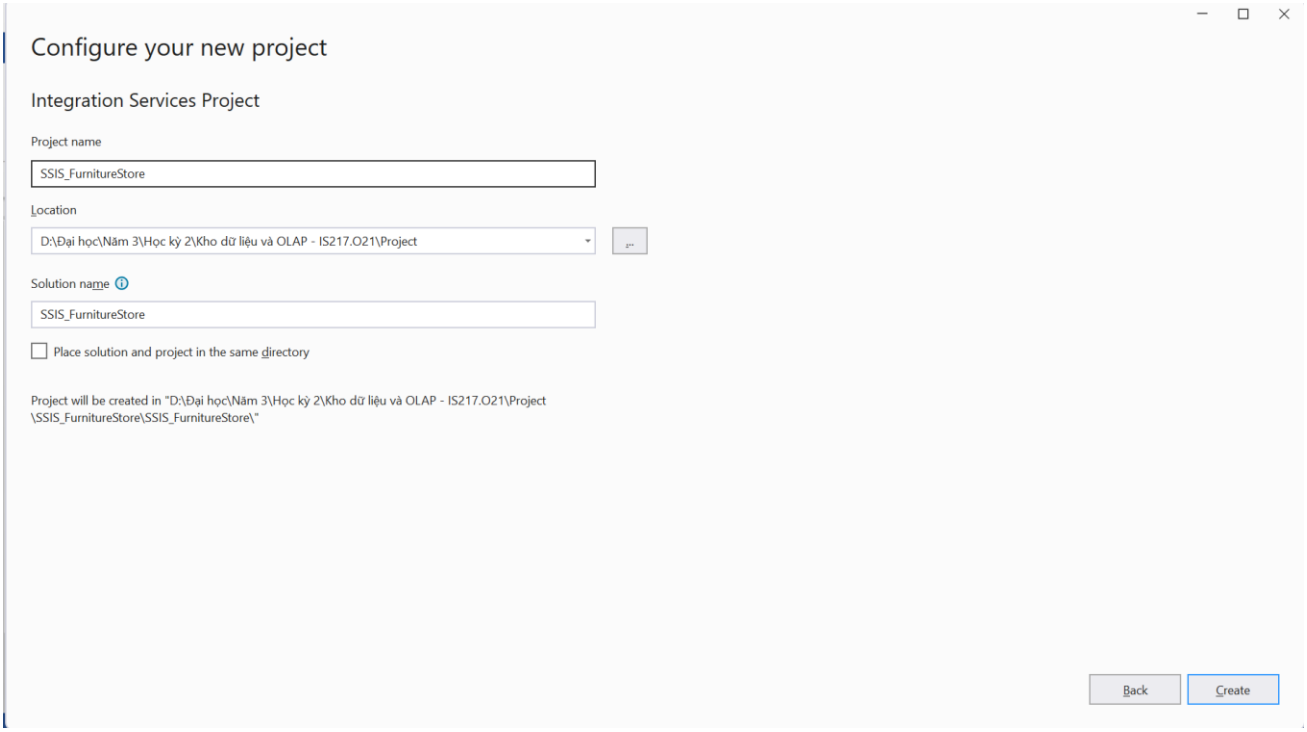
Hình 2.11 Tạo mới project trong Visual Studio Community 2022

Bước 2: Chọn **Integration Services Project** và chọn **Next**



Hình 2.12 Chọn mẫu Integration Services Project

Bước 3: Đặt tên cho project và nhấn **Create** để tiến hành tạo project.



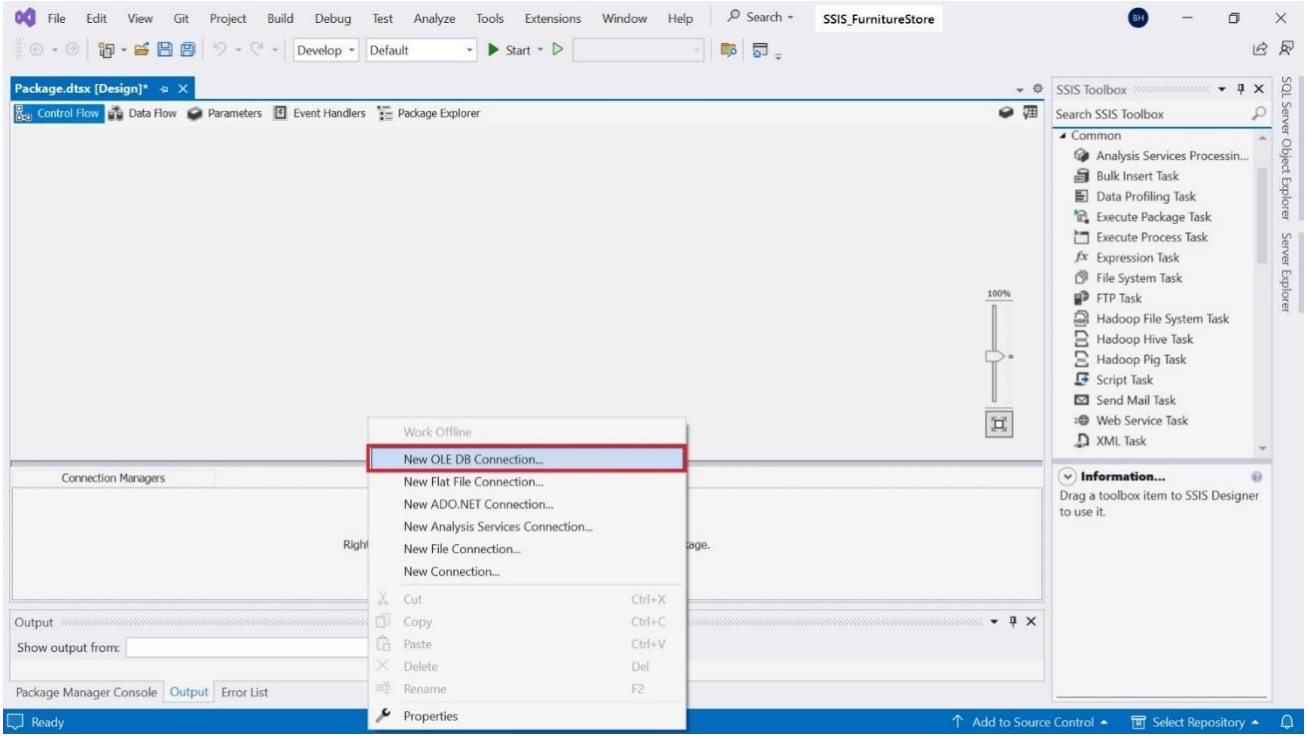
Hình 2.13 Đặt tên cho project và xác nhận tạo project

2.2. Tạo bảng các bảng Dim và bảng Fact

2.2.1. Tạo kết nối đến Server

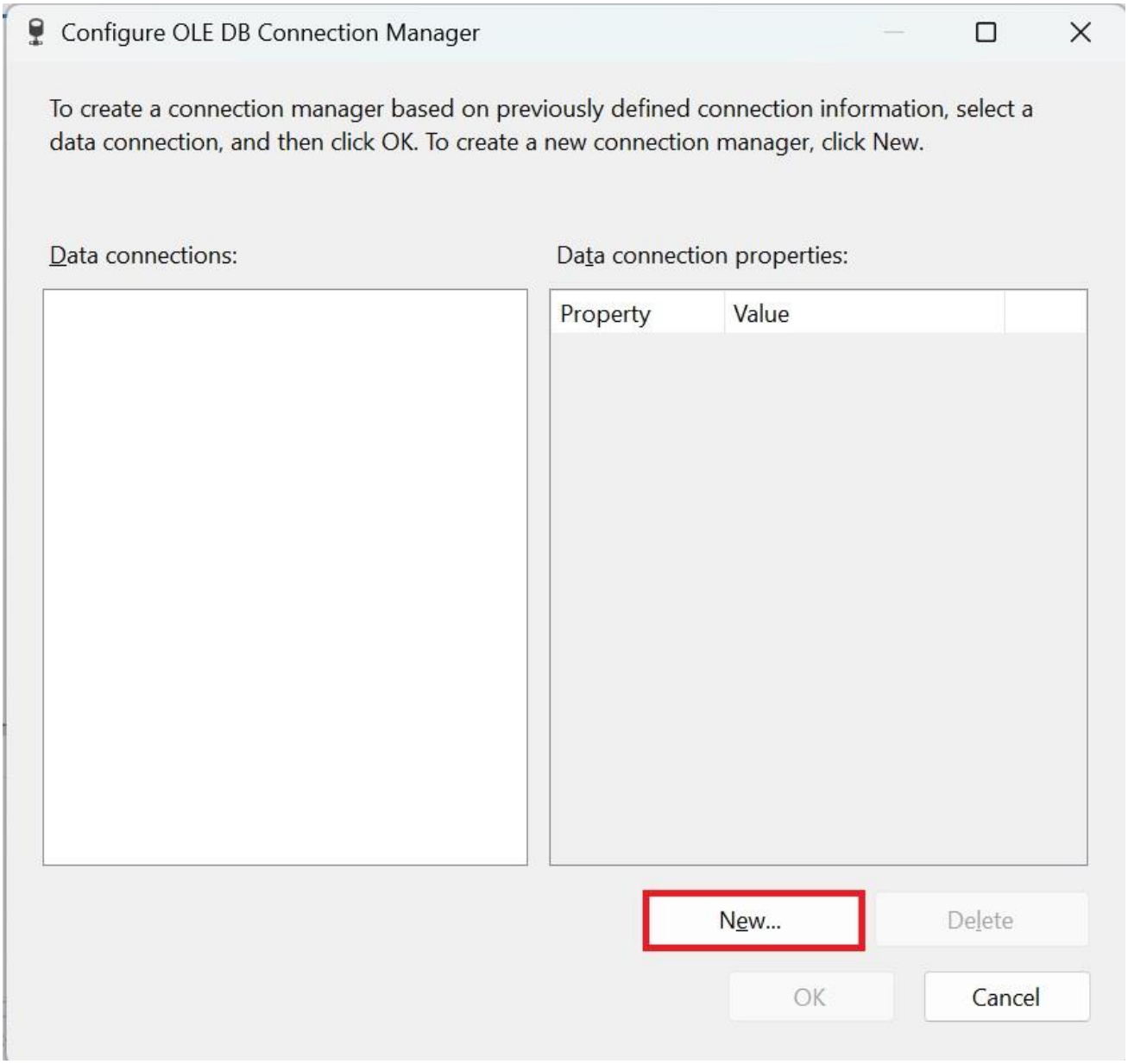
Trước khi tiến hành việc xử lý và tách dữ liệu ra các bảng khác nhau thì ta cần phải tạo kết nối đến với Server.

Bước 1: Trong vùng **Connection Managers** ta nhấn chuột phải và chọn New OLE DB Destination.



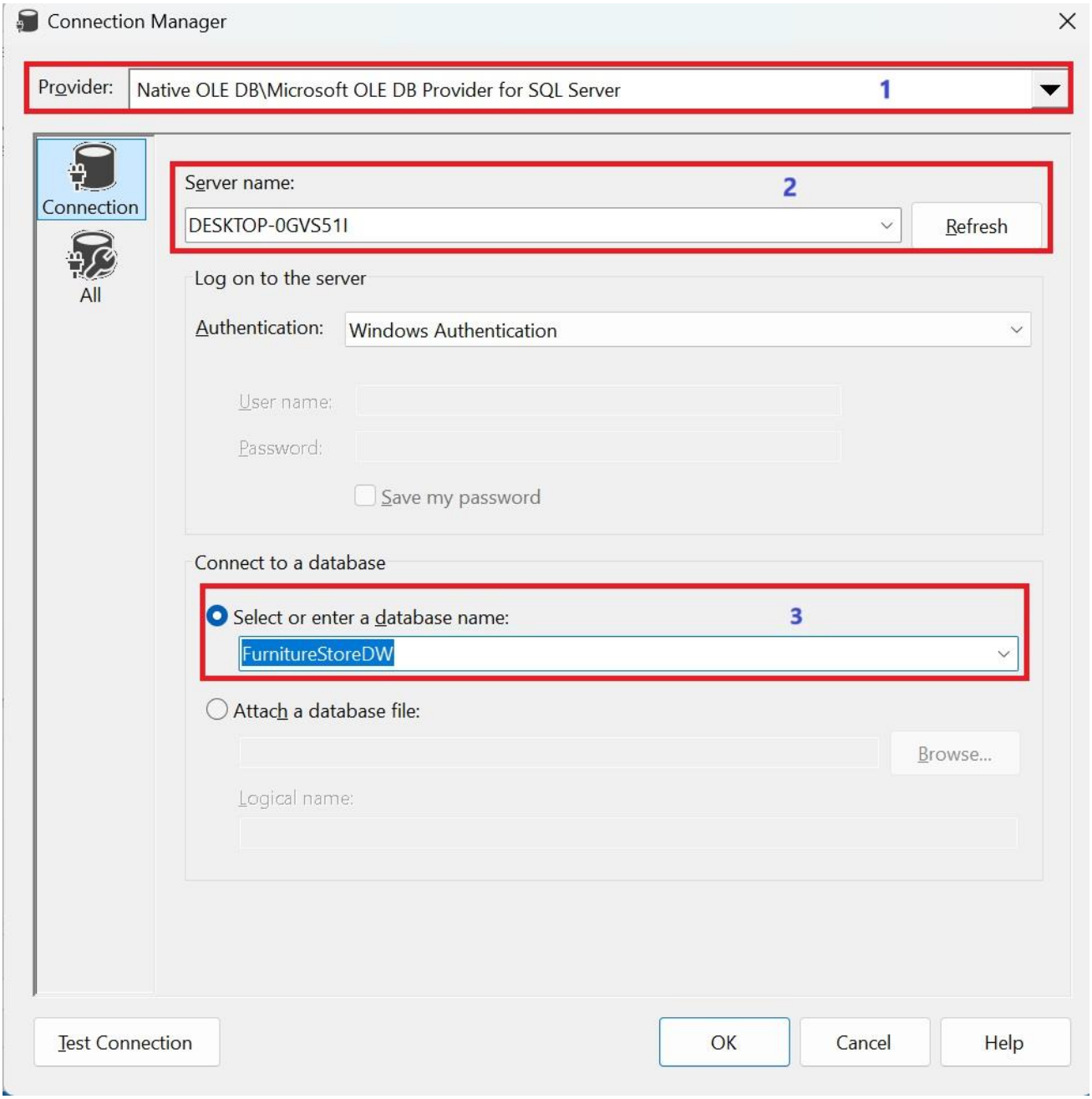
Hình 2.14 Tạo kết nối tới Server trong cửa sổ Connection Managers (1)

Bước 2: Nhấn vào "**New**" để tạo một Data Connection mới.



Hình 2.15 Tạo kết nối tới Server trong cửa sổ Connection Managers (2)

Bước 3: Tiến hành cài đặt Connection Manager theo thứ tự như sau: Chọn **Provider** → Điền tên **Server name** → Chọn kết cơ sở dữ liệu muốn kết nối tới → Nhấn **OK**. Thứ tự cài đặt có thể được thực hiện theo ảnh **2.16.**

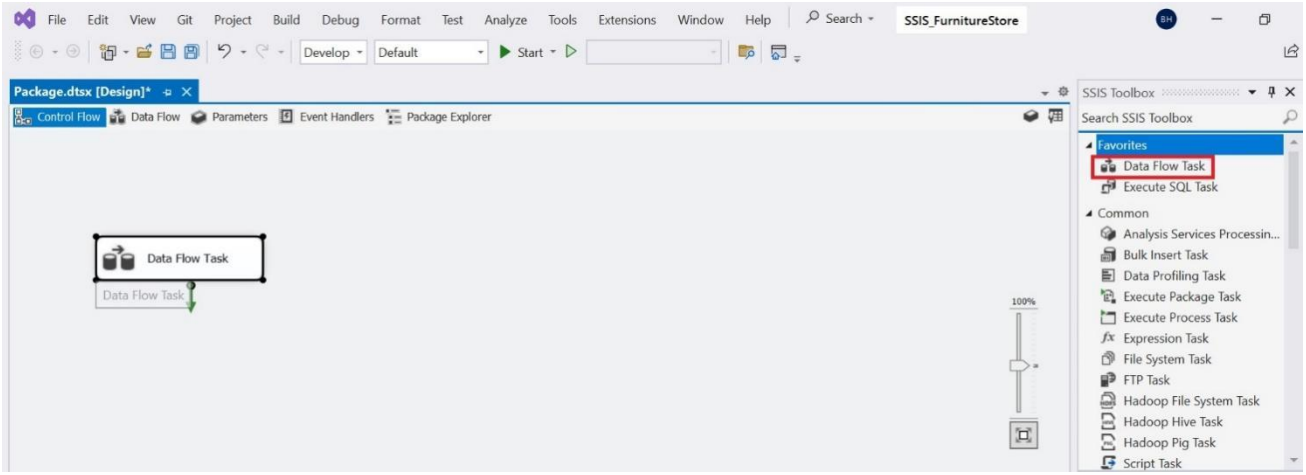


Hình 2.16 Tạo kết nối tới Server trong cửa sổ Connection Managers (3)

2.2.2. Tạo các bảng Dim và Fact

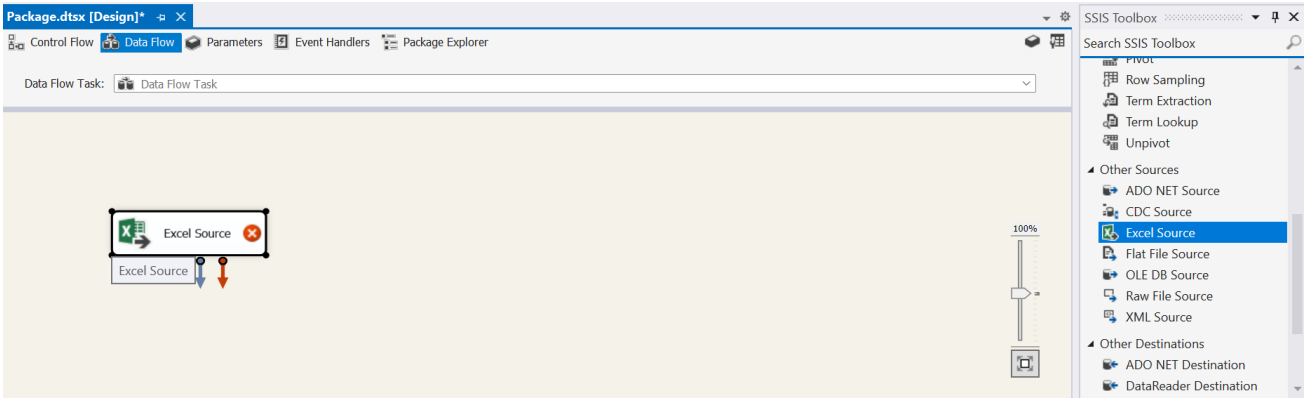
Trước khi có thể tạo các bảng Dim và bảng Fact thì ta cần tải dữ liệu gốc từ file Excel vào Data Flow:

Bước 1: Tạo một Data Flow mới bằng cách kéo thành phần **Data Flow Task** từ bên cửa sổ SSIS Toolbox vùng làm việc chính của project.



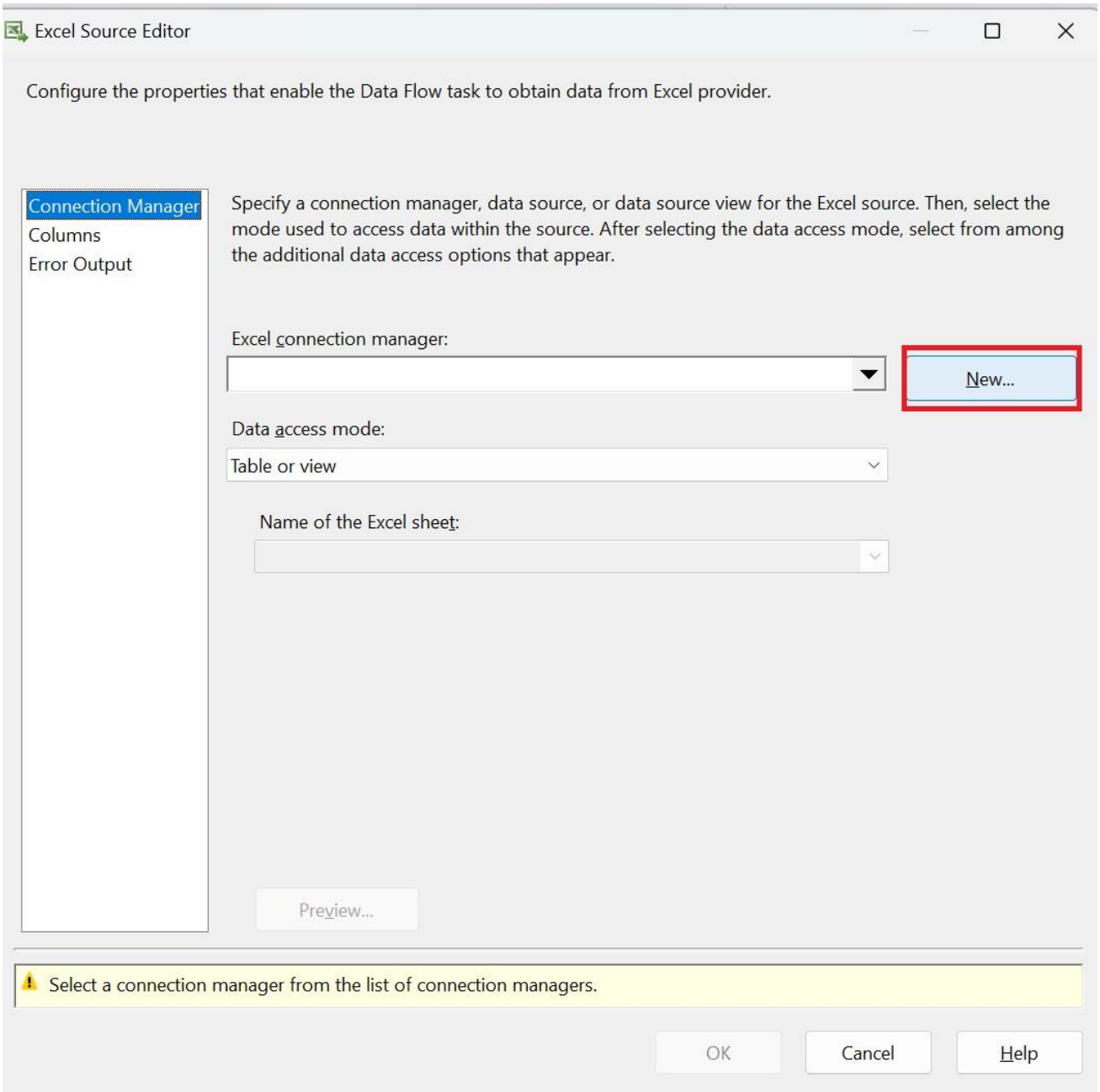
Hình 2.17 Tạo Data Flow Task mới

Bước 2: Khi nhấn đúp vào **Data Flow Task** vừa tạo thì IDE sẽ chuyển qua màn hình **Data Flow**. Ở đây vì nhóm muốn tải dữ liệu file Excel vì vậy nhóm sẽ sử dụng thành phần **Excel Source** của **SSIS Toolbox.**



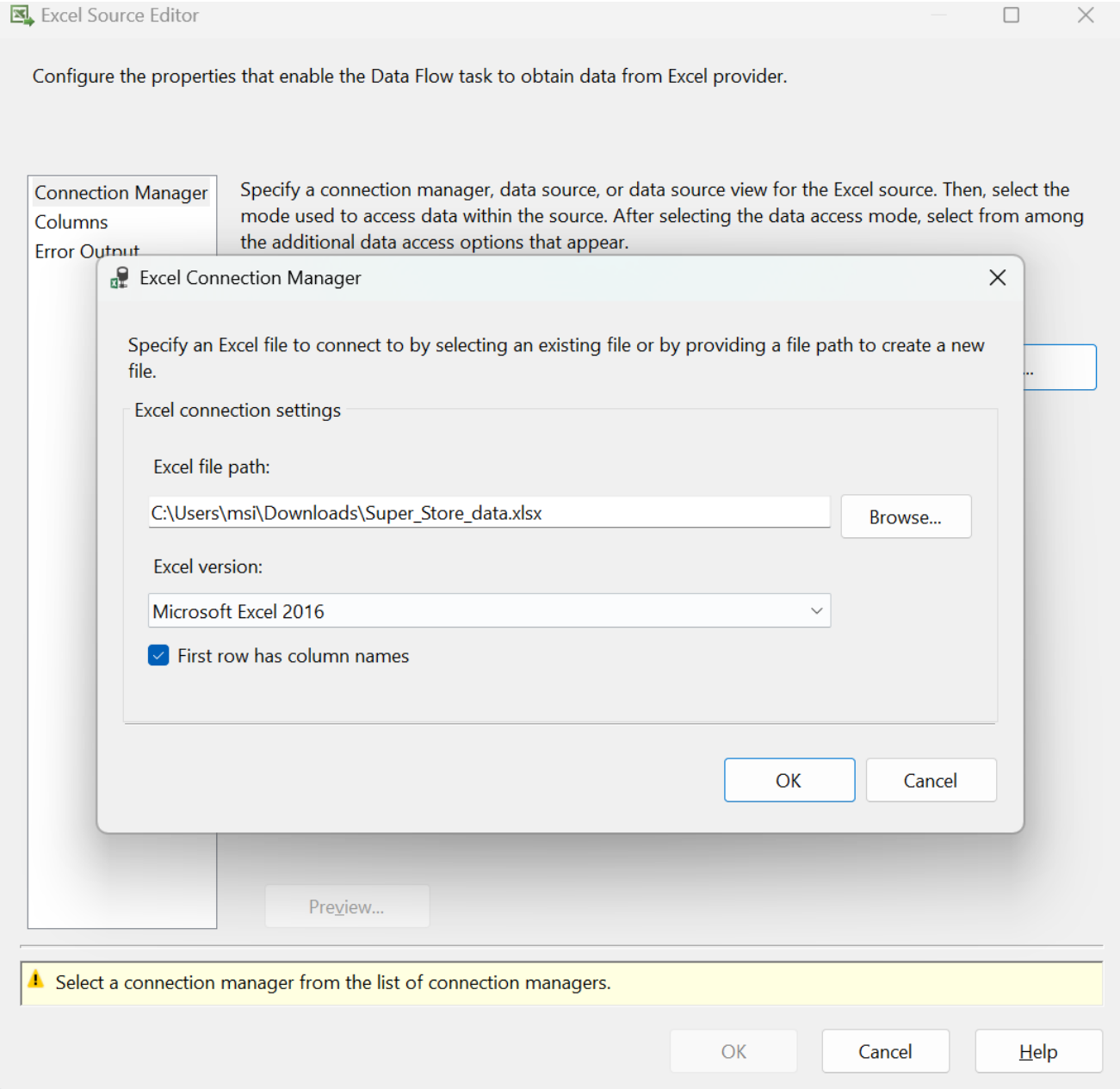
Hình 2.18 Sử dụng Excel Source để tải dữ liệu từ file Excel

Bước 3: Nhấn đúp vào Excel Source mới vừa tạo để mở hộp thoại Editor. Trong hộp thoại thì nhấn New để tạo mới một Connection Manager.



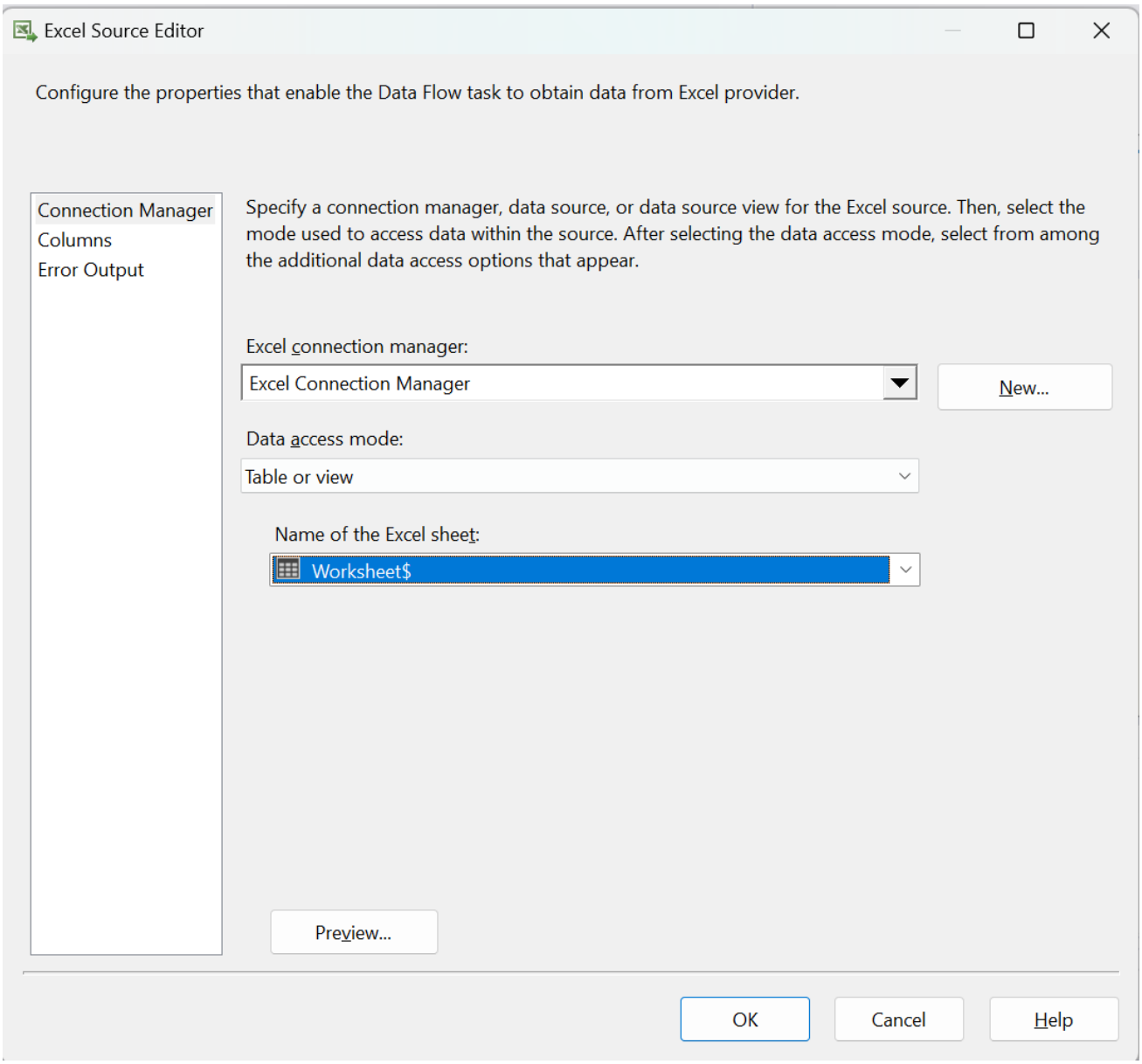
Hình 2.19 Hộp thoại Editor của Excel Source

Bước 4: Khi hộp thoại tạo mới mở lên thì ta tiến hành chọn đường dẫn tới file Excel chứa dữ liệu gốc trong ô **Excel file path** và chọn phiên bản trong **Excel version** và nhấn **OK.**



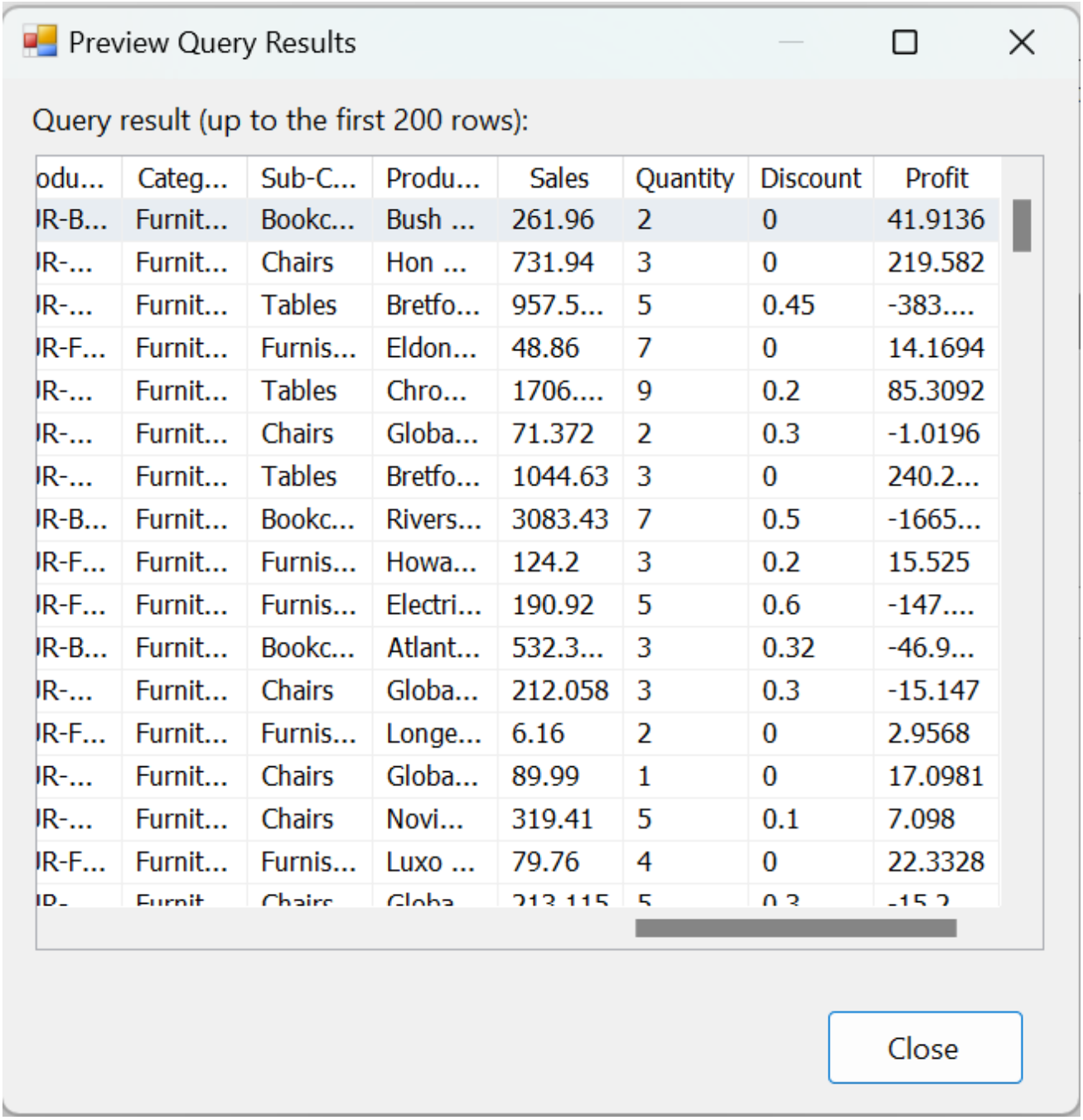
Hình 2.20 Tạo mới một Excel Connection

Bước 5: Chọn Excel sheet chứa dữ liệu.



Hình 2.21 Chọn Excel Sheet

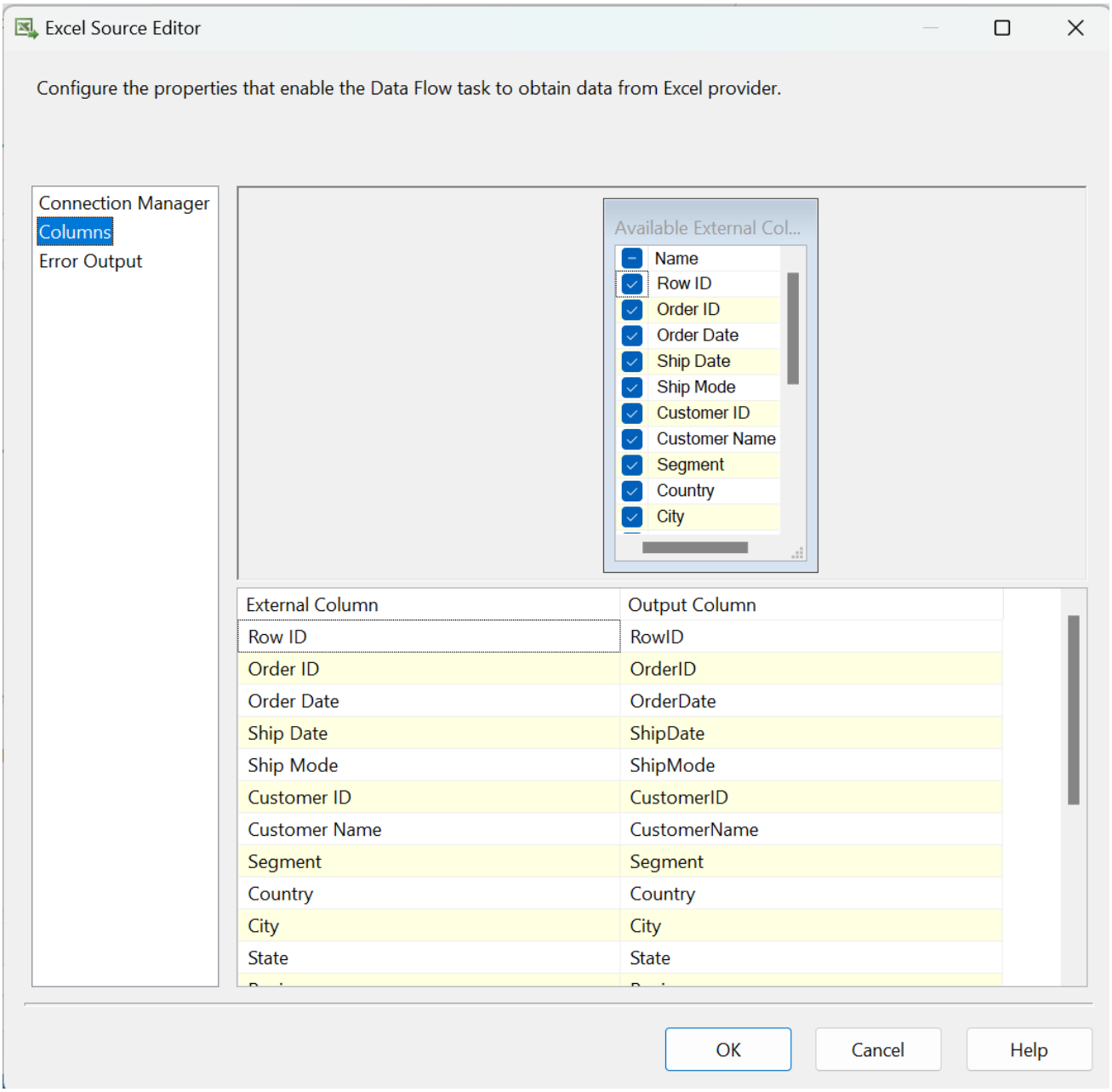
Bước 6: Nhấn vào nút **Preview…** để xem lại các cột dữ liệu và dữ liệu có trong file Excel.



odu...	Categ...	Sub-C...	Produ...	Sales	Quantity	Discount	Profit
IR-B...	Furnit...	Bookc...	Bush ...	261.96	2	0	41.9136
IR-...	Furnit...	Chairs	Hon ...	731.94	3	0	219.582
IR-...	Furnit...	Tables	Bretfo...	957.5...	5	0.45	-383....
IR-F...	Furnit...	Furnis...	Eldon...	48.86	7	0	14.1694
IR-...	Furnit...	Tables	Chro...	1706....	9	0.2	85.3092
IR-...	Furnit...	Chairs	Globa...	71.372	2	0.3	-1.0196
IR-...	Furnit...	Tables	Bretfo...	1044.63	3	0	240.2...
IR-B...	Furnit...	Bookc...	Rivers...	3083.43	7	0.5	-1665...
IR-F...	Furnit...	Furnis...	Howa...	124.2	3	0.2	15.525
IR-F...	Furnit...	Furnis...	Electri...	190.92	5	0.6	-147....
IR-B...	Furnit...	Bookc...	Atlant...	532.3...	3	0.32	-46.9...
IR-...	Furnit...	Chairs	Globa...	212.058	3	0.3	-15.147
IR-F...	Furnit...	Furnis...	Longe...	6.16	2	0	2.9568
IR-...	Furnit...	Chairs	Globa...	89.99	1	0	17.0981
IR-...	Furnit...	Chairs	Novi...	319.41	5	0.1	7.098
IR-F...	Furnit...	Furnis...	Luxo ...	79.76	4	0	22.3328

Hình 2.22 Xem trước dữ liệu được tải lên

Bước 7: Chuyển qua tab **Columns** để chọn các cột dữ liệu mà chúng ta muốn sử dụng để thực hiện trong suốt quá trình **SSIS**, đối với nhóm thì nhóm sẽ không sử dụng **Postal Code** và **Category** nên sẽ bỏ chọn hai ô này. Đồng thời, chúng ta sẽ đặt **alias** cho **Sub-Category** từ **Sub-Category** thành **Category**.

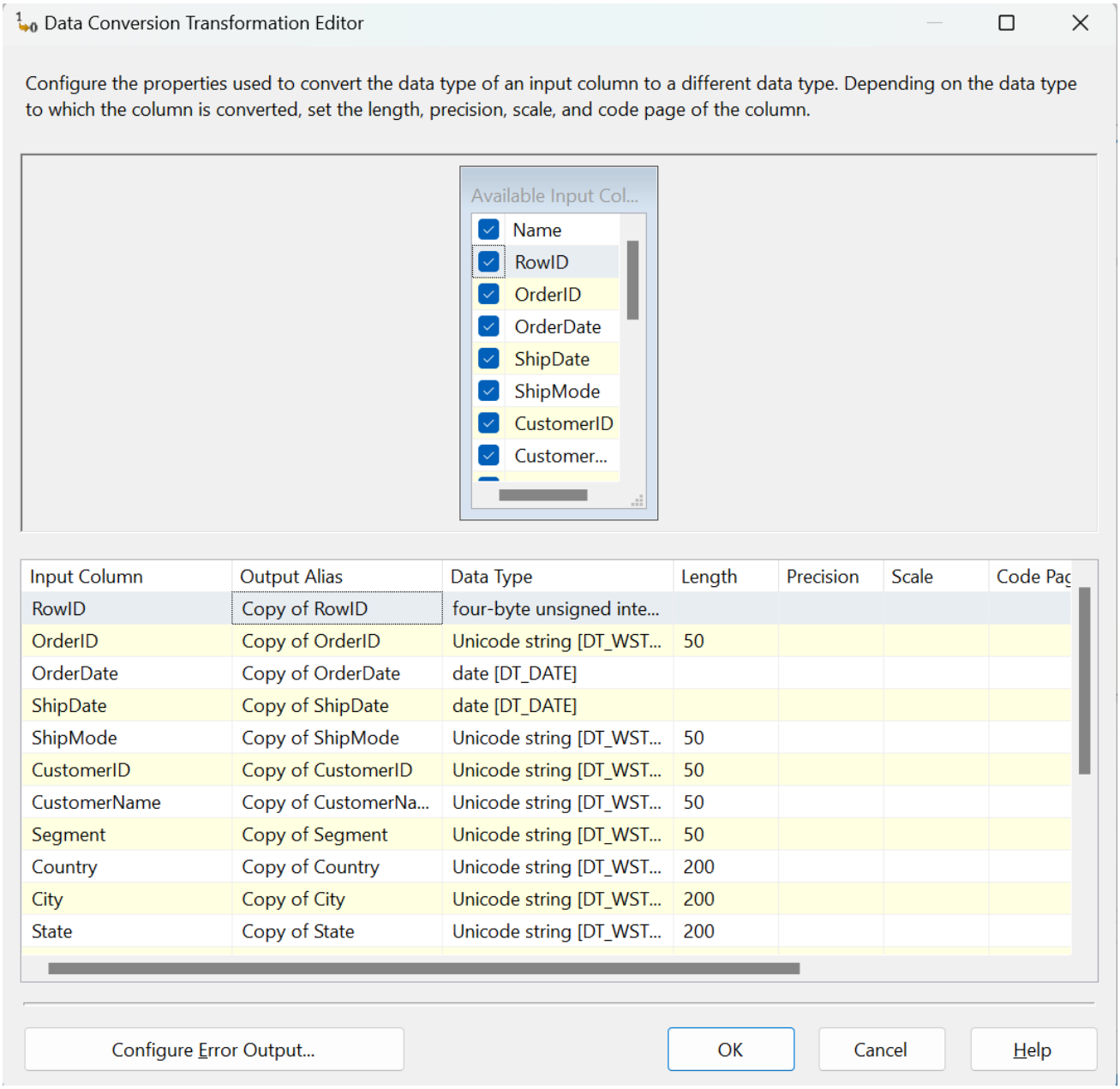


Hình 2.23 Chọn những cột dữ liệu cần sử dụng cho quá trình SSIS

*** Lưu ý:** Vì Output Column sẽ mặc định lấy tên giống với External Column vì vậy để tránh việc có khoảng cách trong tên của các cột dữ liệu sẽ khiến cho việc có truy xuất dữ liệu SQL có thể lỗi nên nhóm đã quyết định sẽ xoá khoảng trắng trong tên của các cột dữ liệu.

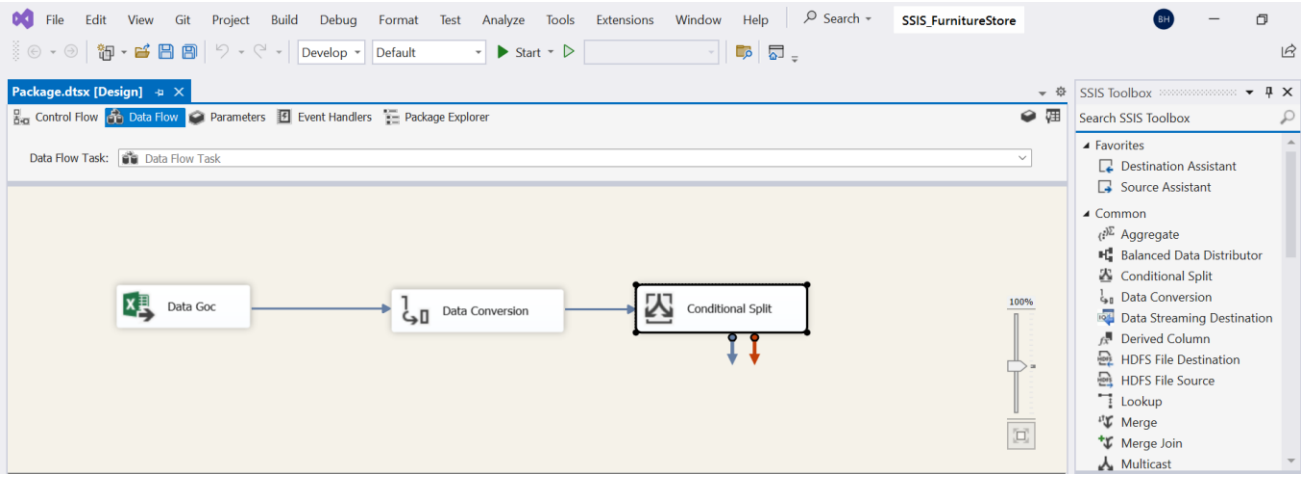
Bước 8: Sau khi đã kiểm tra đã kiểm tra tất cả các thông tin thì nhấn **OK** để hoàn tất quá trình tải dữ liệu vào **Excel File Source**.

Bước 9: Tạo thành phần **Data Conversion** để giúp chuyển đổi kiểu dữ liệu của các cột dữ liệu nhằm tránh những các lỗi về định dạng và giúp cho quá trình thao tác với các dữ liệu dễ dàng hơn cho các bước sau. Sau khi tạo thành phần thì chúng ta sẽ kết nối **Excel File Source** với **Data Conversion** bằng cách kéo mũi tên **Excel File Source** nối với **Data Conversion**.

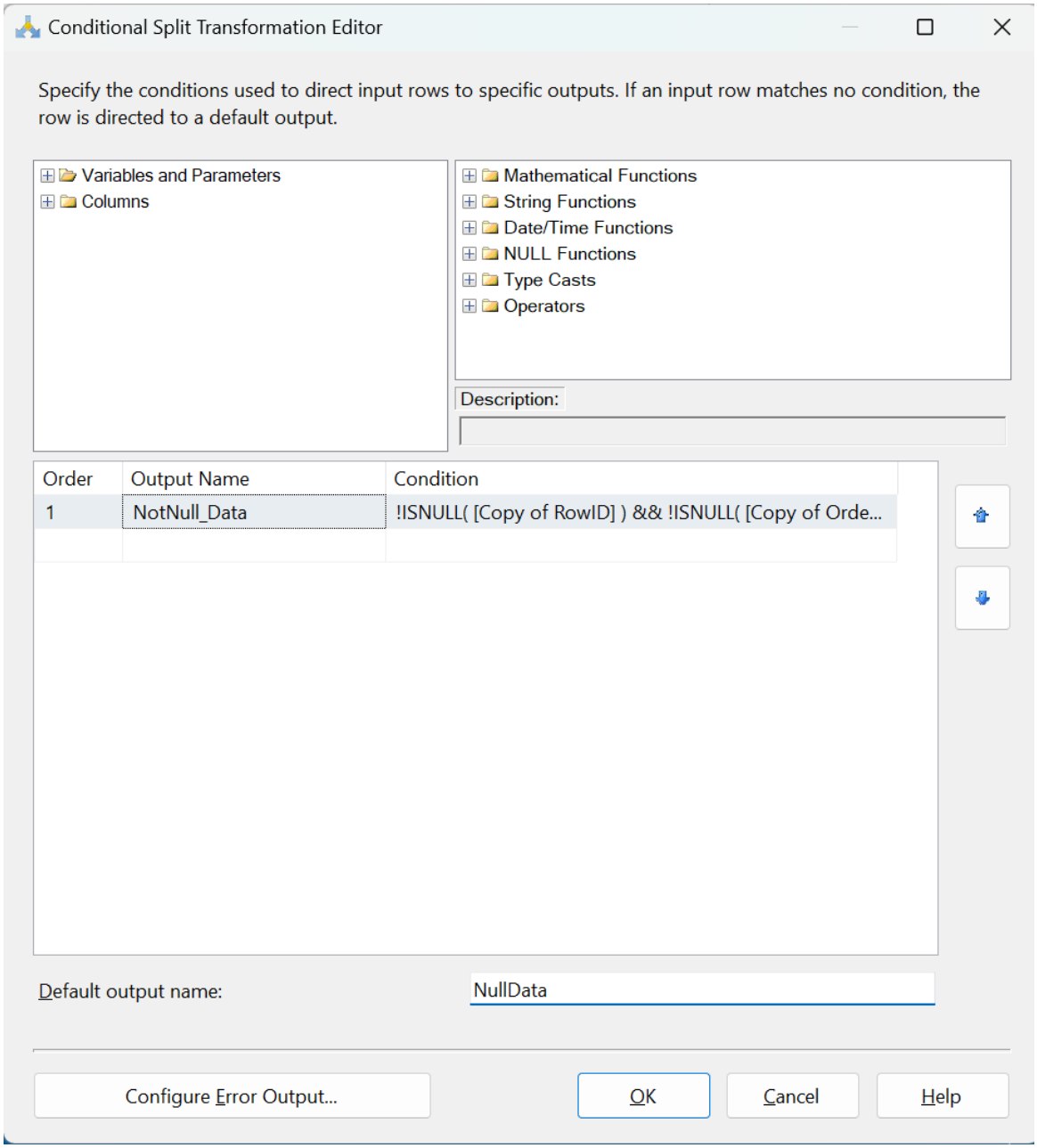


Hình 2.24 Chuyển kiểu dữ liệu của các cột bằng Data Conversion

Bước 10: Dùng công cụ **Conditional Split** và kết nối với **Data Conversion** để bắt đầu cắt dữ liệu có điều kiện, lọc ra những dòng NULL ra khỏi dữ liệu trước khi đưa dữ liệu vào kho dữ liệu. Những dữ liệu không bị NULL thì ta sẽ đặt tên của **Output Name** là **NotNull_Data**

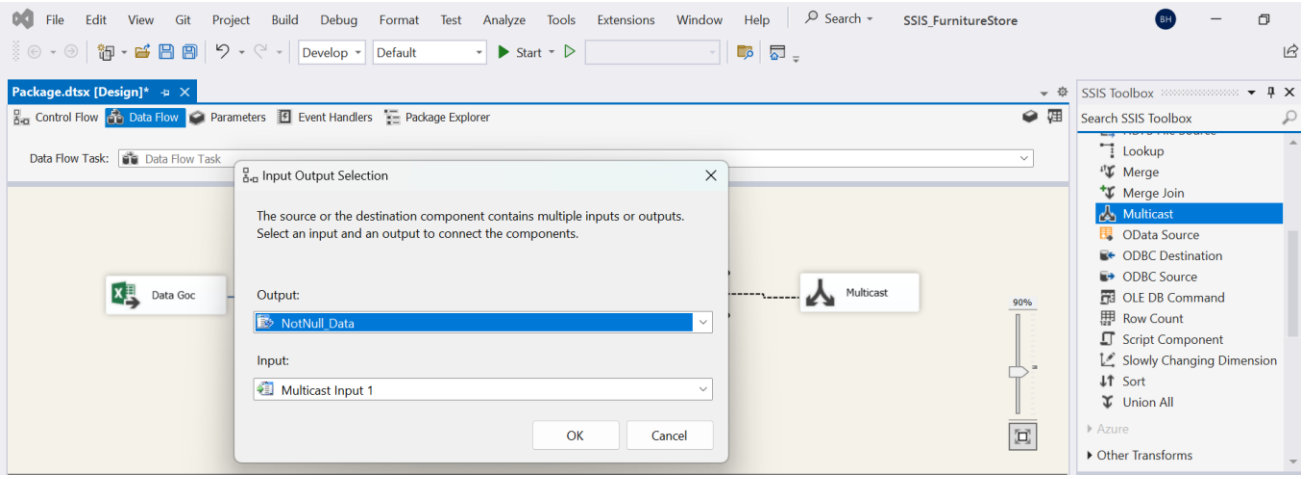


Hình 2.25 Kết nối Data Conversion tới Conditional Split



Hình 2.26 Lọc ra những dòng dữ liệu bị NULL

Bước 11: Tạo **Multicast** để có thể phân tán dữ liệu vừa được xử lý thành nhiều phiên bản đến các bảng Dimension và Fact. Sau đó, chúng ta sẽ nối **Conditional Split** đến với **Multicast** để kết nối

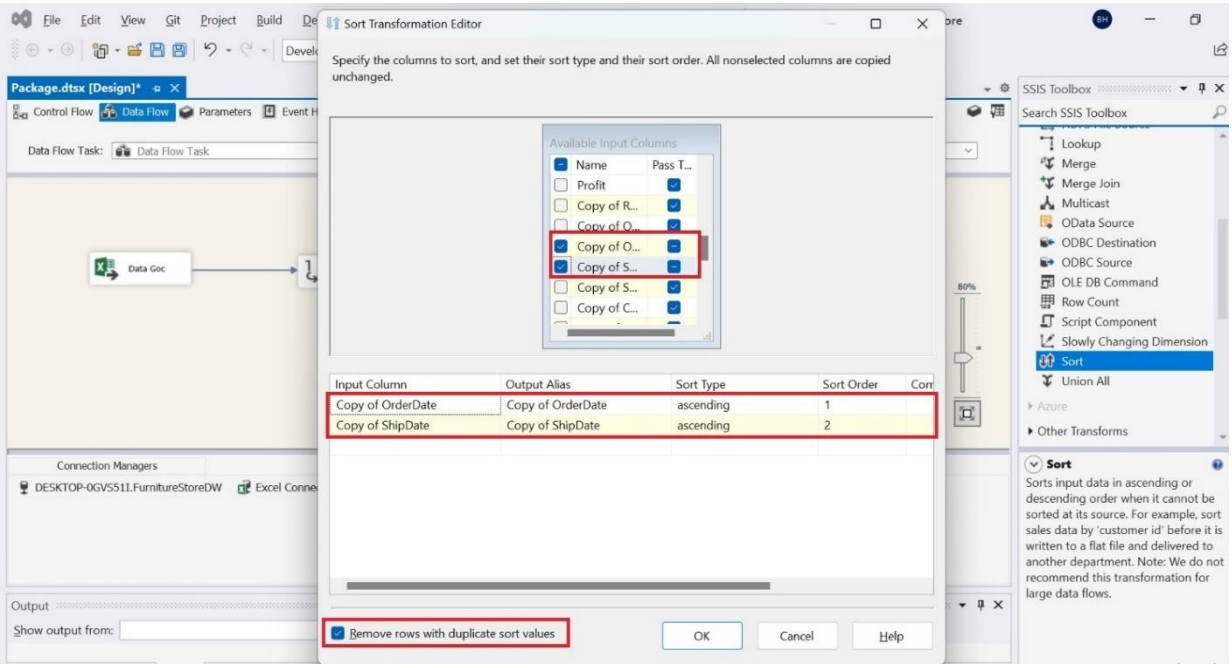


Hình 2.27 Sử dụng Multicast để phân tán dữ liệu

Bảng Dim_Date

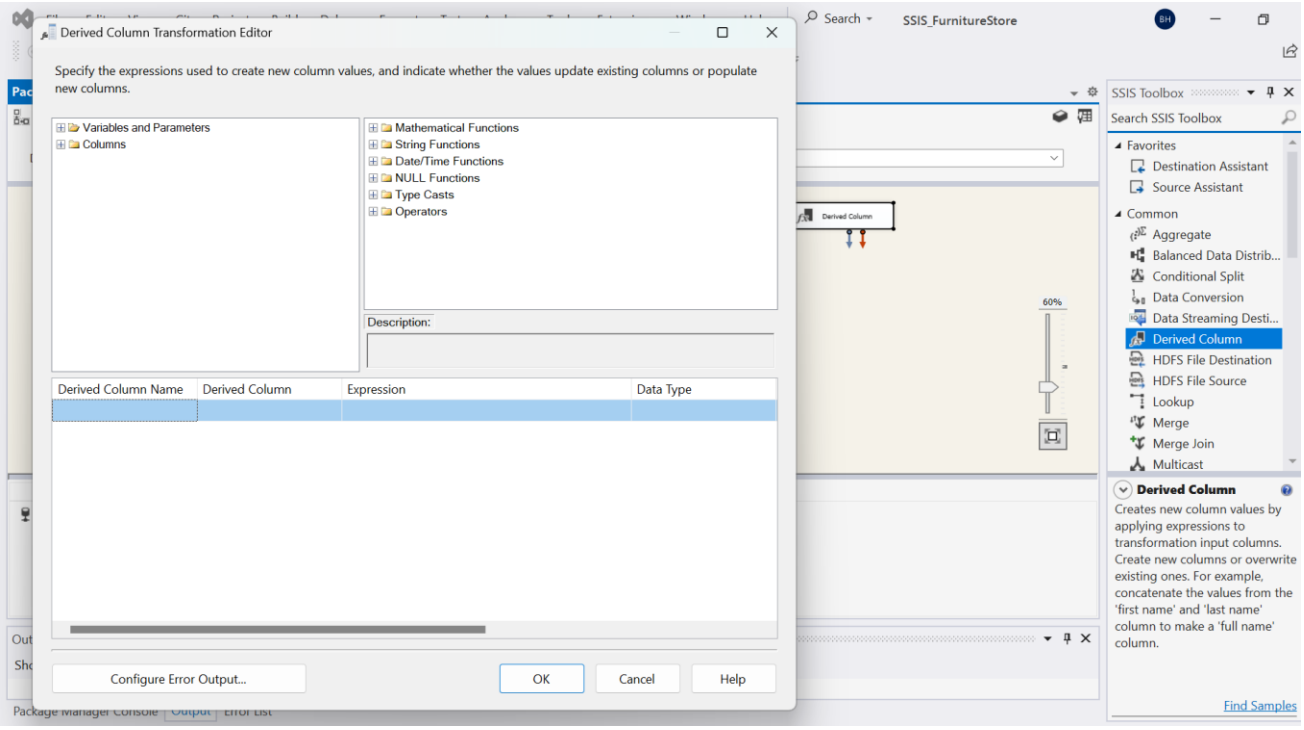
Bước 1: Tạo mới thành phần **Sort** của SSIS Toolbox và đặt tên là Sort_Date để lấy ra các cột dữ liệu cần thiết cho **Dim_Date**. Nhấn đúp **Sort_Date** để mở hộp thoại **Edit**.

Hai dữ liệu mà ta cần cho **Dim_Date** đó chính là **OrderDate** và **ShipDate** vì vậy ta sẽ chọn vào ô đầu tiên của hai dòng **Copy of OrderDate** và **Copy of ShipDate**(dữ liệu đã được xử lý) theo thứ tự vừa nêu. Sau đó, ta chọn **Remove rows with duplicate sort values** để xoá các dòng dữ liệu trùng nhau → Chọn **OK** để hoàn thành quá trình chọn.



Hình 2.28 Chọn dữ liệu cần thiết cho Dim_Date và xoá dòng trùng

Bước 2: Thêm thành phần **Derived Column** và nhấn đúp để mở hộp thoại **Edit**.

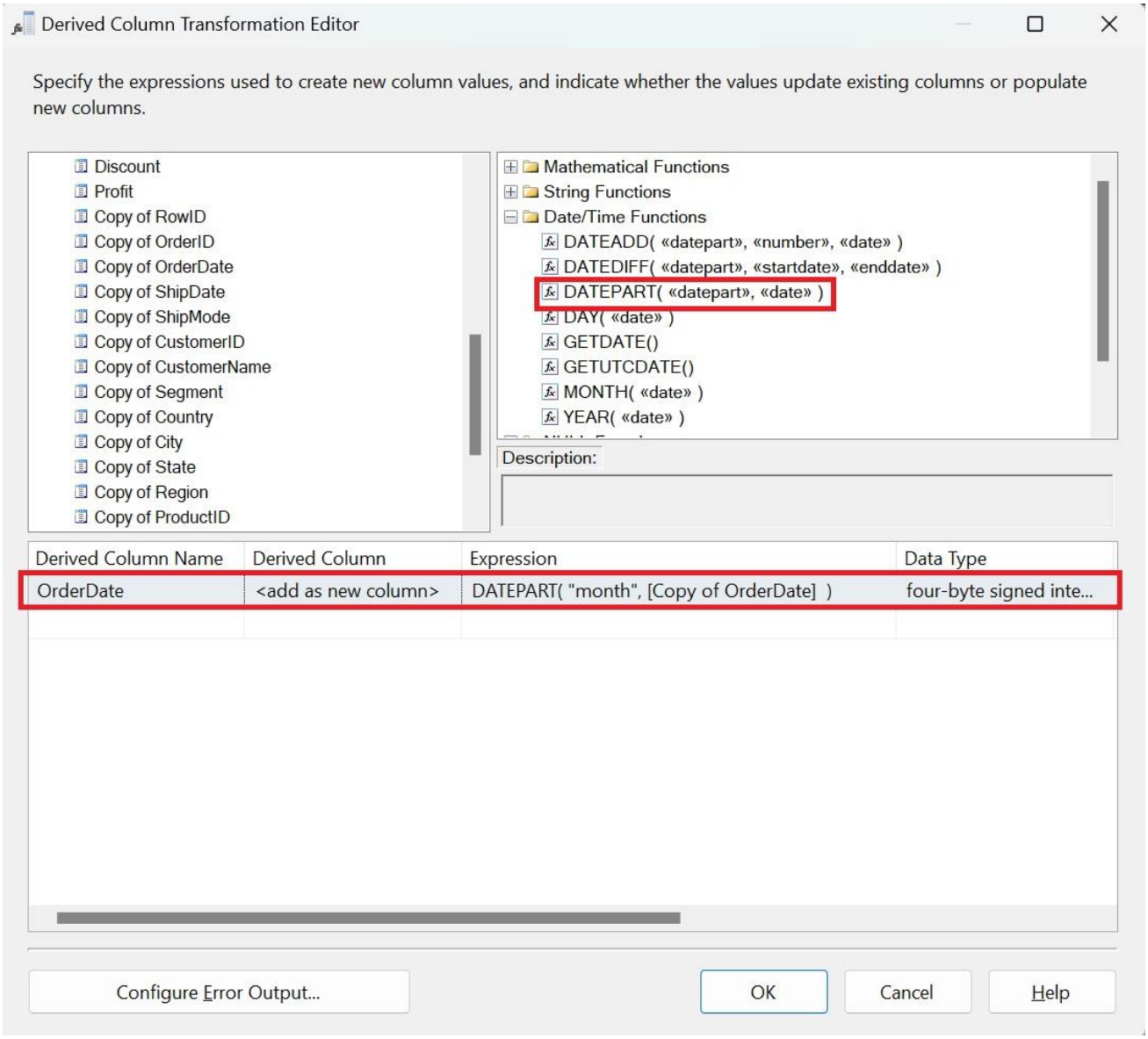


Hình 2.29 Sử dụng Derived Column xử lý dữ liệu trước đưa dữ liệu vào Dim_Date

Bước 3: Chia dữ liệu từ cột **Copy of OrderDate** có kiểu dữ liệu **dd/MM/yyyy** thành các cột OrderMonth, OrderQuarter, OrderYear

Ở đây chúng ta sẽ chọn phương thức **DATEPART(<kiểu dữ liệu thời gian>, cột dữ liệu)** để tiến hành tách các thuộc tính cần thiết của dữ liệu ngày tháng. Ví dụ, nếu như ta muốn lấy tháng của **Copy of OrderDate** thì câu lệnh sẽ như sau:

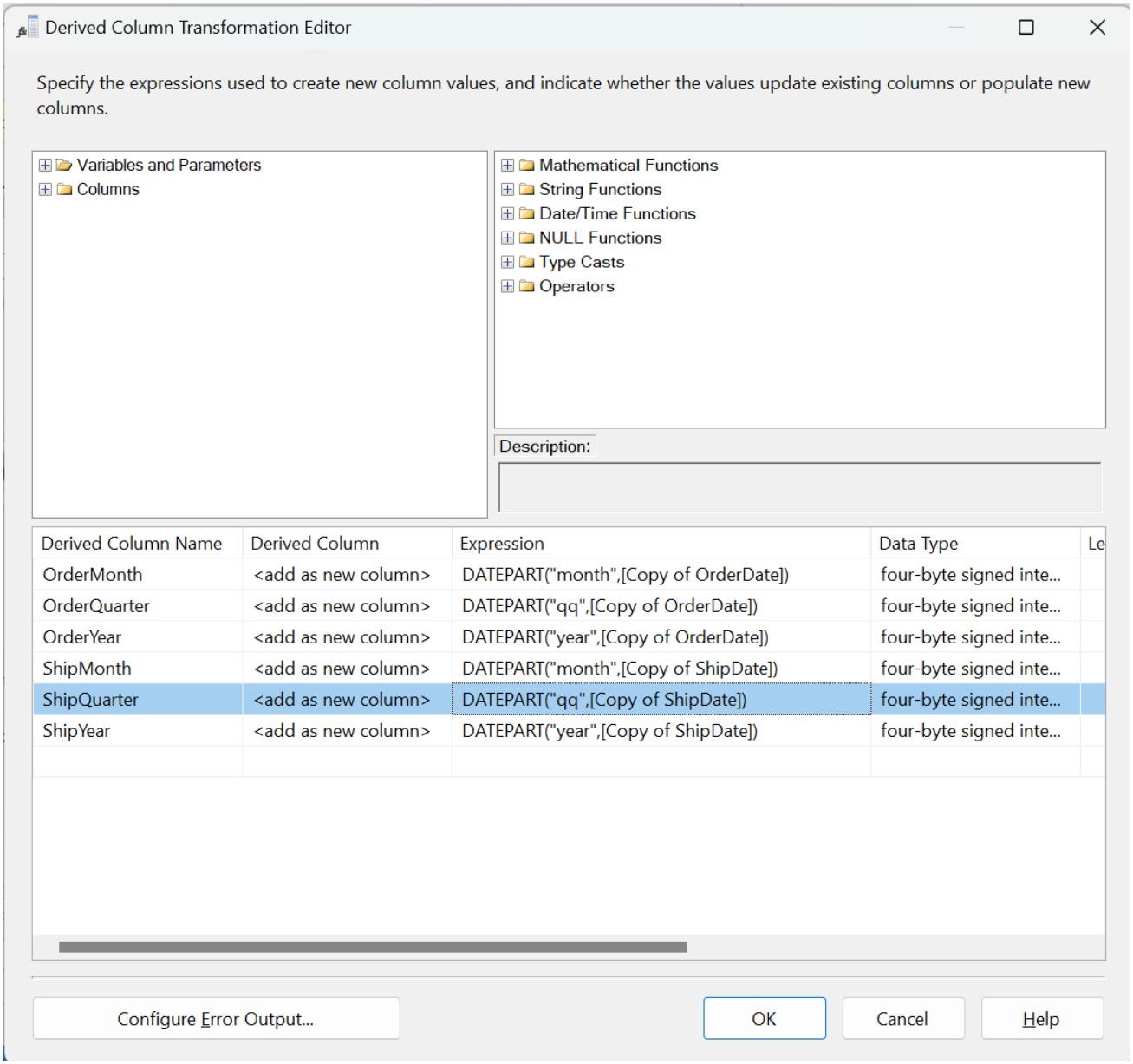
DATEPART(“**month**”, [**Copy Of OrderDate**])



Hình 2.30 Sử dụng hàm DATEPART để tách các thành phần của dữ liệu ngày tháng

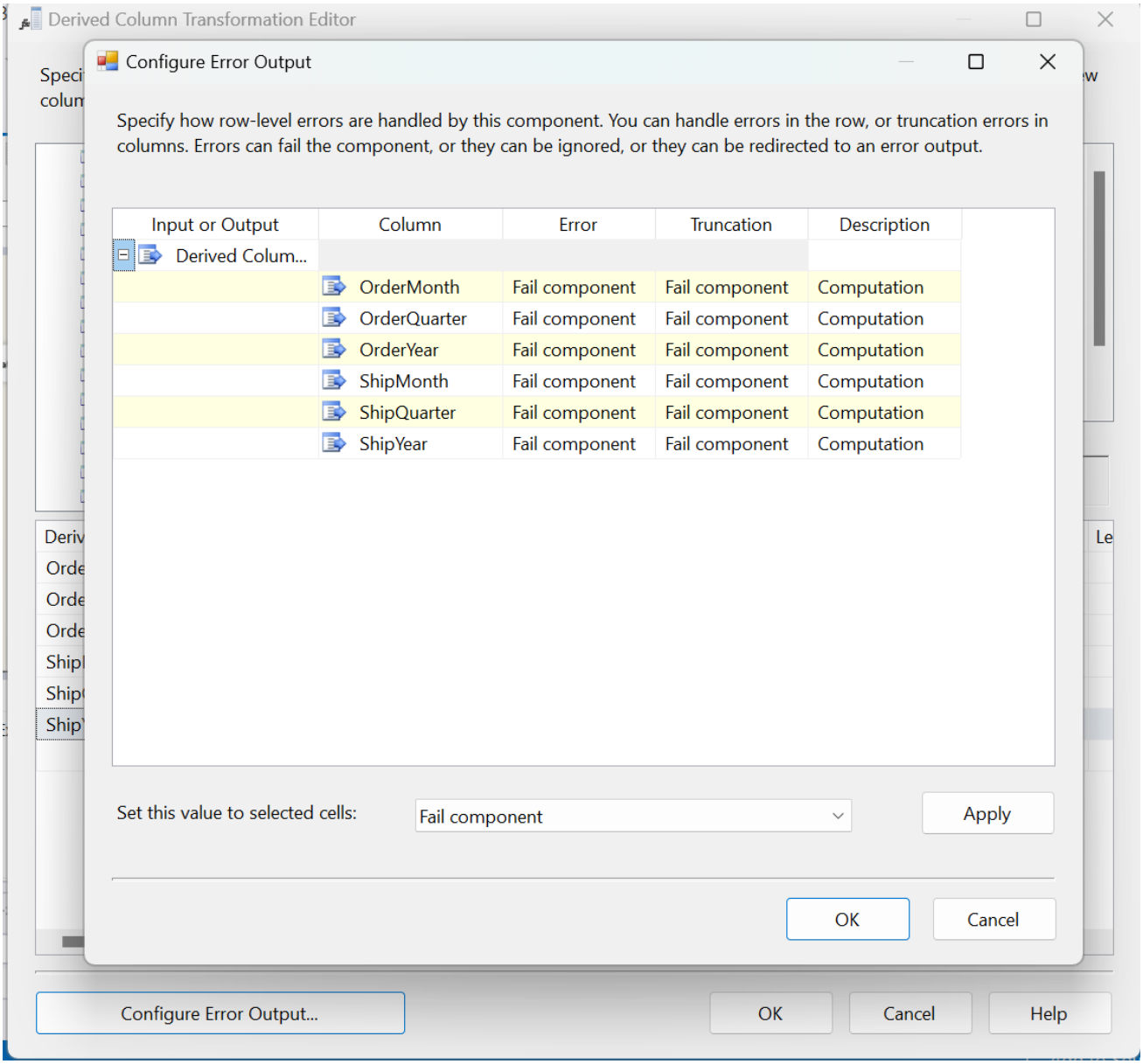
Bước 4: Tương tự với các cột **OrderQuarter, OrderYear** của cột **Copy of OrderDate** và các cột **ShipMonth, ShipQuarter, ShipYear** của cột **Copy of ShipDate**.

Tiếp theo, nhấn vào nút **Configure Error Output…**



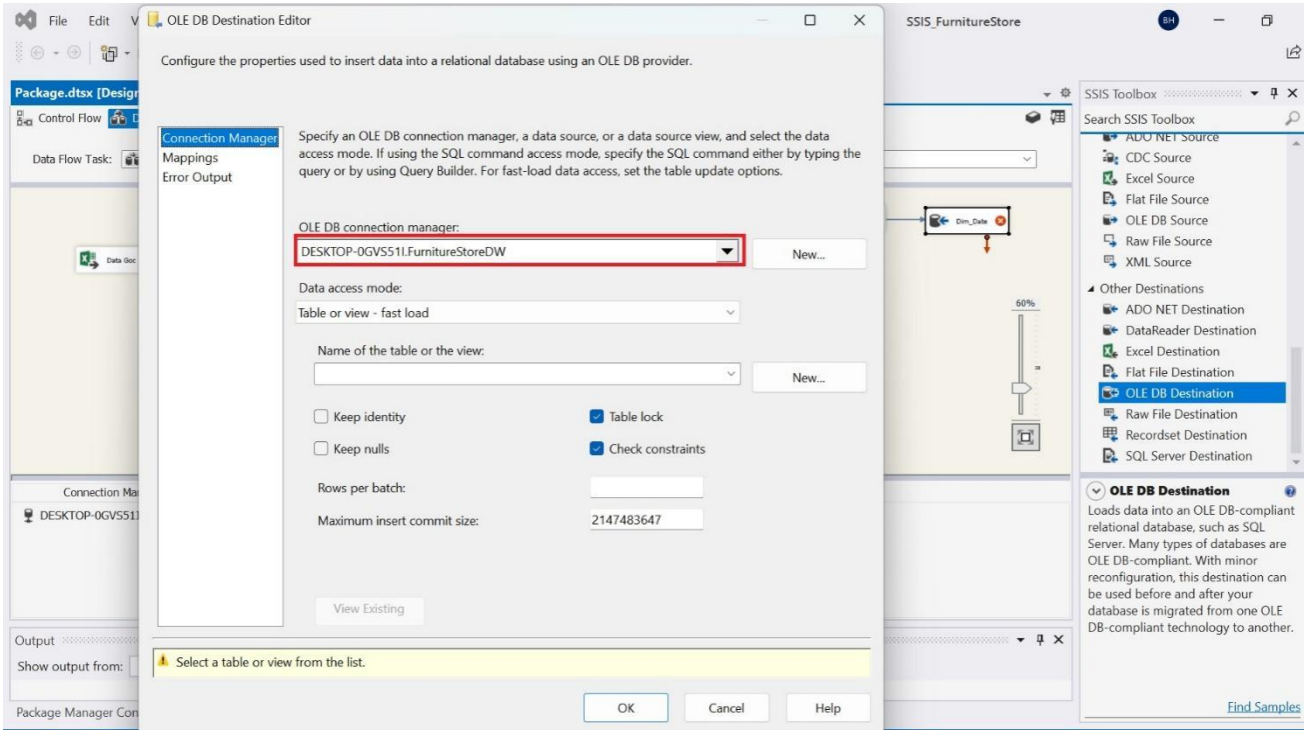
Hình 2.30 Áp dụng hàm DATEPART để tách thành các cột cần thiết

Bước 5: Trong hộp thoại **Configure Error Output** thì ta có thể thấy từ 2 cột **Copy of OrderDate** và **Copy of ShipDate** đã được chia thành 6 cột tương ứng với lược đồ dữ liệu của bảng **Dim_Date** mà ta đã đề ra. Nhấn **OK** để hoàn tất quá trình chia cột.



Hình 2.31 Mở hộp thoại Configure Error Output để kiểm tra các cột vừa tách

Bước 6: Tạo một **OLE DB Destination**. Đây là nơi chúng ta định nghĩa một nơi để chứa những dữ liệu ta đã xử lý trong quá trình SSIS. Ở đây thì chúng ta sẽ đặt tên cho thành phần này là **Dim_Date** và nhấn đúp vào **Dim_Date** để mở hộp thoại **Editor** và chọn kết nối mà chúng ta đã tạo ở bước **Tạo kết nối đến Server** để kết nối với **FurnitureStoreDW** trong mục **OLE DB Connection Manager**.



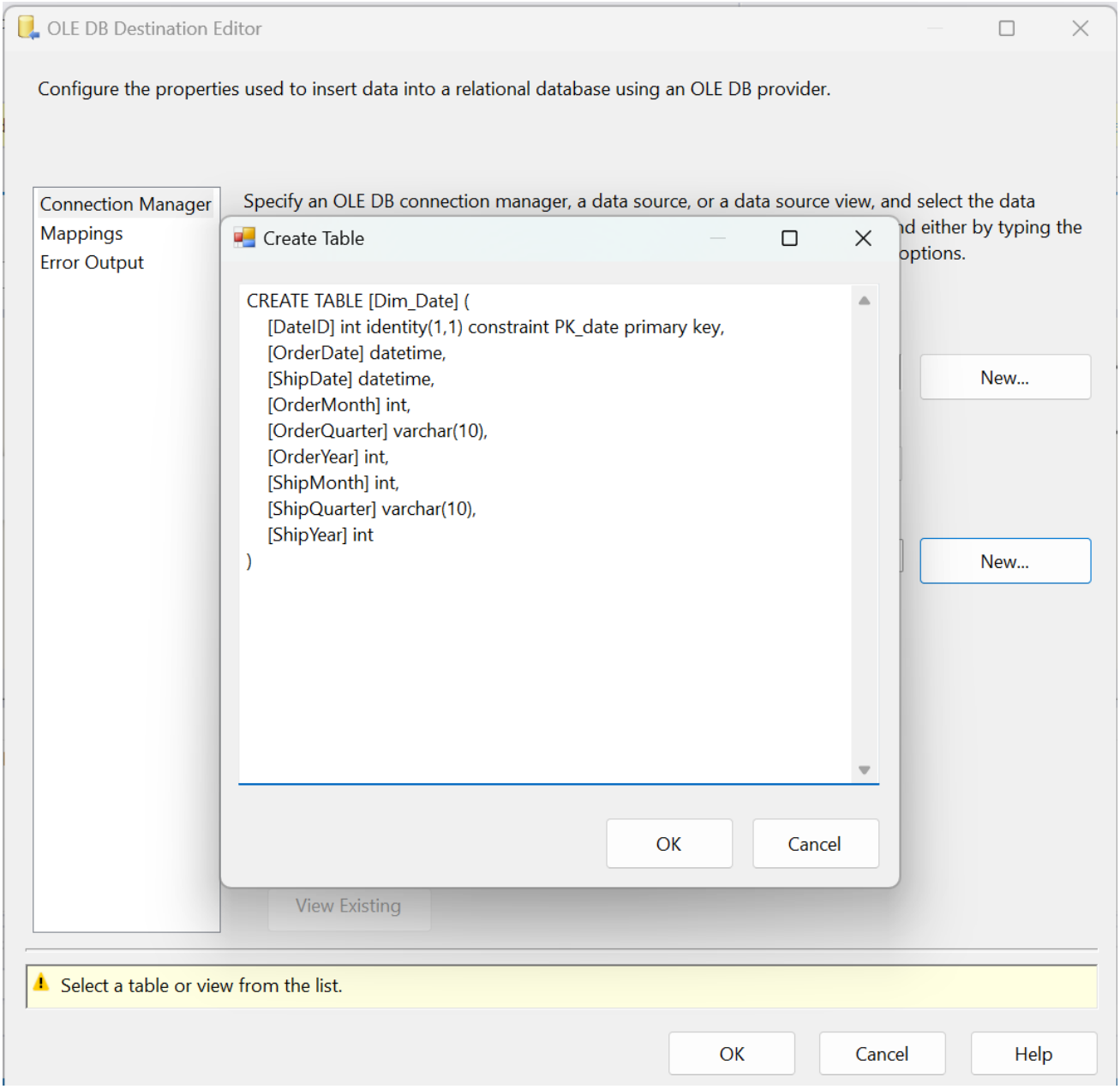
Hình 2.32 Chọn kết nối đã được thiết lập

Bước 7: Sau khi đã chọn kết nối thì chọn **New** để tiến hành tạo mới bảng **Dim_Date**.

Nội dung câu lệnh SQL tạo bảng **Dim_Date** như sau:

```
CREATE TABLE [Dim_Date] (

    [DateID] int identity(1,1) constraint PK_date primary key,

    [OrderDate] datetime,

    [ShipDate] datetime,

    [OrderMonth] int,

    [OrderQuarter] varchar(10),

    [OrderYear] int,

    [ShipMonth] int,
```

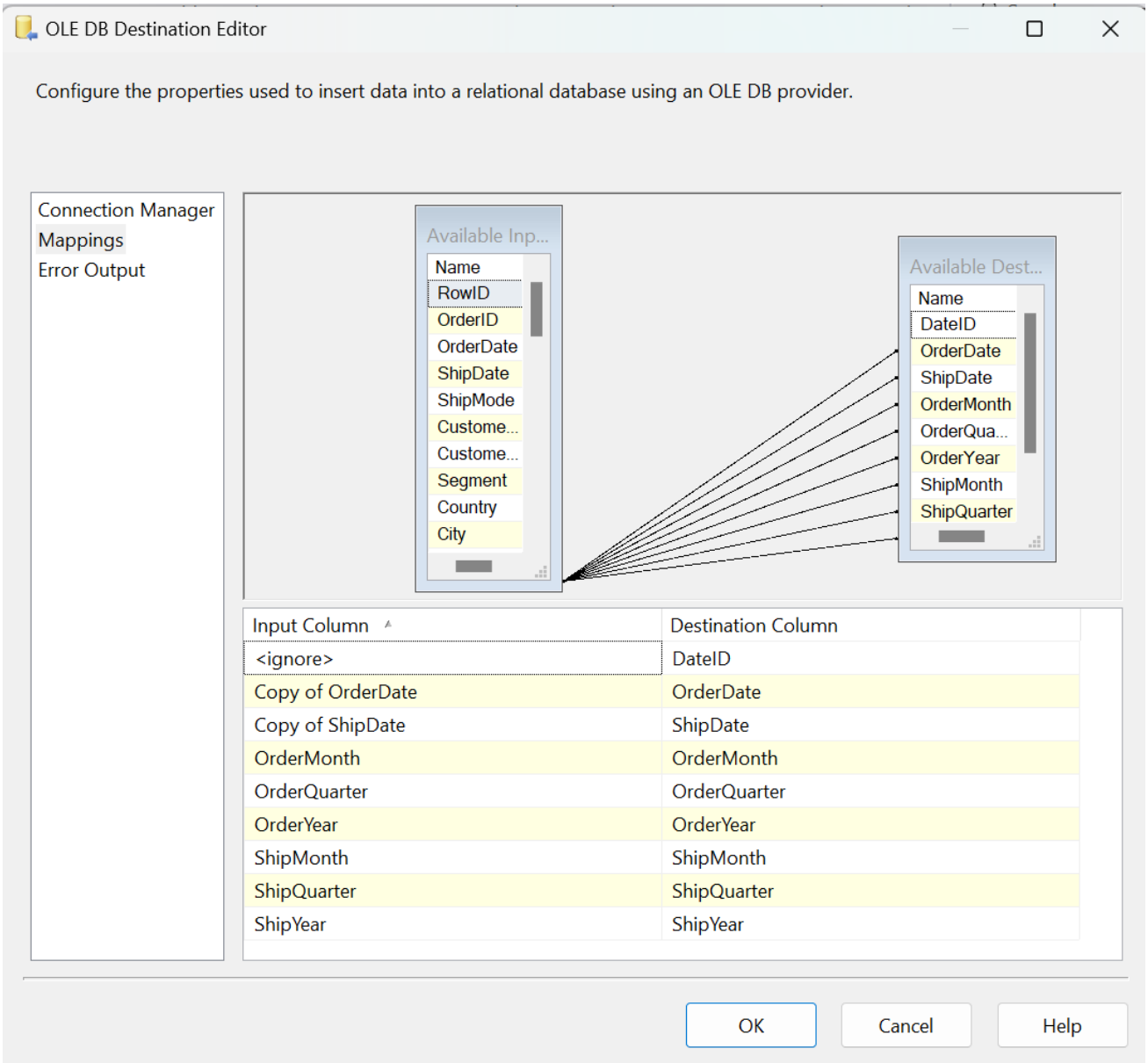
```
    [ShipQuarter] varchar(10),

    [ShipYear] int

)
```



Hình 2.33 Tạo mới bảng bằng hộp thoại Create Table

Sau khi đã dán câu lệnh vào hộp thoại **Create Table** thì nhấn **OK** để hoàn thành tạo bảng.

Bước 8: Tiếp đến, ta nhấn vào mục **Mappings** để tiến hành ánh xạ các dữ liệu đã xử lý đến các cột của bảng **Dim_Date** vừa mới tạo.

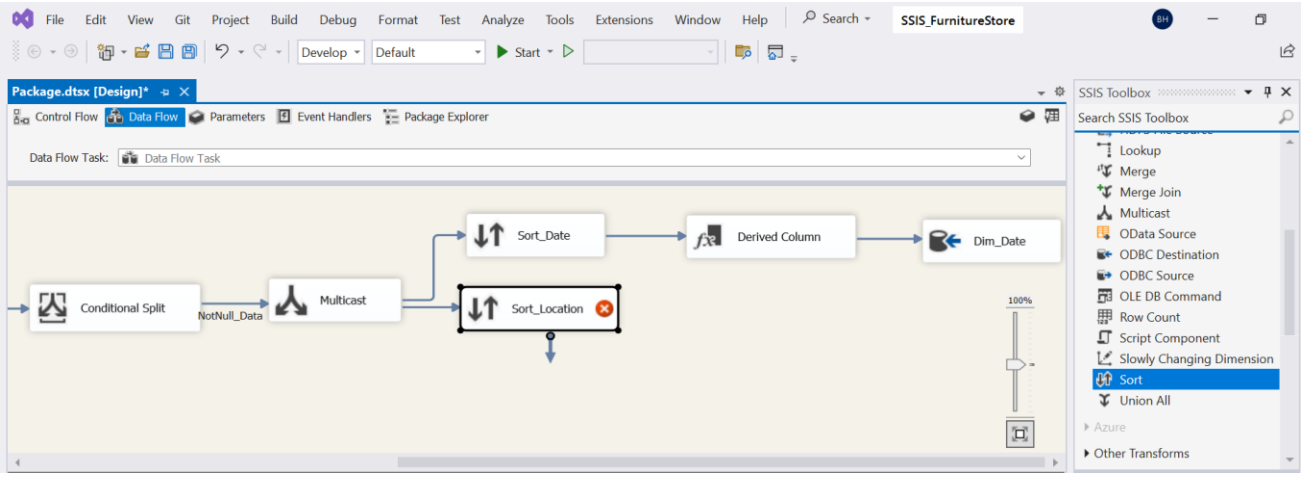


Hình 2.34 Ánh xạ các cột của dữ liệu nguồn đến các cột của dữ liệu đích

Nhấn **OK** để hoàn tất thiết lập cho bảng Dim_Date.

Bảng Dim_Location

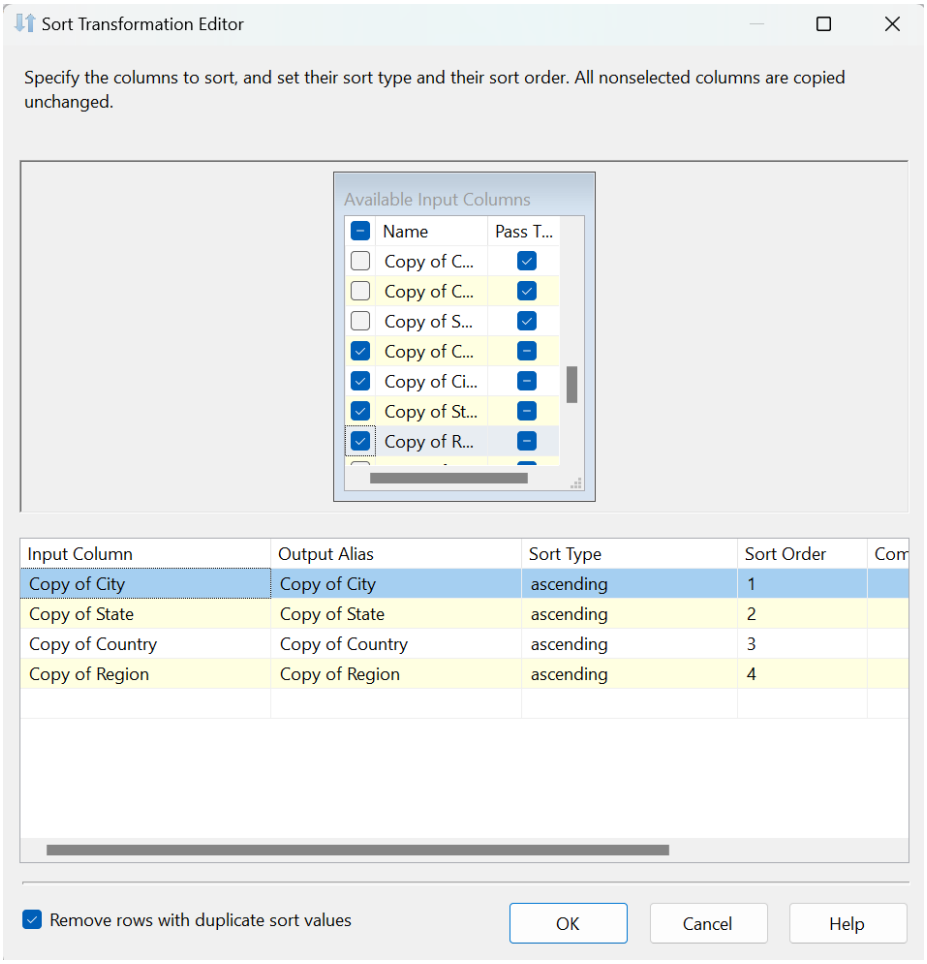
Bước 1: Tạo mới thành phần **Sort** của SSIS Toolbox và đặt tên là **Sort_Location** để lấy ra các cột dữ liệu cần thiết cho **Dim_Location**.



Hình 2.35 Tạo mới thành phần Sort để lấy những cột cần thiết

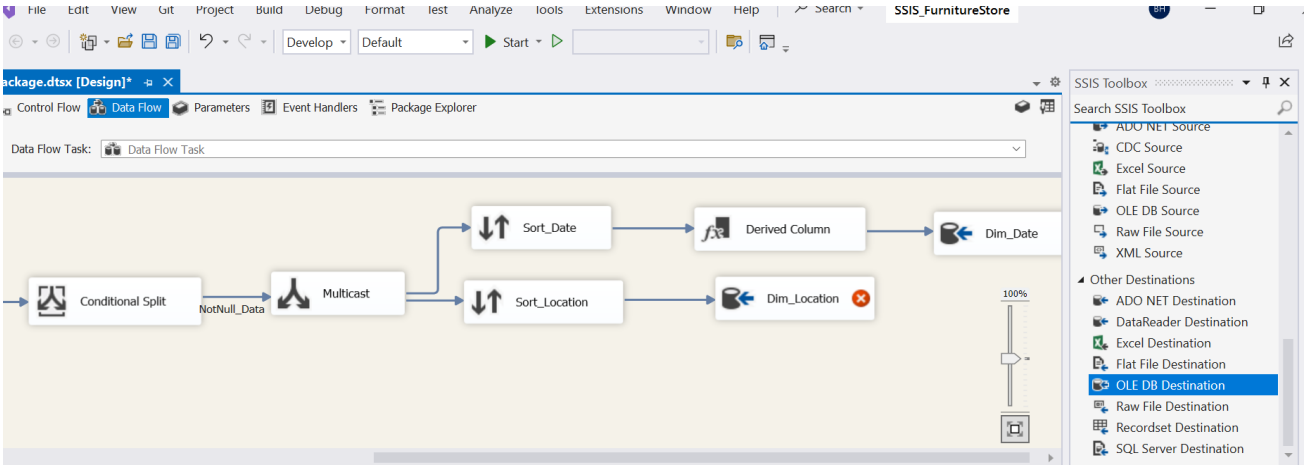
Bước 2: Nhấn đúp **Sort_Location** để mở hộp thoại **Edit**. Ở đây, ta sẽ chọn lần lượt các cột **City, State, Country,** và **Region** làm các cột dữ liệu để đưa vào bảng **Dim_Location.**

Sau đó, ta chọn **Remove rows with duplicate sort values** để xoá đi các dòng dữ liệu trùng nhau và chọn **OK**.



Hình 2.36 Chọn dữ liệu cần thiết cho Dim_Location và xoá dòng trùng

Bước 3: Tạo một **OLE DB Destination** để đưa dữ liệu gốc đã qua xử lý và sắp xếp vào trong kho dữ liệu **FurnitureStoreDW** và đưa vào bảng **Dim_Location**. Đặt tên cho thành phần **OLE DB Destination** này là **Dim_Location.**

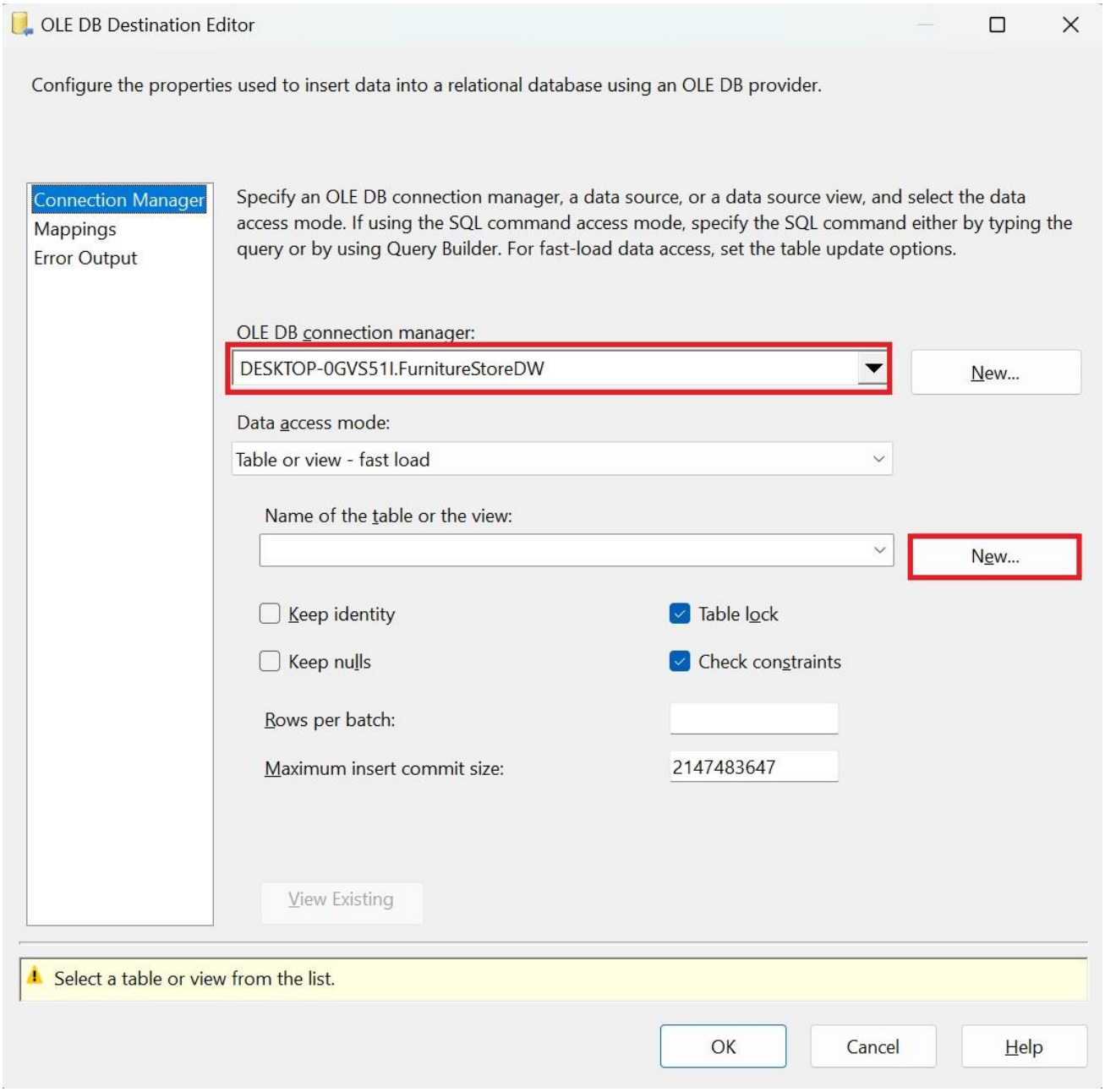


Hình 2.37 Tạo thành phần OLE DB Destination

Bước 4: Tiếp đến, chọn **Connection** đến kho dữ liệu đã được tạo và chọn **New…** đã tạo bảng **Dim_Location**.

Đây là câu lệnh SQL để tạo bảng **Dim_Location**:

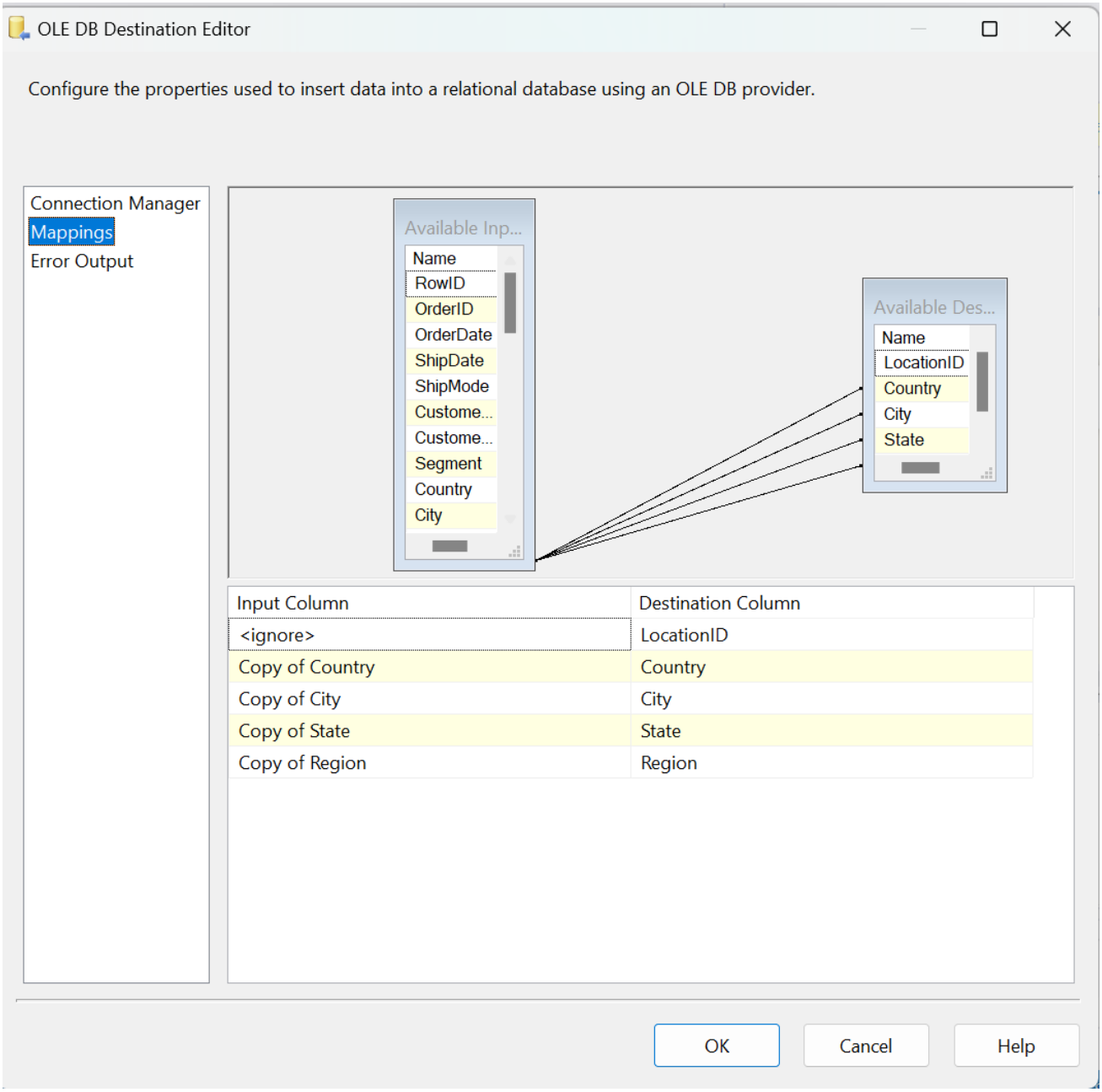
```
CREATE TABLE [Dim_Location] (

    [LocationID] int identity(1,1) constraint PK_location primary key,

    [Country] nvarchar(200),

    [City] nvarchar(200),

    [State] nvarchar(200),

    [Region] nvarchar(200)

)
```



Hình 2.38 Chọn connection và New… để tạo bảng Dim_Location

Sau khi đã dán câu lệnh vào hộp thoại **Create Tabl**e thì nhấn **OK** để hoàn thành tạo bảng.

Bước 5: Tiếp đến, ta nhấn vào mục **Mappings** để tiến hành ánh xạ các dữ liệu đã xử lý đến các cột của bảng **Dim_Location** vừa mới tạo.

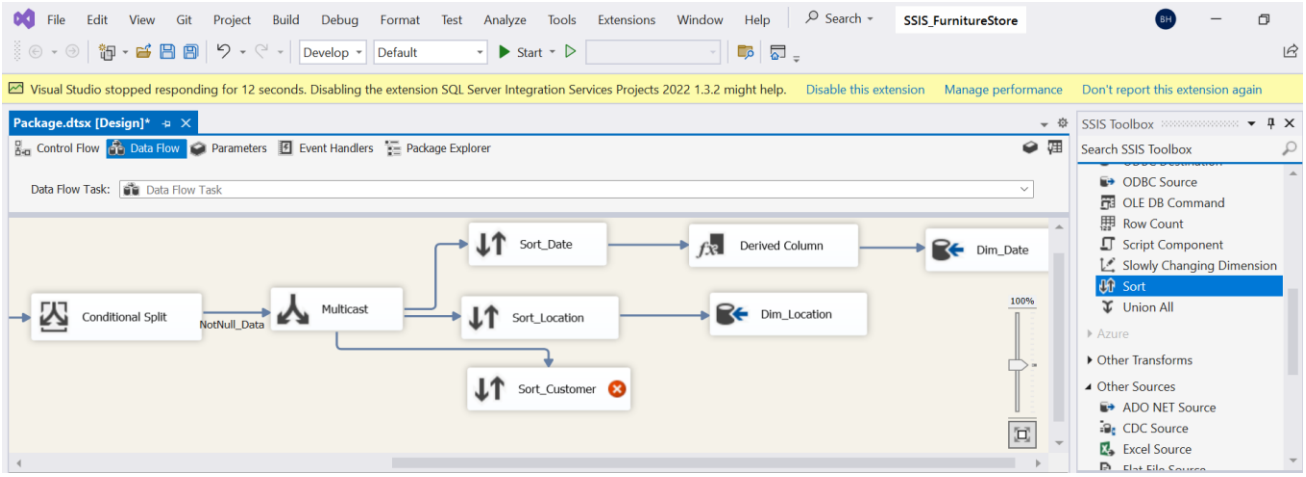


Hình 2.39 Ánh xạ các cột dữ liệu nguồn tới các cột bảng Dim_Location

Nhấn **OK** để hoàn tất thiết lập cho bảng Dim_Location.

Bảng Dim_Customer

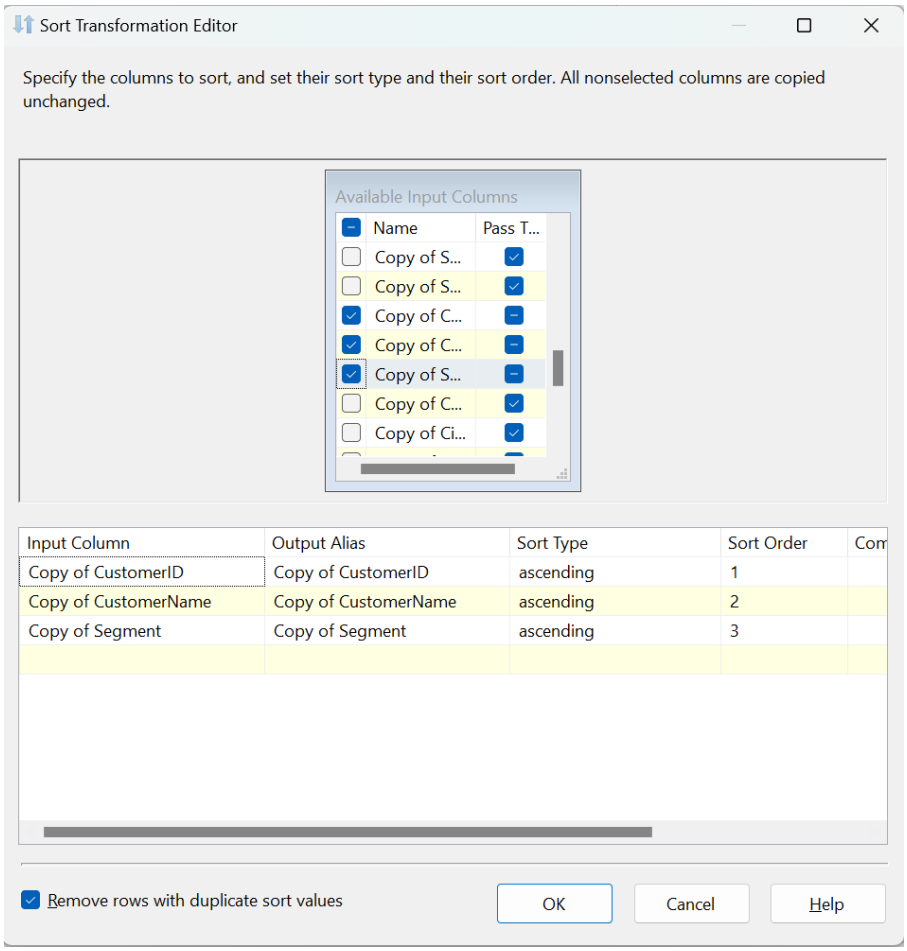
Bước 1: Tạo mới thành phần **Sort** của SSIS Toolbox và đặt tên là **Sort_Customer** để lấy ra các cột dữ liệu cần thiết cho **Dim_Customer**.



Hình 2.40 Tạo mới thành phần Sort để lấy những cột cần thiết

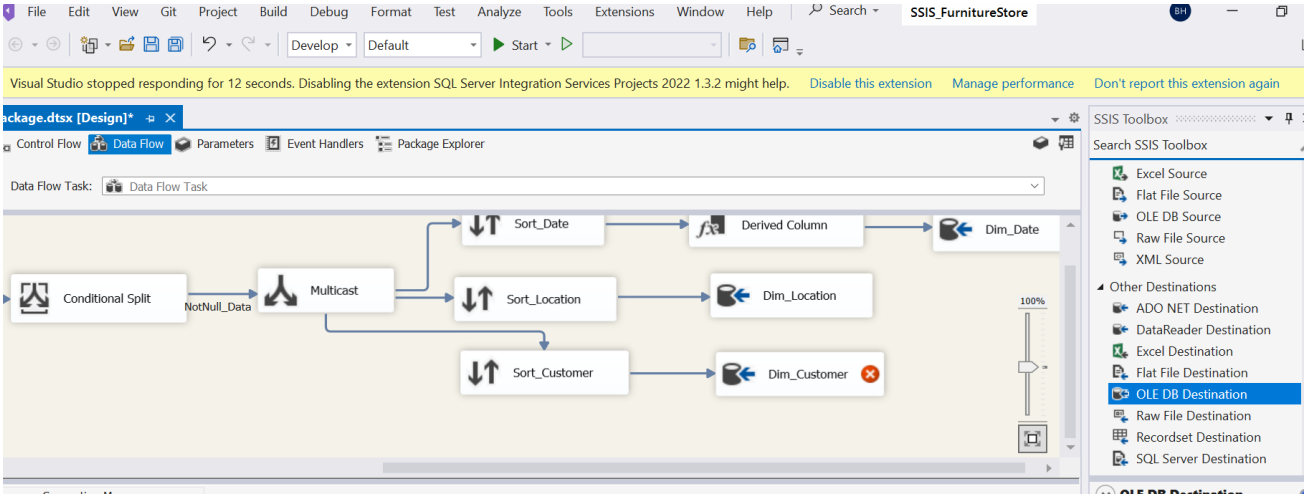
Bước 2: Nhấn đúp vào **Sort_Customer** để mở hộp thoại **Edit**. Ở đây, ta sẽ chọn lần lượt các cột **Copy of CustomerID, Copy of CustomerName, Copy of Segment** làm các cột dữ liệu để đưa vào bảng **Dim_Customer.**

Sau đó, ta chọn **Remove rows with duplicate sort values** để xoá đi các dòng dữ liệu trùng nhau và chọn **OK**.



Hình 2.41 Chọn dữ liệu cần thiết cho Dim_Customer và xoá dòng trùng

Bước 3: Tạo một **OLE DB Destination** để đưa dữ liệu gốc đã qua xử lý và sắp xếp vào trong kho dữ liệu **FurnitureStoreDW** và đưa vào bảng **Dim_Customer**. Đặt tên cho thành phần **OLE DB Destination** này là **Dim_Customer.**

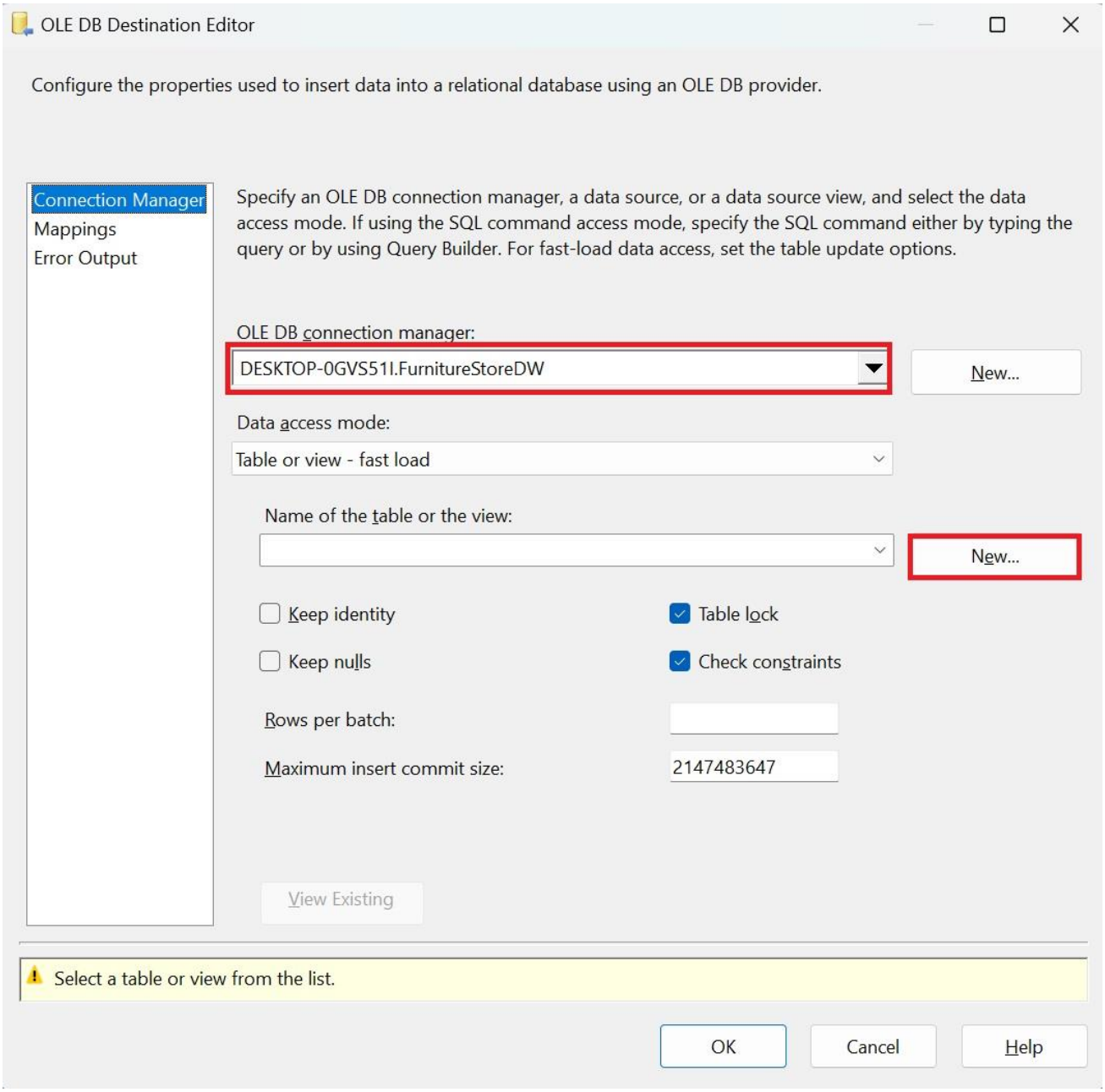


Hình 2.42 Tạo thành phần OLE DB Destination

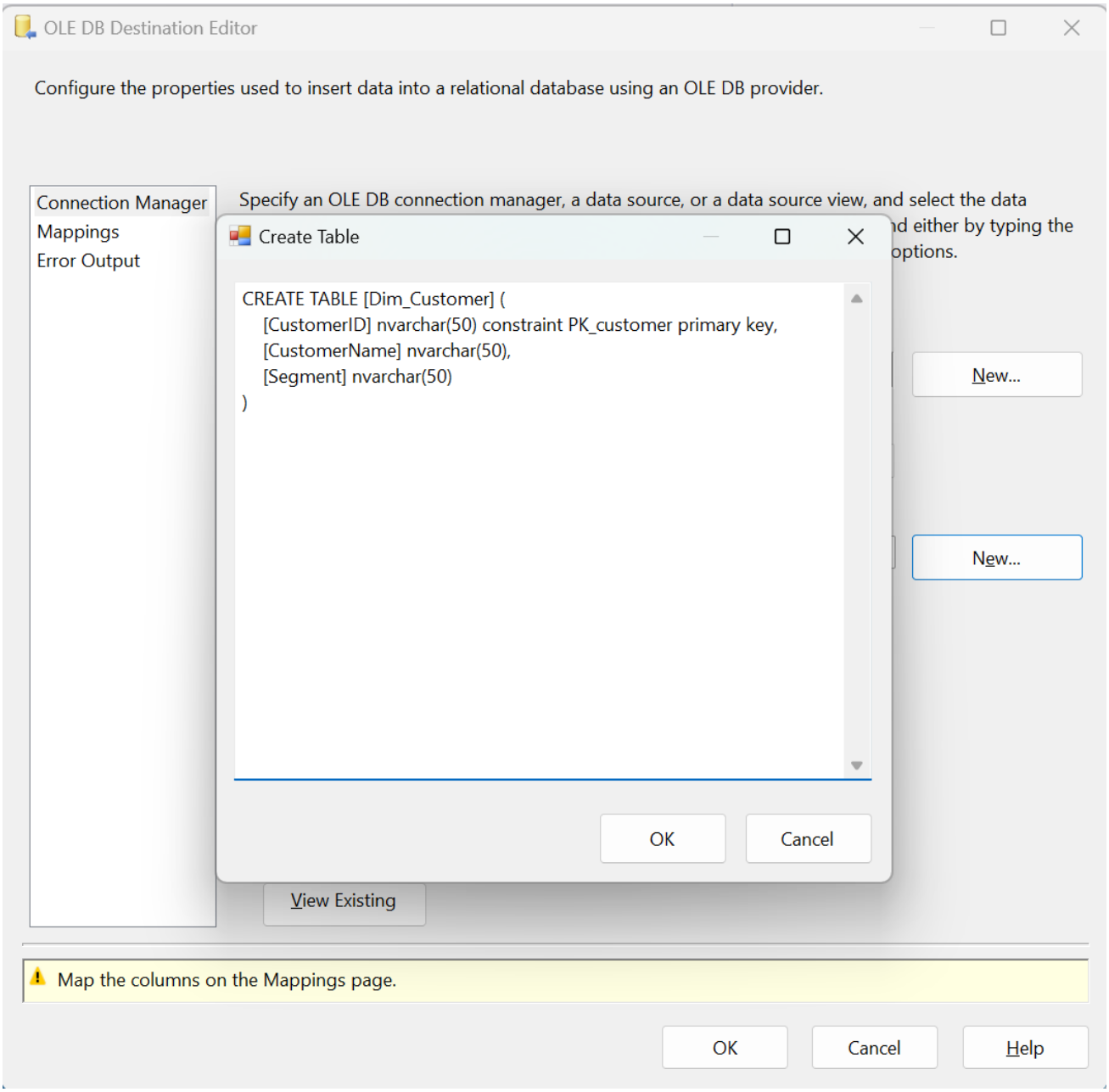
Bước 4: Tiếp đến, chọn **Connection** đến kho dữ liệu đã được tạo và chọn **New…** để tạo bảng **Dim_Customer**.

Đây là câu lệnh SQL để tạo bảng **Dim_Customer**:

```
CREATE TABLE [Dim_Customer] (

    [CustomerID] nvarchar(50) constraint PK_customer primary key,

    [CustomerName] nvarchar(50),

    [Segment] nvarchar(50)

)
```



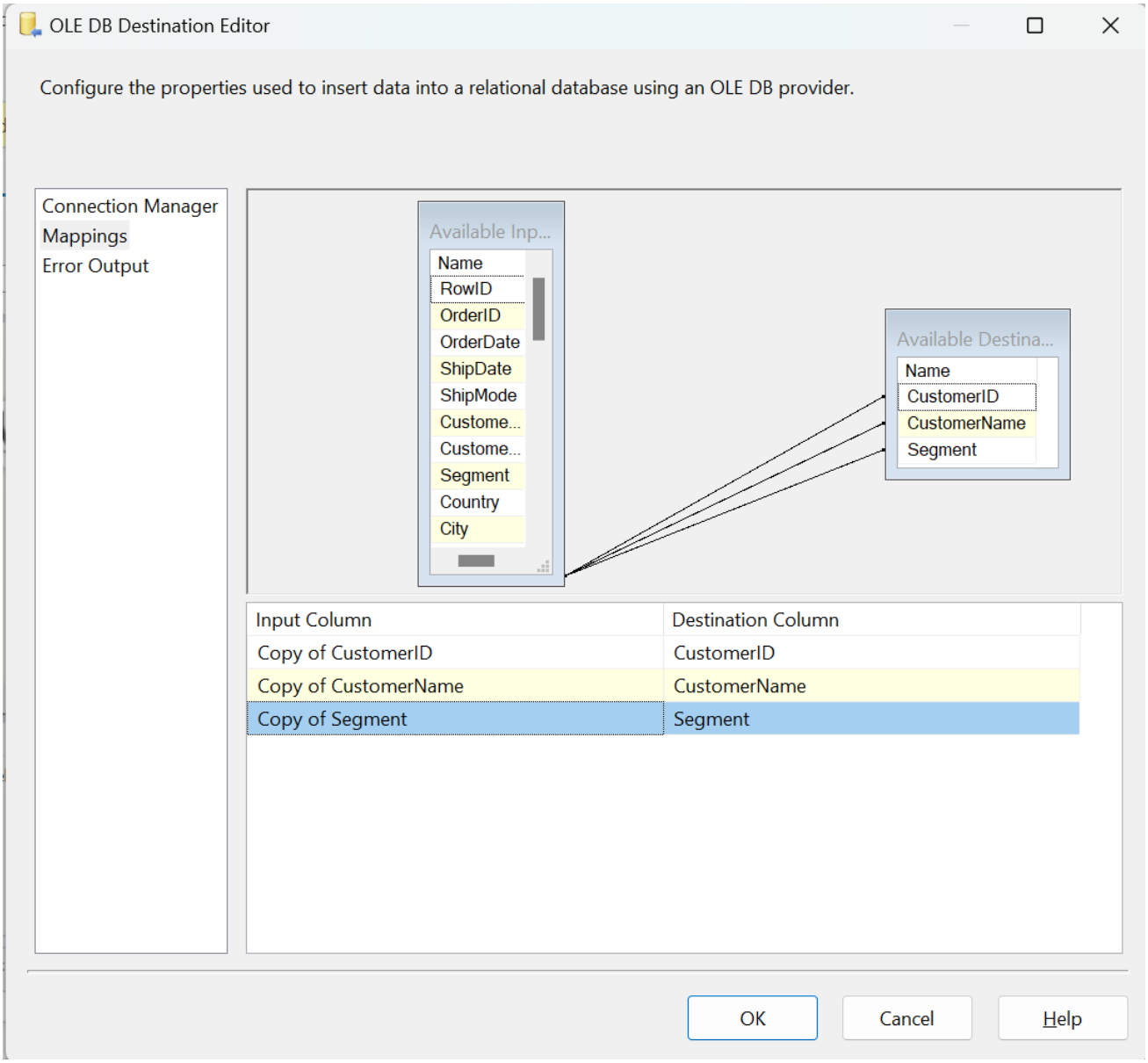
Hình 2.43 Chọn connection và New… để tạo bảng Dim_Customer



Hình 2.44 Dán câu lệnh SQL tạo bảng Dim_Customer vào hộp thoại Create Table

Sau khi đã dán câu lệnh vào hộp thoại **Create Tabl**e thì nhấn **OK** để hoàn thành tạo bảng.

Bước 5: Tiếp đến, ta nhấn vào mục **Mappings** để tiến hành ánh xạ các dữ liệu đã xử lý đến các cột của bảng **Dim_Customer** vừa mới tạo.

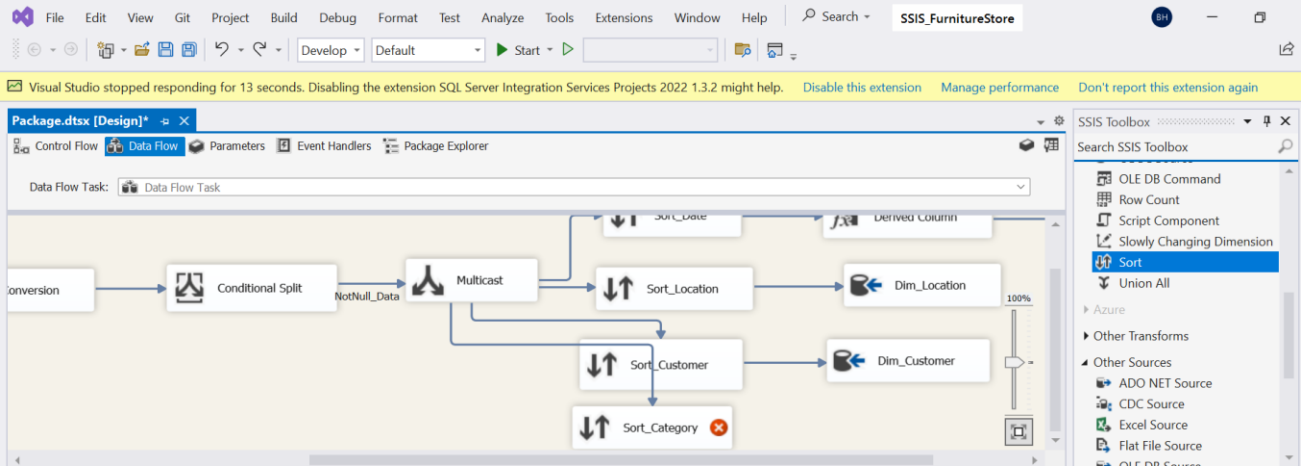


Hình 2.45 Ánh xạ các cột dữ liệu nguồn tới các cột bảng Dim_Customer

Nhấn **OK** để hoàn tất thiết lập cho bảng **Dim_Customer**.

Bảng Dim_Category

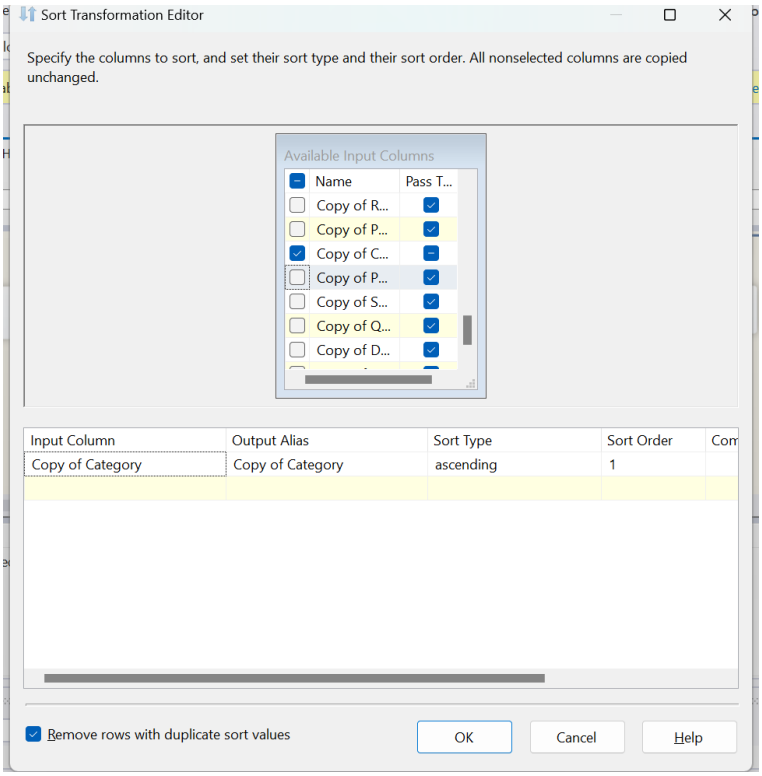
Bước 1: Tạo mới thành phần **Sort** của SSIS Toolbox và đặt tên là **Sort_Category** để lấy ra các cột dữ liệu cần thiết cho **Dim_Category**.



Hình 2.46 Tạo mới thành phần Sort để lấy những cột cần thiết

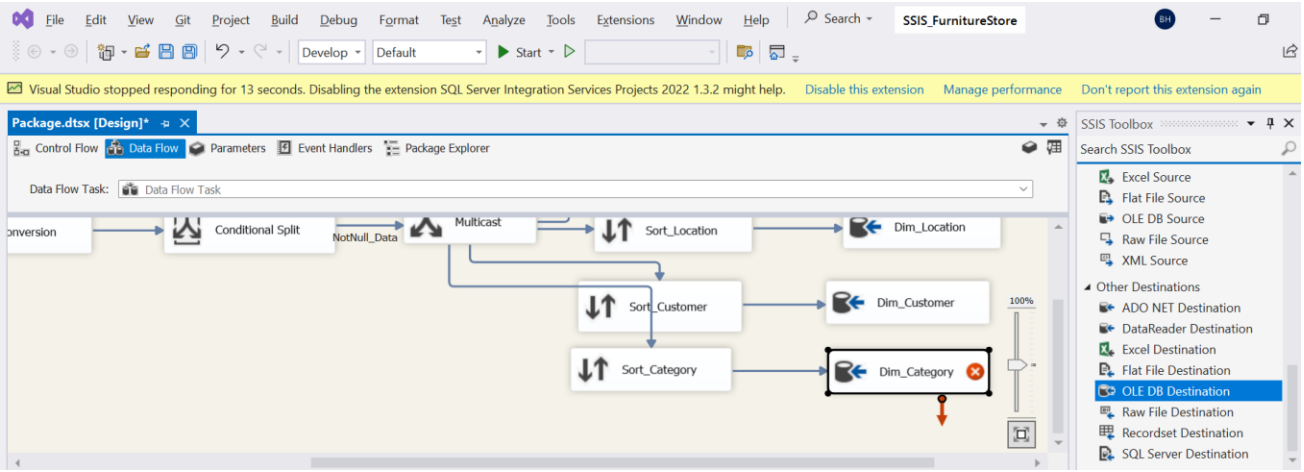
Bước 2: Nhấn đúp vào **Sort_Category** để mở hộp thoại **Edit**. Ở đây, ta sẽ chọn lần lượt các cột **Copy of Category** làm các cột dữ liệu để đưa vào bảng **Dim_Category.**

Sau đó, ta chọn **Remove rows with duplicate sort values** để xoá đi các dòng dữ liệu trùng nhau và chọn **OK**.



Hình 2.47 Chọn dữ liệu cần thiết cho Dim_Category và xoá dòng trùng

Bước 3: Tạo một **OLE DB Destination** để đưa dữ liệu gốc đã qua xử lý và sắp xếp vào trong kho dữ liệu **FurnitureStoreDW** và đưa vào bảng **Dim_Category**. Đặt tên cho thành phần **OLE DB Destination** này là **Dim_Category.**

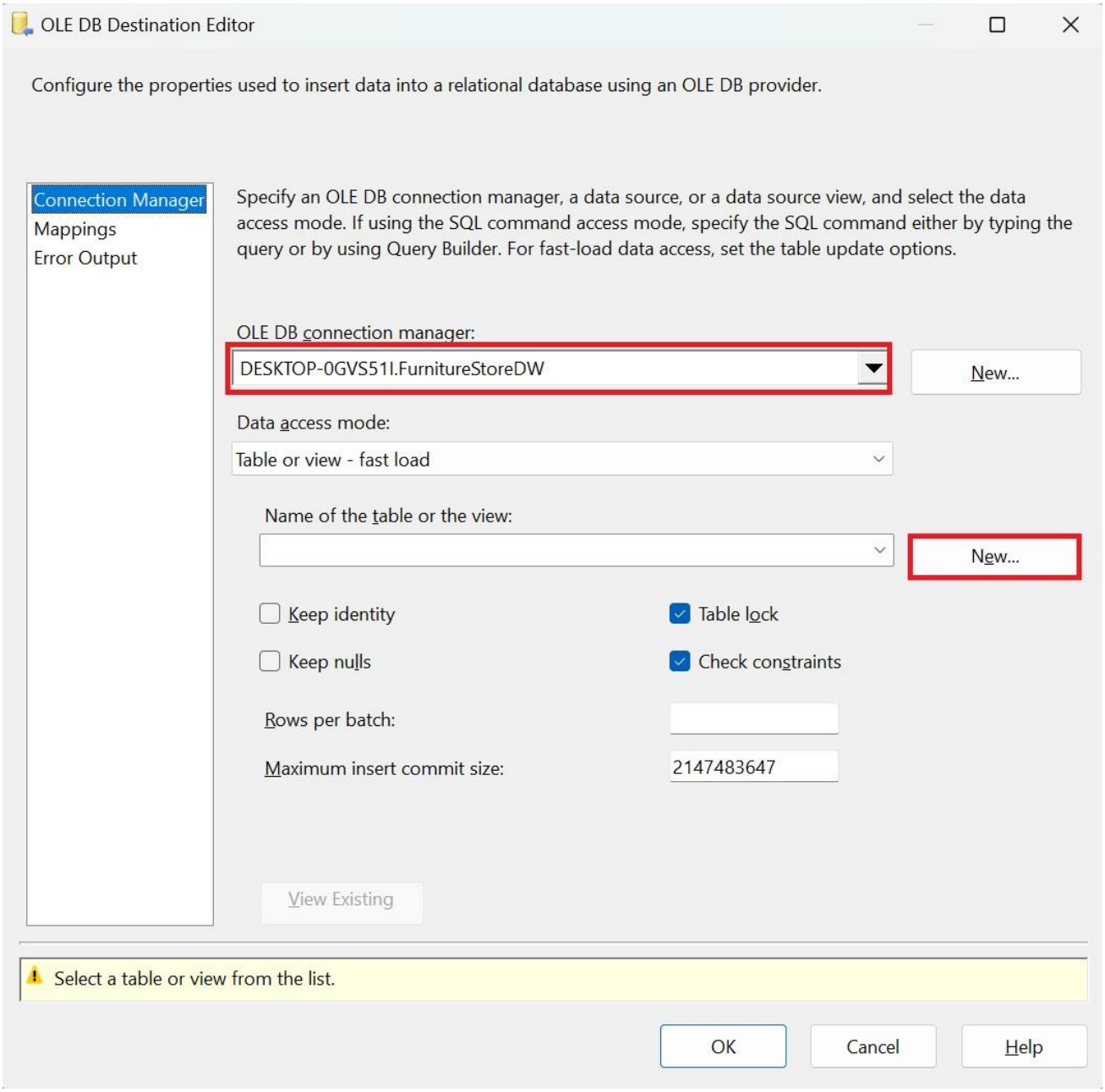


Hình 2.48 Tạo thành phần OLE DB Destination

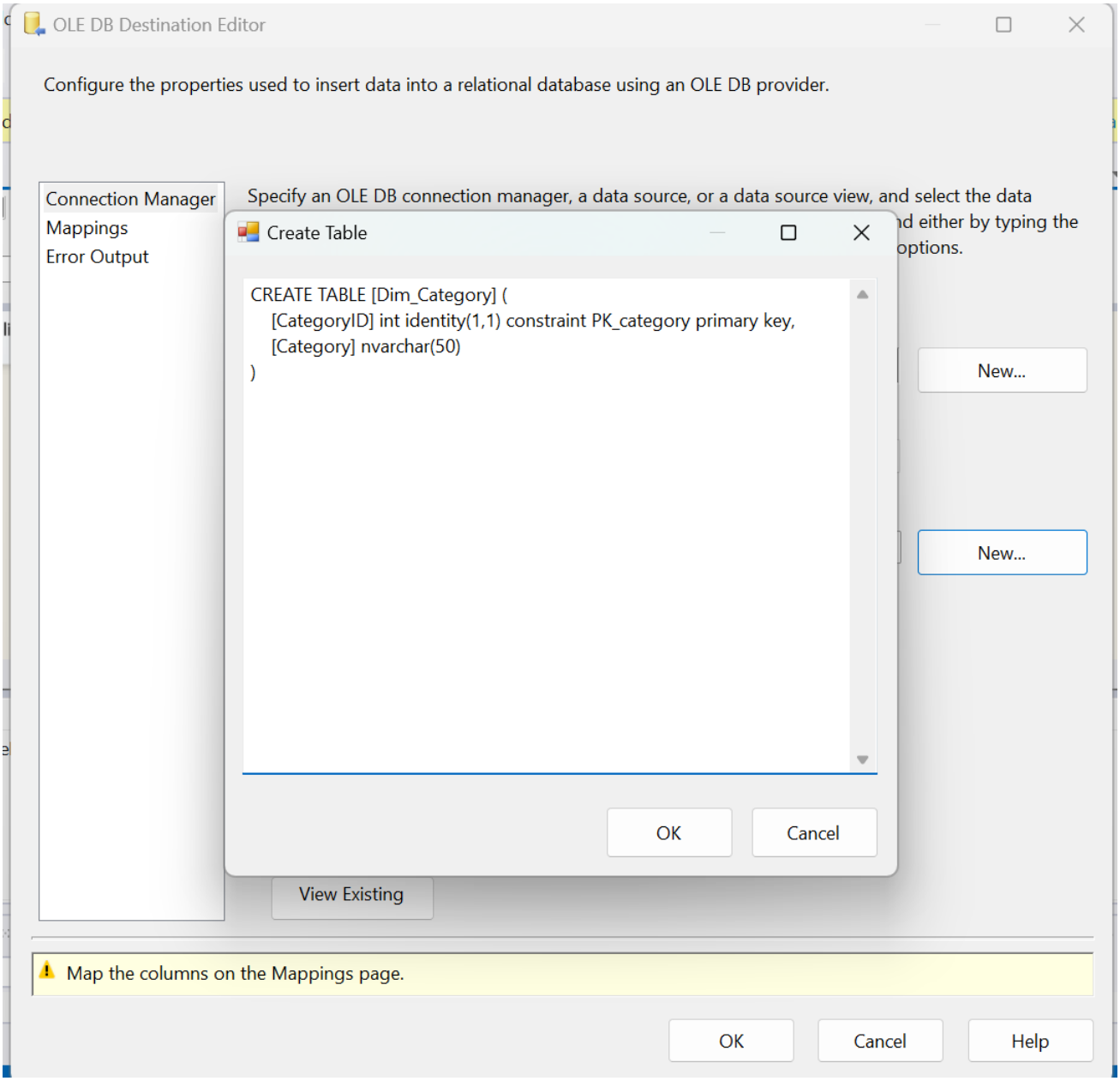
Bước 4: Tiếp đến, chọn **Connection** đến kho dữ liệu đã được tạo và chọn **New…** để tạo bảng **Dim_Category**.

Đây là câu lệnh SQL để tạo bảng **Dim_Category**:

```
CREATE TABLE [Dim_Category] (

    [CategoryID] int identity(1,1) constraint PK_category primary key,

    [Category] nvarchar(50)

)
```



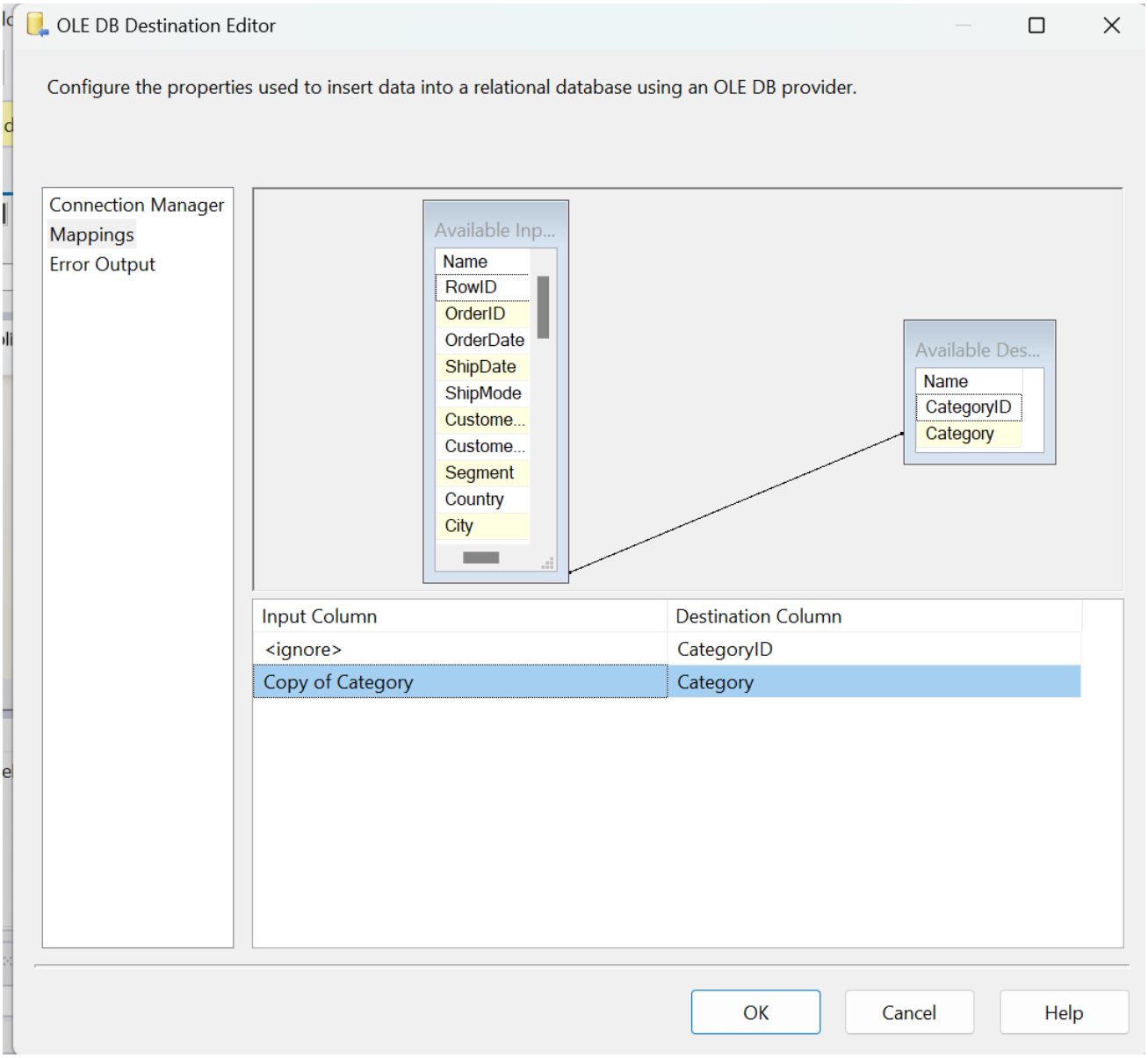
Hình 2.49 Chọn connection và New… để tạo bảng Dim_Category



Hình 2.50 Dán câu lệnh SQL tạo bảng Dim_Category vào hộp thoại Create Table

Sau khi đã dán câu lệnh vào hộp thoại **Create Tabl**e thì nhấn **OK** để hoàn thành tạo bảng.

Bước 5: Tiếp đến, ta nhấn vào mục **Mappings** để tiến hành ánh xạ các dữ liệu đã xử lý đến các cột của bảng **Dim_Category** vừa mới tạo.



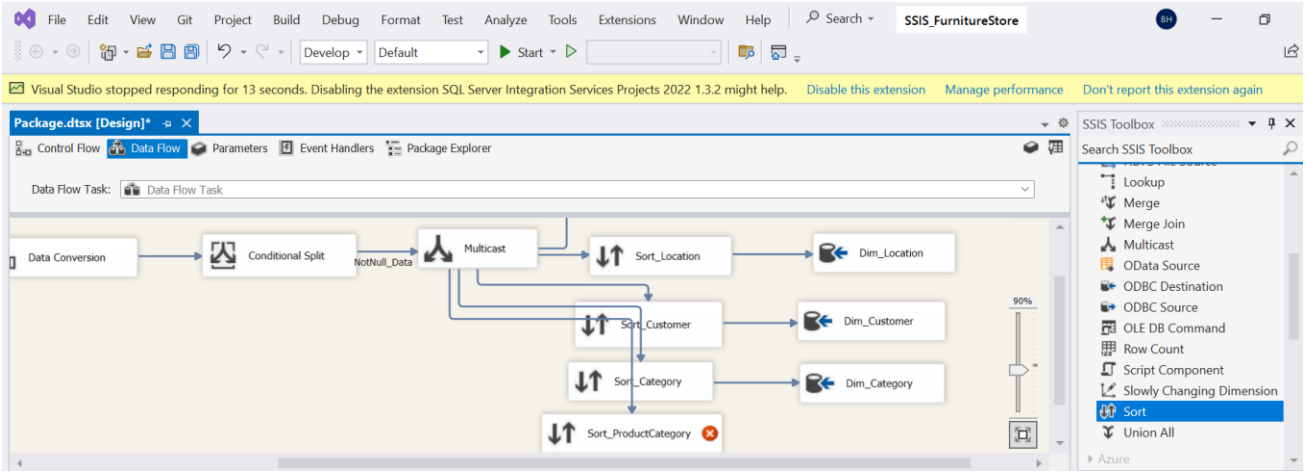
Hình 2.51 Ánh xạ các cột dữ liệu nguồn tới các cột bảng Dim_Category

Nhấn **OK** để hoàn tất thiết lập cho bảng **Dim_Category**.

Bảng Dim_ProductCategory

Bảng **Dim_ProductCategory** là bảng dùng để kết với bảng **Dim_Product** vì mô hình trong mô hình bông tuyết thì **Dim_Category** không có quan hệ trực tiếp với bảng **Fact** mà chỉ có quan hệ trực tiếp với **Dim_Product**, vì vậy **Dim_ProductCategory** sẽ được tách ra để kết với **Dim_Category** và sau đó sẽ được chuyển đổi lại thành **Dim_Product**.

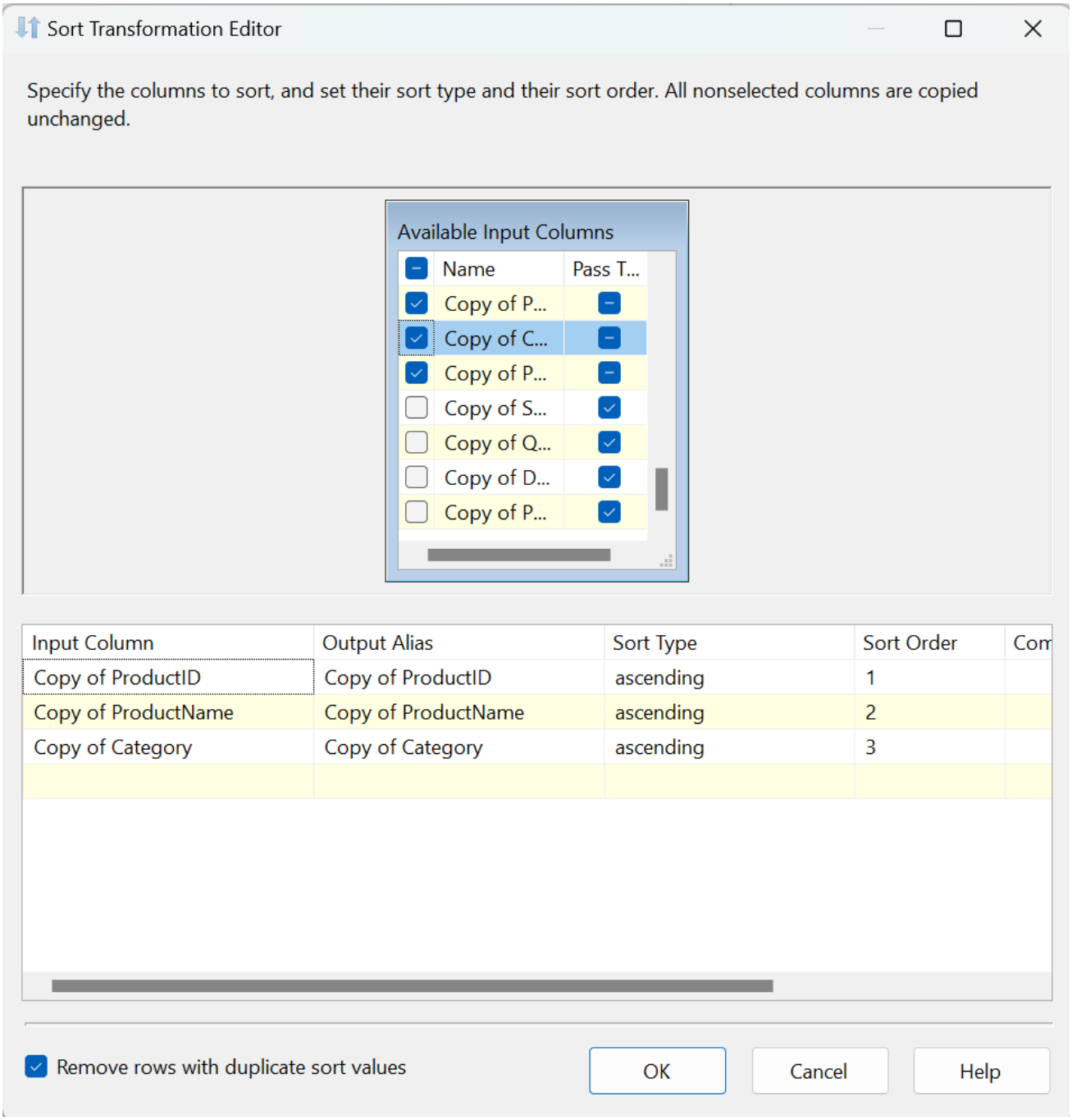
Bước 1: Tạo mới thành phần **Sort** của **SSIS Toolbox** và đặt tên là **Sort_ProductCategory** để lấy ra các cột dữ liệu cần thiết cho **Dim_ProductCategory**.



Hình 2.52 Tạo mới thành phần Sort để lấy những cột cần thiết

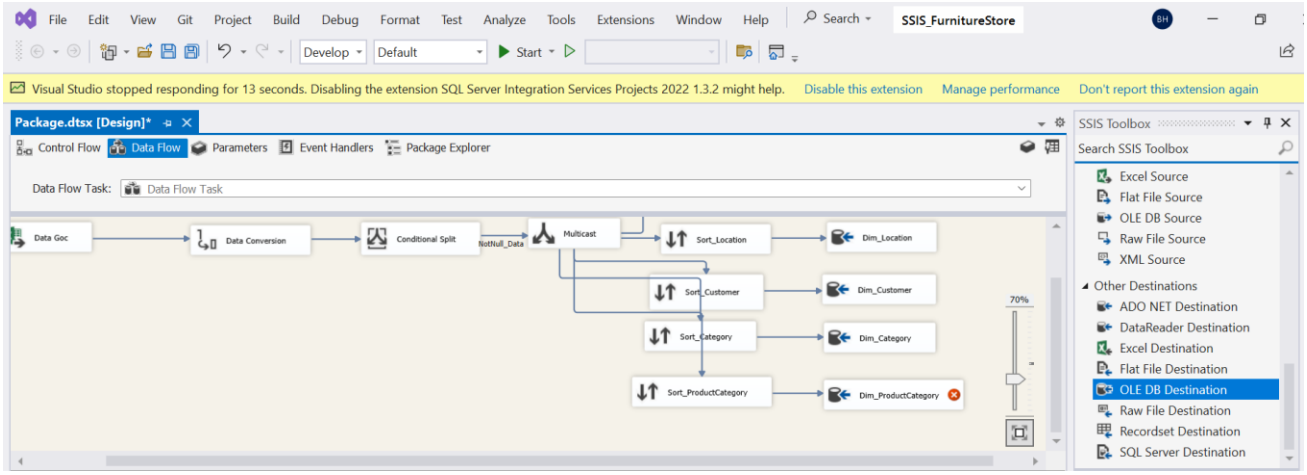
Bước 2: Nhấn đúp vào **Sort_ProductCategory** để mở hộp thoại **Edit**. Ở đây, ta sẽ chọn lần lượt các cột **Copy of ProductID, Copy of ProductName, Copy of Category** làm các cột dữ liệu để đưa vào bảng **Dim_ProductCategory.** Trong đó cột Copy of Category sẽ là cột dùng để kết với bảng **Dim_Category** để tạo ra bảng **Dim_Product**.

Sau đó, ta chọn **Remove rows with duplicate sort values** để xoá đi các dòng dữ liệu trùng nhau và chọn **OK**.



Hình 2.53 Chọn dữ liệu cần thiết cho Dim_Category và xoá dòng trùng

Bước 3: Tạo một **OLE DB Destination** để đưa dữ liệu gốc đã qua xử lý và sắp xếp vào trong kho dữ liệu **FurnitureStoreDW** và đưa vào bảng **Dim_ProductCategory**. Đặt tên cho thành phần **OLE DB Destination** này là **Dim_ProductCategory.**

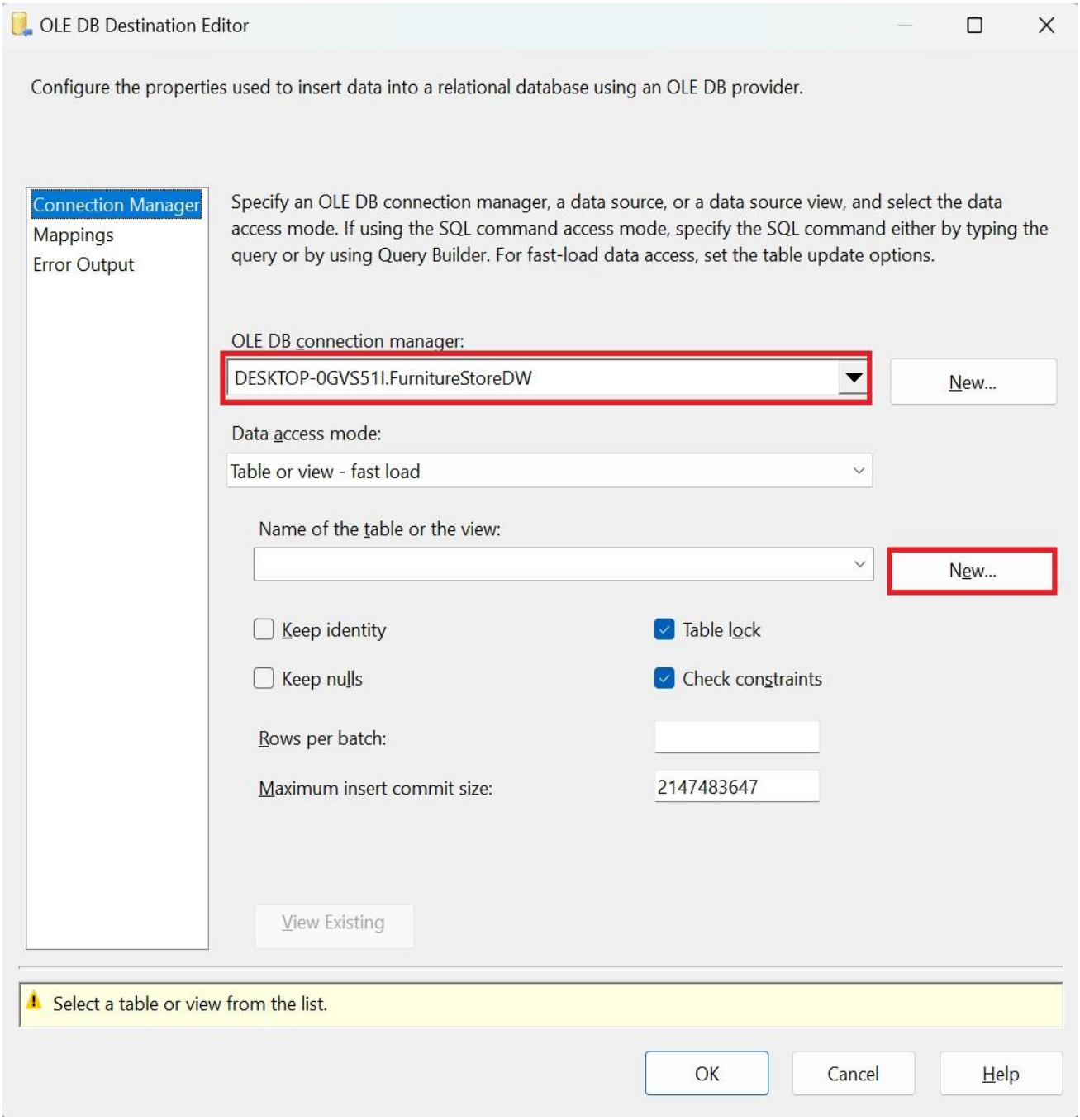


Hình 2.54 Tạo thành phần OLE DB Destination

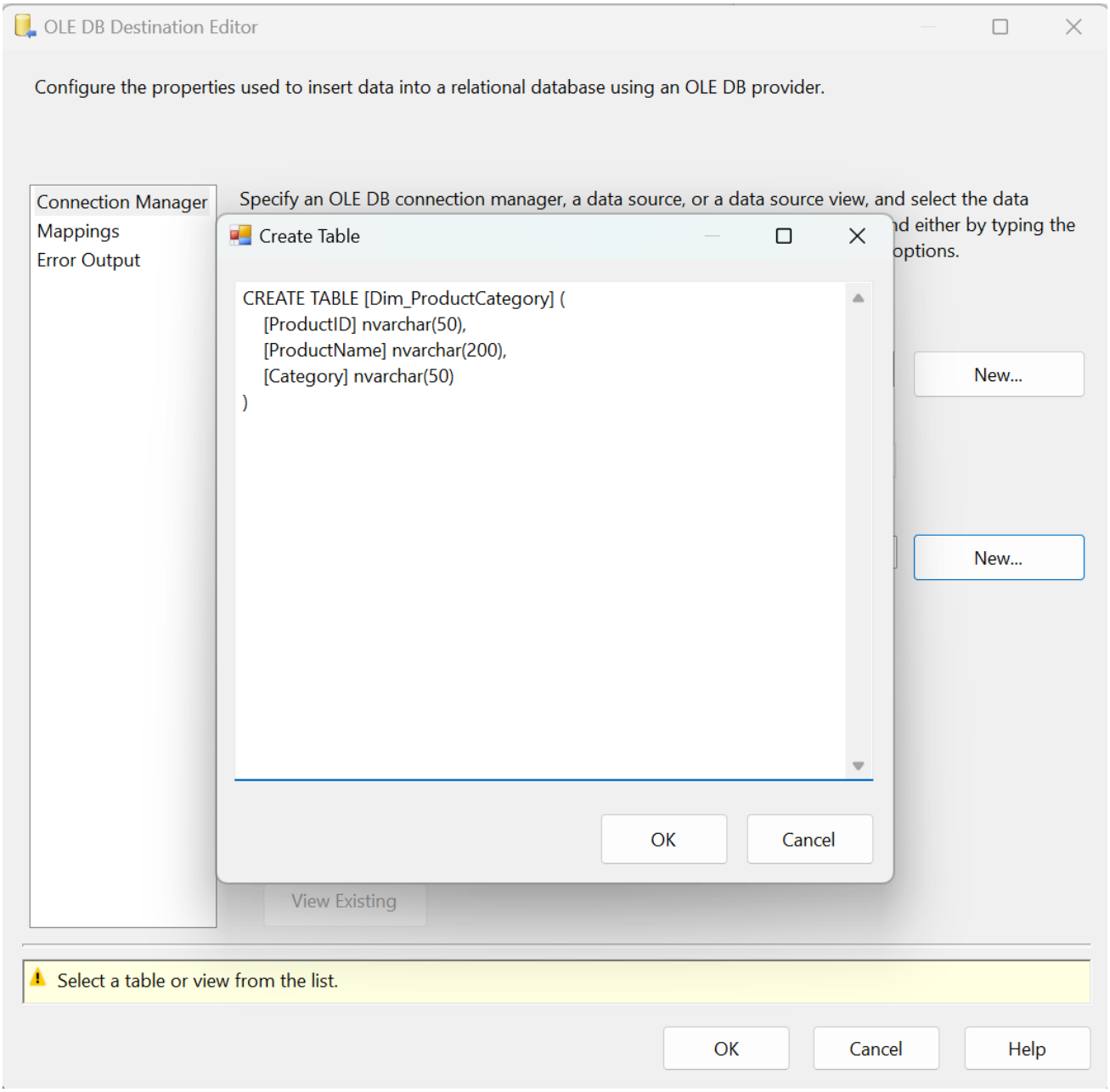
Bước 4: Tiếp đến, chọn **Connection** đến kho dữ liệu đã được tạo và chọn **New…** để tạo bảng **Dim_Category**.

Đây là câu lệnh SQL để tạo bảng **Dim_ProductCategory**:

```
CREATE TABLE [Dim_ProductCategory] (

    [ProductID] nvarchar(50),

    [ProductName] nvarchar(200),

    [Category] nvarchar(50)

)
```



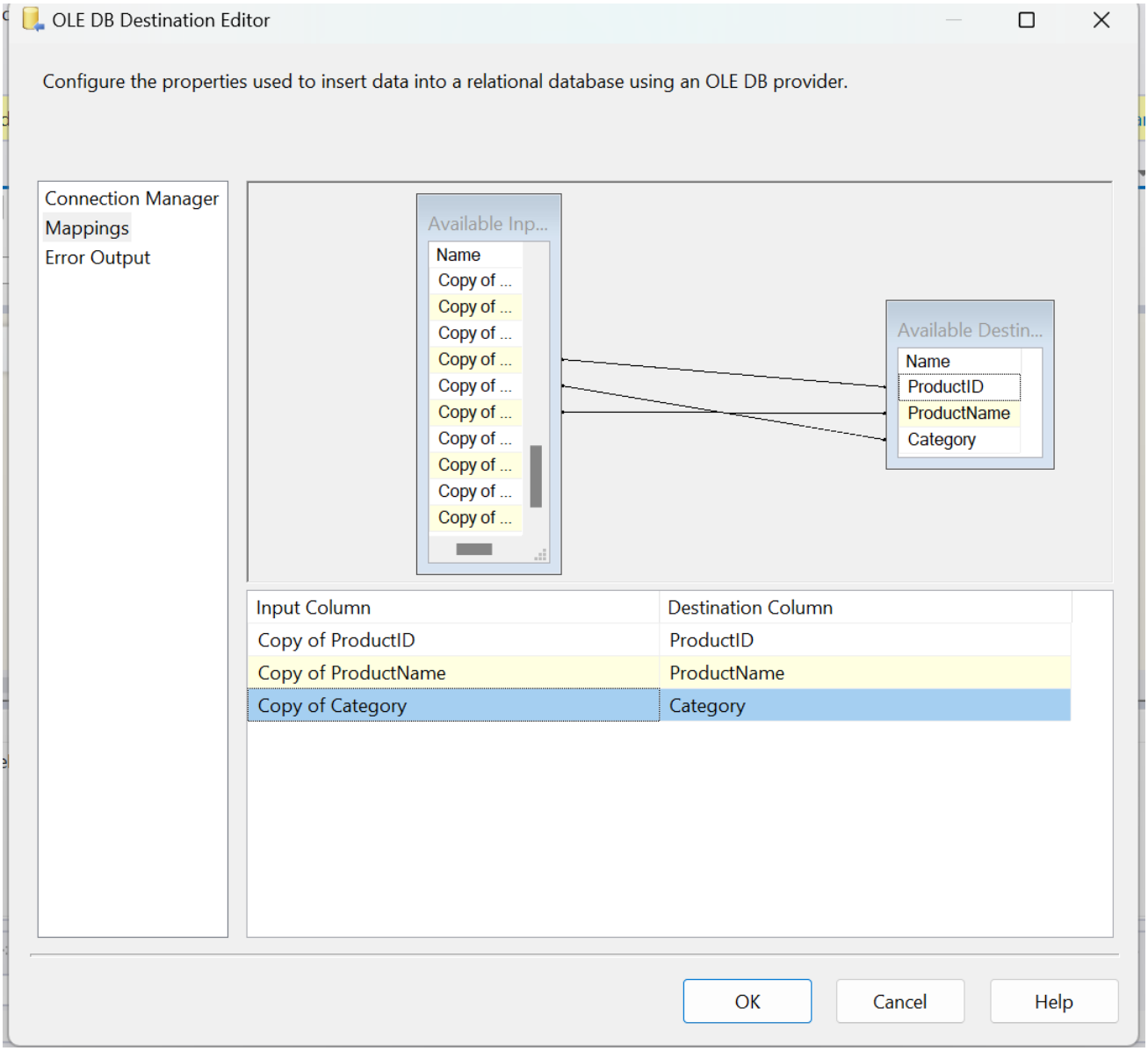
Hình 2.55 Chọn connection và New… để tạo bảng Dim_ProductCategory



Hình 2.56 Dán câu lệnh SQL tạo bảng Dim_ProductCategory vào hộp thoại Create Table

Sau khi đã dán câu lệnh vào hộp thoại **Create Tabl**e thì nhấn **OK** để hoàn thành tạo bảng.

Bước 5: Tiếp đến, ta nhấn vào mục **Mappings** để tiến hành ánh xạ các dữ liệu đã xử lý đến các cột của bảng **Dim_ProductCategory** vừa mới tạo.

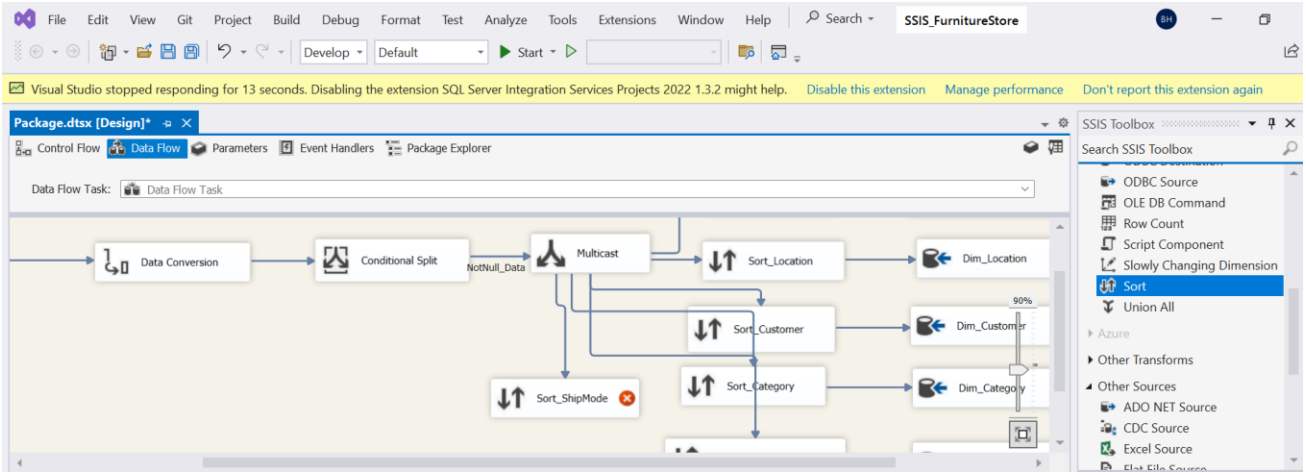


Hình 2.57 Ánh xạ các cột dữ liệu nguồn tới các cột bảng Dim_ProductCategory

Nhấn **OK** để hoàn tất thiết lập cho bảng **Dim_ProductCategory**.

Bảng Dim_ShipMode

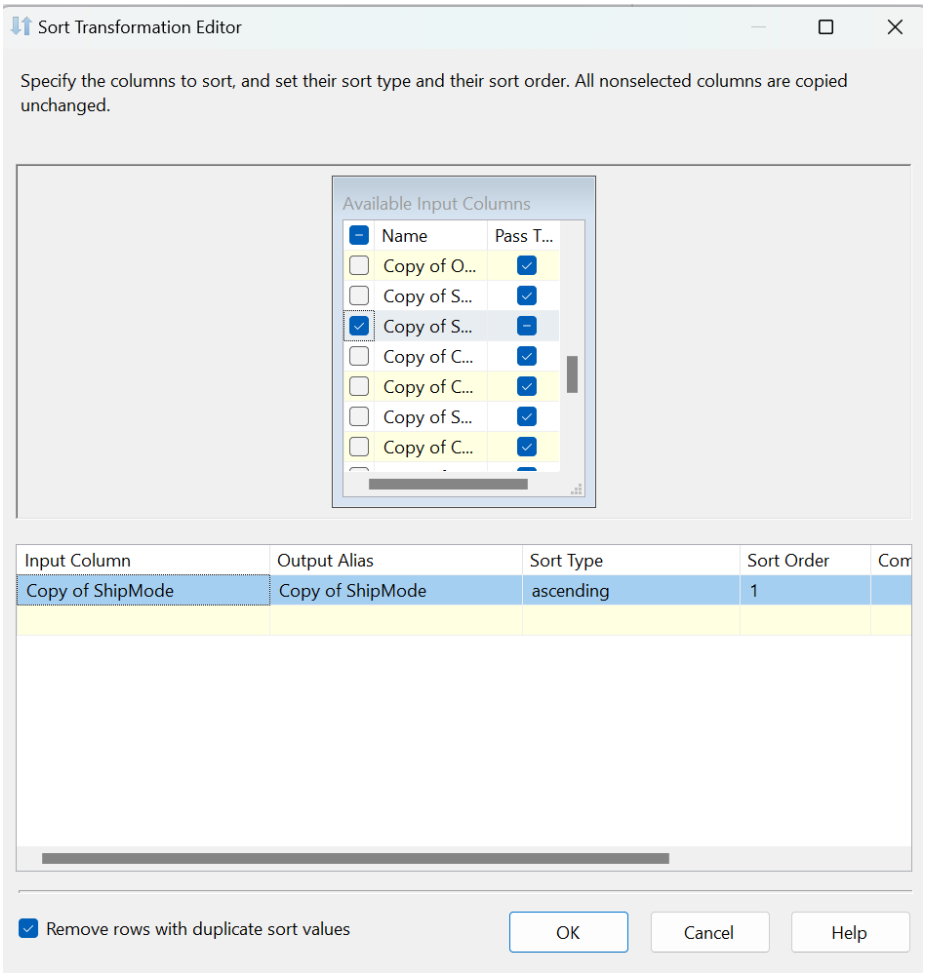
Bước 1: Tạo mới thành phần **Sort** của **SSIS Toolbox** và đặt tên là **Sort_ShipMode** để lấy ra các cột dữ liệu cần thiết cho **Dim_ShipMode**.



Hình 2.58 Tạo mới thành phần Sort để lấy những cột cần thiết

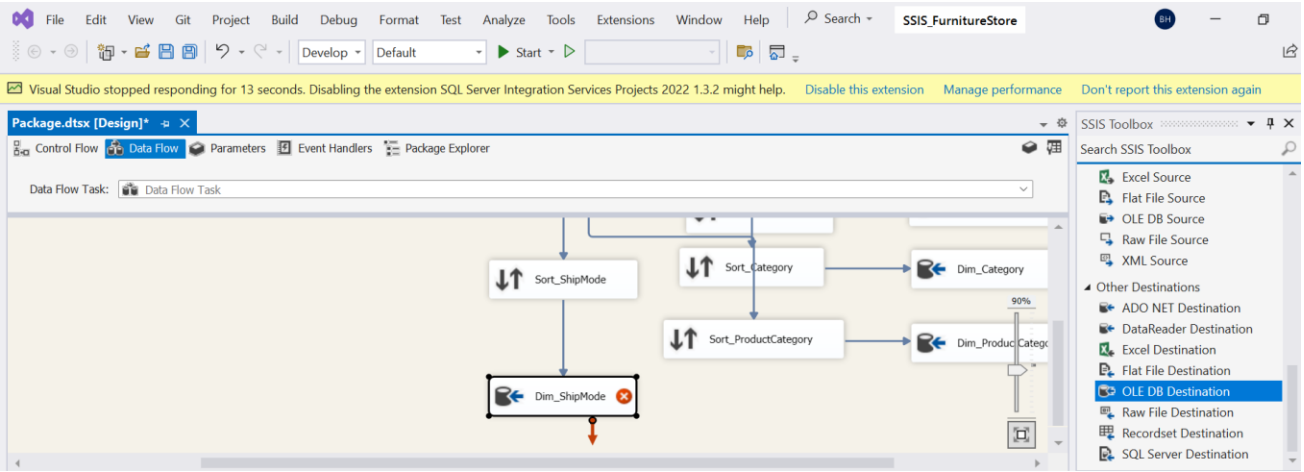
Bước 2: Nhấn đúp vào **Sort_ShipMode** để mở hộp thoại **Edit**. Ở đây, ta sẽ chọn cột **Copy of ShipMode** làm các cột dữ liệu để đưa vào bảng **Dim_ShipMode.**

Sau đó, ta chọn **Remove rows with duplicate sort values** để xoá đi các dòng dữ liệu trùng nhau và chọn **OK**.



Hình 2.59 Chọn dữ liệu cần thiết cho Dim_ShipMode và xoá dòng trùng

Bước 3: Tạo một **OLE DB Destination** để đưa dữ liệu gốc đã qua xử lý và sắp xếp vào trong kho dữ liệu **FurnitureStoreDW** và đưa vào bảng **Dim_ShipMode**. Đặt tên cho thành phần **OLE DB Destination** này là **Dim_ShipMode.**

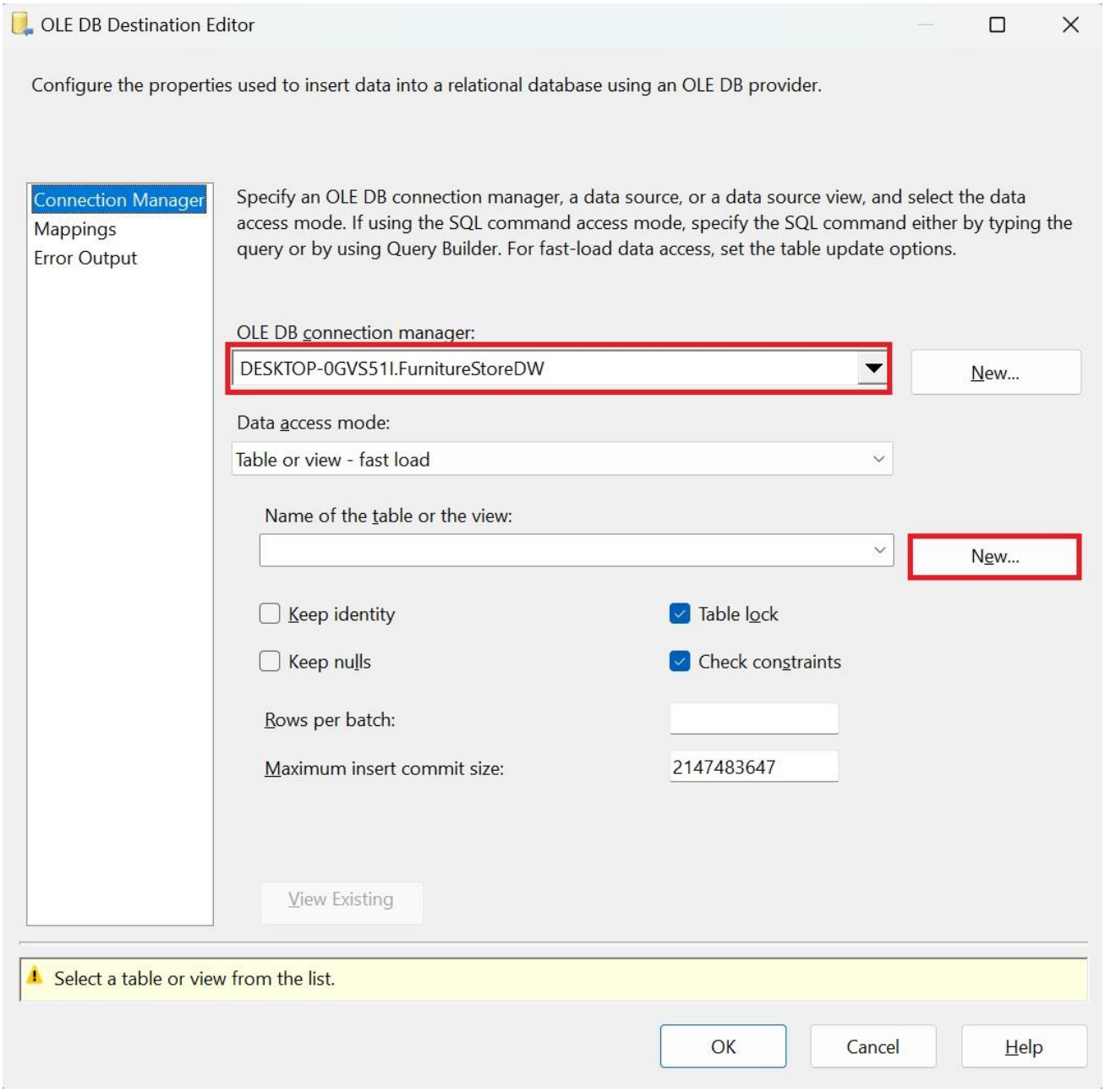


Hình 2.60 Tạo thành phần OLE DB Destination

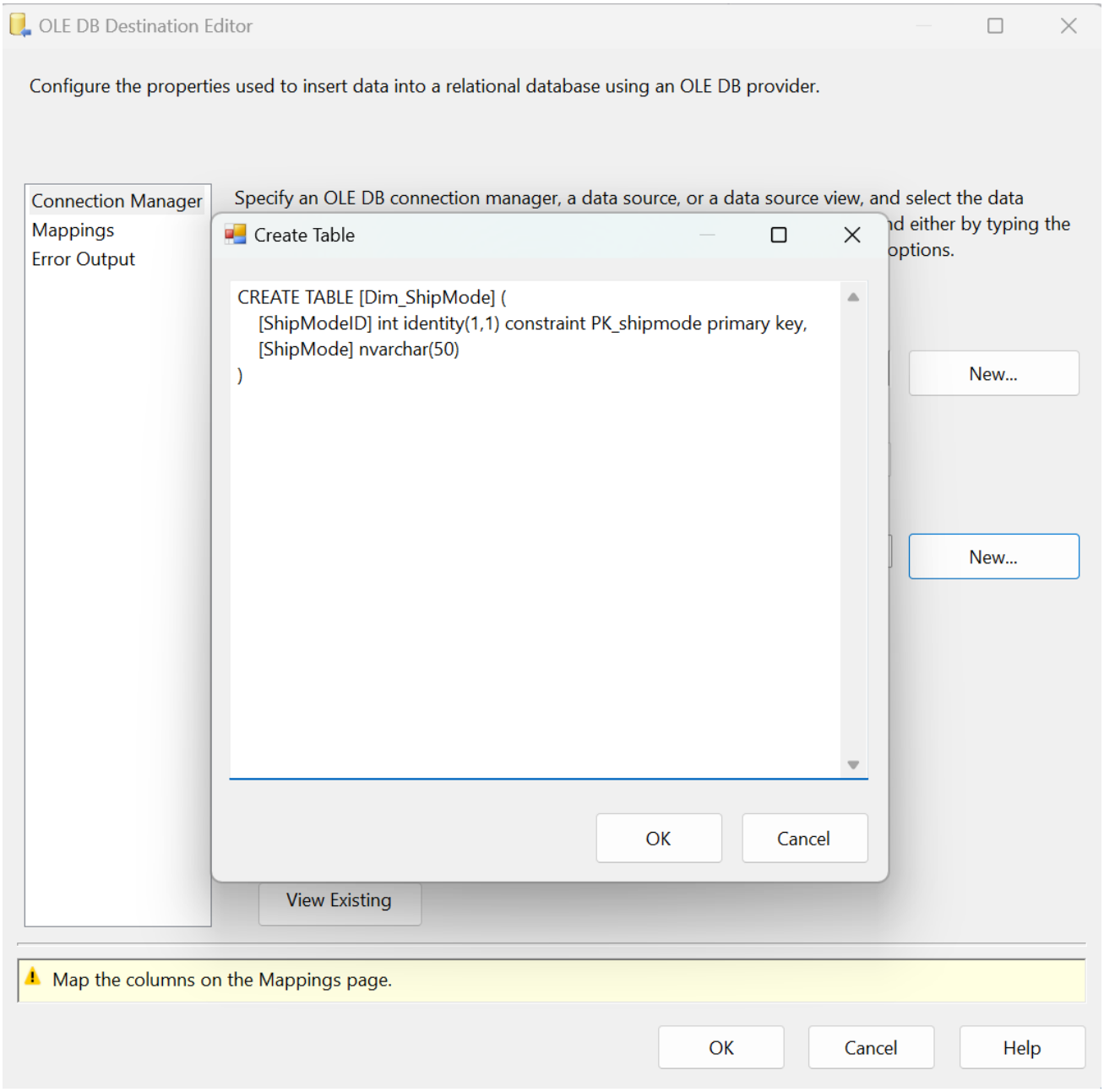
Bước 4: Tiếp đến, chọn **Connection** đến kho dữ liệu đã được tạo và chọn **New…** để tạo bảng **Dim_ShipMode**.

Đây là câu lệnh SQL để tạo bảng **Dim_ShipMode**:

```
CREATE TABLE [Dim_ShipMode] (

    [ShipModeID] int identity(1,1) constraint PK_shipmode primary key,

    [ShipMode] nvarchar(50)

)
```



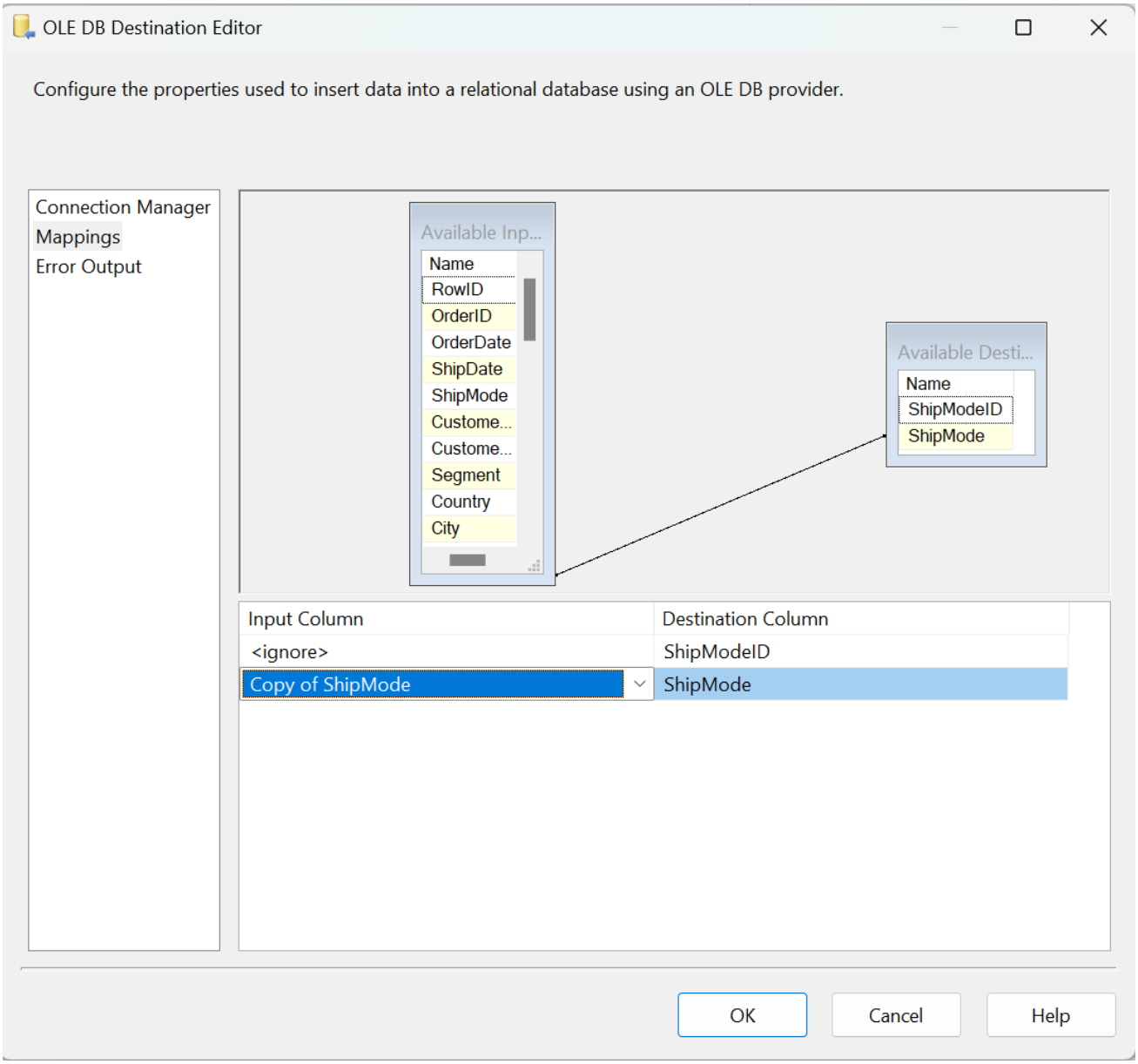
Hình 2.61 Chọn connection và New… để tạo bảng Dim_Shipmode



Hình 2.62 Dán câu lệnh SQL tạo bảng Dim_ShipMode vào hộp thoại Create Table

Sau khi đã dán câu lệnh vào hộp thoại **Create Tabl**e thì nhấn **OK** để hoàn thành tạo bảng.

Bước 5: Tiếp đến, ta nhấn vào mục **Mappings** để tiến hành ánh xạ các dữ liệu đã xử lý đến các cột của bảng **Dim_ShipMode** vừa mới tạo.



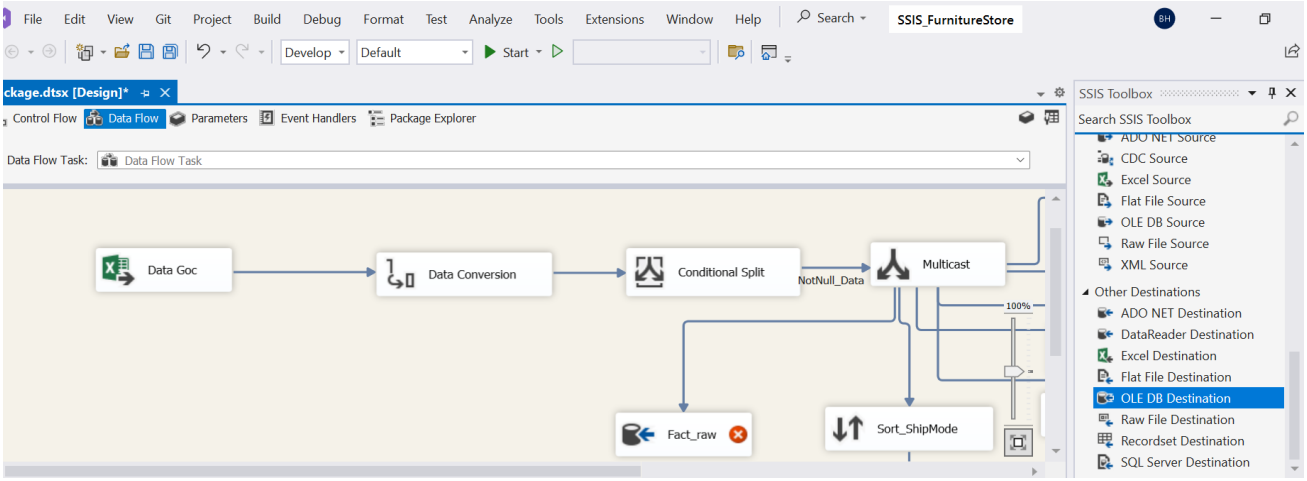
Hình 2.63 Ánh xạ các cột dữ liệu nguồn tới các cột bảng Dim_ShipMode

Nhấn **OK** để hoàn tất thiết lập cho bảng **Dim_ShipMode**.

Bảng Fact_raw

Bảng **Fact_raw** là bảng dữ liệu sơ khai ban đầu của bảng **Fact** và bảng sau đó sẽ được merge với các bảng Dimension để tạo ra bảng **Fact**, vì vậy các cột của bảng **Fact_raw** sẽ là tất cả các cột của dữ liệu gốc và chứa tất cả các dòng dữ liệu. Quá trình tạo ra bảng **Fact_raw** sẽ diễn ra như sau:

Bước 1: Tạo mới một thành phần **OLE DB Destination** và đặt tên cho thành phần này là **Fact_raw**. Sau đó, tiến hành kết nối Multicast đến với Fact_raw để tiến hành đổ dữ liệu đã được xử lý vào bảng **Fact_raw**.



Hình 2.64 Tạo mới một OLE DB Destination

Bước 2: Nhấn đúp vào Fact_raw để mở hộp thoại **Edit** → chọn **Connection** kết nối tới server và kho dữ liệu mà ta đã tạo → Chọn **New…** để tiến hành tạo bảng Fact_raw.

Nội dung của câu lệnh SQL tạo bảng Fact_raw như sau:

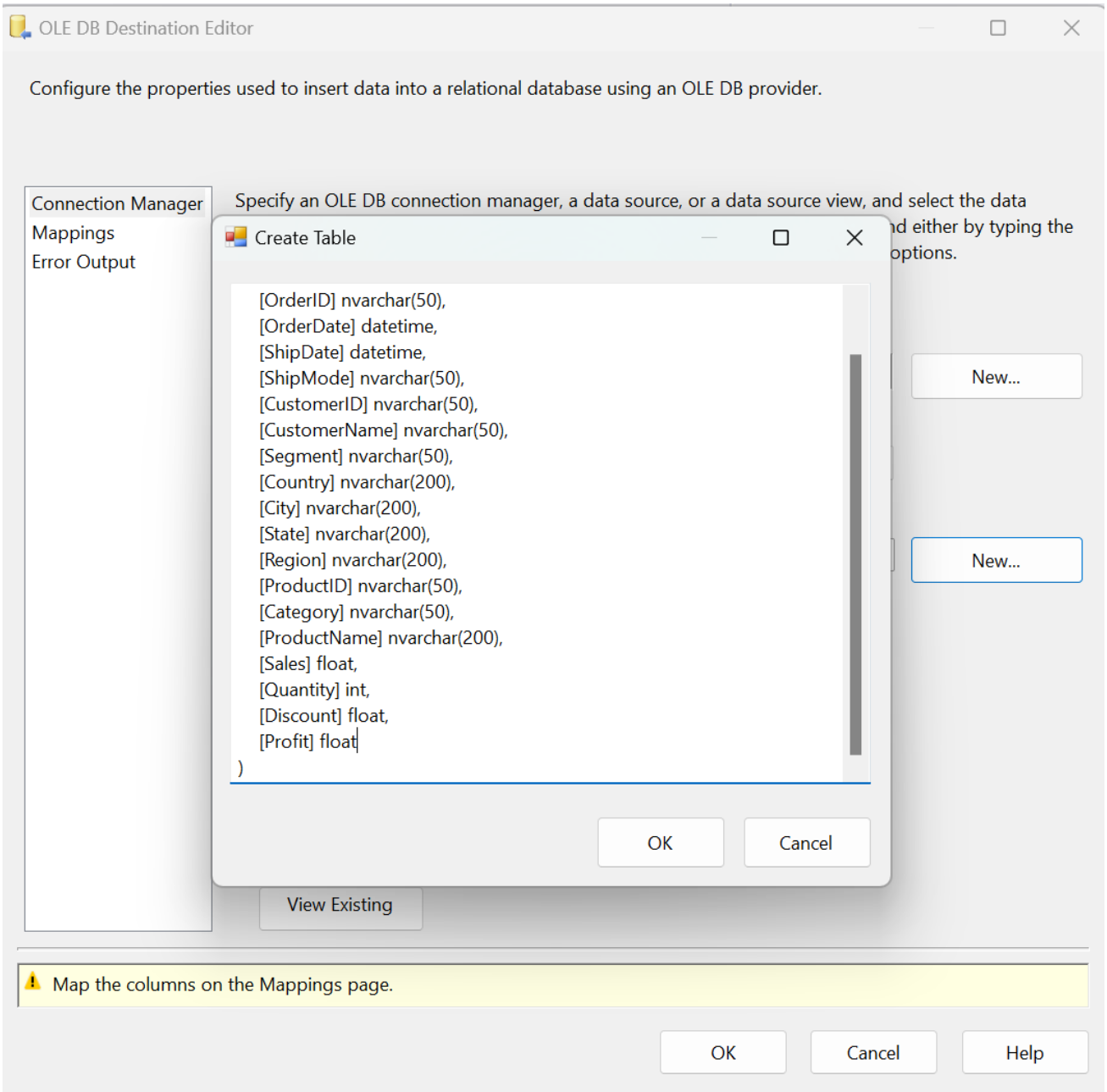
```
CREATE TABLE [Fact_raw] (
    [RowID] int constraint PK_fact primary key,
    [OrderID] nvarchar(50),
    [OrderDate] datetime,
    [ShipDate] datetime,
    [ShipMode] nvarchar(50),
    [CustomerID] nvarchar(50),
    [CustomerName] nvarchar(50),
    [Segment] nvarchar(50),
    [Country] nvarchar(200),
    [City] nvarchar(200),
    [State] nvarchar(200),
    [Region] nvarchar(200),
    [ProductID] nvarchar(50),
    [Category] nvarchar(50),
```

```
    [ProductName] nvarchar(200),
    [Sales] float,
    [Quantity] int,
    [Discount] float,
    [Profit] float
)
```

- **Lưu ý:**

+ Bảng **Fact_raw** sẽ chứa tất cả các cột và dữ liệu của dữ liệu gốc để trong quá trình sau có thể tiến hành Merge với các bảng khác để hoàn thiện bảng Fact.

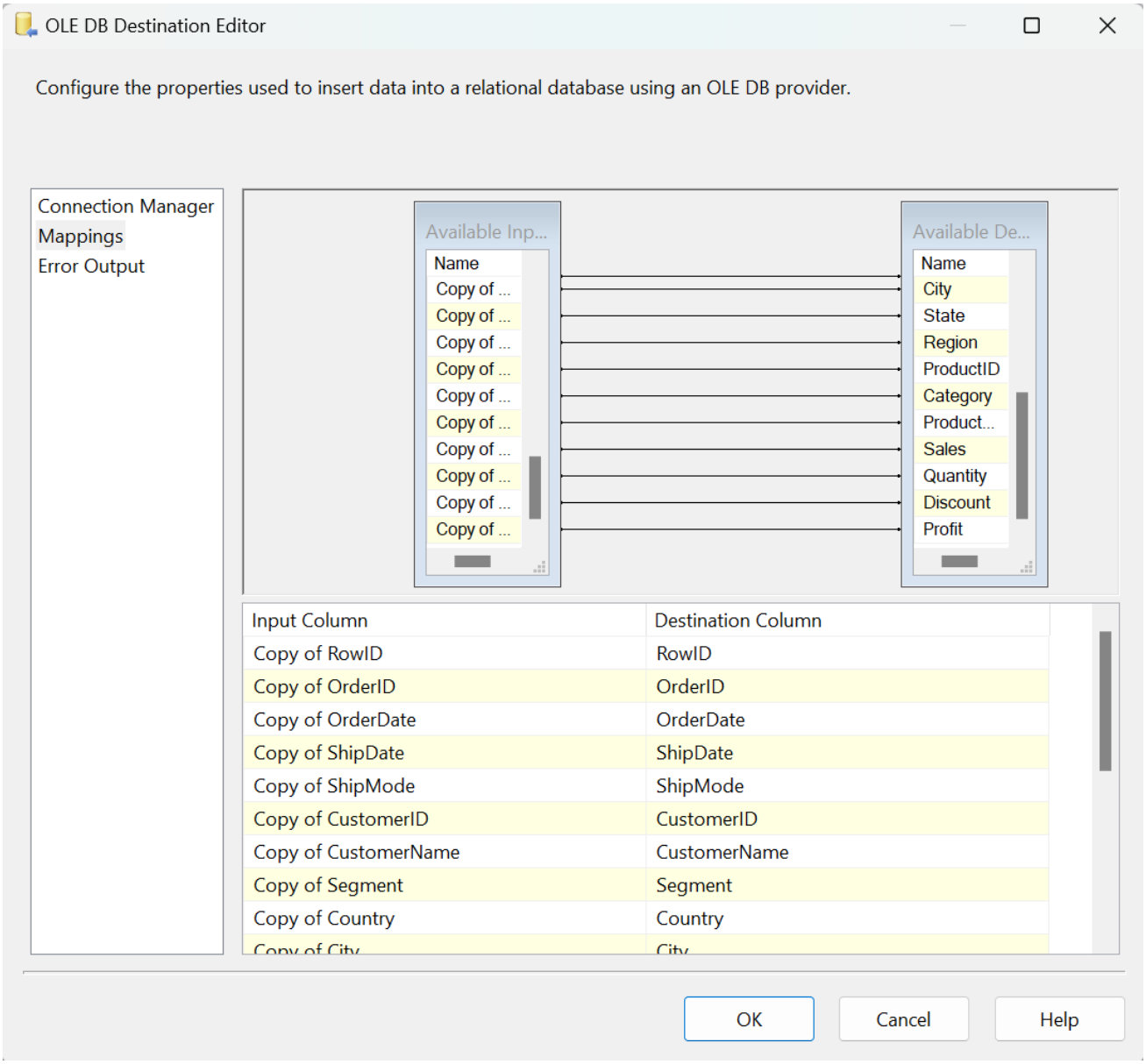
+ Bởi vì **RowID** trong dữ liệu gốc cũng chính là ID của tập dữ liệu nên ta sẽ chọn cột này làm khoá chính của bảng **Fact**.



Hình 2.65 Tạo bảng Fact_raw

Bước 3: Tiếp đến ta chọn mục **Mappings** để tiến hành ánh xạ các cột dữ liệu. Ở đây ta có thể nhận thấy rằng **RowID** cũng chính là ID của tập dữ liệu gốc vì vậy ta sẽ ánh xạ cột **Copy of RowID** trong **Input Column** tới cột **RowID** là khoá chính của bảng **Fact_raw**.

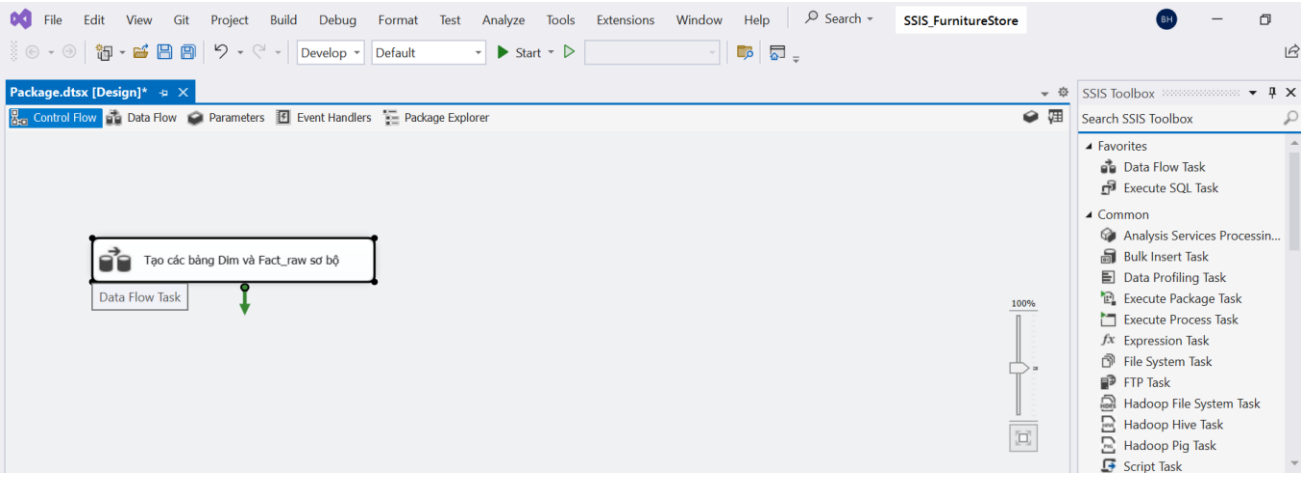
Sau khi đã kiểm tra việc ánh xạ thì nhấn **OK** để hoàn thành quá trình tạo bảng.



Hình 2.66 Ánh xạ các cột của dữ liệu xử lý tới các cột dữ liệu của Fact_raw

Ở các phần sau ta sẽ thực hiện quá trình tạo ra bảng **Dim_Product** cũng như tiến hành kết các bảng **Dimension** với bảng **Fact_raw** bằng cách loại bỏ những dữ liệu bị trùng trong **Fact_raw** đồng thời thực hiện thêm khoá ngoại của các bảng **Dimension** vào bảng **Fact_raw** nhằm thu gọn bảng **Fact_raw**, tạo ra bảng **Fact** cuối cùng để tối ưu hoá cho quá trình phân tích dữ liệu.

Để kết thúc bước này thì ta quay lại tab Control Flow và đặt tên cho Flow ta vừa mới thực hiện là **Tạo các bảng Dim và Fact_raw sơ bộ**.

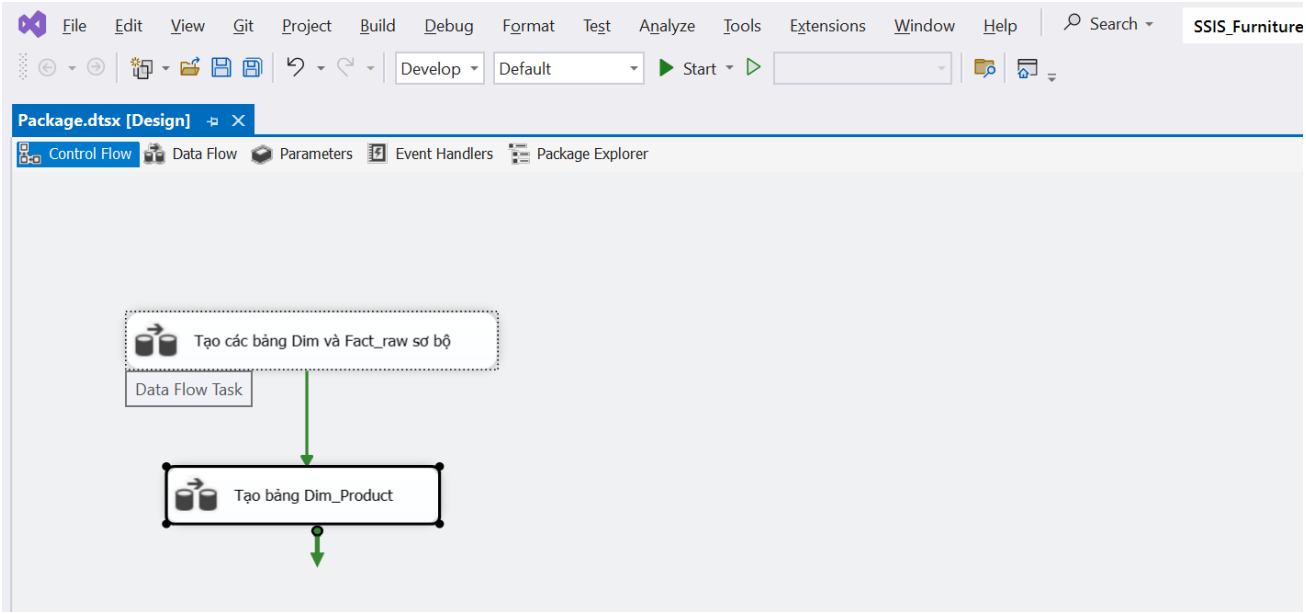


Hình 2.67 Đặt tên cho Control flow vừa cài đặt

Tạo bảng Dim_Product

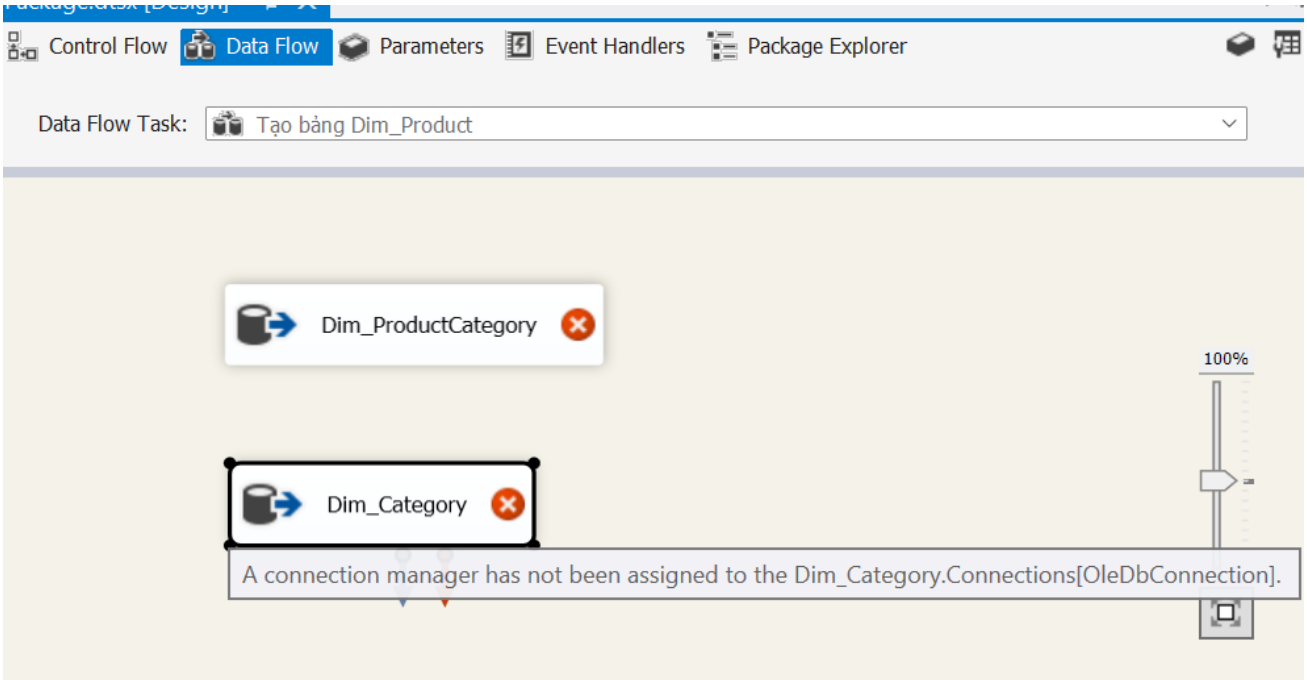
Như đã trình bày ở mục 2.1.2 ở bước tạo **Dim_ProductCategory** thì bảng **Dim_ProductCategory** sẽ được kết với bảng **Dim_Category** trong phần này để tạo ra bảng **Dim_Product** theo như thiết kế sơ đồ dữ liệu bông tuyết ban đầu.

Bước 1: Đầu tiên chúng ta tạo một thành phần **Data Flow Task** trong môi trường thực hiện và đặt tên cho Flow Task này là **Tạo bảng Dim_Product.** Và tiến hành kết nối phần dữ liệu mà ta đã xử lý ở Flow Task Tạo các bảng **Dim và Fact_raw sơ bộ** để có thể có dữ liệu của hai bảng **Dim_ProductCategory** và **Dim_Product** bằng các kéo mũi tên màu đỏ từ **Tạo các bảng Dim** và **Fact_raw sơ bộ** đến **Tạo bảng Dim_Product.**



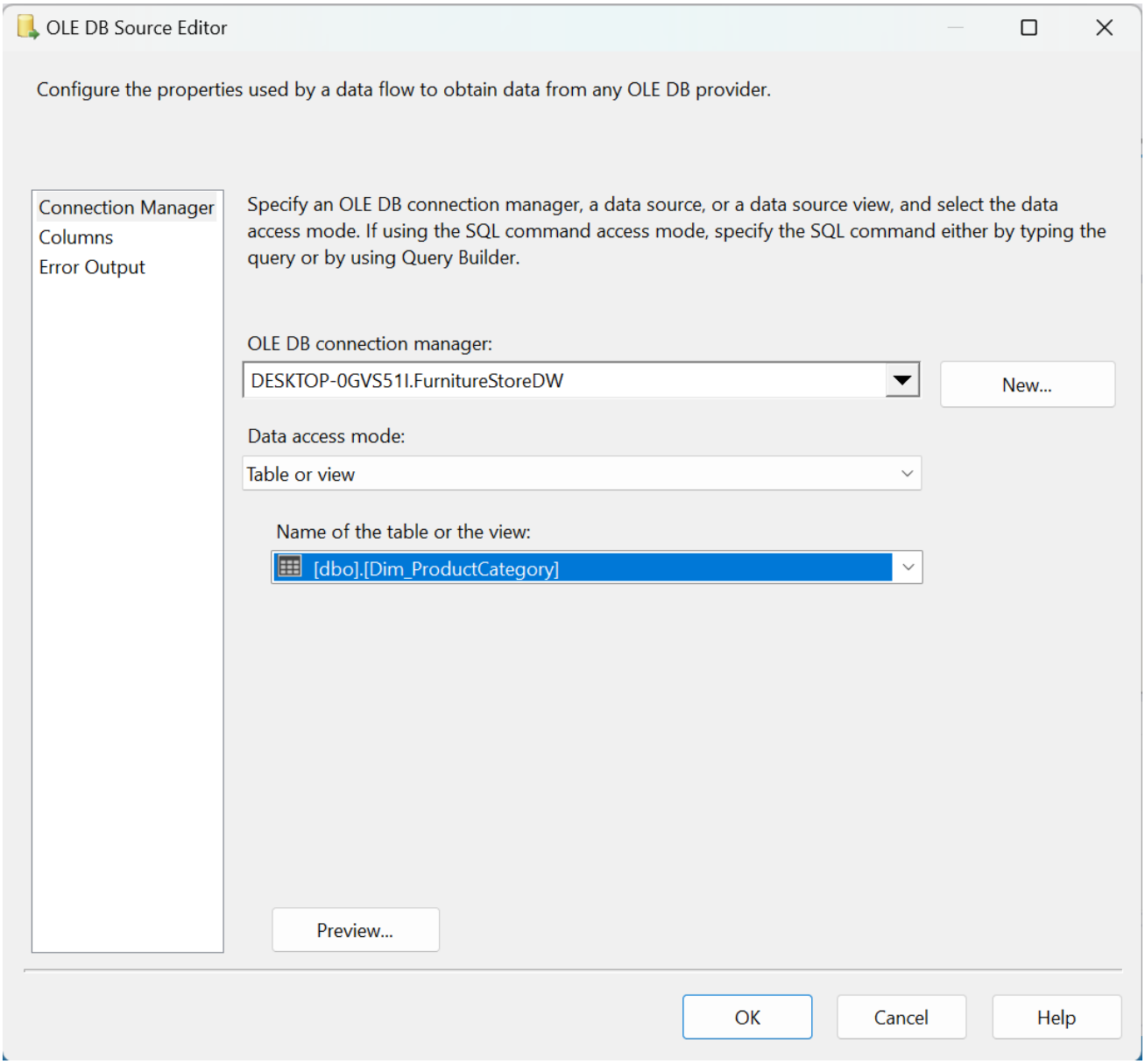
Hình 2.68 Tạo Data Flow Task Dim_Product

Bước 2: Nhấn đúp vào **Tạo bảng Dim_Product** để chuyển qua mục Data Flow và tiến hành cài đặt Flow. Đầu tiên, ta tạo hai thành phần **OLE DB Source** từ SSID Toolbox để có thể lấy cấu trúc và dữ liệu từ hai bảng **Dim_ProductCategory** và **Dim_Category** để có thể tiến hành kết với nhau.



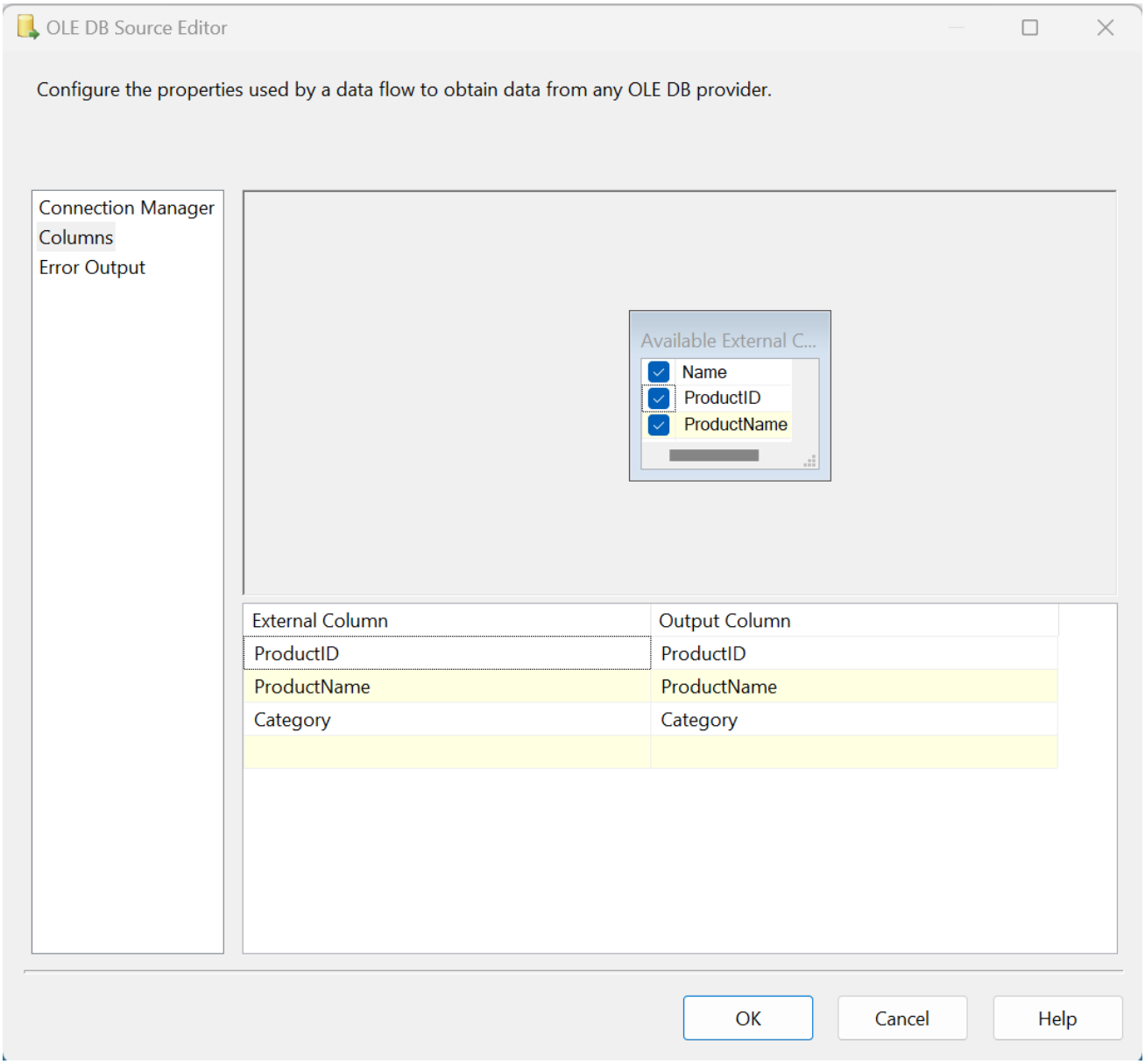
Hình 2. 69 Thêm OLE DB Source

Bước 3: Trong Dim_ProductCategory chúng ta chọn kết nối hiện tại mà ta kết nối tới FurnitureStoreDW và chọn soure bảng Dim_ProductCategory.



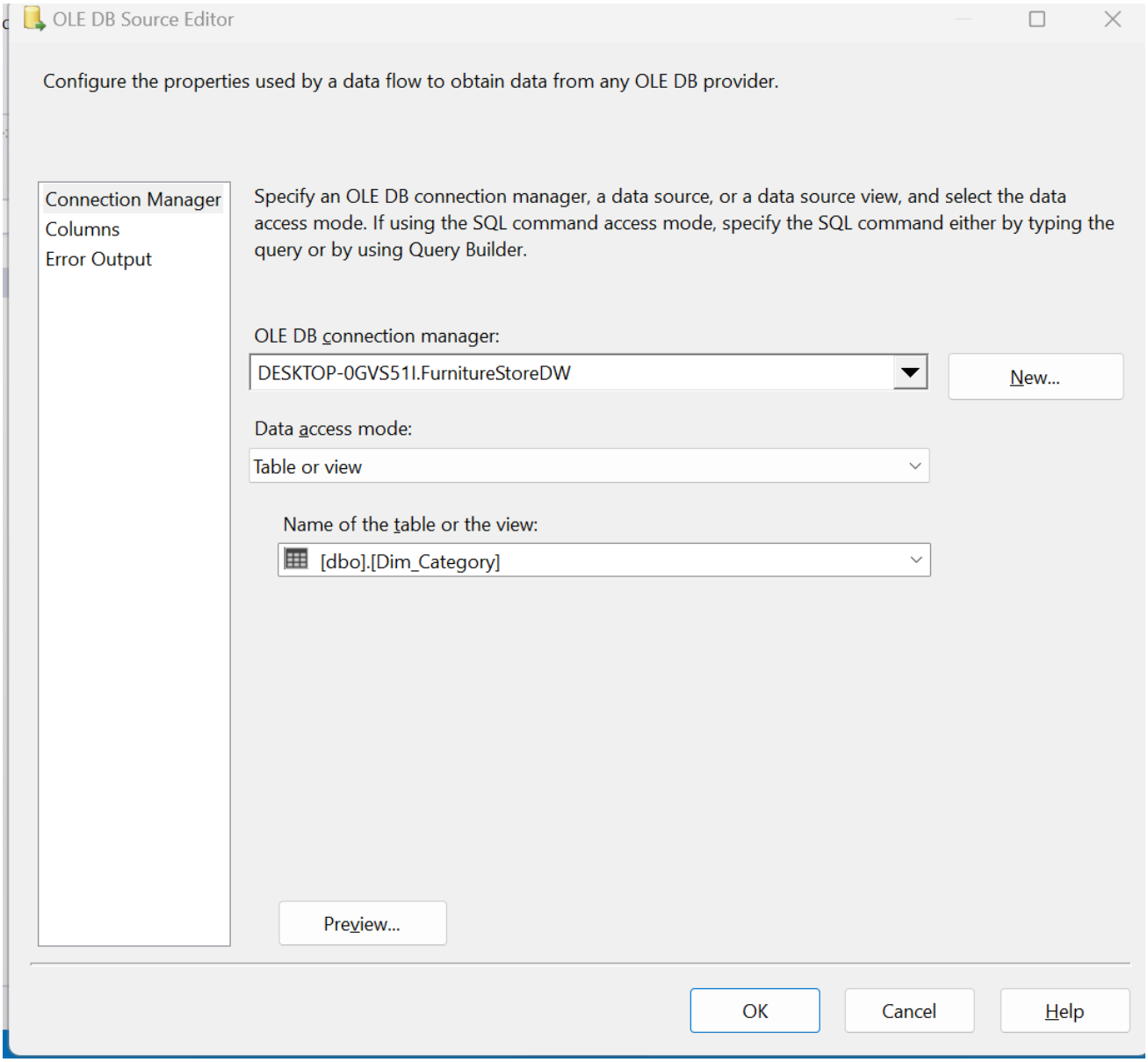
Hình 2.70 Chọn kết nối và bảng nguồn

Bước 4: Ta chuyển sang tab **Columns** để kiểm tra các cột dữ liệu có trong bảng.

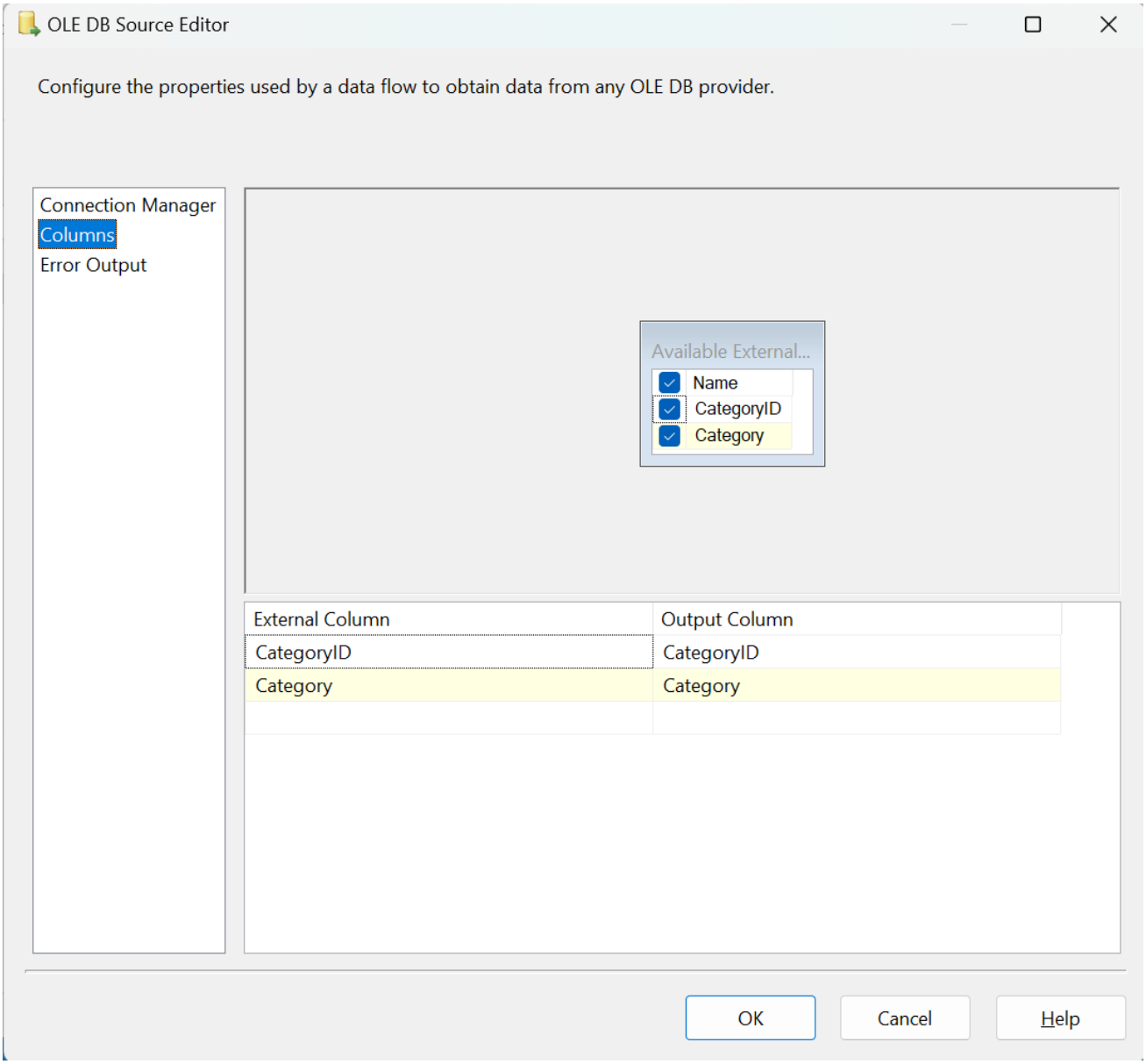


Hình 2.71 Thông tin cột của bảng

Bước 5: Tương tự với **Dim_Category** ta cũng chọn kết nối tới **FurnitureStoreDW** và chọn bảng nguồn là **Dim_Category.**

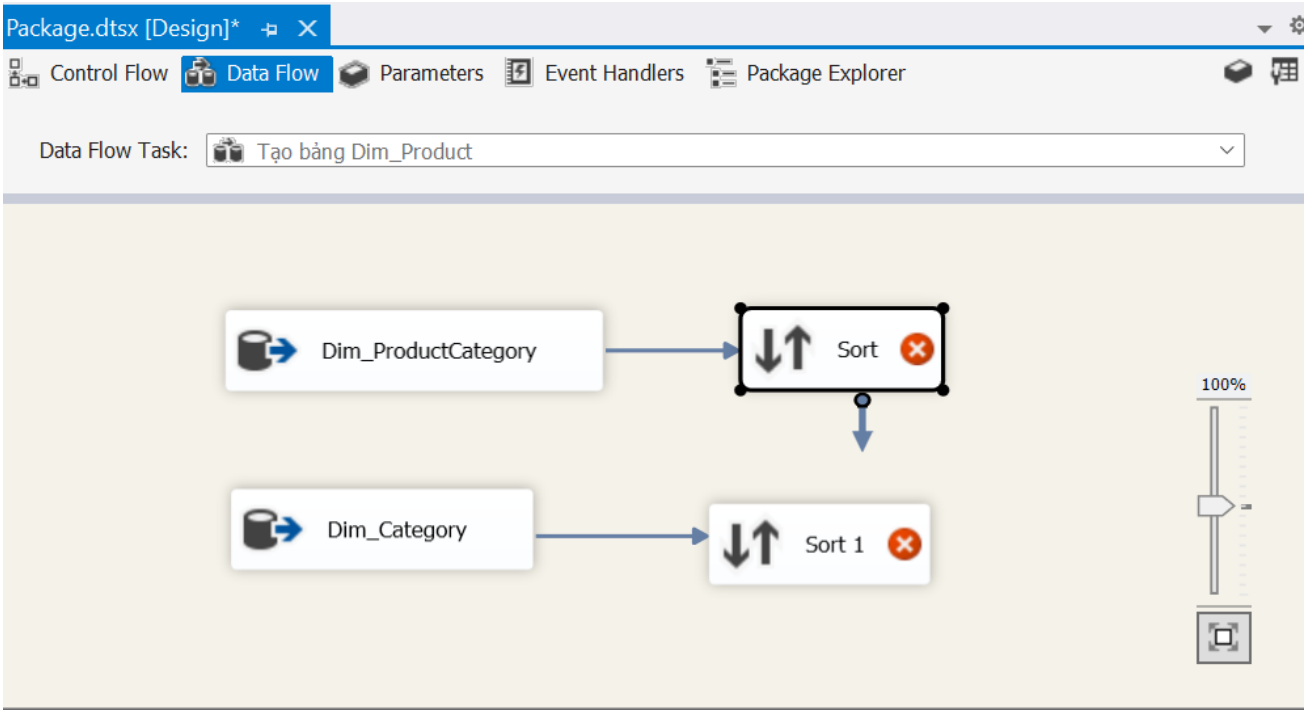


Hình 2.72 Chọn kết nối và bảng nguồn



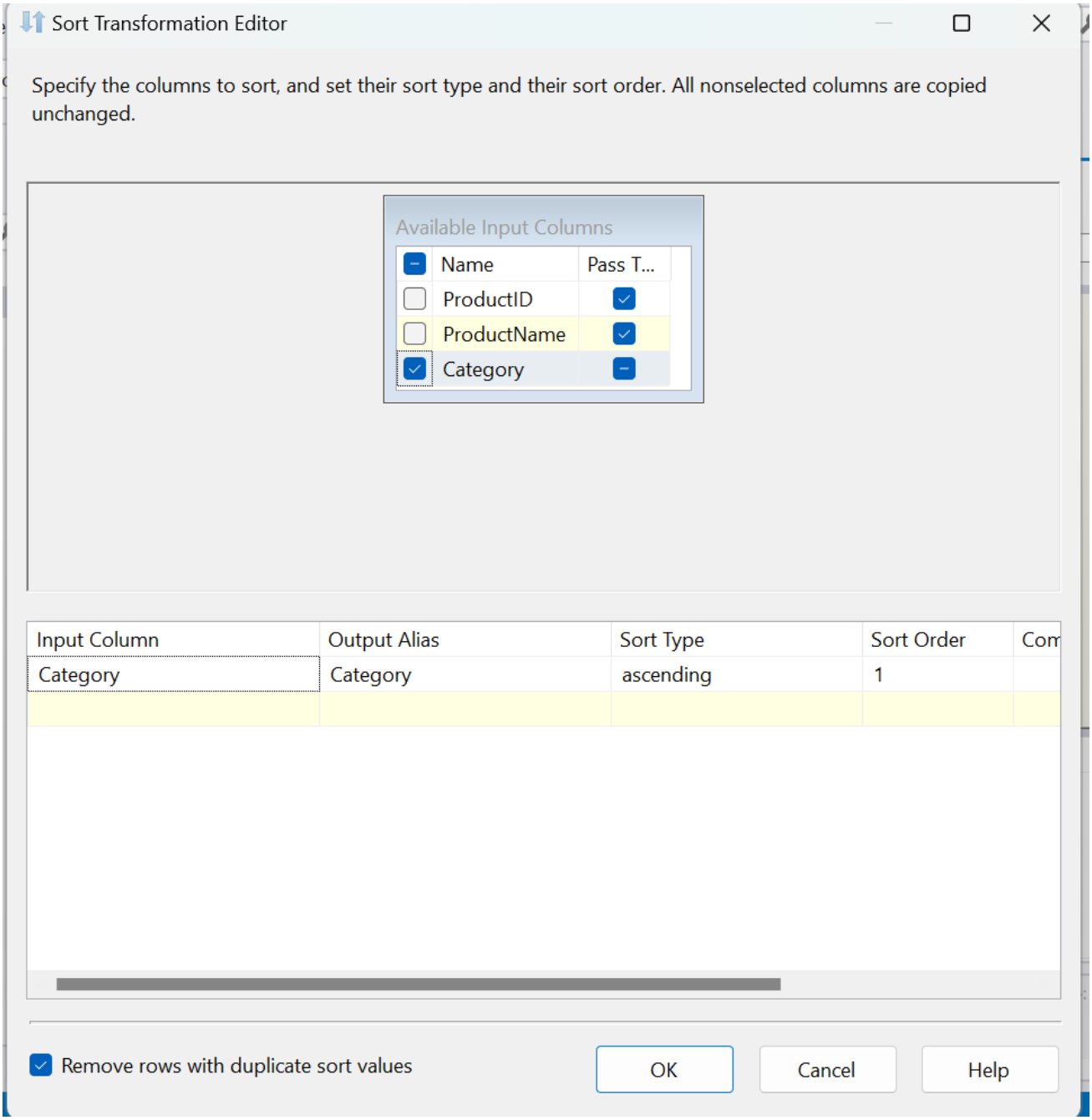
Hình 2.73 Thông tin các cột của bảng

Bước 6: Kéo thả 2 thành phần Sort vào vùng làm việc để tiến hành sắp xếp các thuộc cần kết giữa hai bảng, sắp xếp các thuộc tính kết với thứ tự giống nhau thì việc kết bảng mới có hiệu lực.

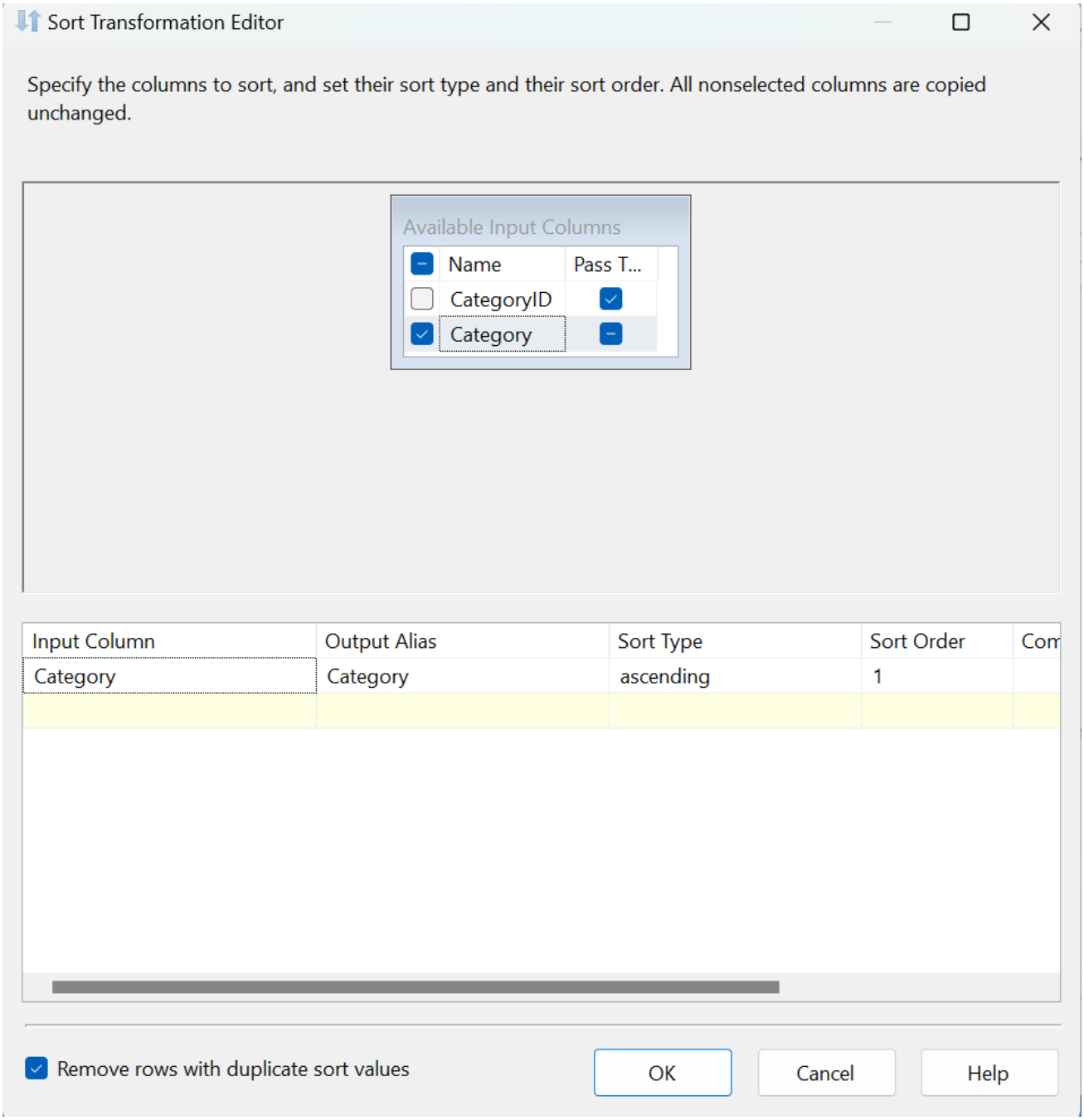


Hình 2.74 Thêm thành phần Sort

Bước 7: Vì thuộc tính kết của hai bảng là **Category** vì vậy ta cần Sort theo thuộc tính **Category** ở cả hai bảng.

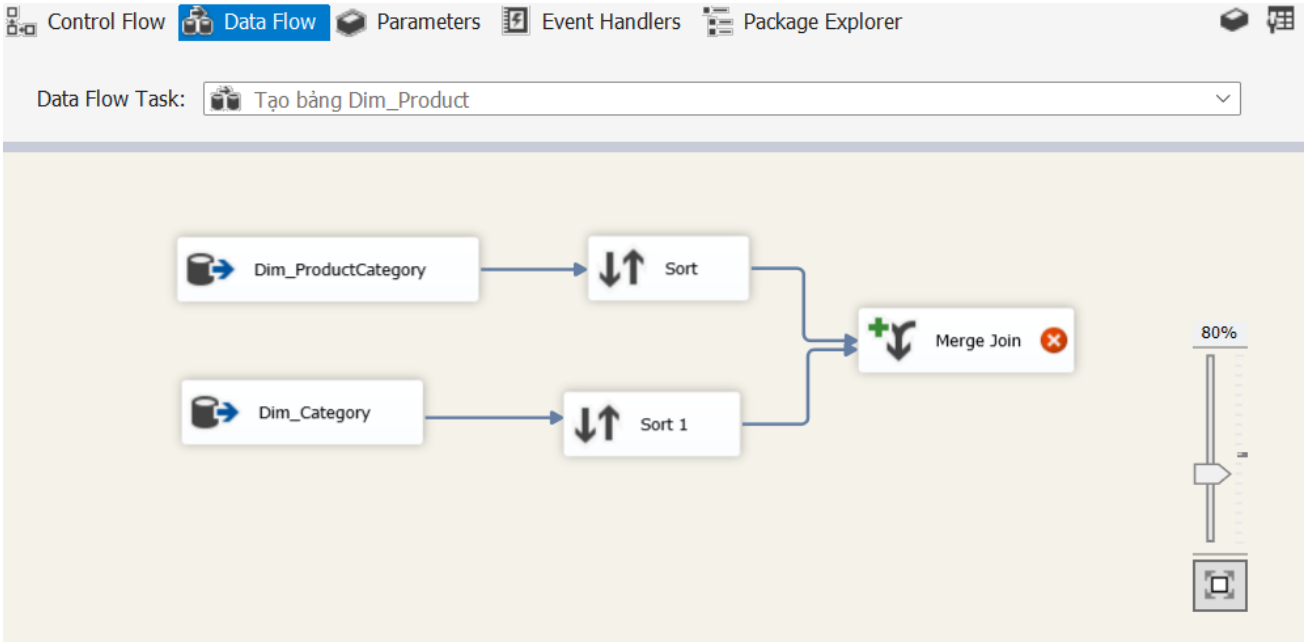


Hình 2.75 Sắp xếp dữ liệu theo thuộc tính kết



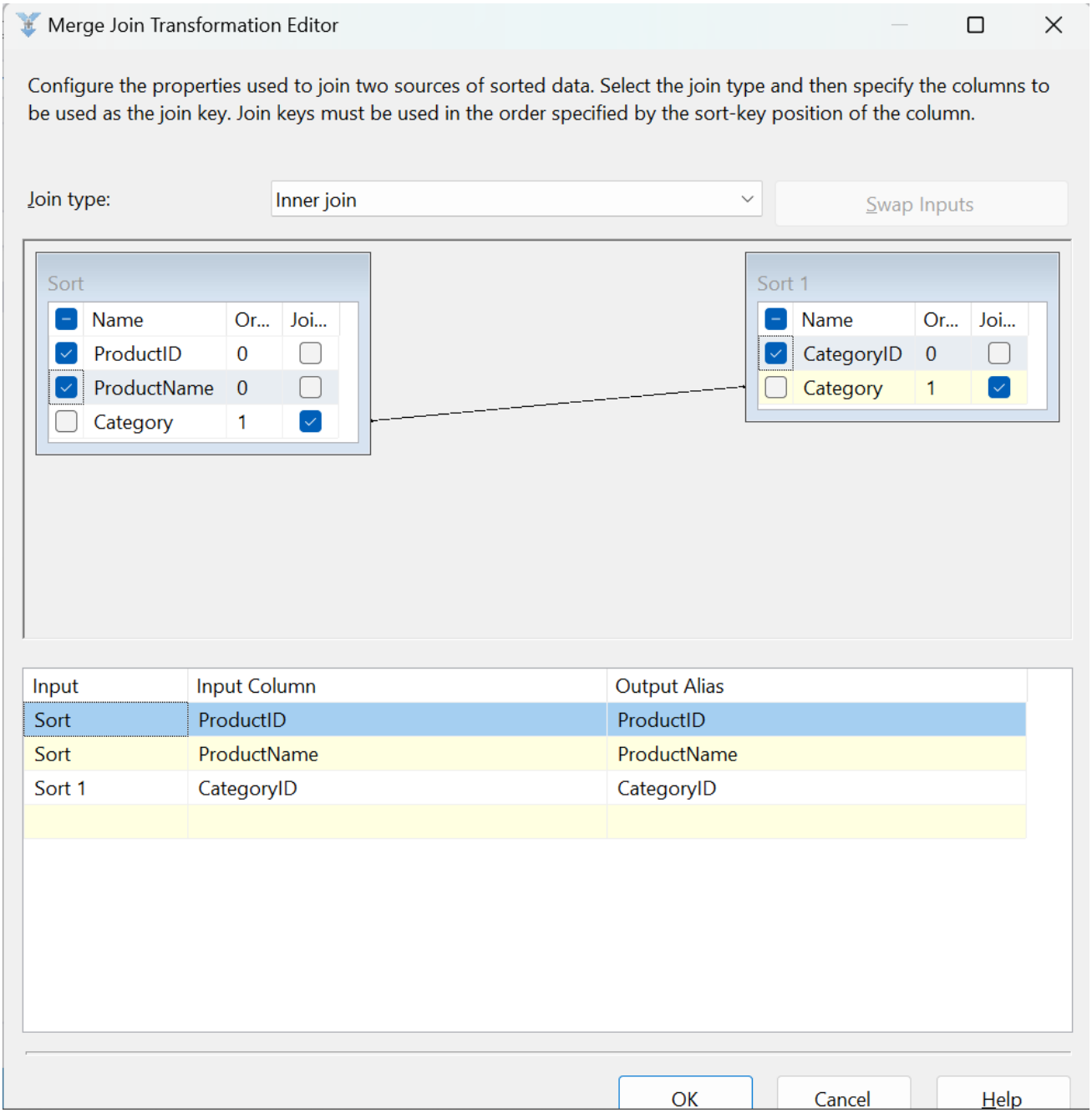
Hình 2.76 Sắp xếp dữ liệu bảng theo thuộc tính kết

Bước 8: Kéo thả một thành phần Merge Join và tiến hành nối 2 thành phần Sort vào để kết hai bảng dữ liệu.



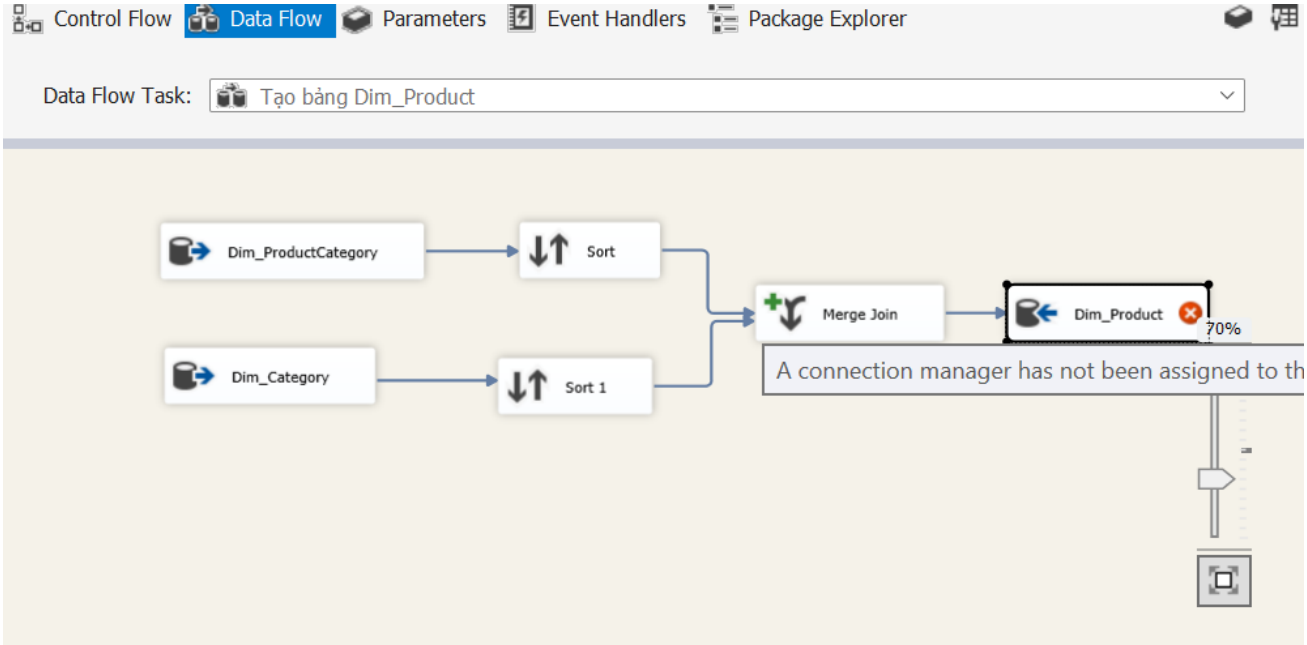
Hình 2.77 Thêm thành phần Merge Join

Bước 9: Nhấn đúp chuột để mở hộp thoại kết bảng, ở đây bảng **Dim_Product** ta cần các thuộc tính: **ProductID**, **ProductName**, **CategoryID**. Vì vậy, ta sẽ tiến hành tích chọn các thuộc tính này làm **InputColumn**.



Hình 2.78 Chọn thuộc tính của bảng kết quả của phép kết và thuộc tính kết

Bước 10: Tiến hành kéo thả một thành phần **OLE DB Destination** vào vùng làm việc và đặt tên là **Dim_Product** và nói thành phần Merger Join tới thành phần này.



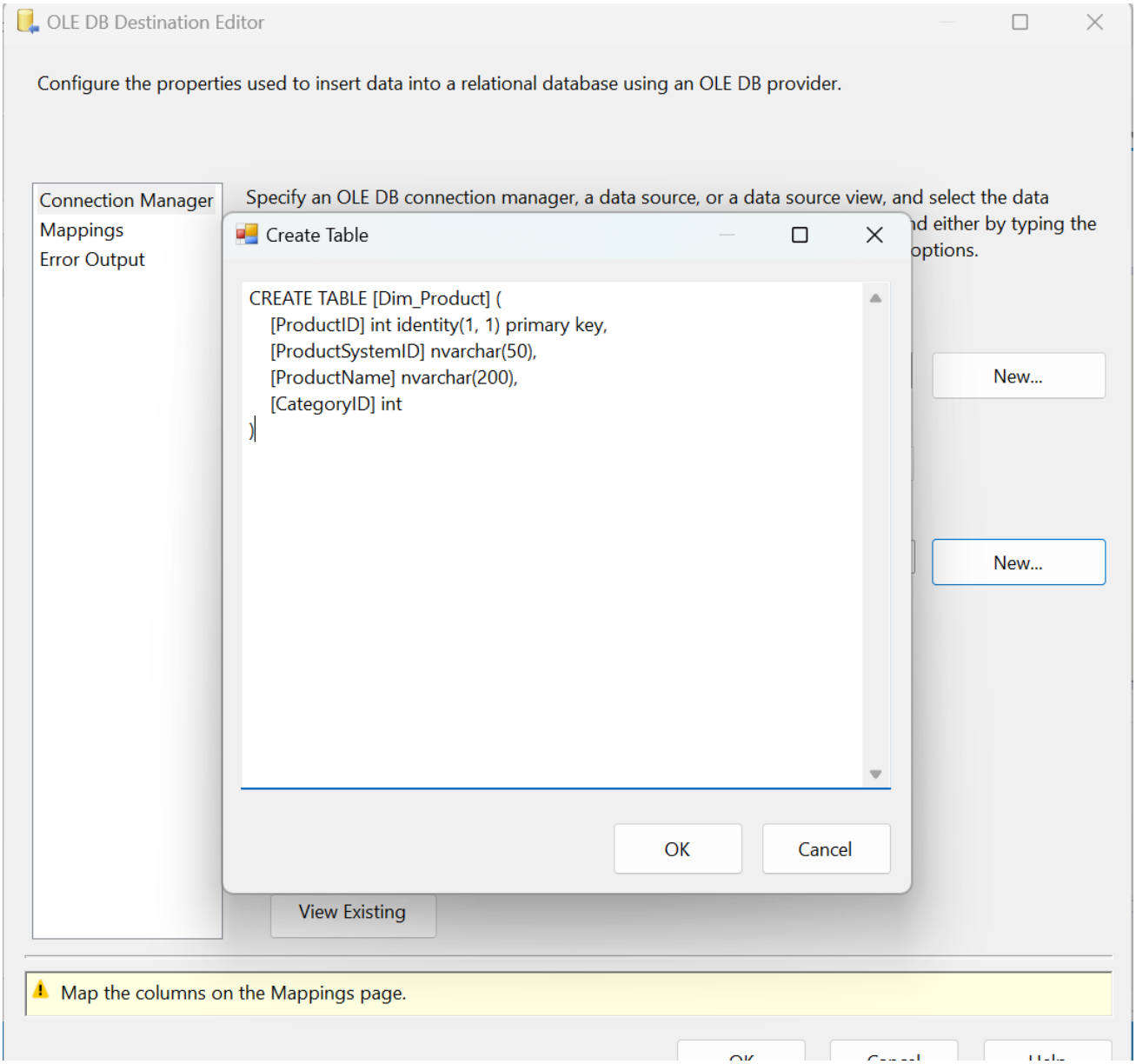
Hình 2.79 Tạo thành phần để đổ dữ liệu Dim_Product

Bước 11: Nhấn đúp chuột voà thành phần **Dim_Product → New** để tạo mới bảng.

Câu lệnh **SQL** tạo **Dim_Product** như sau:

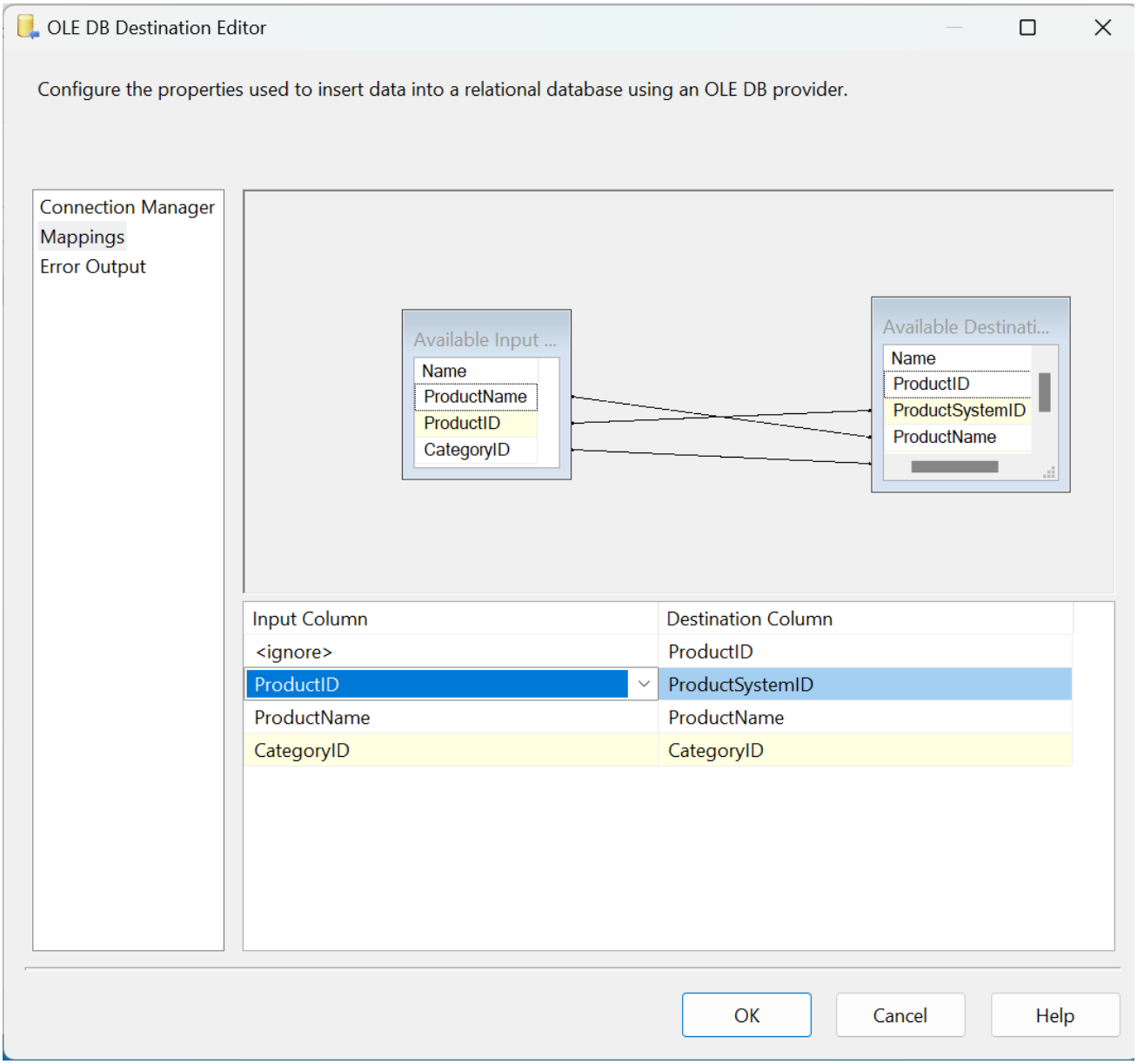
```
CREATE TABLE [Dim_Product] (
    [ProductID] int identity(1, 1) primary key,
    [ProductSystemID] nvarchar(50),
    [ProductName] nvarchar(200),
    [CategoryID] int
)
```

Trong đó, ta tạo thêm một cột **ProductID** là khoá chính của bảng **Dim_Product** trong kho dữ liệu với thuộc tính tự động tăng và **ProductSystemID** là **ID** của **Product** trong dữ liệu ban đầu bởi vì trong dữ liệu gốc thì có những sản phẩm trùng **ID** nhau vì vậy nếu như lấy **ID** đó làm khoá chính thì có thể gây ra lỗi.



Hình 2.80 Câu lệnh tao bảng Dim_Product

Bước 11: Chọn tab Mappings để tiến hành ánh xạ các cột dữ liệu từ **Input Column** sang **Destination Column** như hình minh hoạ sau,

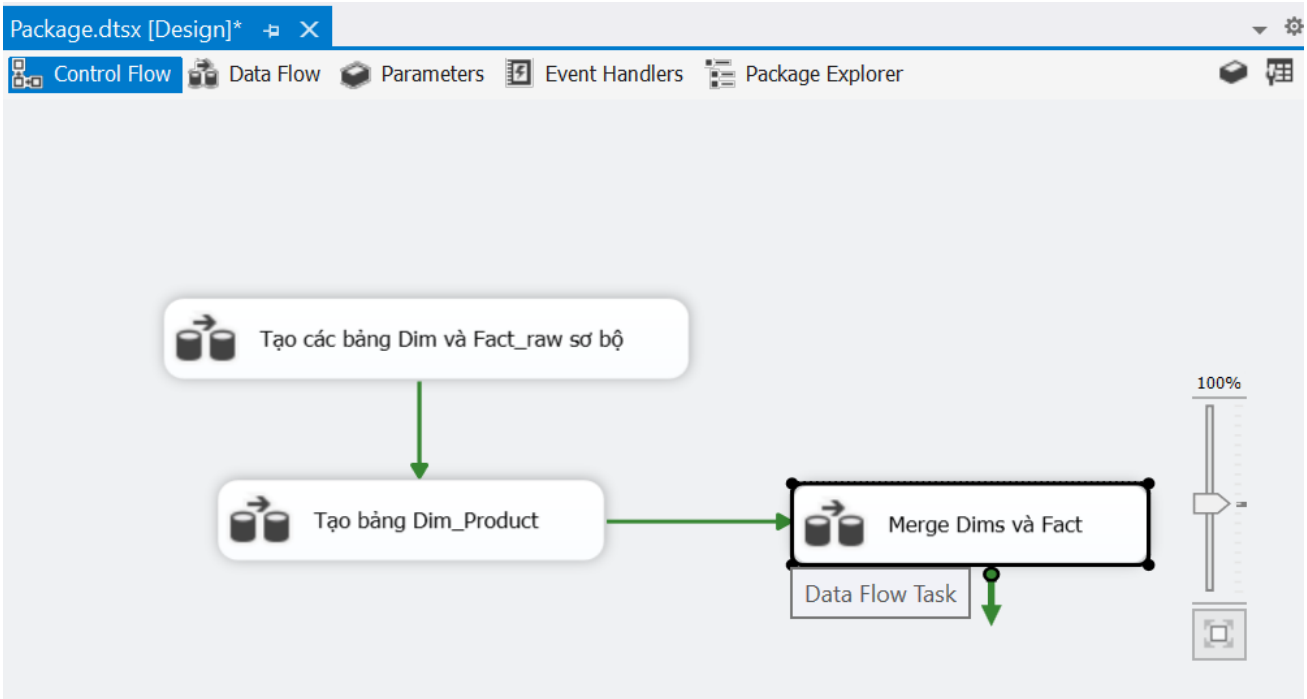


Hình 2.81 Ánh xạ thuộc tính cho bảng Dim_Product

Bước 12: Nhấn **OK** để hoàn thành thiết lập để tạo bảng **Dim_Product.**

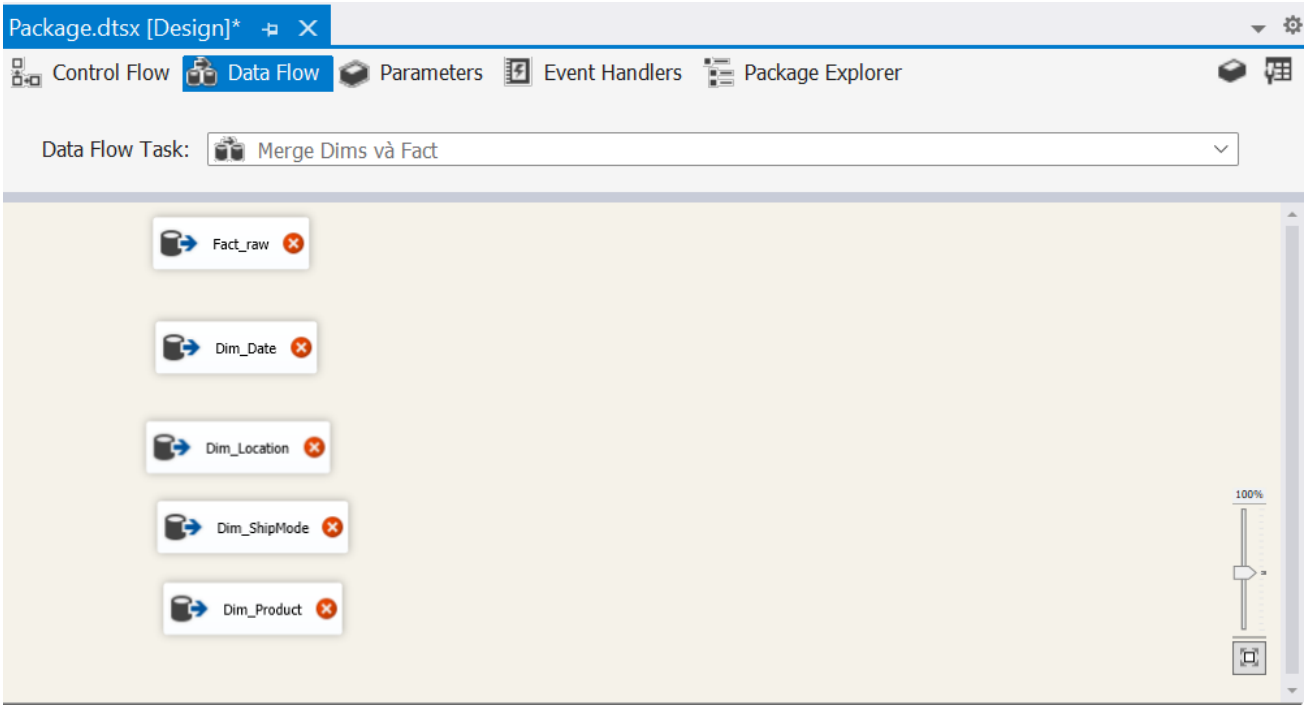
Tạo bảng Fact

Bước 1: Sau khi đã hoàn thành **Data Flow Task** tạo bảng **Dim_Product** thì ta quay lại tab **Data Flow** và ta tạo thêm một **Data Flow Task** mới đặt tên là **Merge Dims và Facts**.



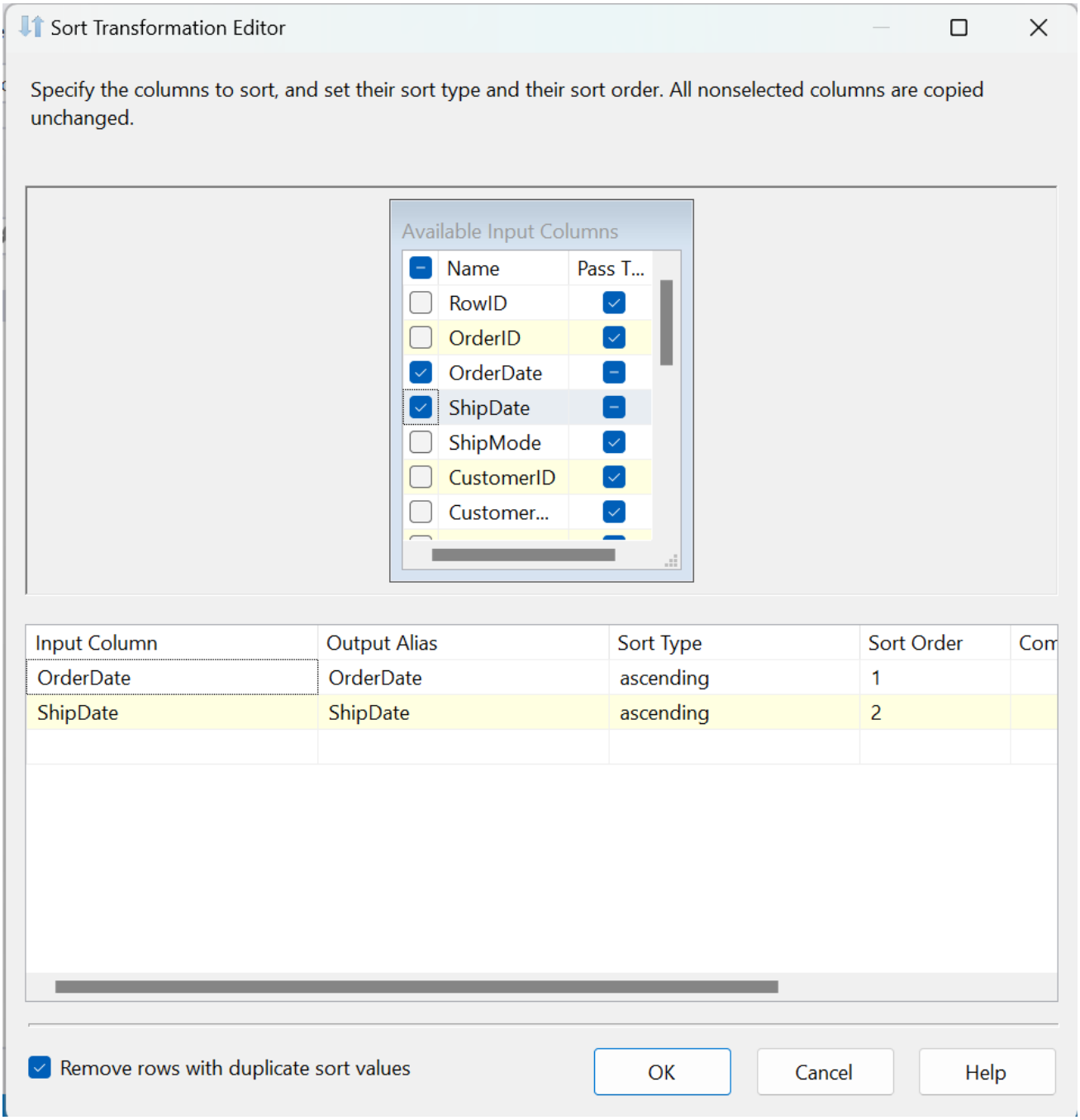
Hình 2.82 Thêm mới Data Flow Task

Bước 2: Kéo thả các thành phần OLE DB Source vào vùng làm việc tương ứng với các bảng cần kết với nhau để tạo ra bảng Fact cuối cùng.



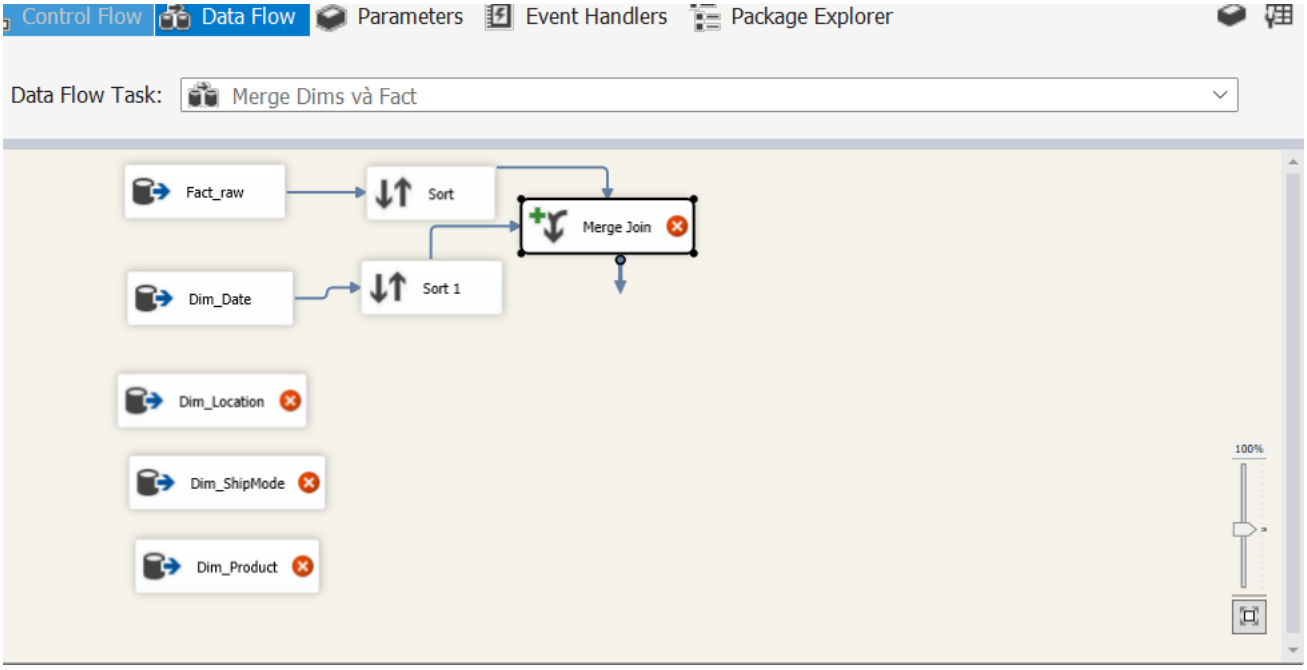
Hình 2.83 Thêm OLE DB Source tương ứng với các bảng cần kết

Bước 3: Kéo thả thành phần Sort để tiến hành sắp xếp dữ liệu **Fact_raw** và **Dim_Date** theo các thuộc tính kết bảng để kết với nhau. Ở đây các thuộc tính cần kết là **OrderDate** và **ShipDate** nên ta sẽ tiến hành sắp xếp dữ liệu hai bảng theo hai thuộc tính này.



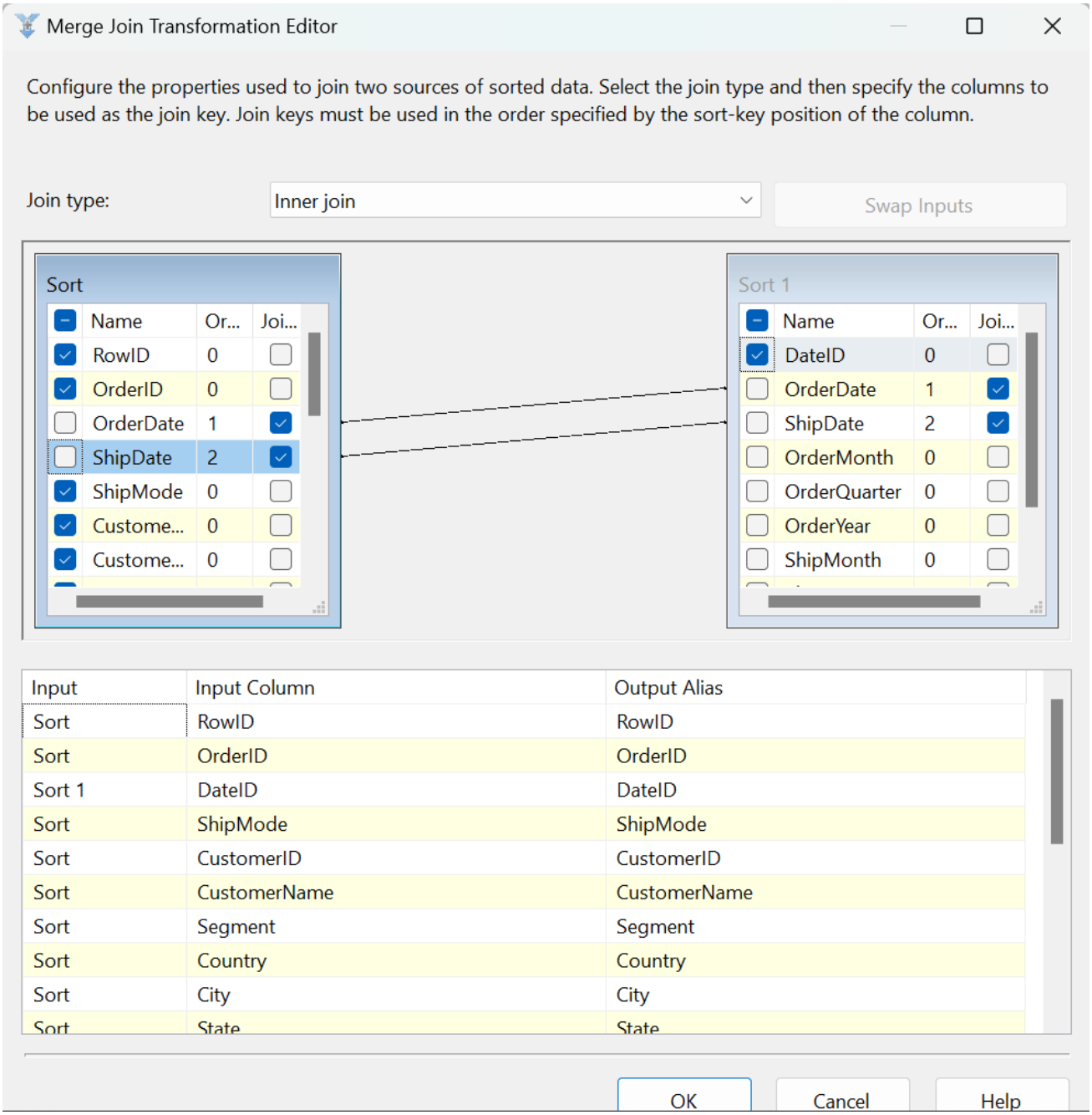
Hình 2. 84 Sắp xếp dữ liệu theo thuộc tính kết

Bước 4: Kéo thả thành phần Merge Join để tiến hành kết bảng **Fact_raw** và **Dim_Date** với nhau.



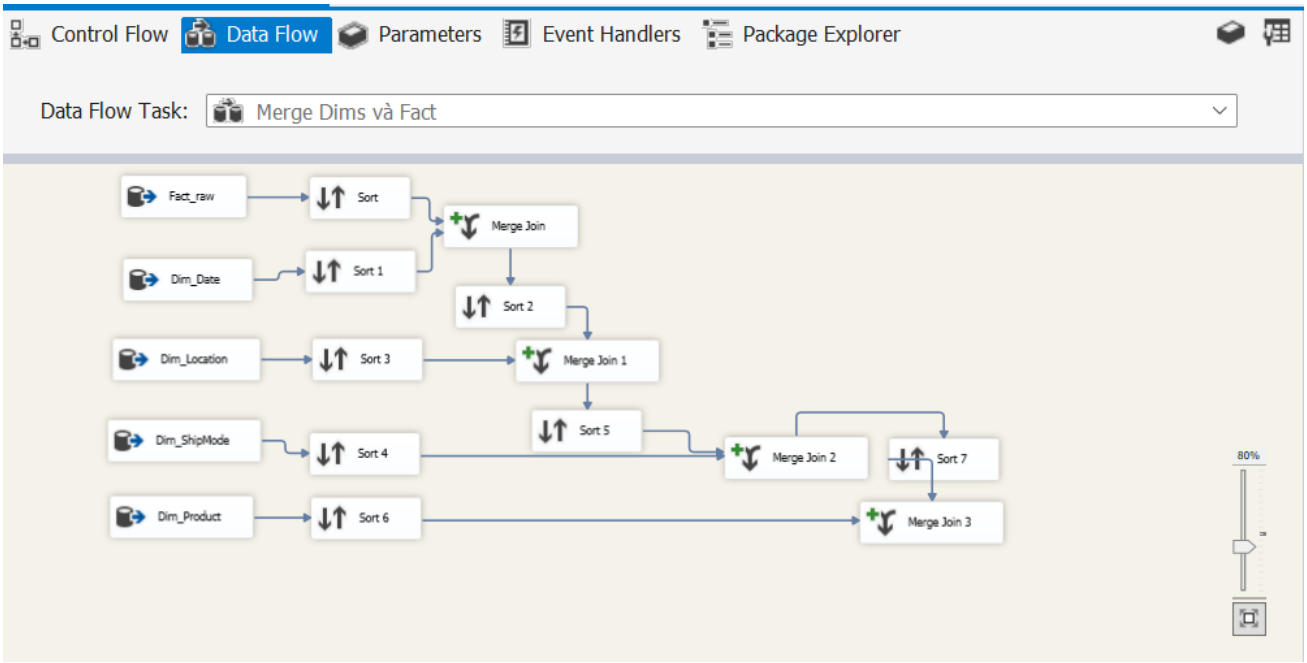
Hình 2.85 Thêm thành phần Merger Join để kết bảng

Bước 5: Nhấn đúp vào thành phần **Merge Join** để tiến hành thực hiện tạo ra bảng kết**.** Ở đây thì ở trong bảng kết sẽ không có thuộc tính **OrderDate** và **ShipDate** nên ta sẽ không tích chọn làm **Input Column**. Bảng kết sẽ bao gồm tất cả thuộc tính của **Fact_raw** trừ hai thuộc tính vừa đề cập và có thêm thuộc tính **DateID**. Cấu hình theo như hình minh hoạ sau.



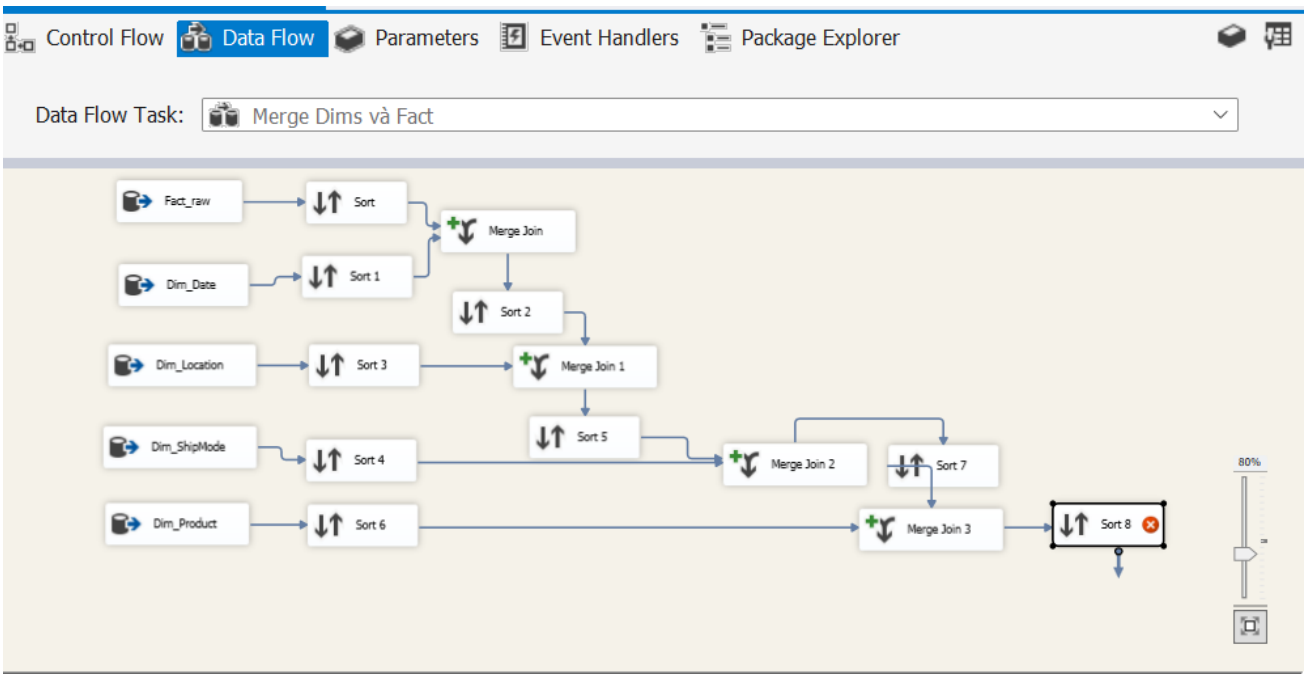
Hình 2.86 Chọn thuộc tính kết và thuộc tính của bảng kết

Bước 6: Tương tự với các bảng Dimension sau đó thì ta cũng cần Sort trước khi có thể thực sự kết hai bảng. Kết quả của quá trình thao tác như hình minh hoạ sau.

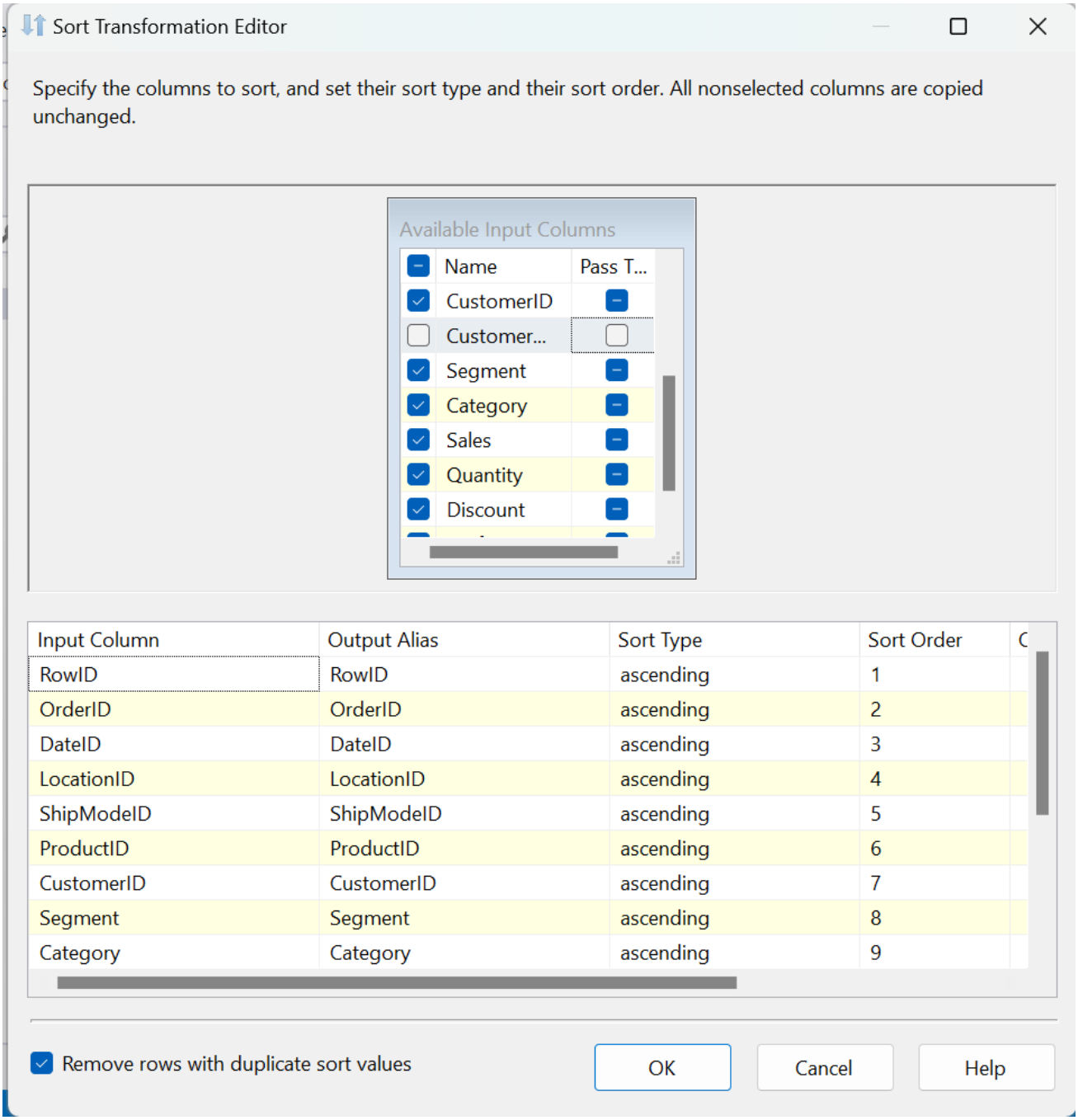


Hình 2.87 Xây dựng mô hình kết bảng

Bước 7: Trước đổ dữ liệu vào bảng **Fact** thì ta tiến hành **Sort** lại một lần nữa để bỏ đi thuộc tính **CustomerName** vì thuộc tính này đã nằm trong bảng **Dim_Customer** nên không cần phải lưu một lần nữa ở trong bảng Fact mà chỉ cần giữ lại thuộc tính **CustomerID**.



Hình 2.88 Sắp xếp dữ liệu trước khi đổ dữ liệu vào bảng



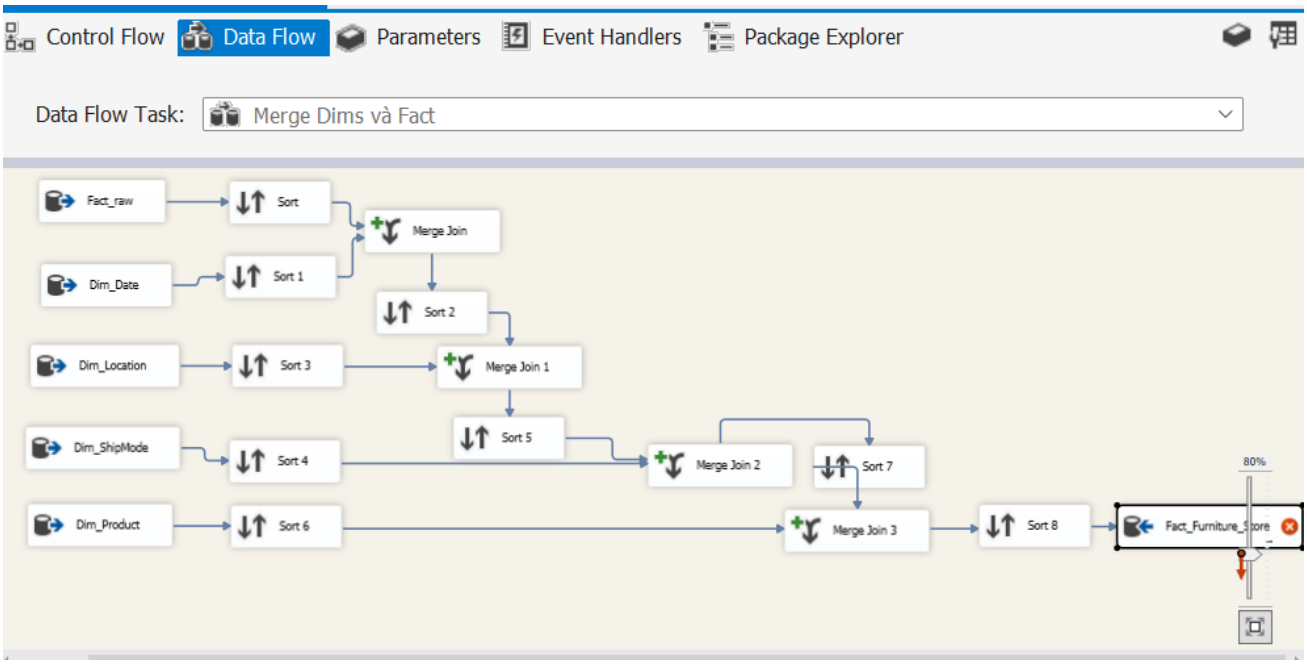
Hình 2. 89 Chọn thuộc cần bỏ và thuộc tính cần giữ lại cho bảng đổ dữ liệu

Bước 8: Tạo một thành phần OLE DB Destination mới và đặt tên là Fact. Tiến hành kết nối tới FurnitureStoreDW → New để tạo mới bảng → Dán câu lệnh SQL tạo bảng Fact.

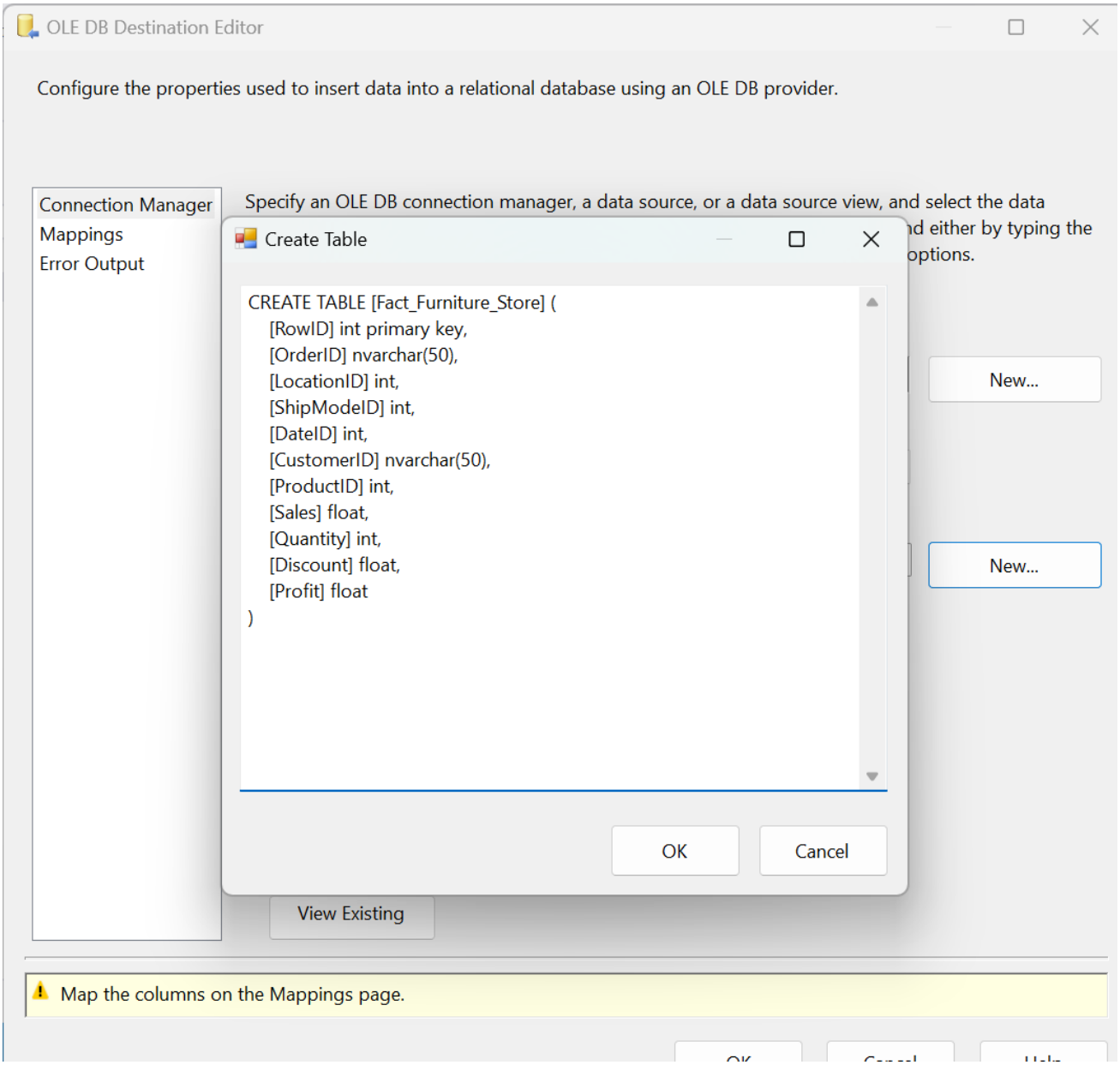
Câu lệnh SQL tạo bảng Fact.

```
CREATE TABLE [Fact_Furniture_Store] (
    [RowID] int primary key,
    [OrderID] nvarchar(50),
    [LocationID] int,
```

```
    [ShipModeID] int,
    [DateID] int,
    [CustomerID] nvarchar(50),
    [ProductID] int,
    [Sales] float,
    [Quantity] int,
    [Discount] float,
    [Profit] float
)
```

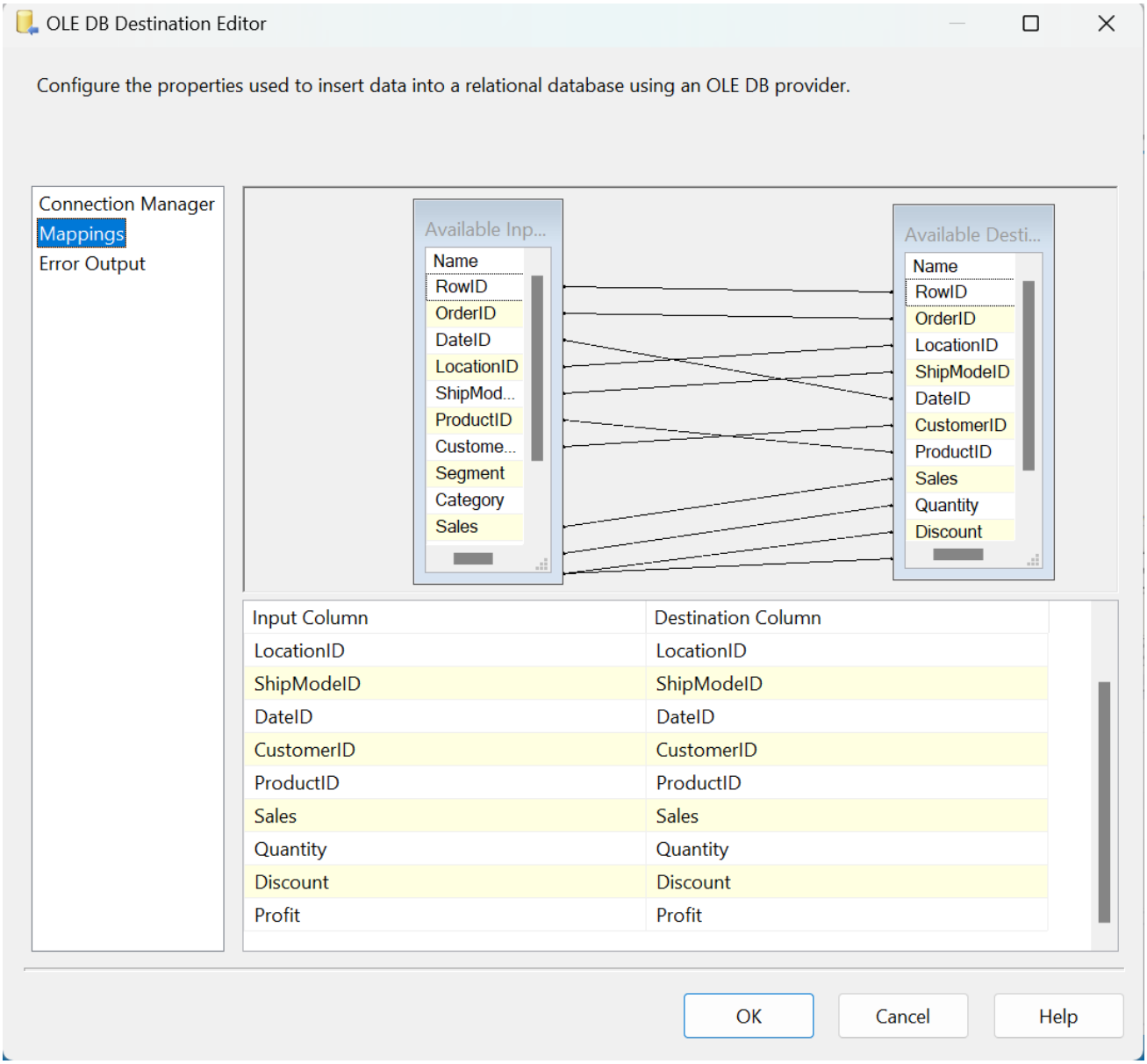


Hình 2.90 Tạo mới thành phần OLE DB Destination để đổ dữ liệu Fact_Furniture_Store



Hình 2.91 Tạo bảng

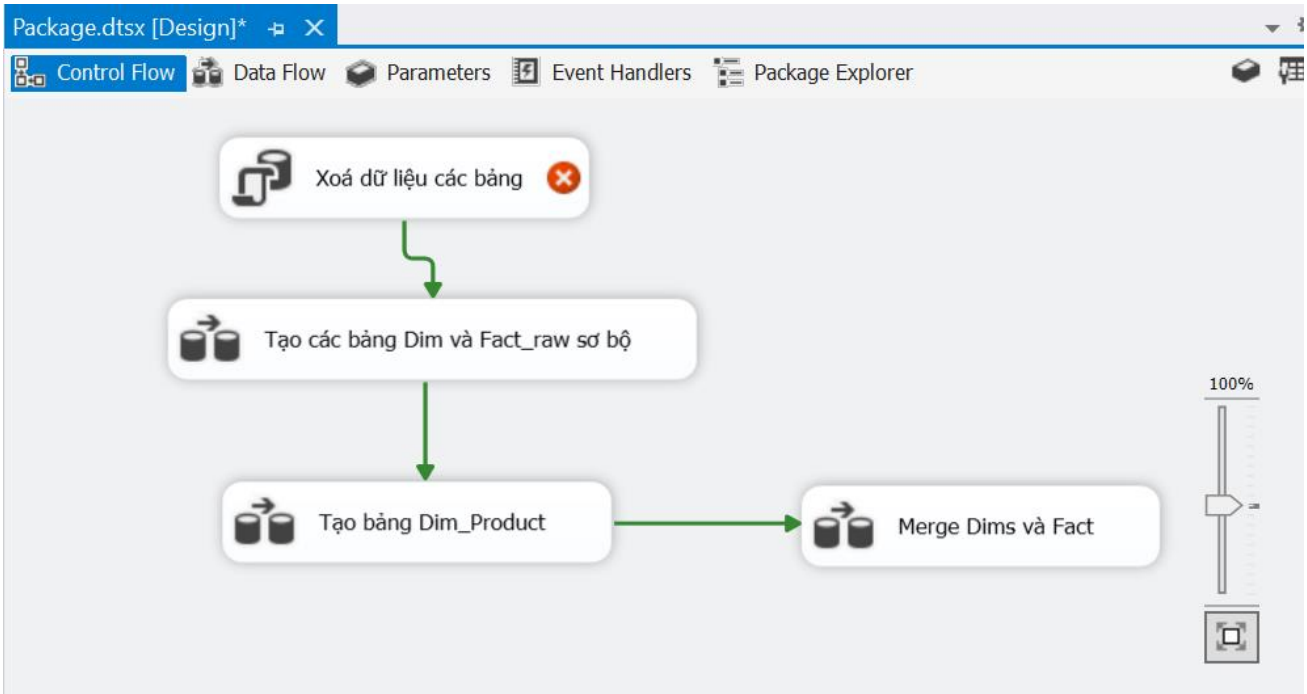
Bước 9: Chọn sang tab Mappings để tiến hành ánh xạ các thuộc tính.



Hình 2.92 Ánh xạ thuộc tính của bảng

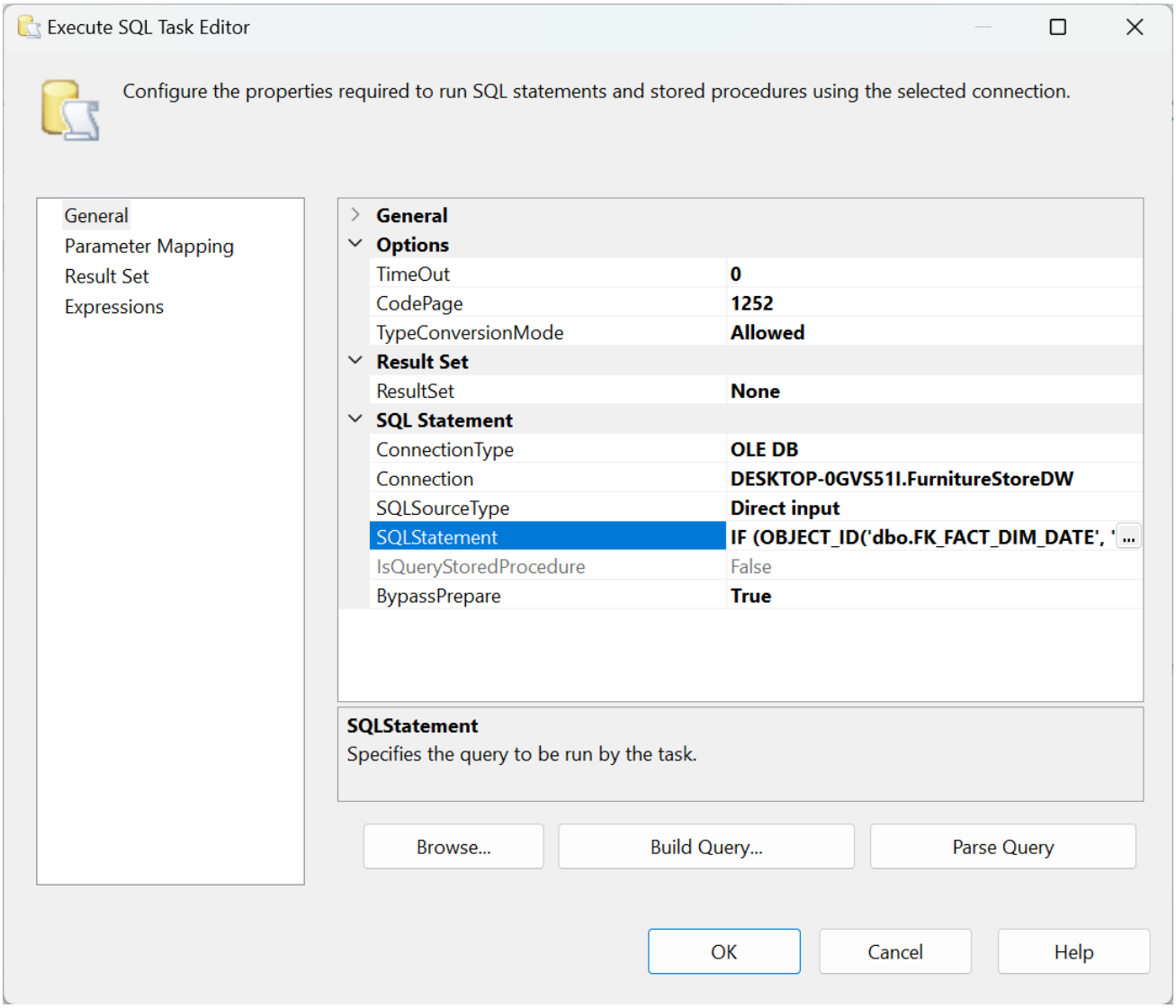
Thêm câu lệnh thêm khoá ngoại và xoá dữ liệu ban đầu

Bước 1: Quay trở lại tab Data Flow Task và thêm mới một Ecxecute SQL Script đặt tên là **Xoá dữ liệu các bảng**.

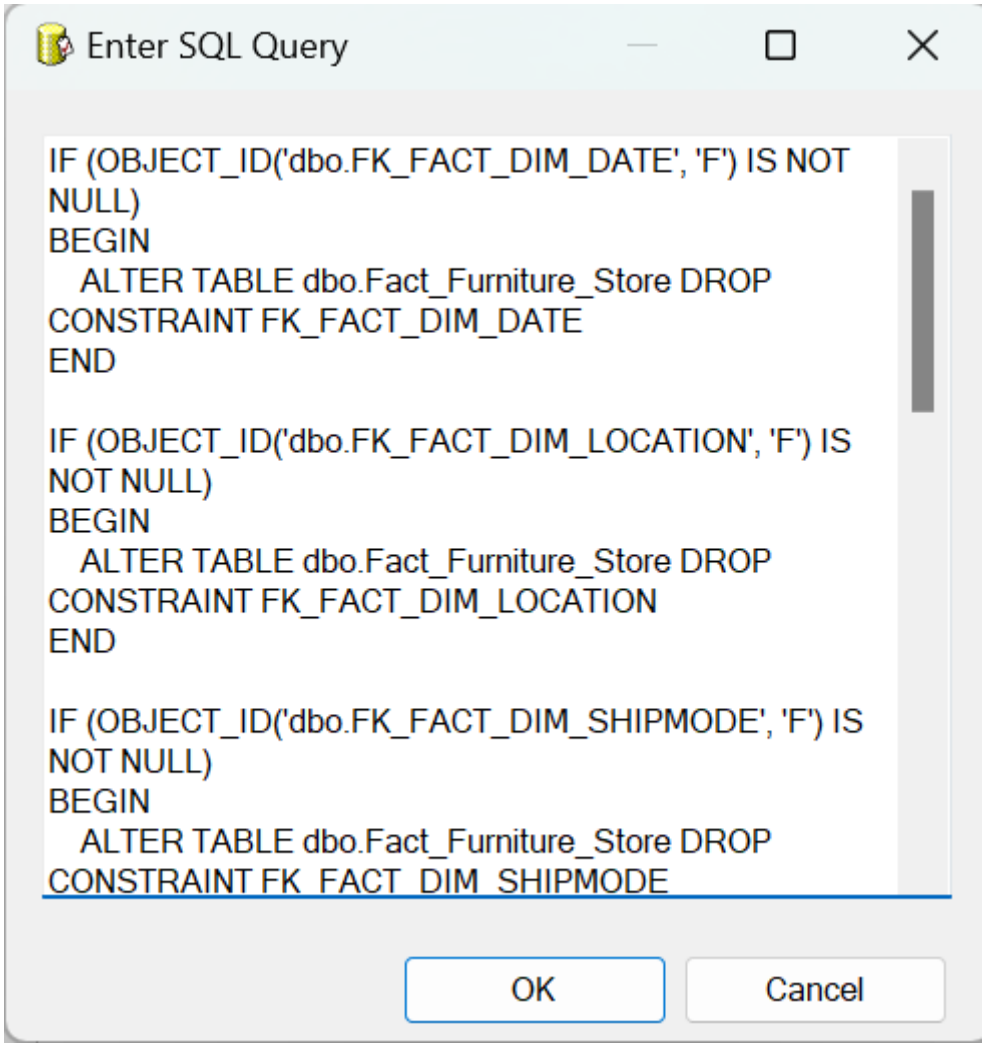


Hình 2.93 Thêm thành phần Execute SQL Task

Bước 2: Nhấn đúp chuột vào thành phần chọn loại kết nối tới **FurnitureStoreDW** và chọn biểu tượng **…** ở mục **SQL Statement** và dán câu lệnh sau vào **SQL Query → OK** 2 lần để hoàn thành.



Hình 2.94 Giao diện cài đặt



Hình 2.95 Dán câu lệnh xoá dữ liệu

Bước 3: Tương tự ta tạo thêm một Execute SQL Task và đặt thêm là **Thêm ràng buộc khoá ngoại.** Thực hiện tương tự như thêm câu lệnh xoá dữ liệu bảng.

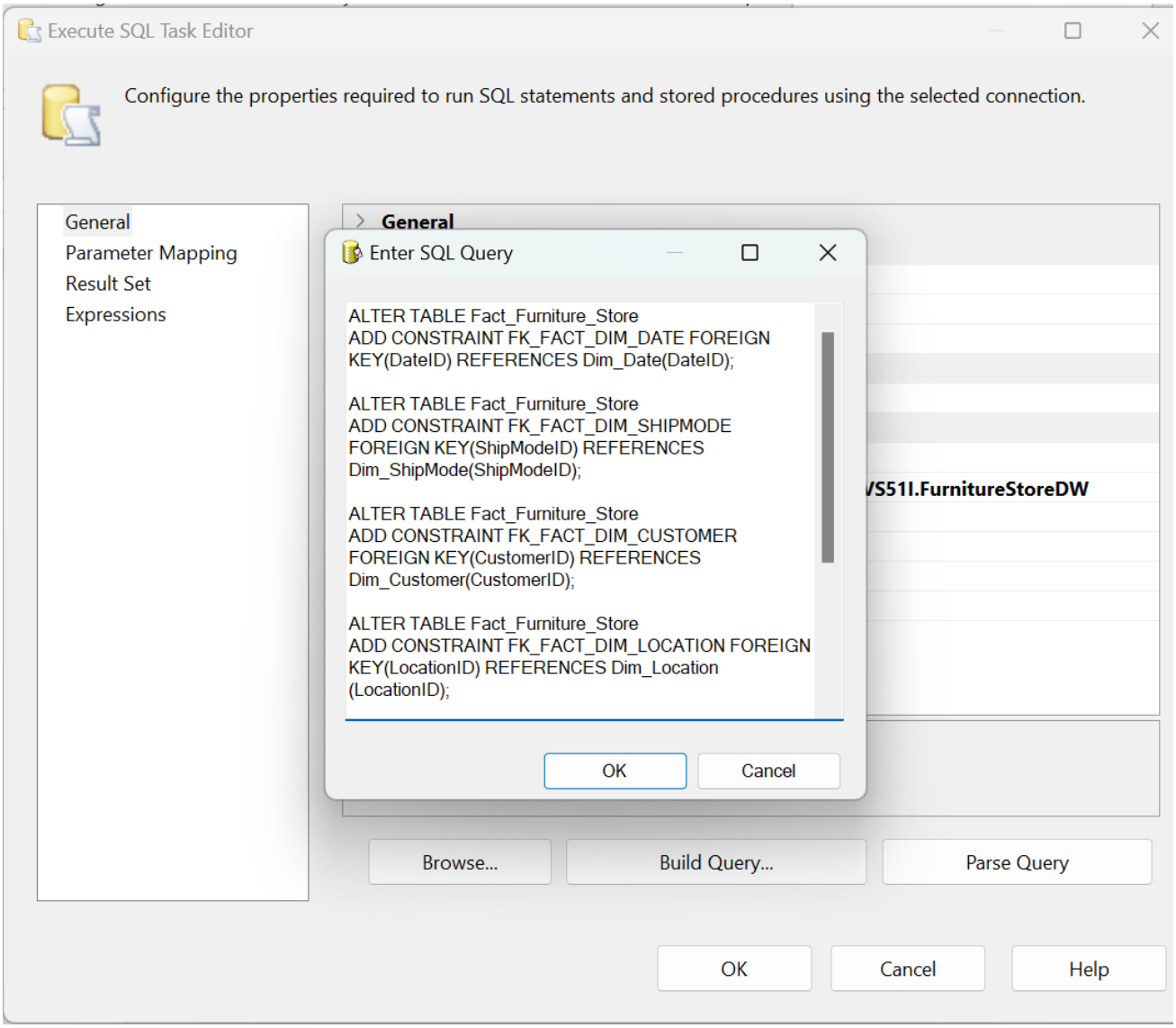
Câu lệnh SQL:

ALTER TABLE **Fact_Furniture_Store**
ADD CONSTRAINT FK_FACT_DIM_DATE FOREIGN KEY(DateID) REFERENCES **Dim_Date**(DateID);

ALTER TABLE **Fact_Furniture_Store**
ADD CONSTRAINT FK_FACT_DIM_SHIPMODE FOREIGN KEY(ShipModeID) REFERENCES **Dim_ShipMode**(ShipModeID);

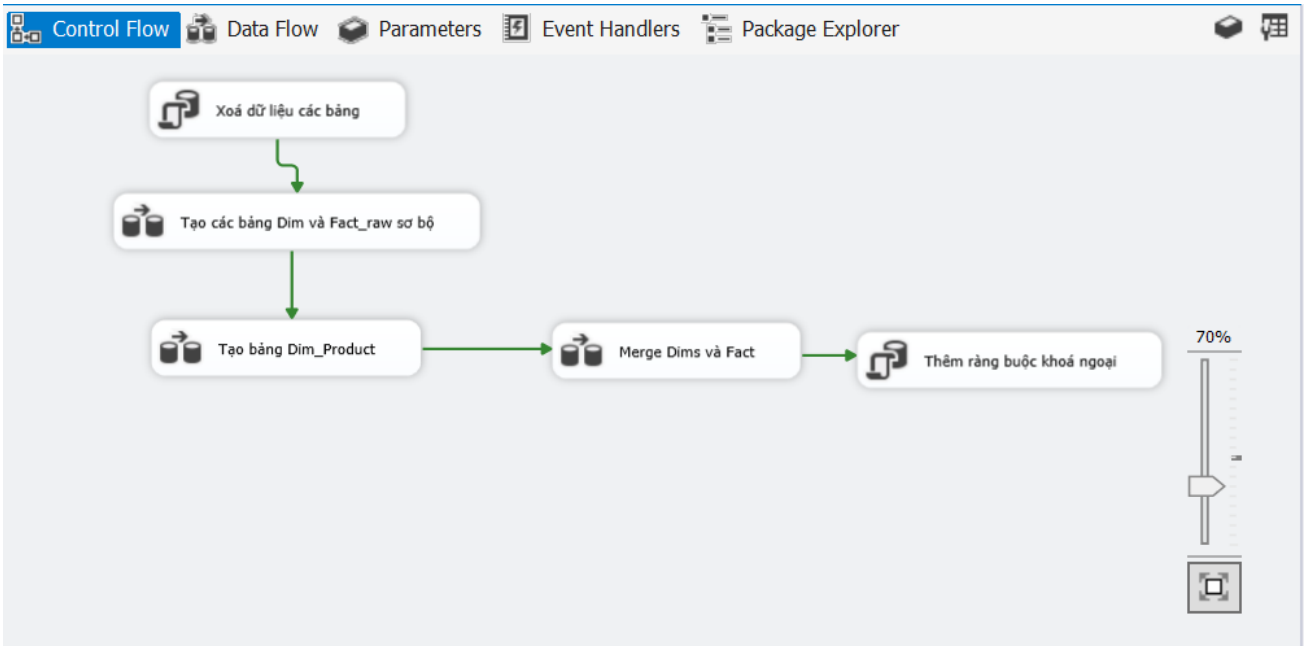
ALTER TABLE **Fact_Furniture_Store**

```
ADD CONSTRAINT FK_FACT_DIM_CUSTOMER FOREIGN KEY(CustomerID) REFERENCES Dim_Customer(CustomerID);

ALTER TABLE Fact_Furniture_Store
ADD CONSTRAINT FK_FACT_DIM_LOCATION FOREIGN KEY(LocationID) REFERENCES Dim_Location(LocationID);

ALTER TABLE Fact_Furniture_Store
ADD CONSTRAINT FK_FACT_DIM_PRODUCT FOREIGN KEY(ProductID) REFERENCES Dim_Product(ProductID);

ALTER TABLE Dim_Product
ADD CONSTRAINT FK_DIM_PRODUCT_DIM_CATEGORY FOREIGN KEY(CategoryID) REFERENCES Dim_Category(CategoryID);
```



Hình 2.96 Dán câu lệnh tạo khoá ngoại

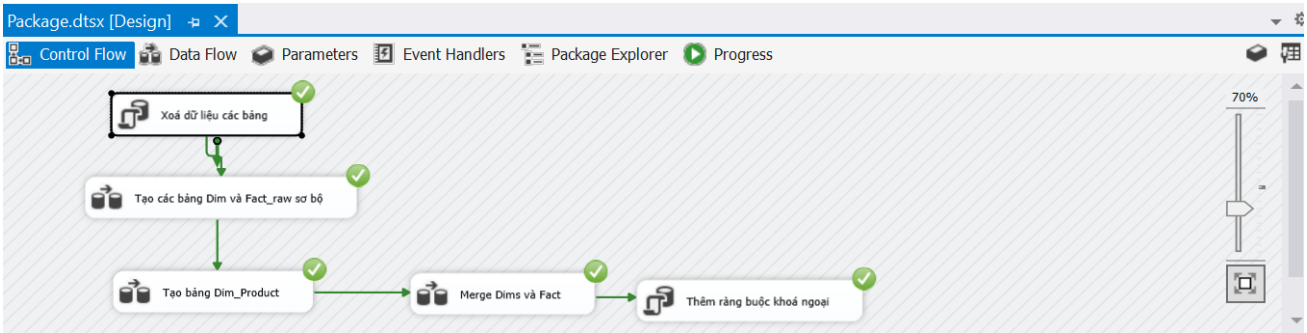
Bước 4: Nhấn **OK** để hoàn thành thiết lập và kéo từ thành phần **Merge Dims** và Fact tới **Thêm ràng buộc khoá ngoại** để tạo luồng dữ liệu.



Hình 2.97 Tổng quan quá trình chạy

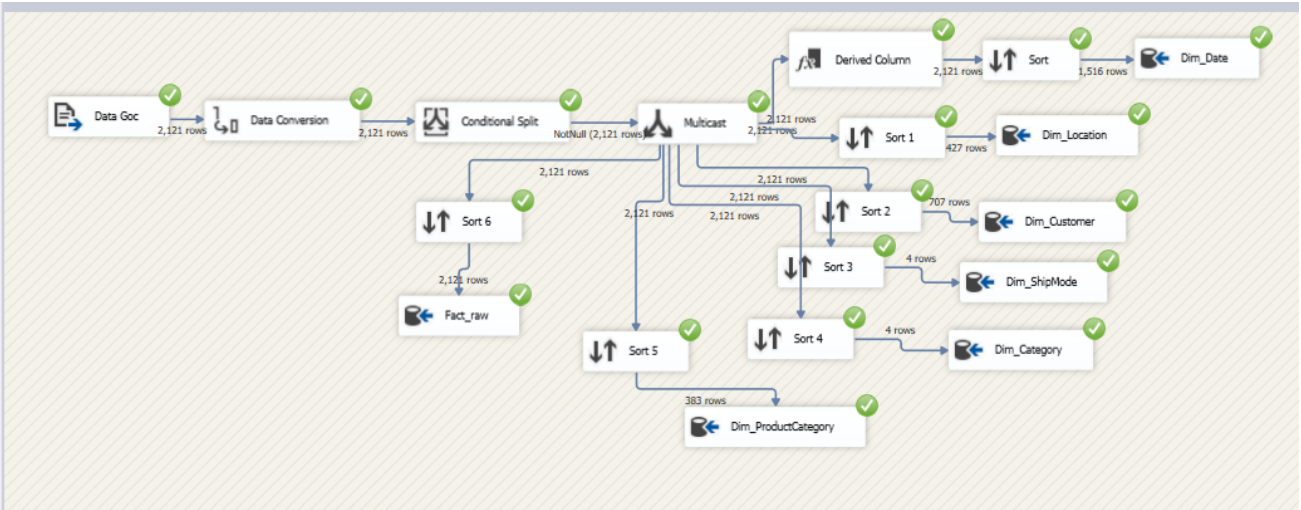
3. Tiến hành chạy dữ liệu

3.1. Kết quả chạy dữ liệu cho dự án



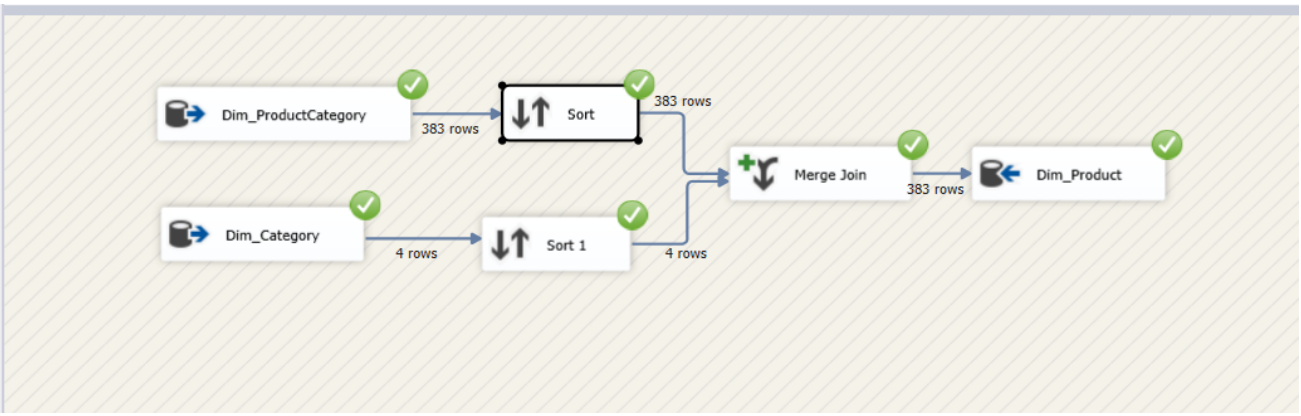
Hình 2.98 Kết quả quá trình chạy(1)

3.2. Tiền xử lý dữ liệu, tạo các bảng Dim và Fact



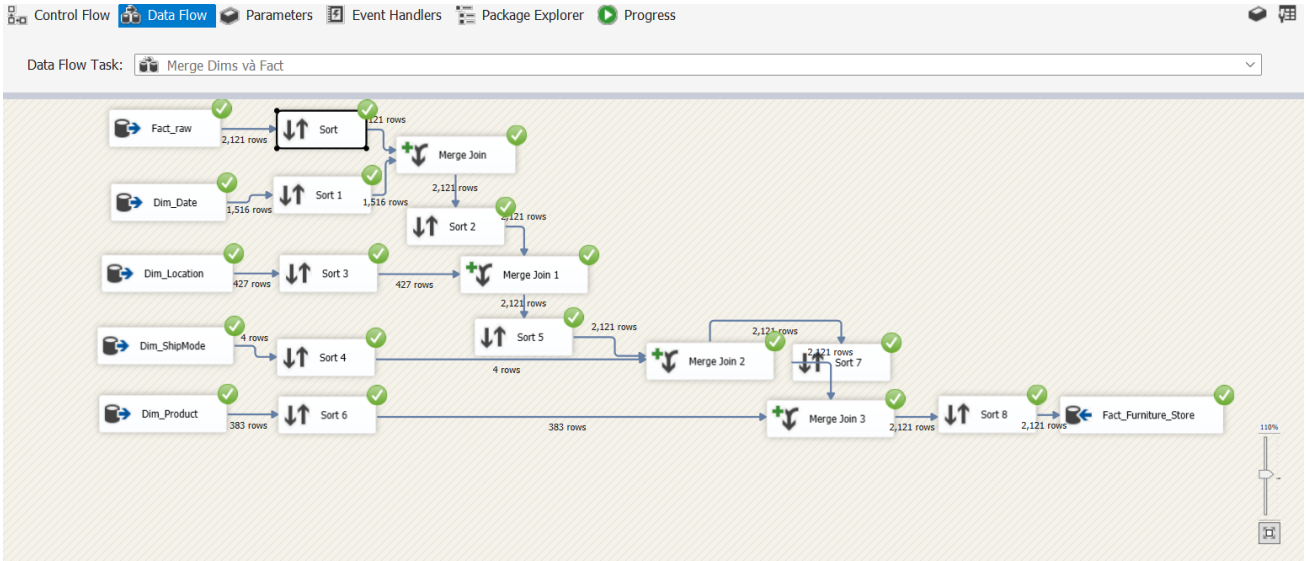
Hình 2.99 Kết quả quá trình chạy(2)

3.3. Tạo bảng Dim_Product



Hình 2.100 Kết quả quá trình chạy(3)

3.4. Merge Dims và Fact



Hình 2.101 Kết quả quá trình chạy(4)

4. Kết quả cơ sở dữ liệu

- Kết quả trong cơ sở dữ liệu tại bảng **Fact_Furniture_Store**:

Results | Messages

	RowID	OrderID	LocationID	ShipModeID	DateID	CustomerID	ProductID	Sales	Quantity	Discount	Profit
1	1	CA-2016-152156	163	3	941	CG-12520	41	261.959991455078	2	0	41.9136009216309
2	2	CA-2016-152156	163	3	941	CG-12520	107	731.940002441406	3	0	219.582000732422
3	4	US-2015-108966	131	4	526	SO-20335	364	957.577514648438	5	0.449999988079071	-383.031005859375
4	6	CA-2014-115812	221	4	96	BH-11710	308	48.8600006103516	7	0	14.1694002151489
5	11	CA-2014-115812	221	4	96	BH-11710	356	1706.18395996094	9	0.200000002980232	85.309196472168
6	24	US-2017-156909	309	3	1234	SF-20065	66	71.3720016479492	2	0.300000011920929	-1.01960003376007
7	25	CA-2015-106320	292	4	515	EB-13870	364	1044.63000488281	3	0	240.264892578125
8	28	US-2015-150630	309	4	500	TB-21520	15	3083.42993164063	7	0.5	-1665.05224609375
9	30	US-2015-150630	309	4	500	TB-21520	207	124.199996948242	3	0.200000002980232	15.5249996185303
10	37	CA-2016-117590	330	1	994	GH-14485	158	190.919998168945	5	0.600000023841858	-147.962997436523
11	39	CA-2015-117415	173	4	631	SN-20710	42	532.399169921875	3	0.319999992847443	-46.9763984680176

Query executed successfully. | DESKTOP-0GVS51I (15.0 RTM) | DESKTOP-0GVS51I\msi (61) | FurnitureStoreDW | 00:00:00 | 1,000 rows

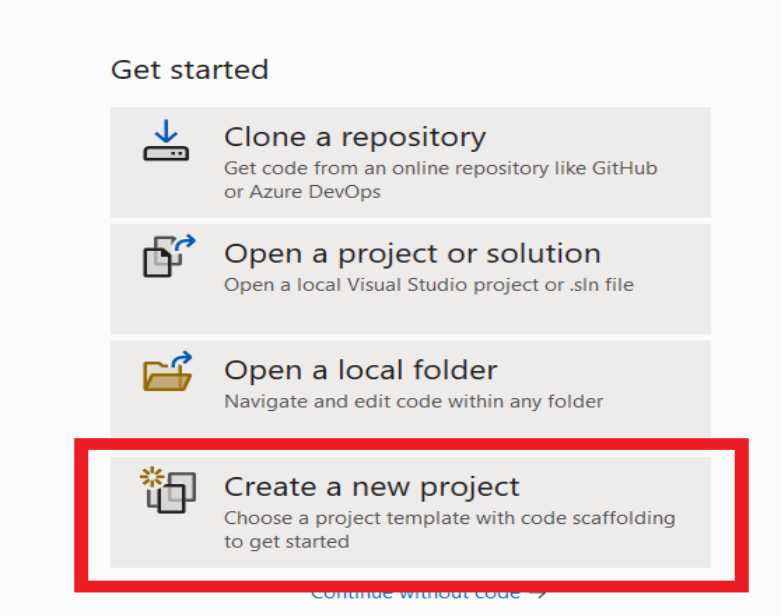
Hình 2.102 Dữ liệu đổ vào cơ sở dữ liệu

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

1. TẠO PROJECT SSAS TRONG VISUAL STUDIO 2022

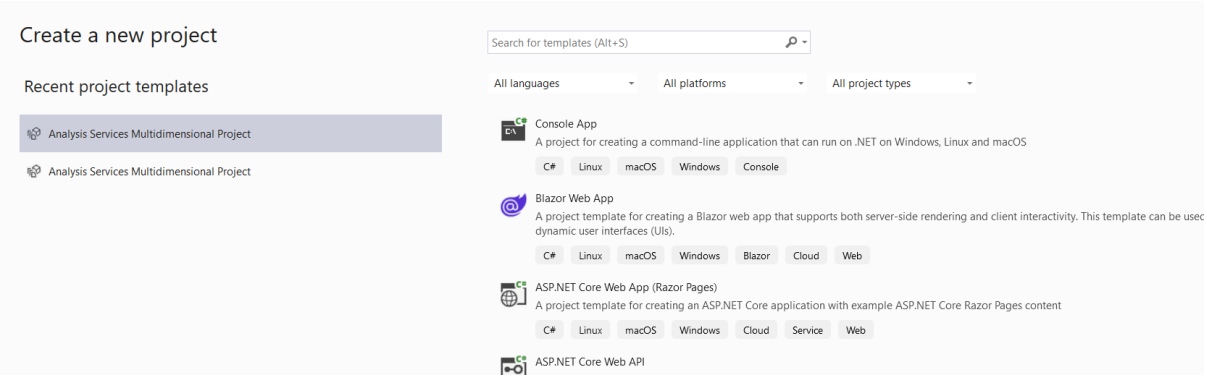
1.1. Tạo mới Project SSAS trong Visual Studio

Bước 1: Vào Visual Studio, chọn Create a new project



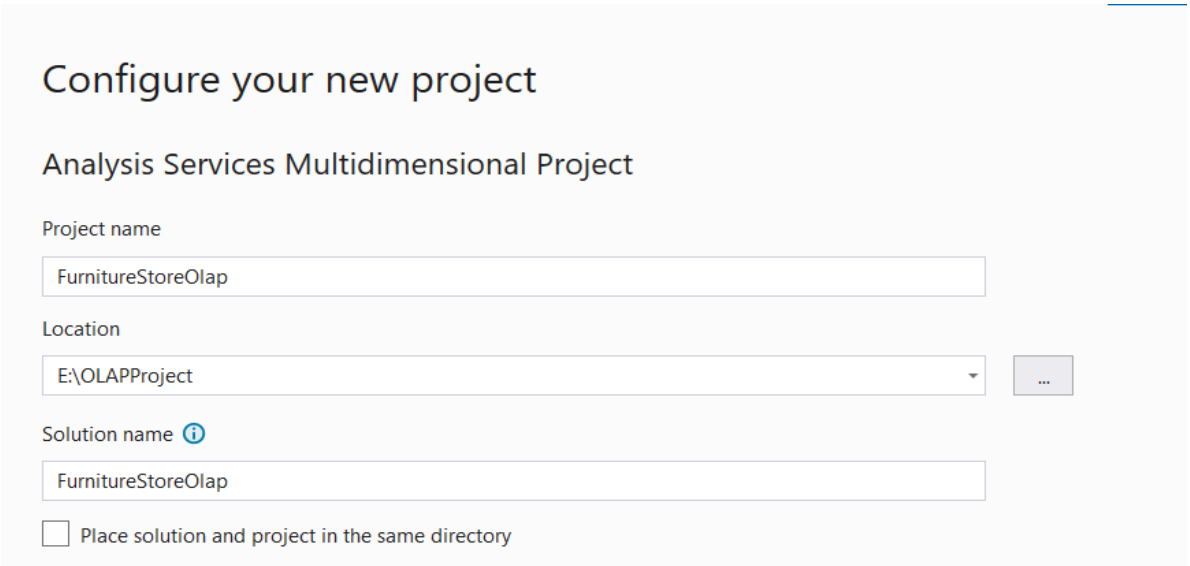
Hình 3. 1 Tạo Project mới bằng Visual Studio 2022

Bước 2: Chọn Analysis Multidimensional Project



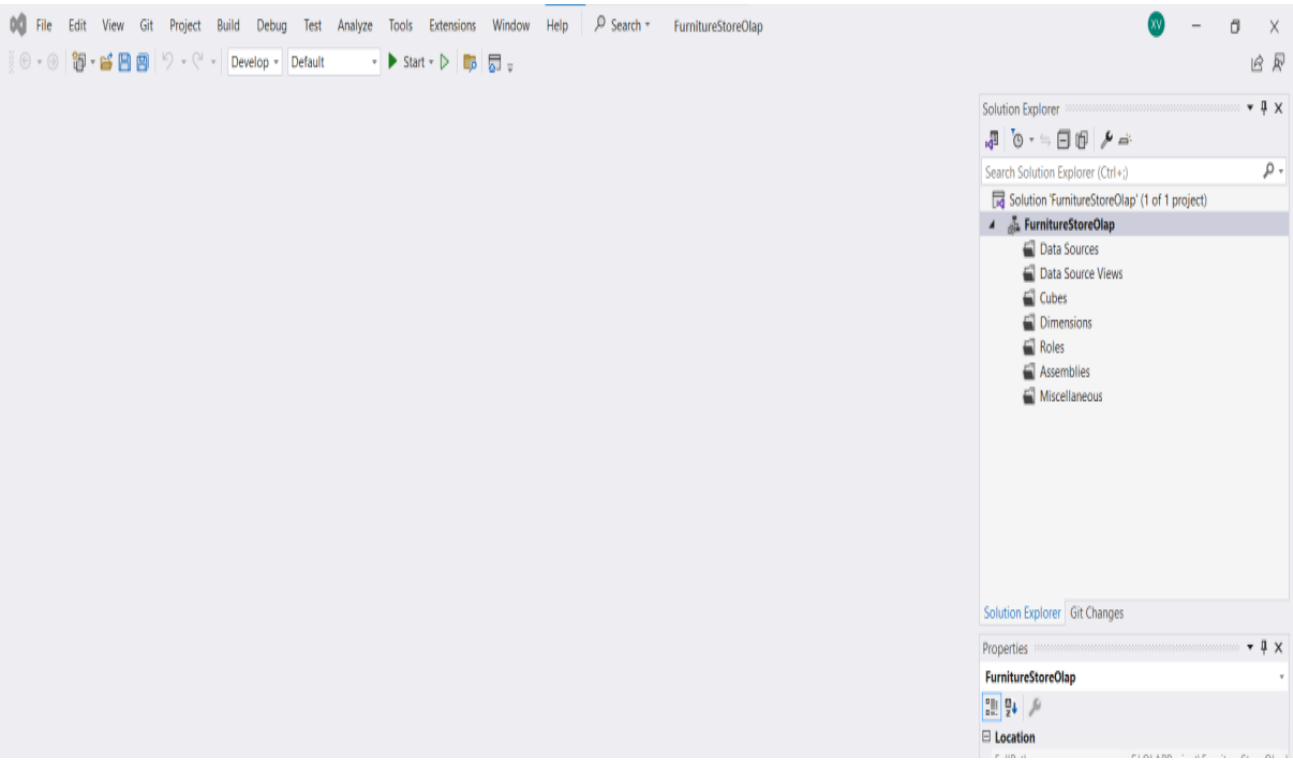
Hình 3.2 Chọn Analysis Multidimensional Project

Bước 3: Đặt tên cho Project và nhấn Create.



Hình 3.3 Đặt tên cho Project

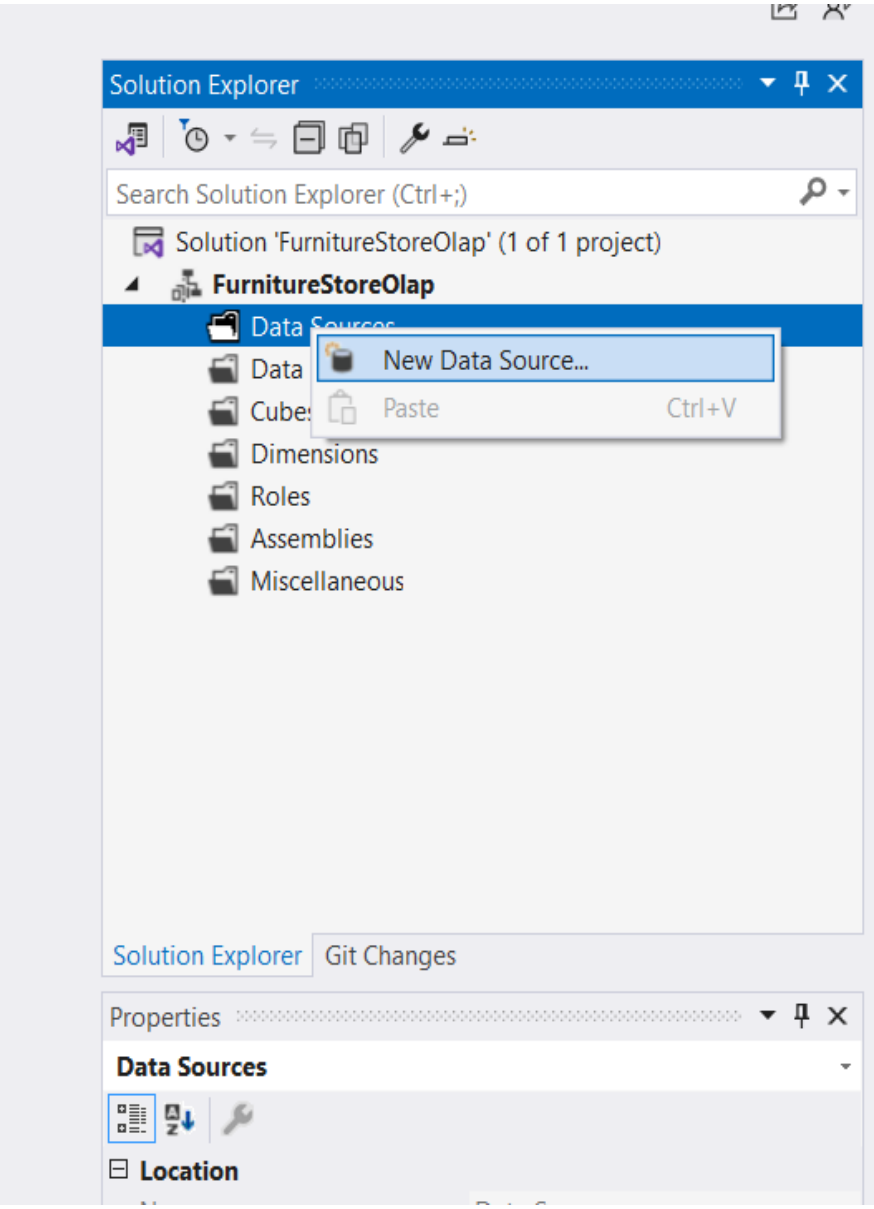
Project SSAS sau khi đã tạo có sẽ chuyển đến màn hình như sau:



Hình 3.4 Project SSAS sau khi đã tạo

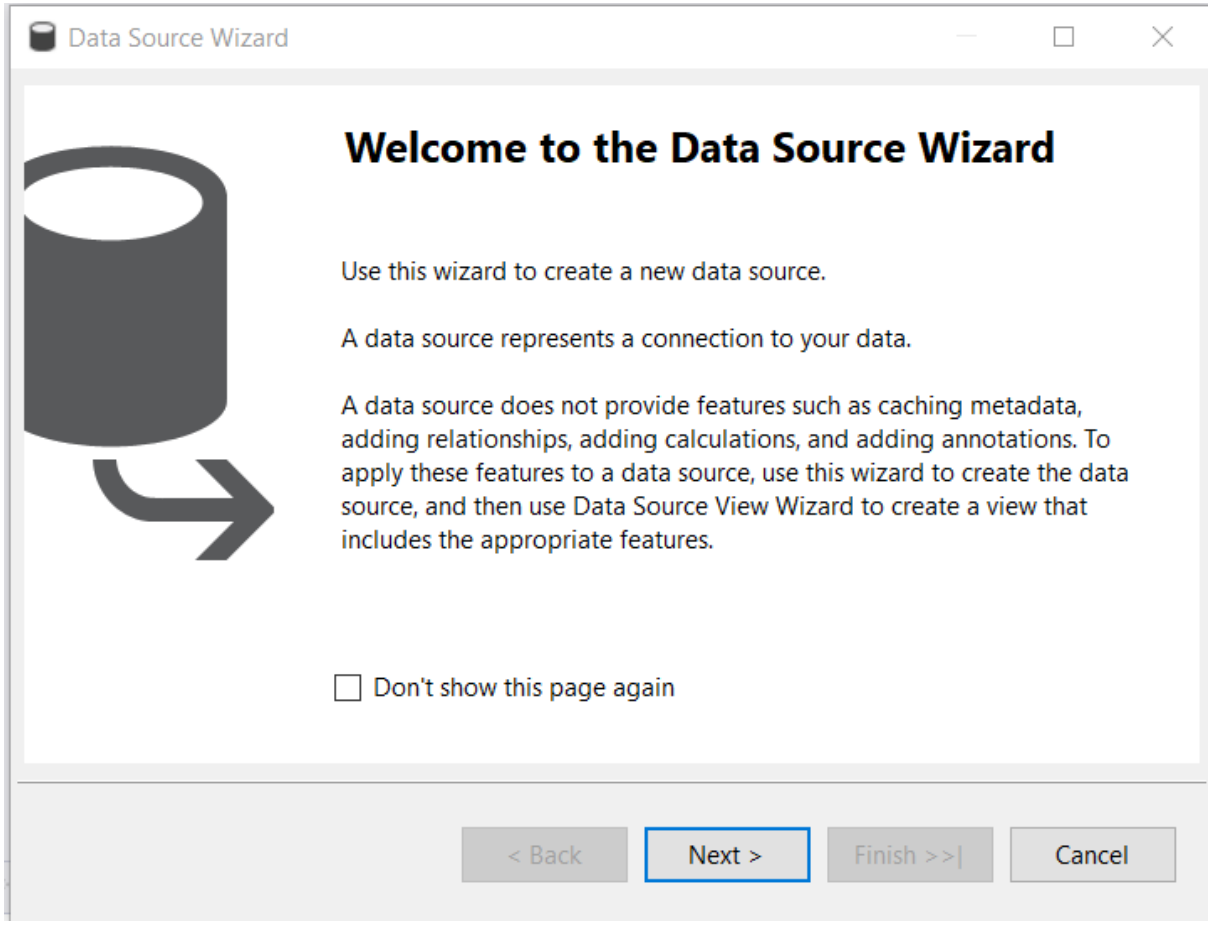
1.2. Tạo DataSource

Bước 1: Click chuột phải vào “Data Souce” chọn “New Data Source”



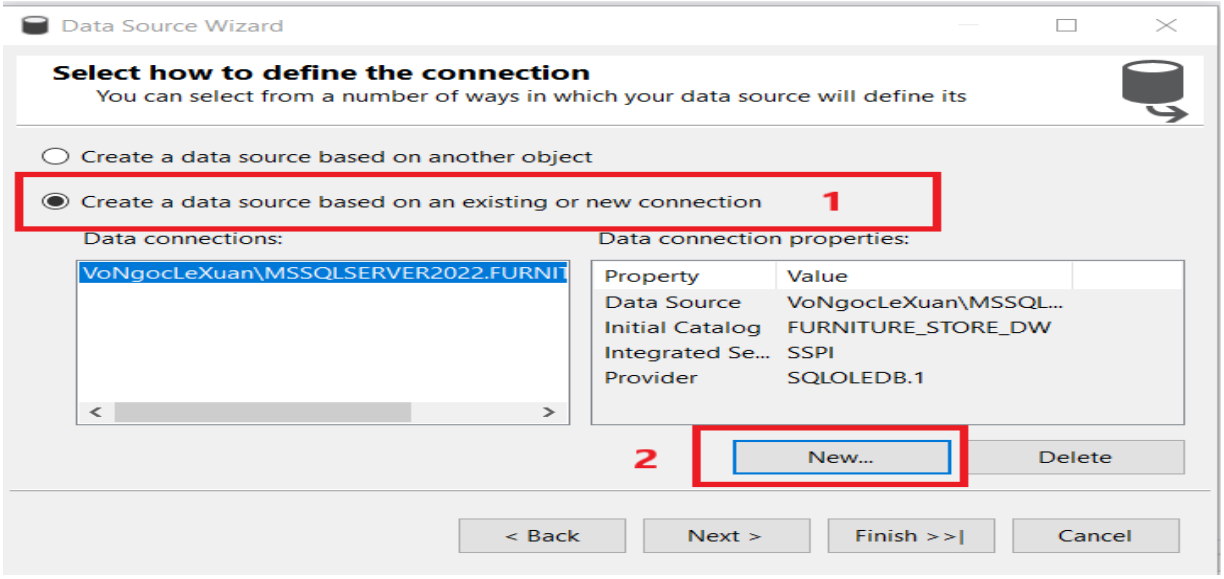
Hình 3.5 Tạo mới Data Source

Bước 2: Chọn "Next"



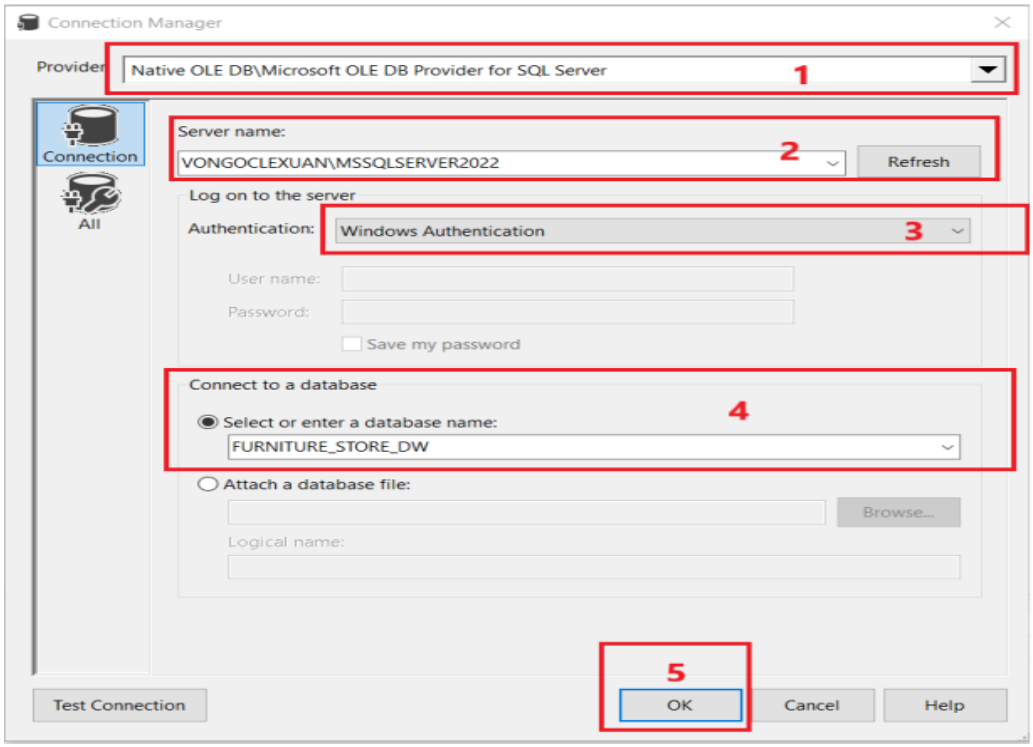
Hình 3.6 Màn hình tạo mới Data Source

Bước 3: Chọn **"Create a data source based on an existing or new connection"** sau đó chọn **"New"**



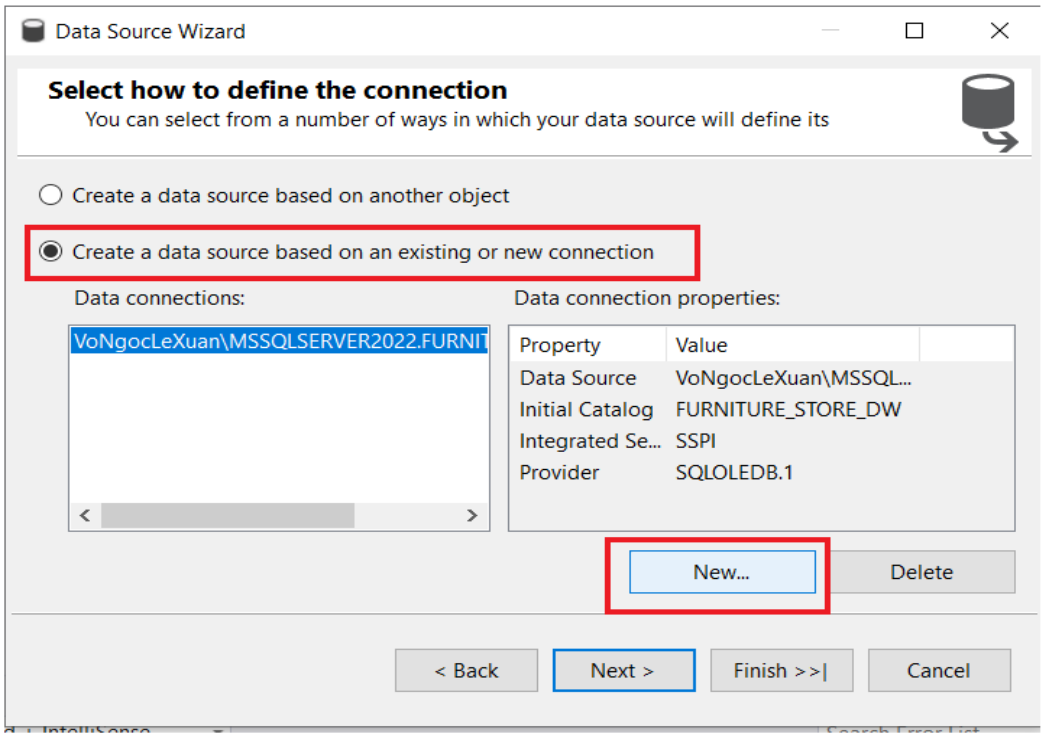
Hình 3.7 Chọn loại Connection để tạo Data Source

Bước 4: Cấu hình Provider sang **"Microsoft OLE DB Provider for SQL Server"**, chuyển Server name sang tên server của SQL Server, Authentication của Windows Authentication và tên database của kho dữ liệu sau, sau đó nhấn **"Ok".**



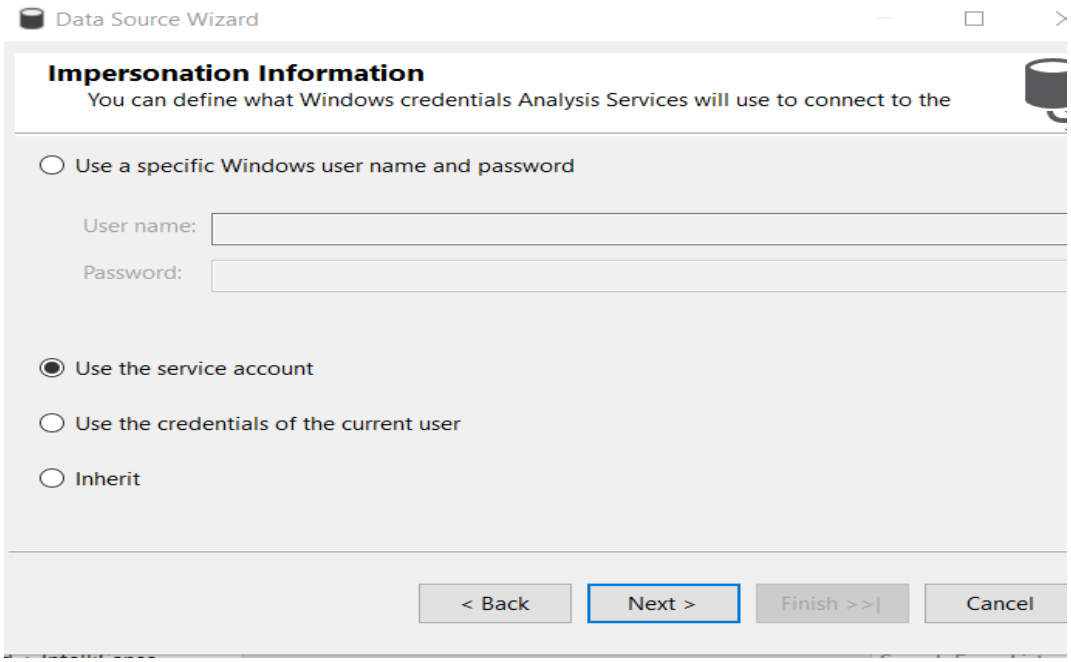
Hình 3.8 Cấu hình Provider, Server name, Authentication, tên Database

Bước 5: Chọn "Create a data source based on an existing or new connection" sau đó chọn "New".



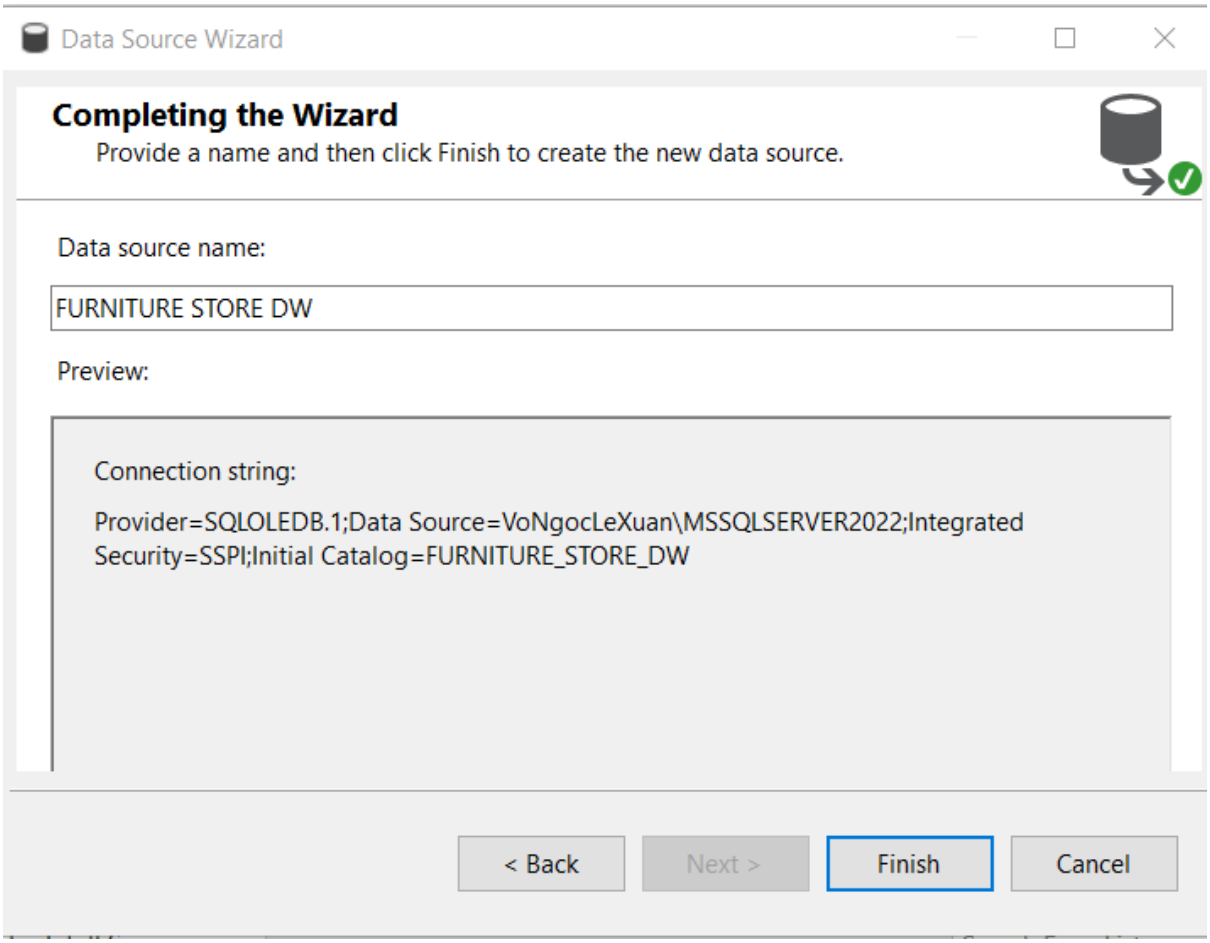
Hình 3.9 Chọn loại Connection để tạo Data Source

Bước 6: Chọn **"Use the service acccount"** và sau đó chọn **"Next"**



Hình 3.10 Cấu hình Impersonal Information

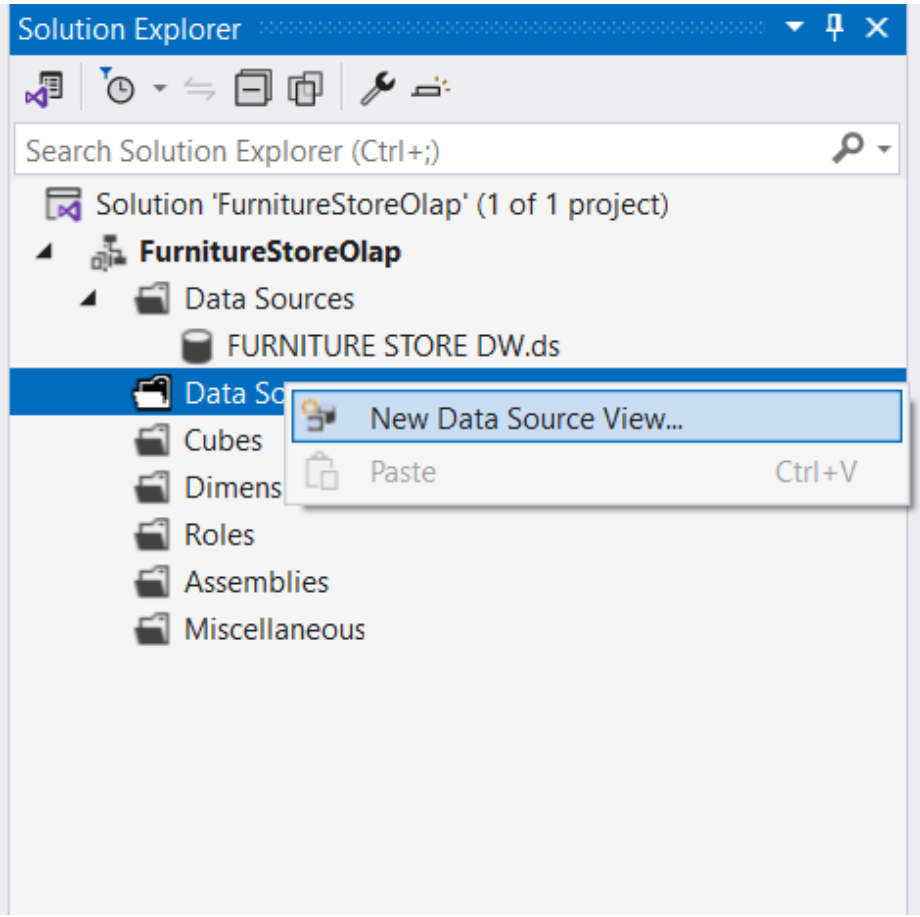
Bước 7: Đặt tên cho Data Source sau đó nhấn **"Finish"**



Hình 3.11 Đặt tên cho Data Source

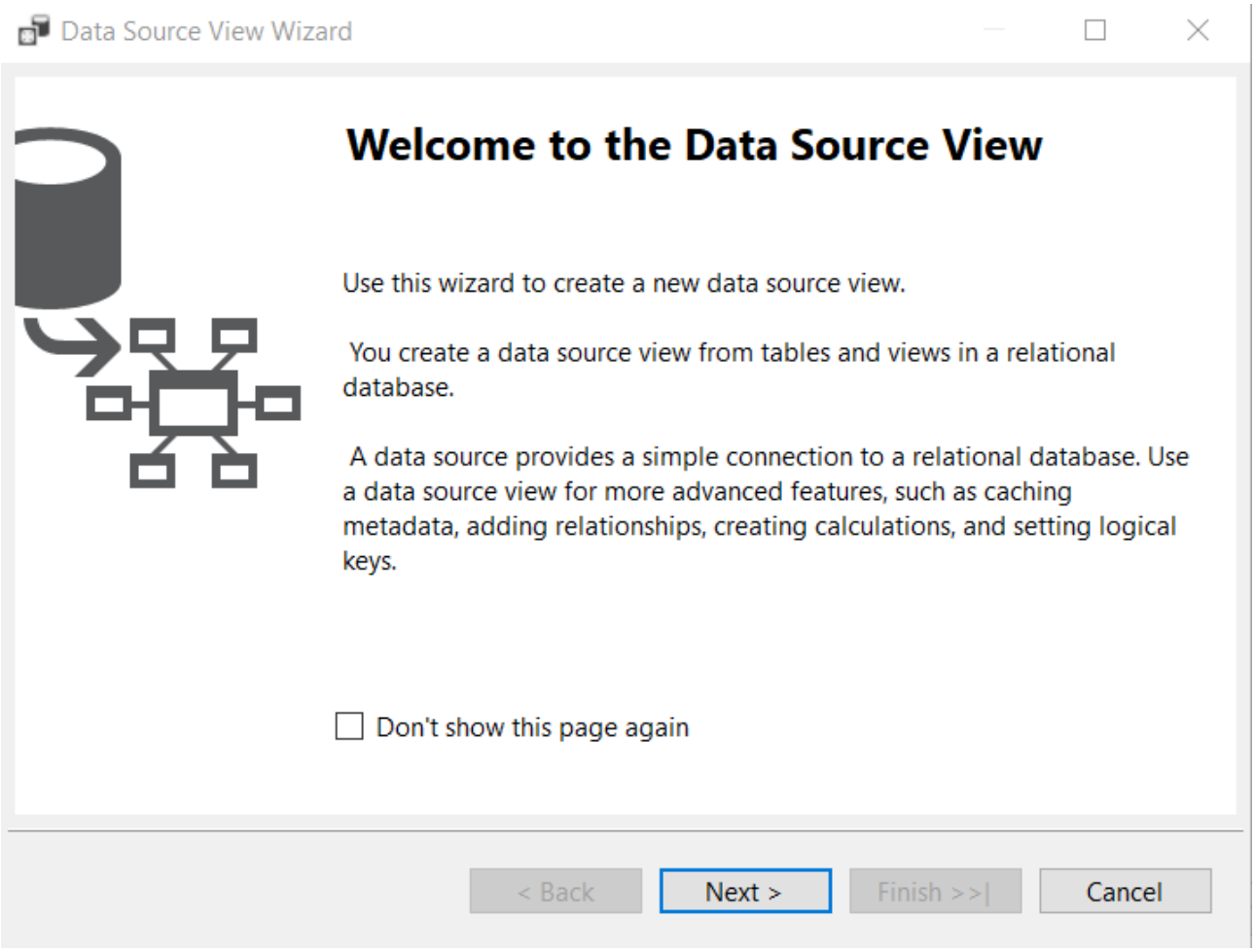
1.3. Tạo Data Source Views

Bước 1: Ở mục **Data Source View**, chọn "**New Data Source View".**



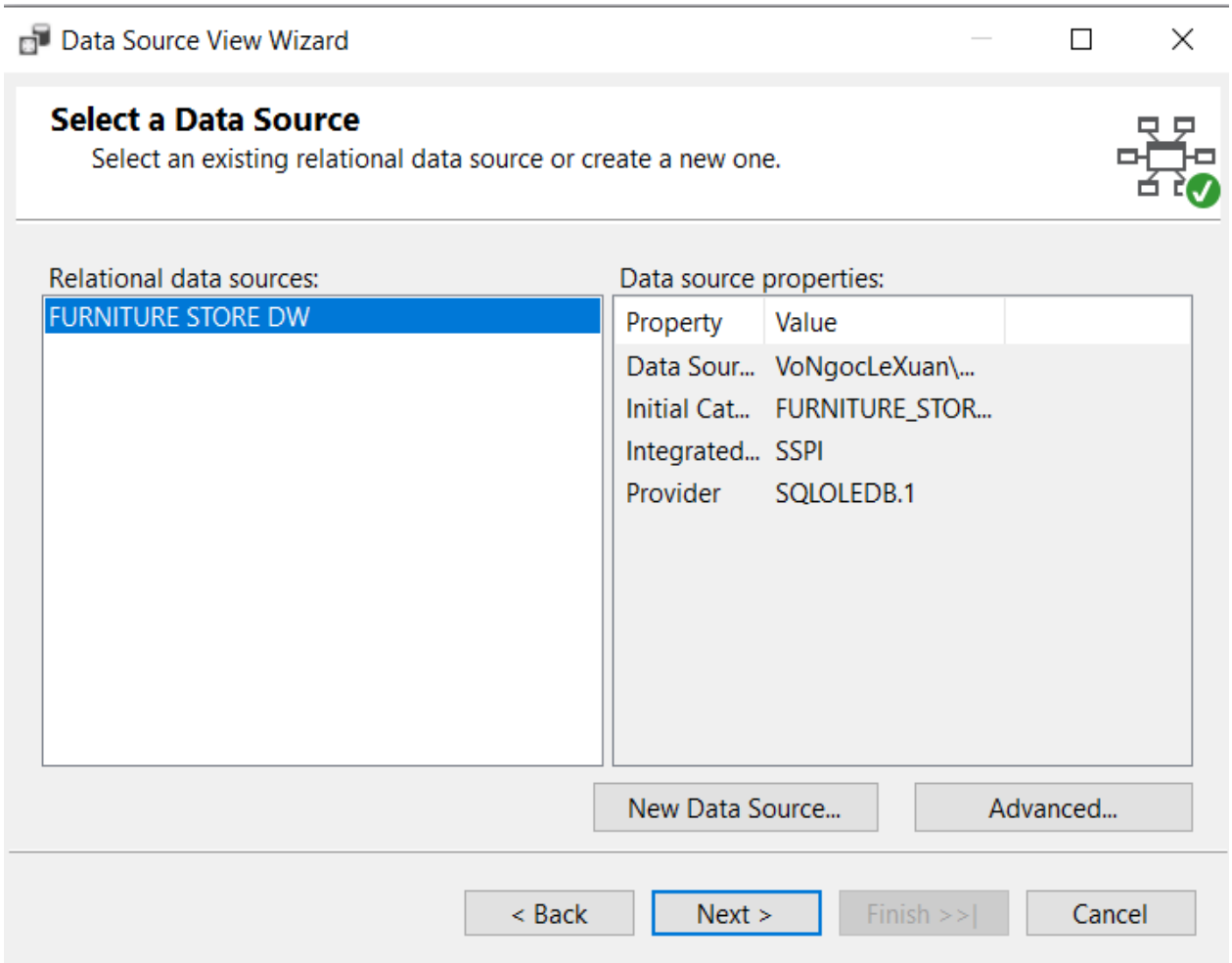
Hình 3. 12 Tạo Data Source View

Bước 2: Ở cửa sổ welcome pop-up, chọn “Next”.



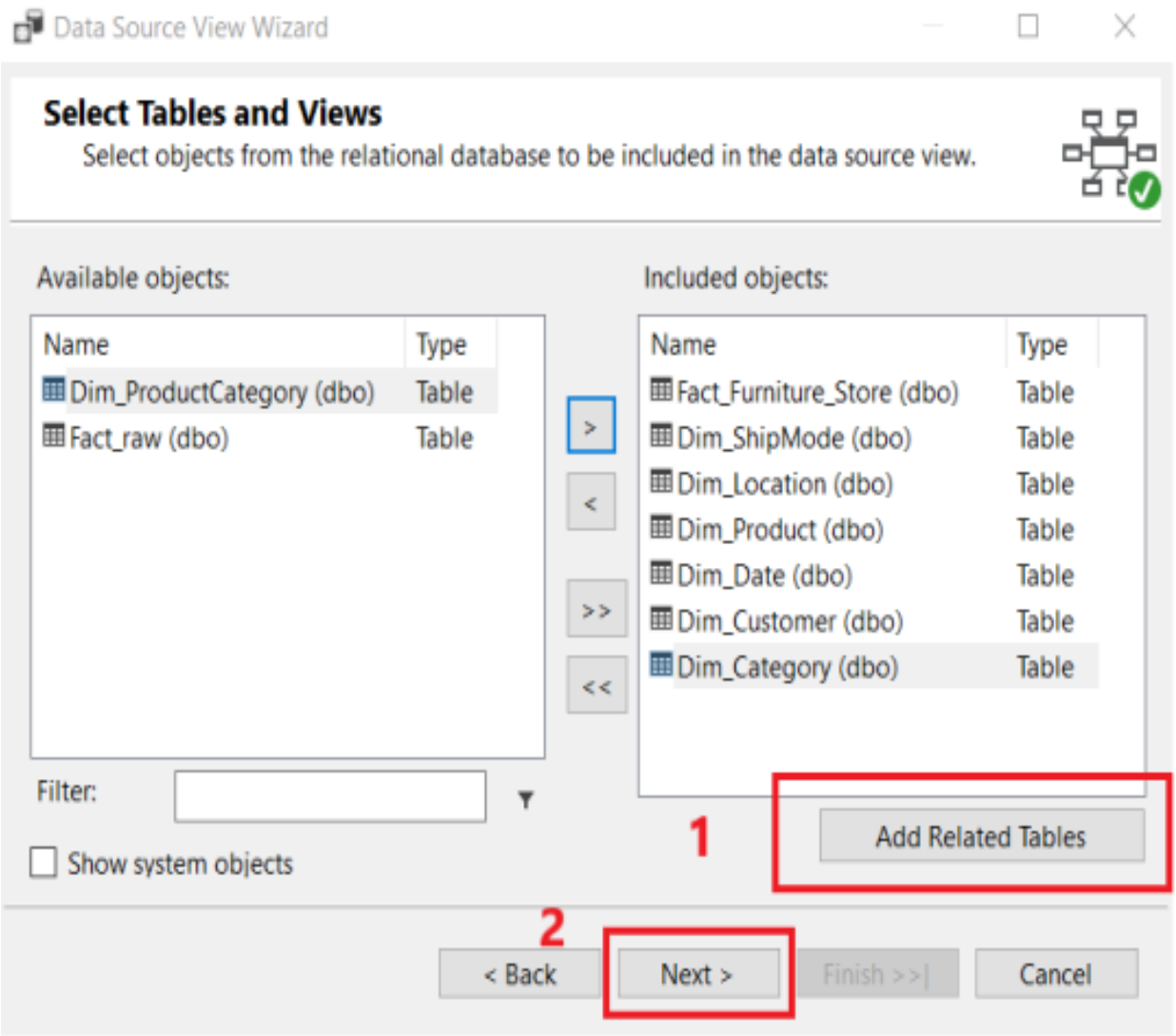
Hình 3. 13 Màn hình tạo Data Source View

Bước 3: Chọn kho dữ liệu **Furniture Store DW** và nhấn "Next"



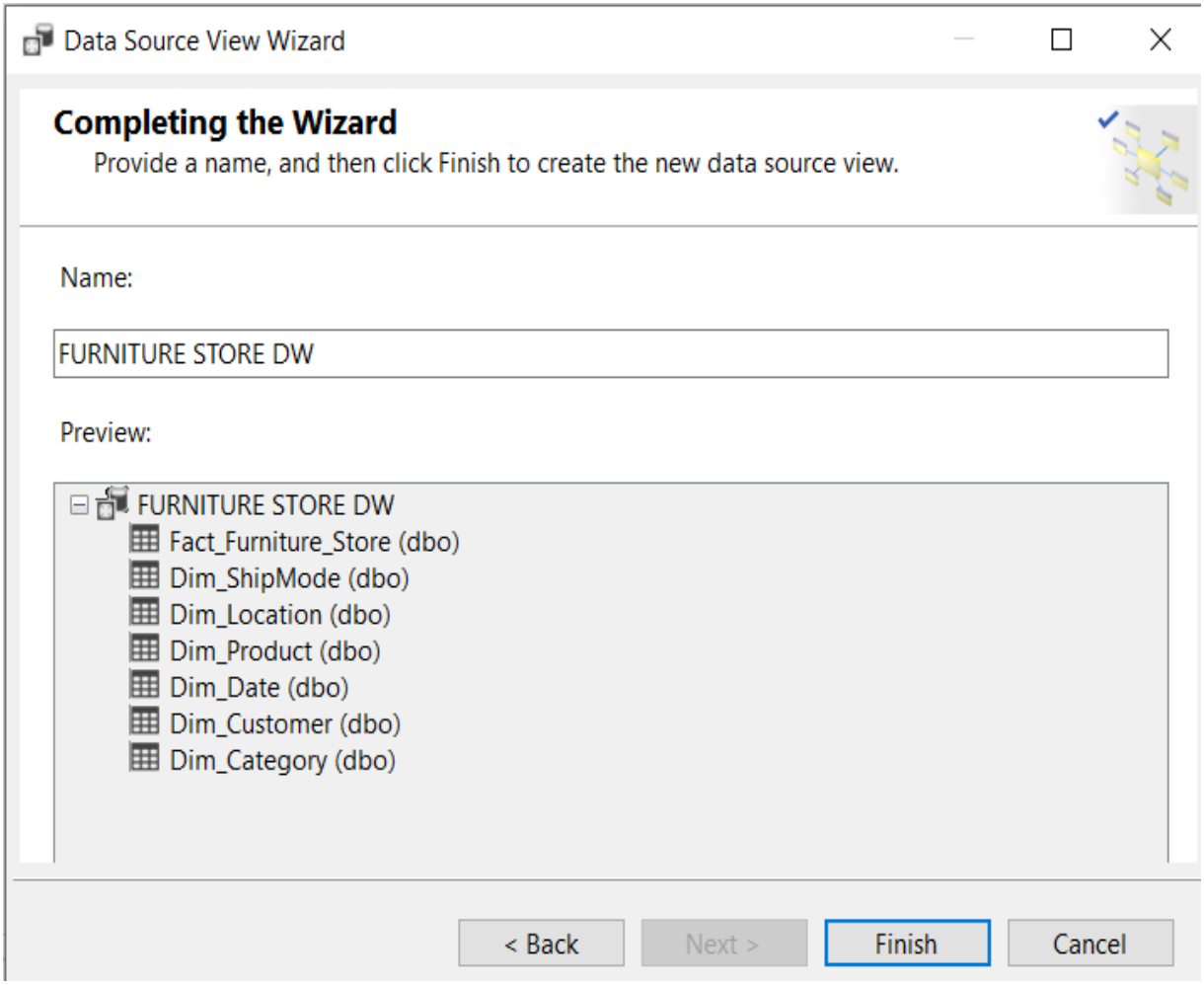
Hình 3.14 Chọn kho dữ liệu cho Data Source

Bước 4: Chọn bảng Fact và Dim cho quá trình phân tích sau đó nhấn "Next".



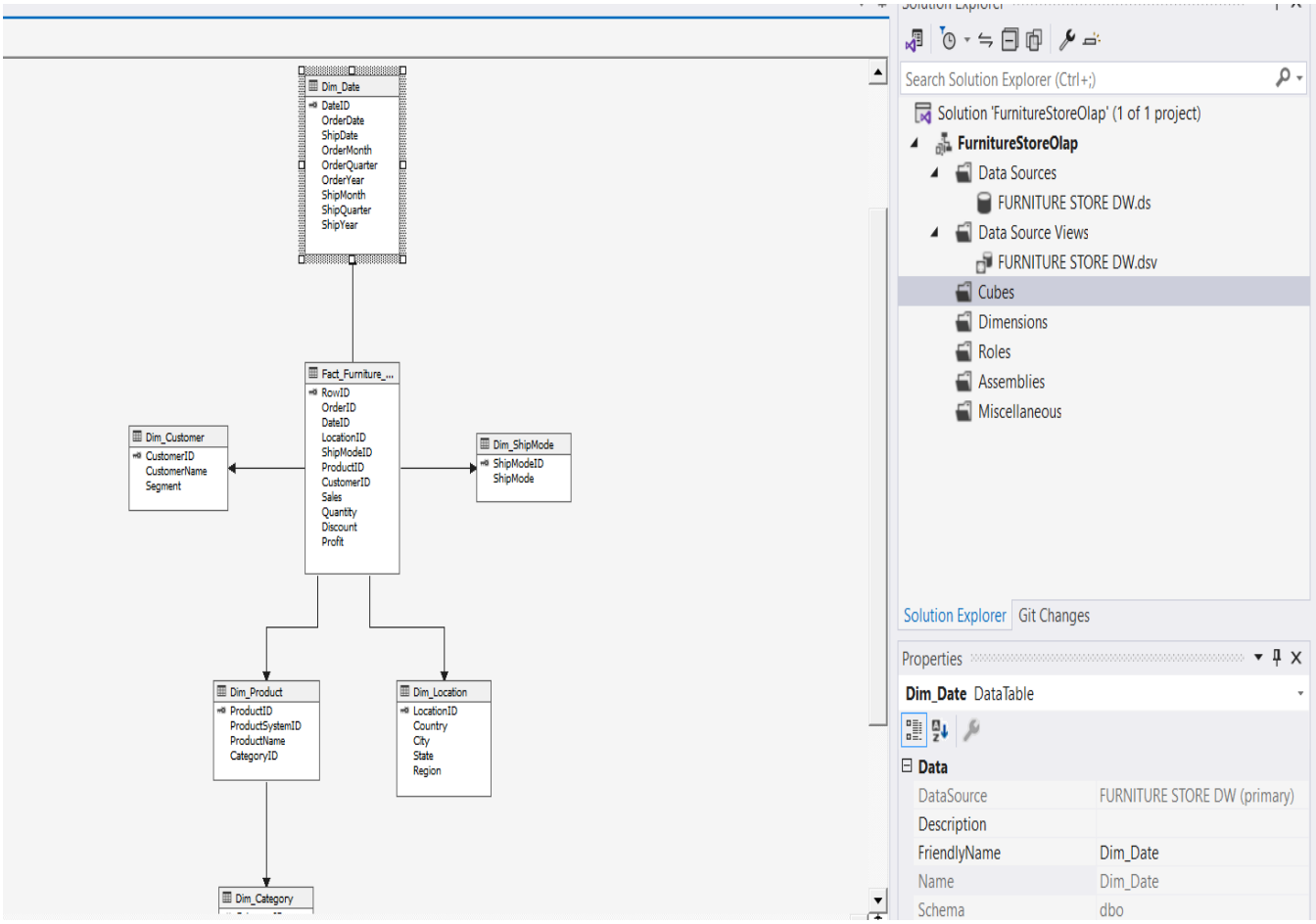
Hình 3.15 Chọn bảng Fact và Dim cho quá trình phân tích

Bước 5: Xem lại các bảng Fact và Dim sau đó nhấn **"Finish"** để hoàn thành quá trình tạo Data Source View



Hình 3.16 Kiểm tra bảng Fact và Dim

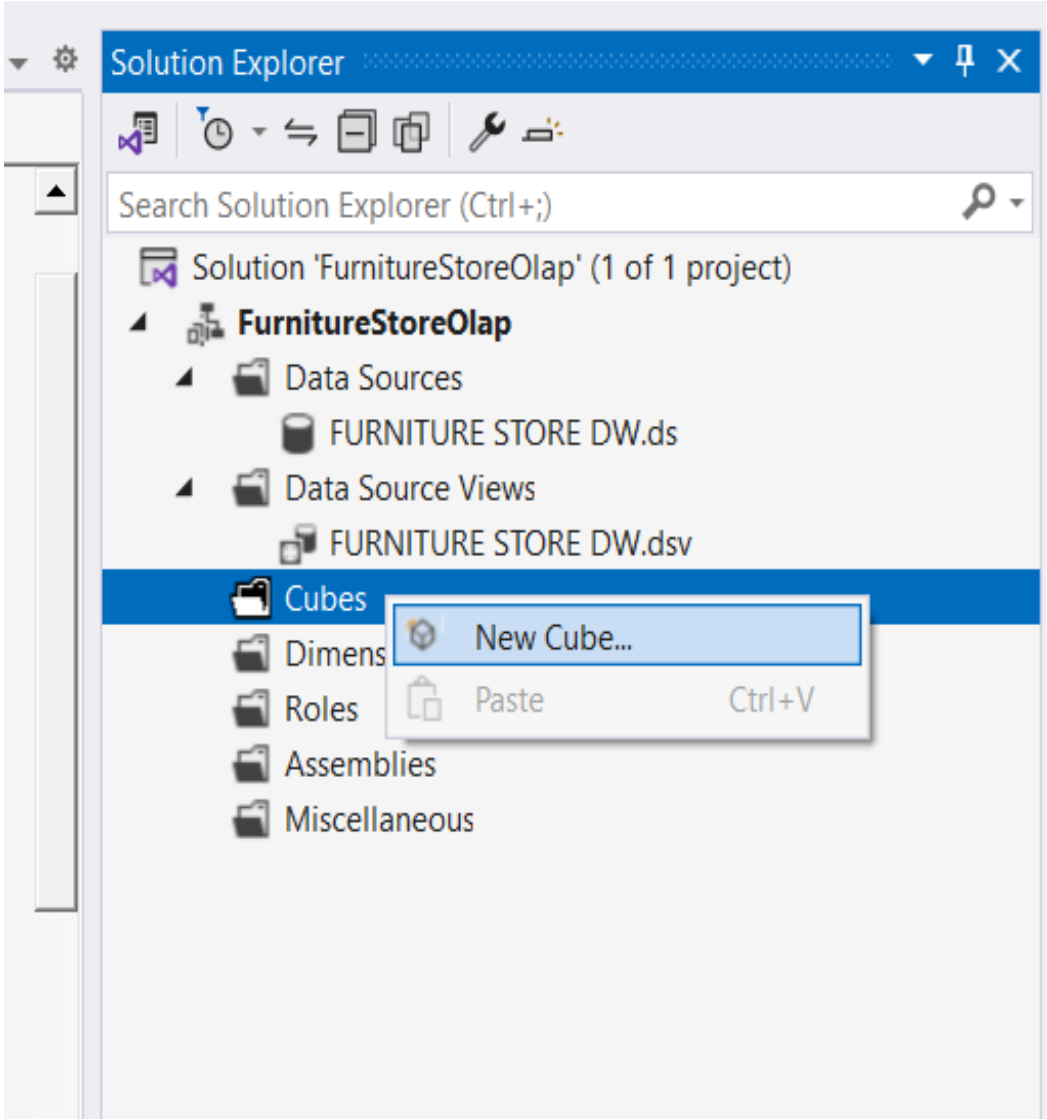
Bước 6: Sau khi hoàn tất việc tạo Data Source View, Visual Studio sẽ chuyển đến ảnh như sau:



Hình 3. 17 Màn hình sau khi hoàn tất tạo Data Source

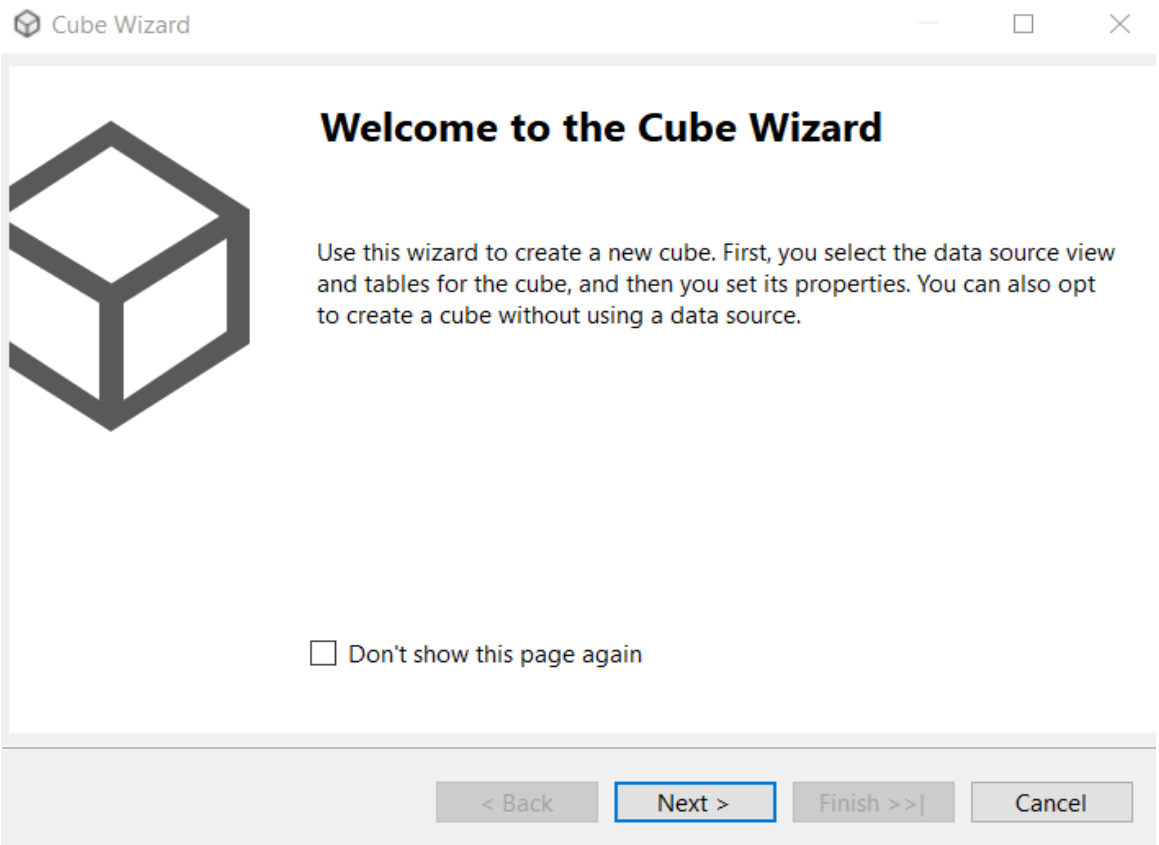
1.4. Xác định các chiều và độ đo cho kho dữ liệu

Bước 1: Ở mục **Cubes**, chọn "New Cube"



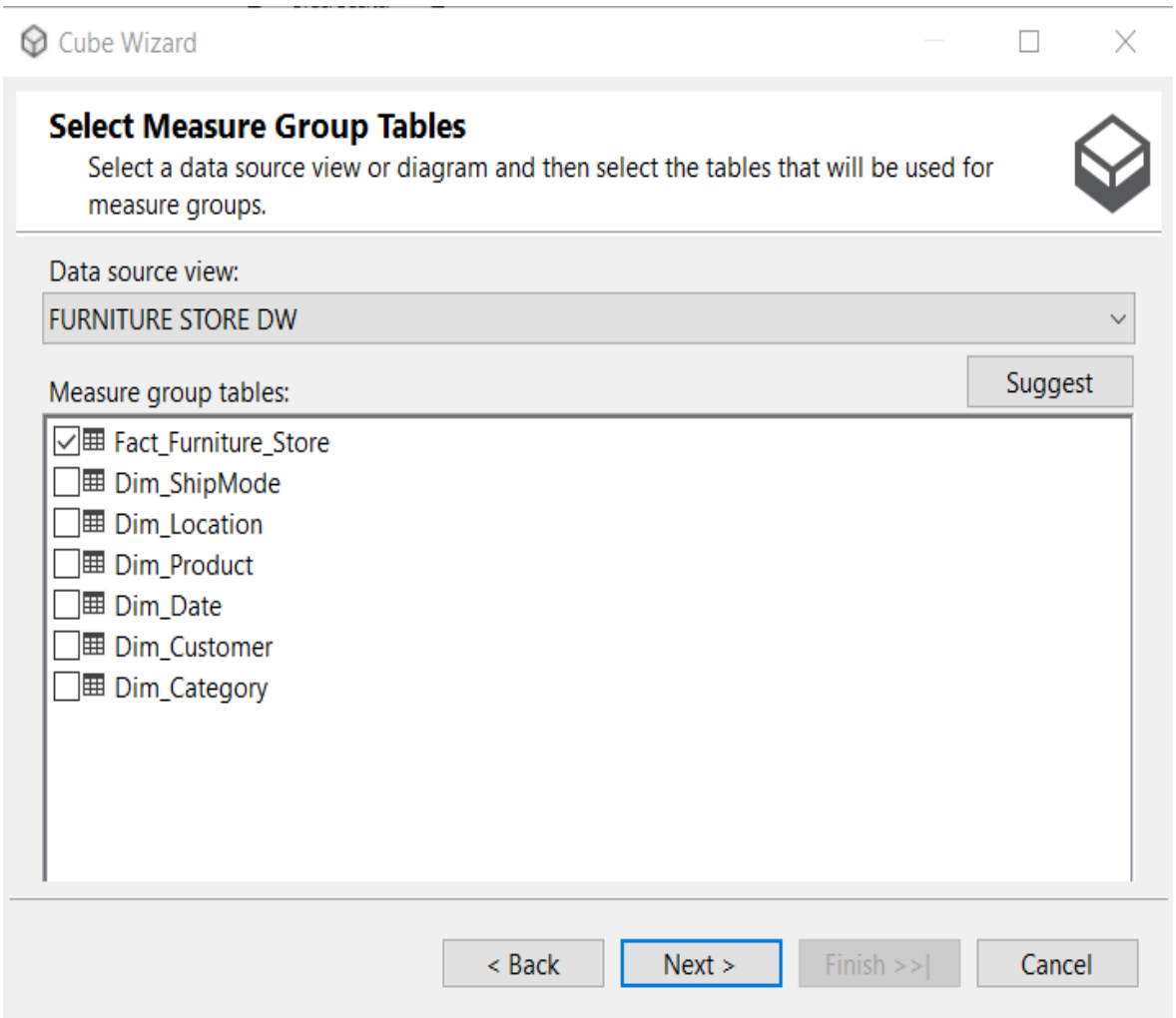
Hình 3.18 Tạo Cube

Bước 2: Ở welcome pop-up, chọn “Next”



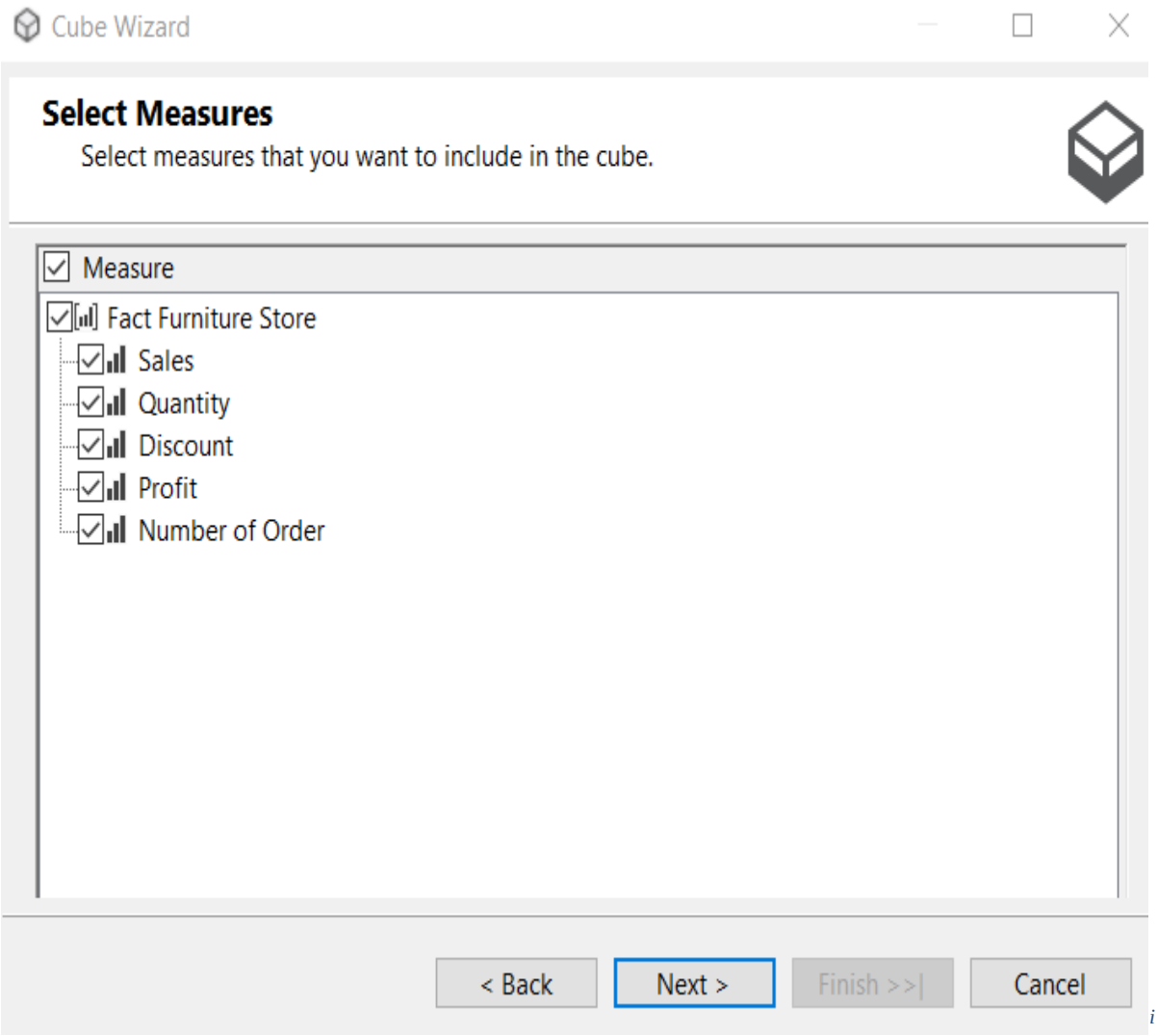
Hình 3. 19 Màn hình tạo Cube

Bước 3: Chọn Bảng **Fact_Furniture_Store** chứa độ đo của kho dữ liệu.



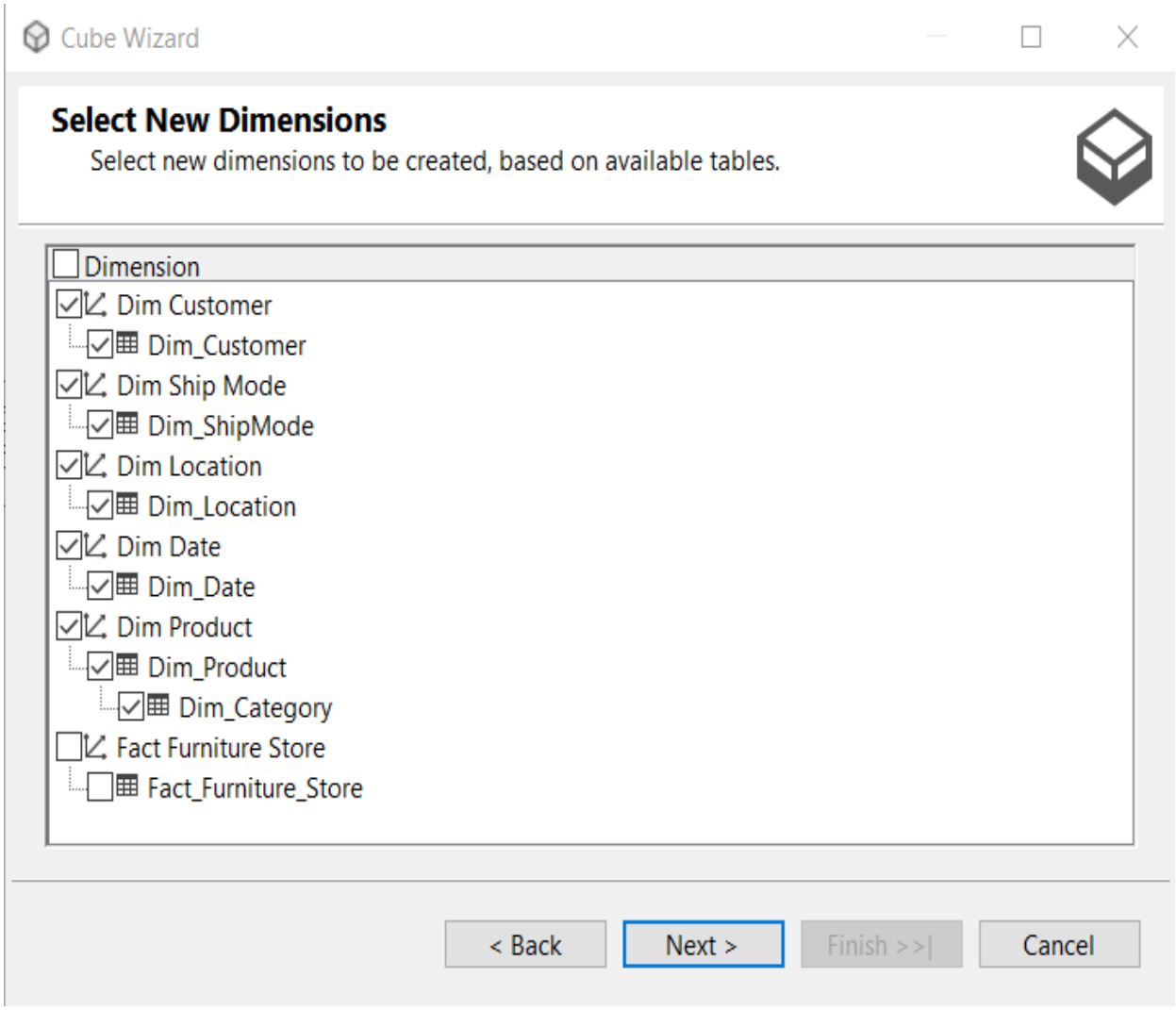
Hình 3.20 Chọn bảng Fact để tạo Cube

Bước 4: Ở bước này, có thể chính sửa tên của các độ đo theo ý muốn sau đó chọn "Next".



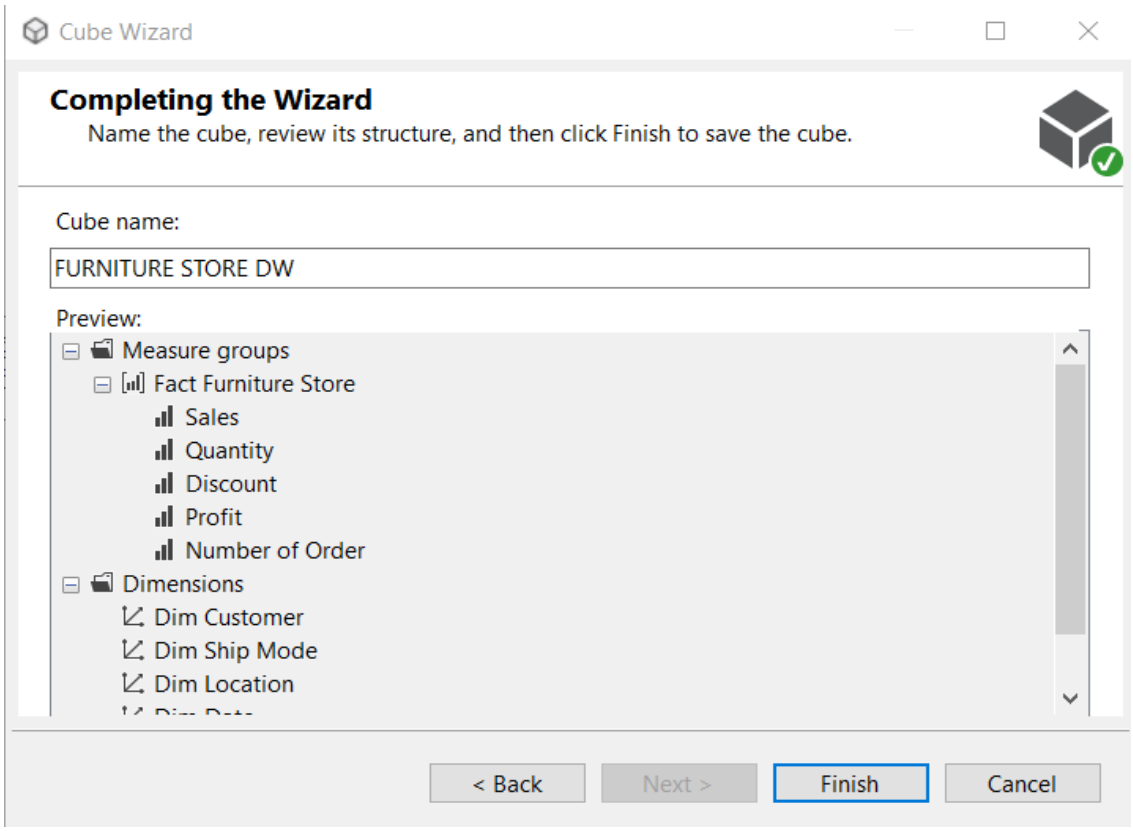
Hình 3.21 Chọn độ đo sử dụng trong Cube

Bước 5: Chọn các bảng chiều cho kho, ở đây, bỏ chọn bảng **Fact_Furniture_Store**



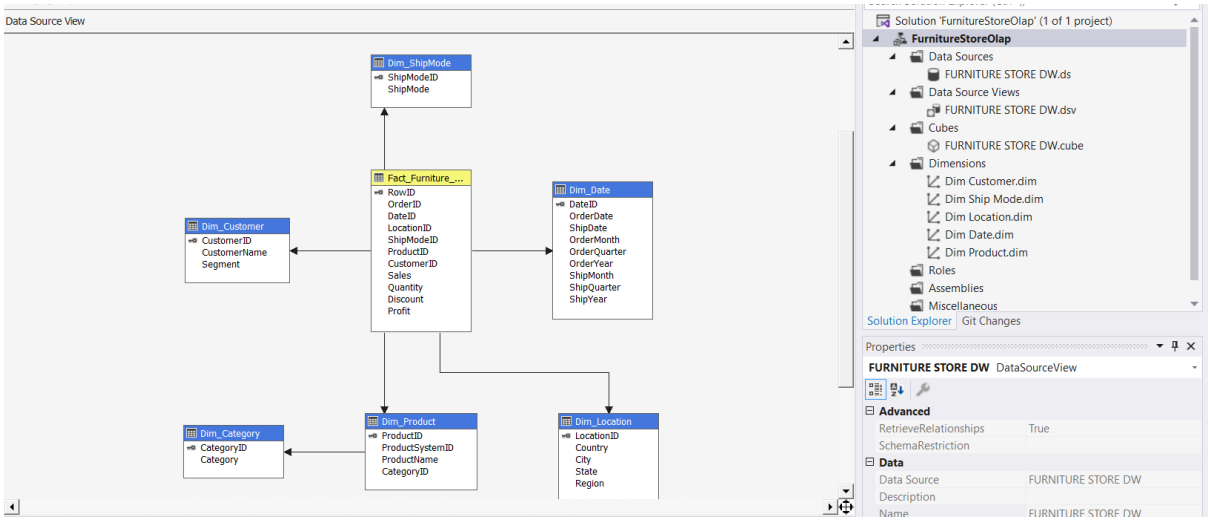
Hình 3.22 Chọn các bảng chiều cho Cube

Bước 6: Xem lại các bảng chiều và độ đo cho kho dữ liệu, sau đó nhấn **"Finish"** để hoàn tất



Hình 3.23 Đặt tên cho Cube và xem lại danh sách các độ đo, bảng chiều

Bước 7: Sau khi hoàn tất việc tạo chiều và độ đo cho kho, Visual Studio sẽ dẫn đến màn hình sau.



Hình 3.24 Mô hình khối dữ liệu

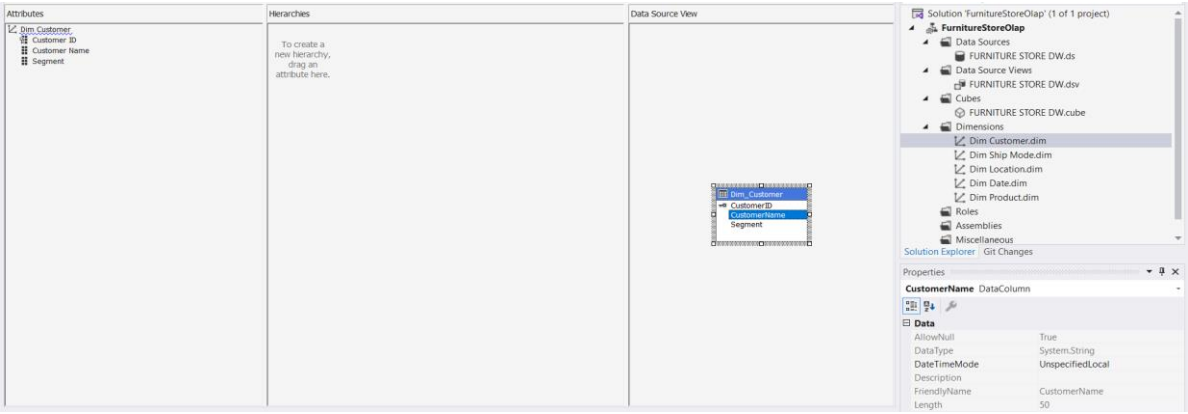
1.5. Thêm thuộc tính vào Dimensions

1.5.1. Dim Customer

Bước 1: Ở mục **Dimensions**, chọn **Dim Customer.dim**, sau đó kéo thả các thuộc tính cần thiết vào **Dim Customer.**

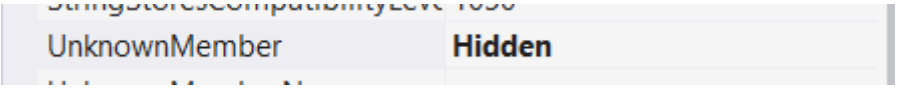
Hình 3.25 Kéo thuộc tính cần thiết vào Dim Customer

Bước 2: Kéo các thuộc tính cho đến khi được như hình



Hình 3.26 Hoàn thiện việc kéo thuộc tính cần thiết cho Dim Customer.

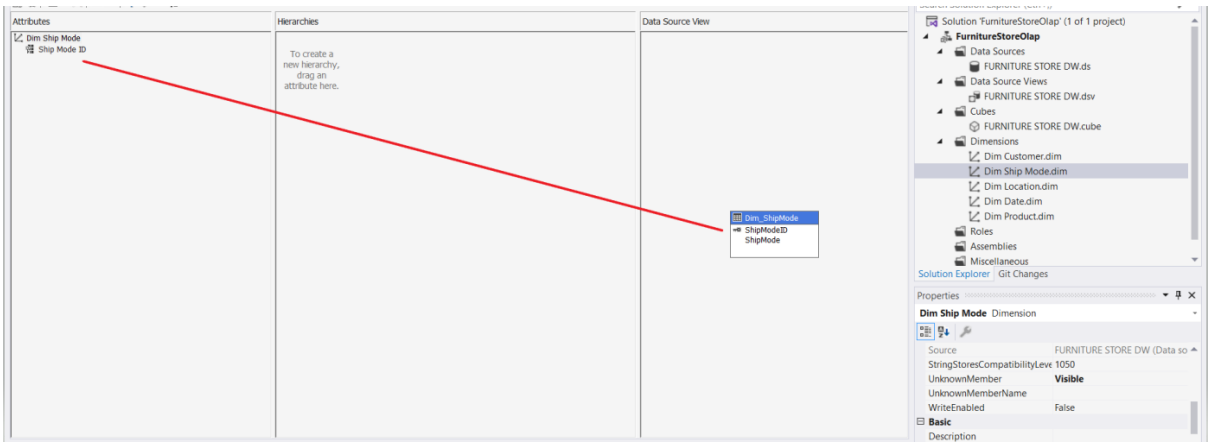
Bước 3: Chỉnh "UnknownMember" từ **Visible** sang **Hidden**



Hình 3.27 Cài đặt thuộc tính UnkownMember thành Hidden

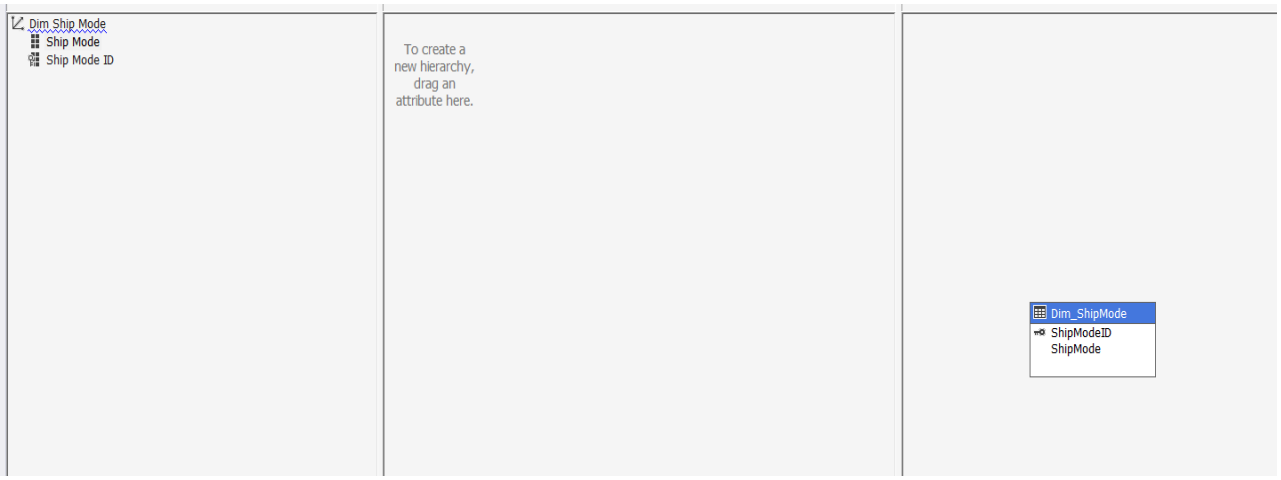
1.5.2. Dim Ship Mode

Bước 1: Ở mục **Dimensions**, chọn **Dim Ship Mode.dim**, sau đó kéo thả các thuộc tính cần thiết vào **Dim Ship Mode.**



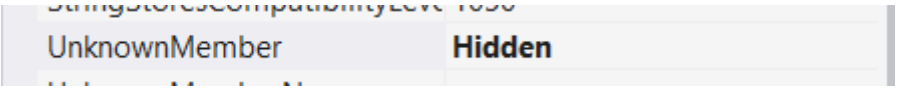
Hình 3.28 Kéo thả các thuộc tính cần thiết vào Dim Ship Mode

Bước 2: Kéo các thuộc tính cho đến khi được như hình.



Hình 3.29 Hoàn thiện việc kéo các thuộc tính cần thiết cho Dim Ship Mode

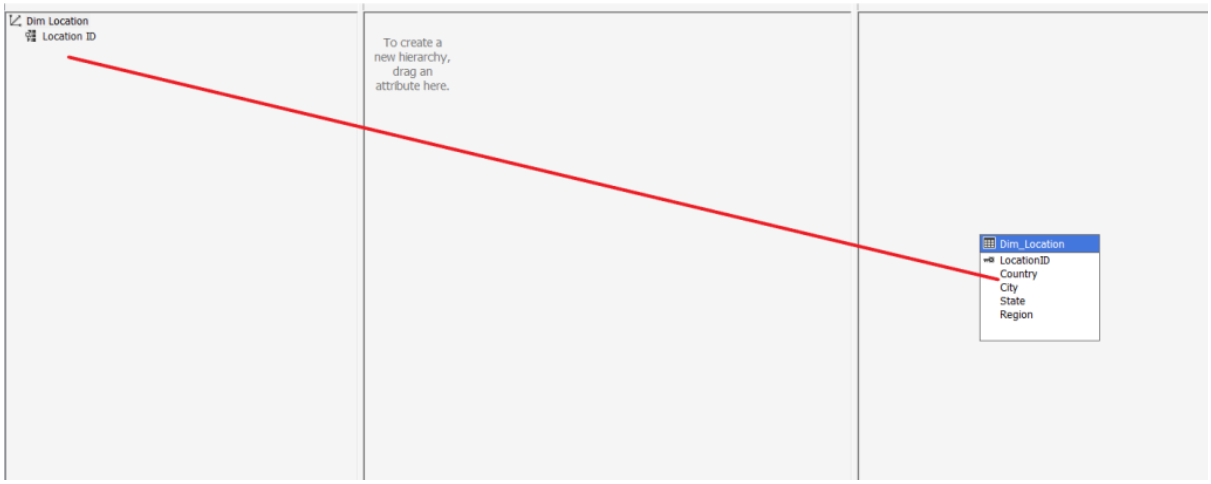
Bước 3: Chỉnh "UnknownMember" từ **Visible** sang **Hidden.**



Hình 3.30 Cài đặt thuộc tính UnknownMember thành Hidden

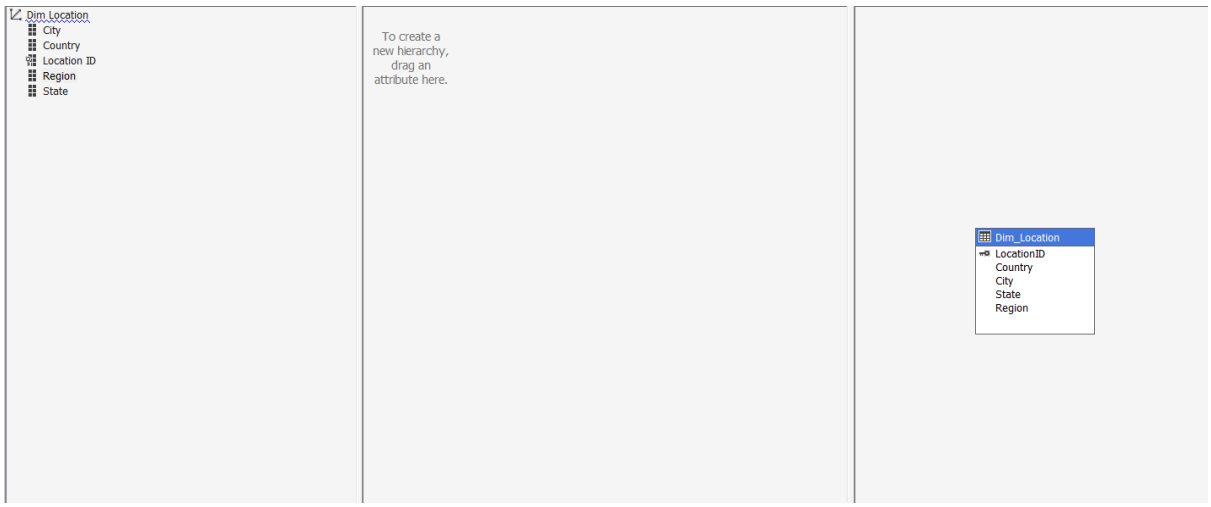
1.5.3. Dim Location

Bước 1: Ở mục **Dimensions**, chọn **Dim Location.dim**, sau đó kéo thả các thuộc tính cần thiết vào **Dim Location.**



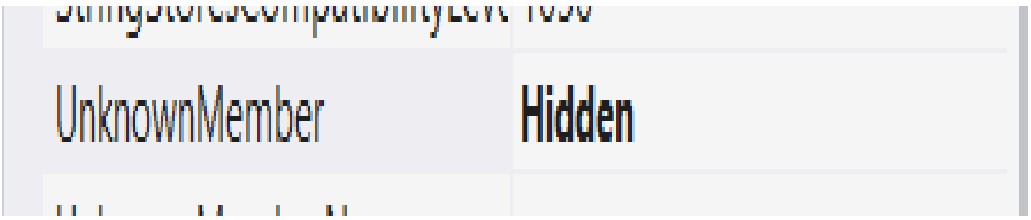
Hình 3.31 Kéo thả các thuộc tính cần thiết vào Dim Location

Bước 2: Kéo các thuộc tính cho đến khi được như hình.



Hình 3.32 Hoàn thiện việc kéo thả các thuộc tính cần thiết vào Dim Location

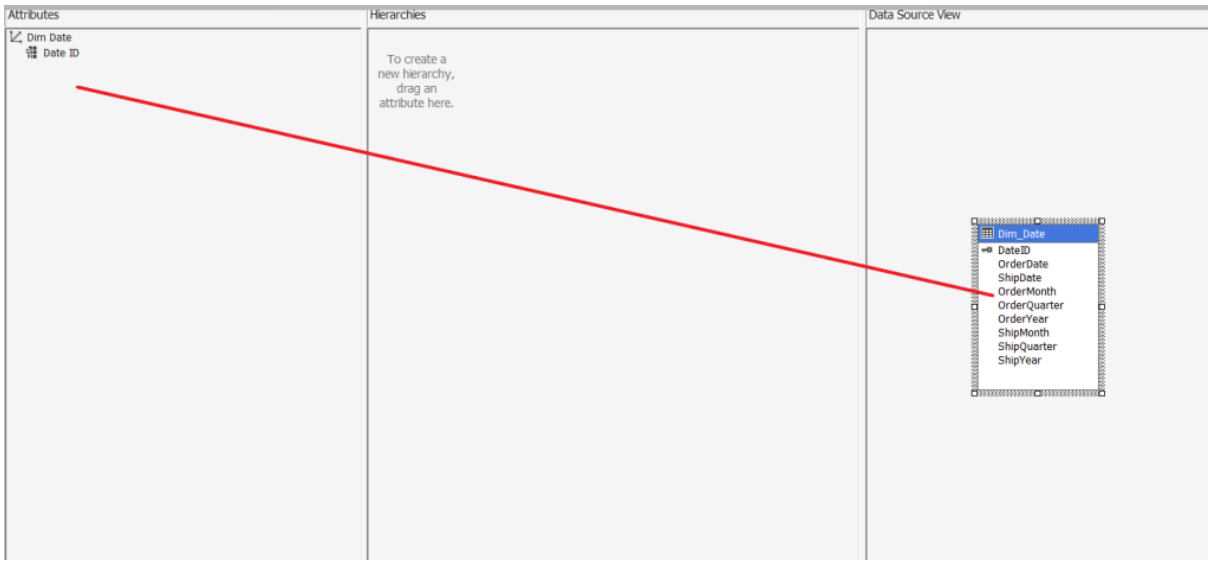
Bước 3: Chỉnh "UnknownMember" từ **Visible** sang **Hidden**



Hình 3.33 Cài đặt thuộc tính UnknownMember thành Hidden

1.5.4. Dim Date

Bước 1: Ở mục **Dimensions**, chọn **Dim Date.dim**, sau đó kéo thả các thuộc tính cần thiết vào **Dim Date**.

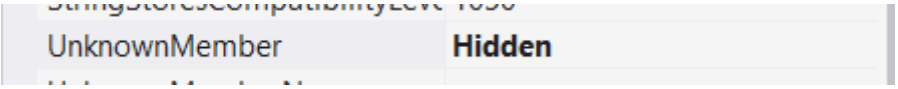


Hình 3.34 Kéo thả các thuộc tính cần thiết vào Dim Date

Bước 2: Kéo các thuộc tính cho đến khi được như hình.

Hình 3.35 Hoàn thiện việc kéo thả các thuộc tính cần thiết vào Dim Date

Bước 3: Chỉnh "UnknownMember" từ **Visible** sang **Hidden**



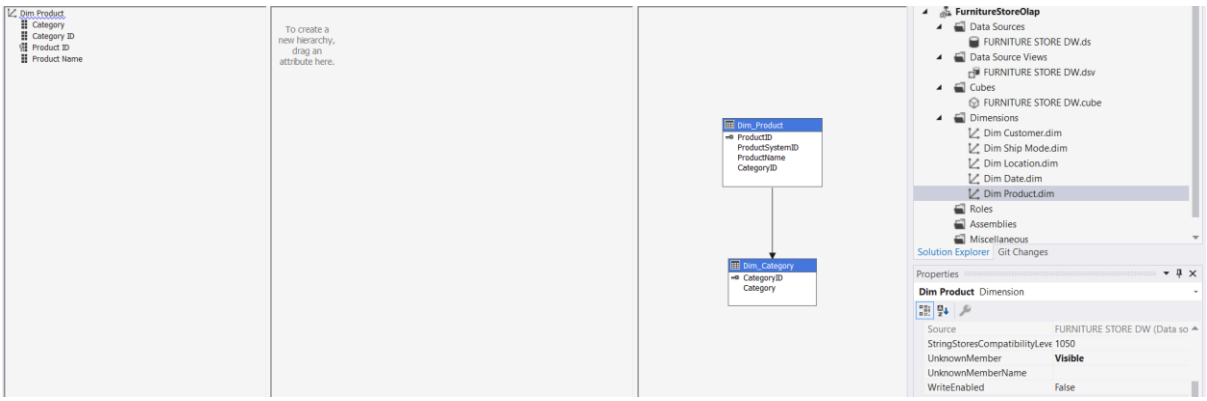
Hình 3.36 Cài đặt thuộc tính UnknownMember thành Hidden

1.5.5. Dim Product

Bước 1: Ở mục **Dimensions**, chọn **Dim Product.dim**, sau đó kéo thả các thuộc tính cần thiết vào **Dim Product**.

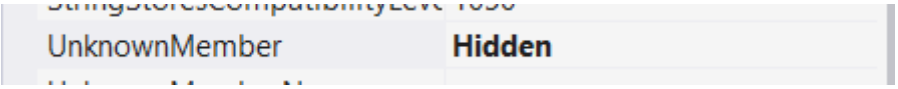
Hình 3.37 Kéo thả các thuộc tính cần thiết vào Dim Product

Bước 2: Kéo các thuộc tính cho đến khi được như hình.



Hình 3.38 Hoàn thiện việc kéo thả các thuộc tính vào Dim Product

Bước 3: Chỉnh "UnknownMember" từ **Visible** sang **Hidden**



Hình 3.39 Cài đặt thuộc tính UnknownMember thành Hidden

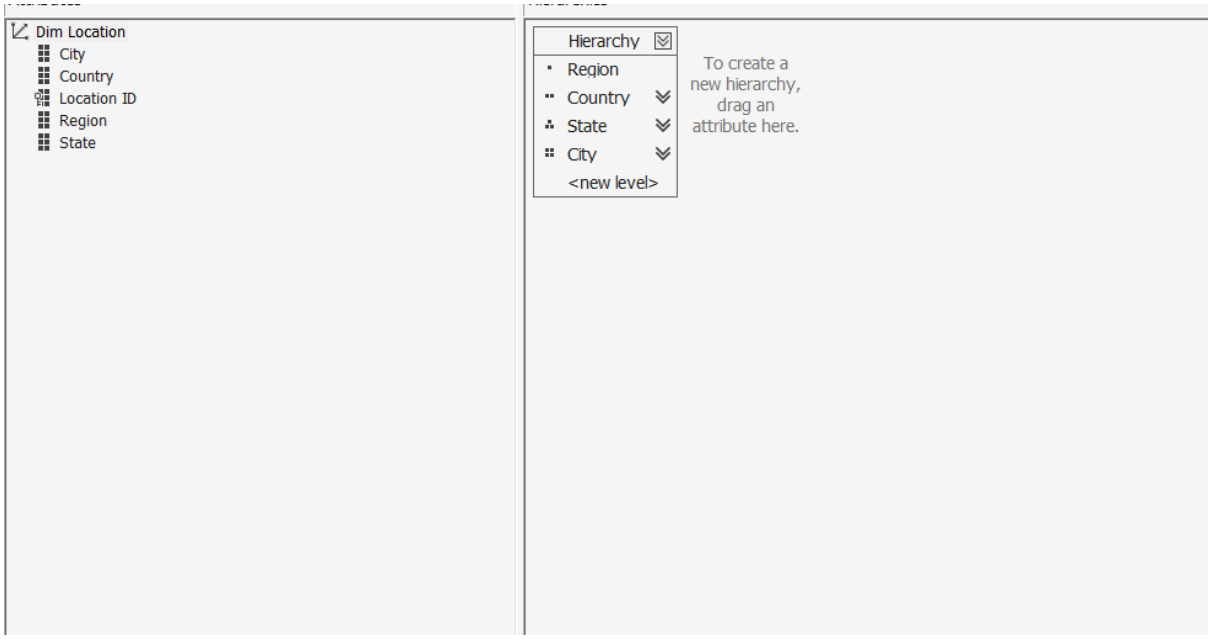
1.6. Tạo Hierachy

1.6.1. Dim Location

Bước 1: Ở "Dimension Structure", kéo các thuộc tính cần thiết để phân cấp từ Dim **Location** sang mục "Hierachies".

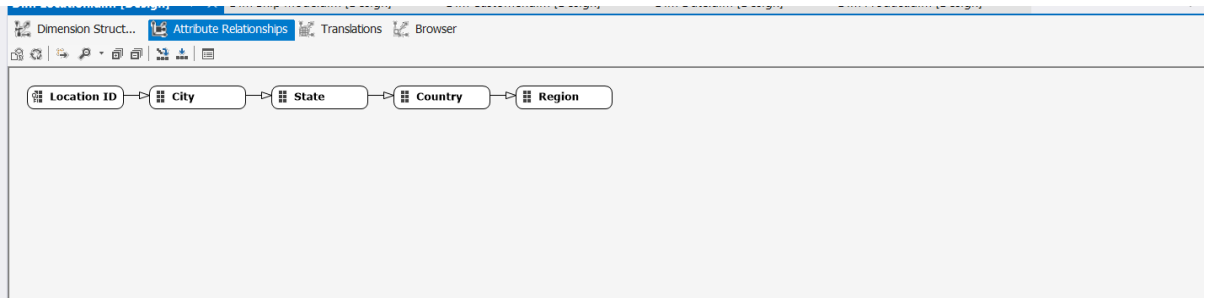
Hình 3.40 Kéo thả các thuộc tính phân cấp vào Hierachies

Bước 2: Kéo lần lượt theo thứ tự **Region→ Country → State → City.**



Hình 3.41 Phân cấp thuộc tính được tạo

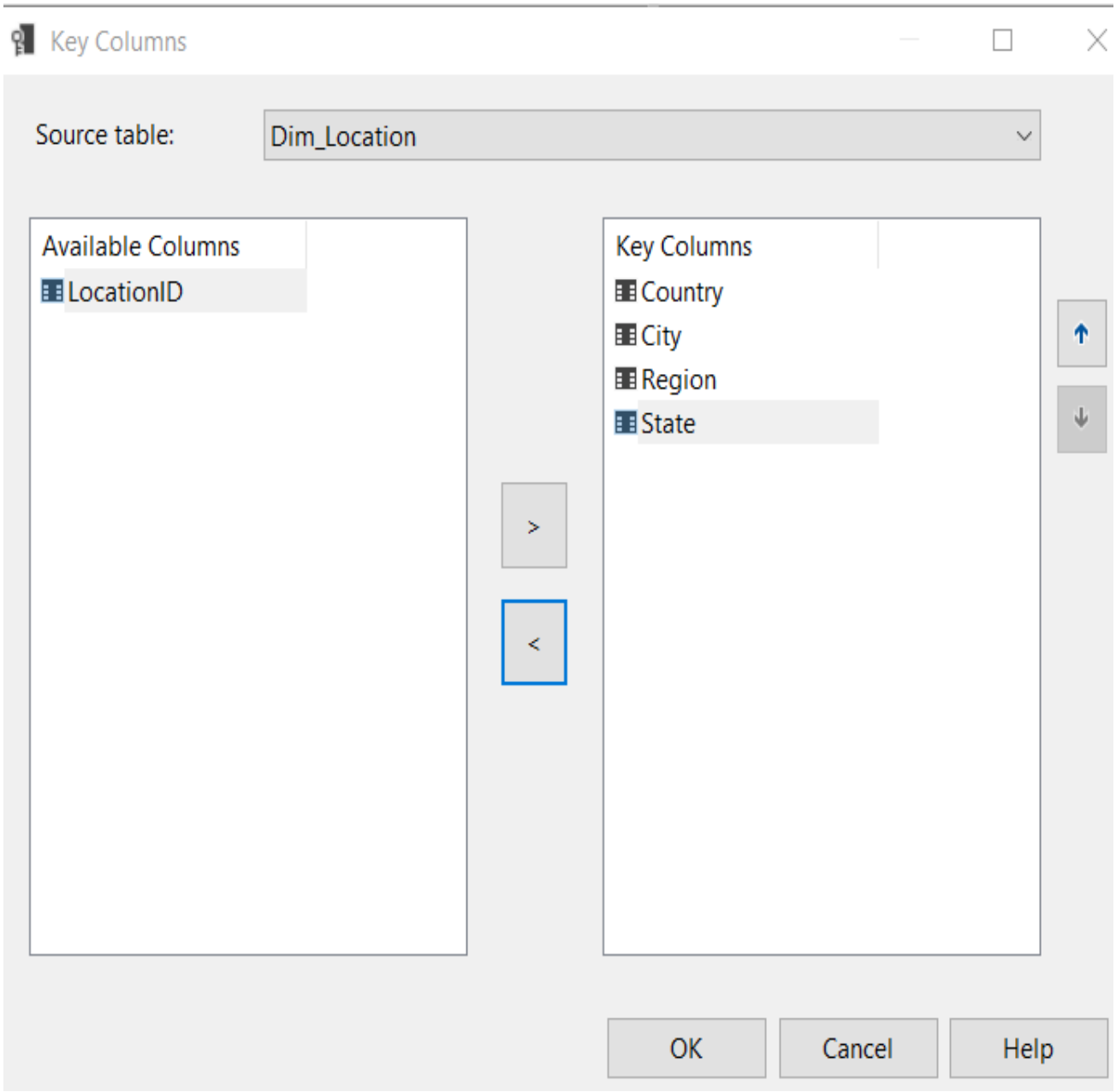
Bước 3: Ở mục **Attribute Relationship**, tiến hành định nghĩa lại mối quan hệ giữa các thuộc tính như hình sau.



Hình 3.42 Định nghĩa mối quan hệ giữa các thuộc tính

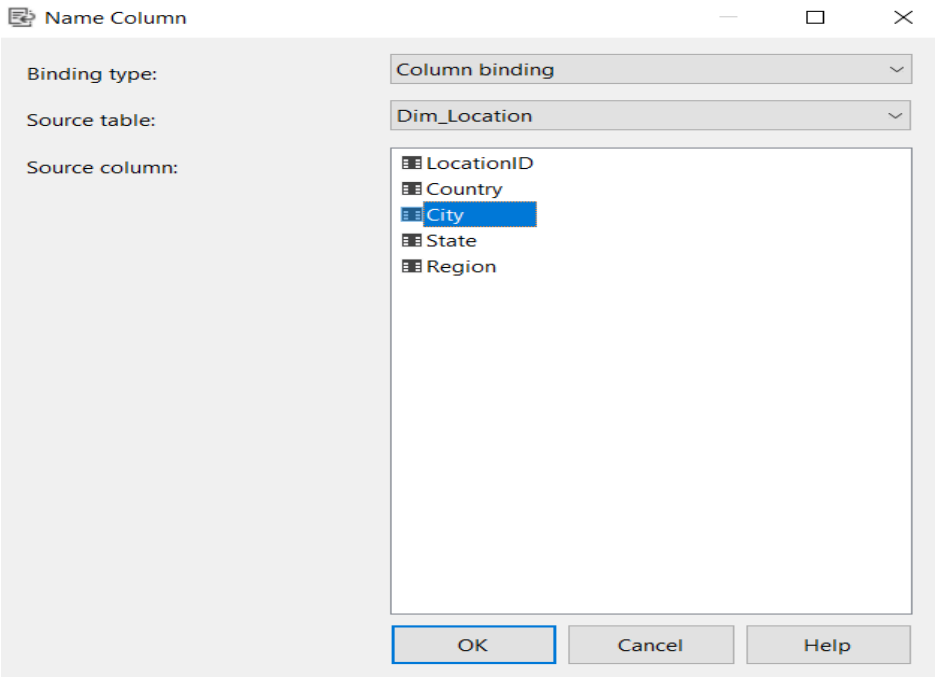
Bước 4: Định nghĩa lại **Key Columns** và **Name Column** cho **Attribute** của phân cấp địa chỉ giao hàng:.

Bước 4a: Ở thuộc tính "City", vì nó là thuộc tính cấp nhỏ nhất nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là "State", "Country", "Region", sau đó nhấn "Ok".



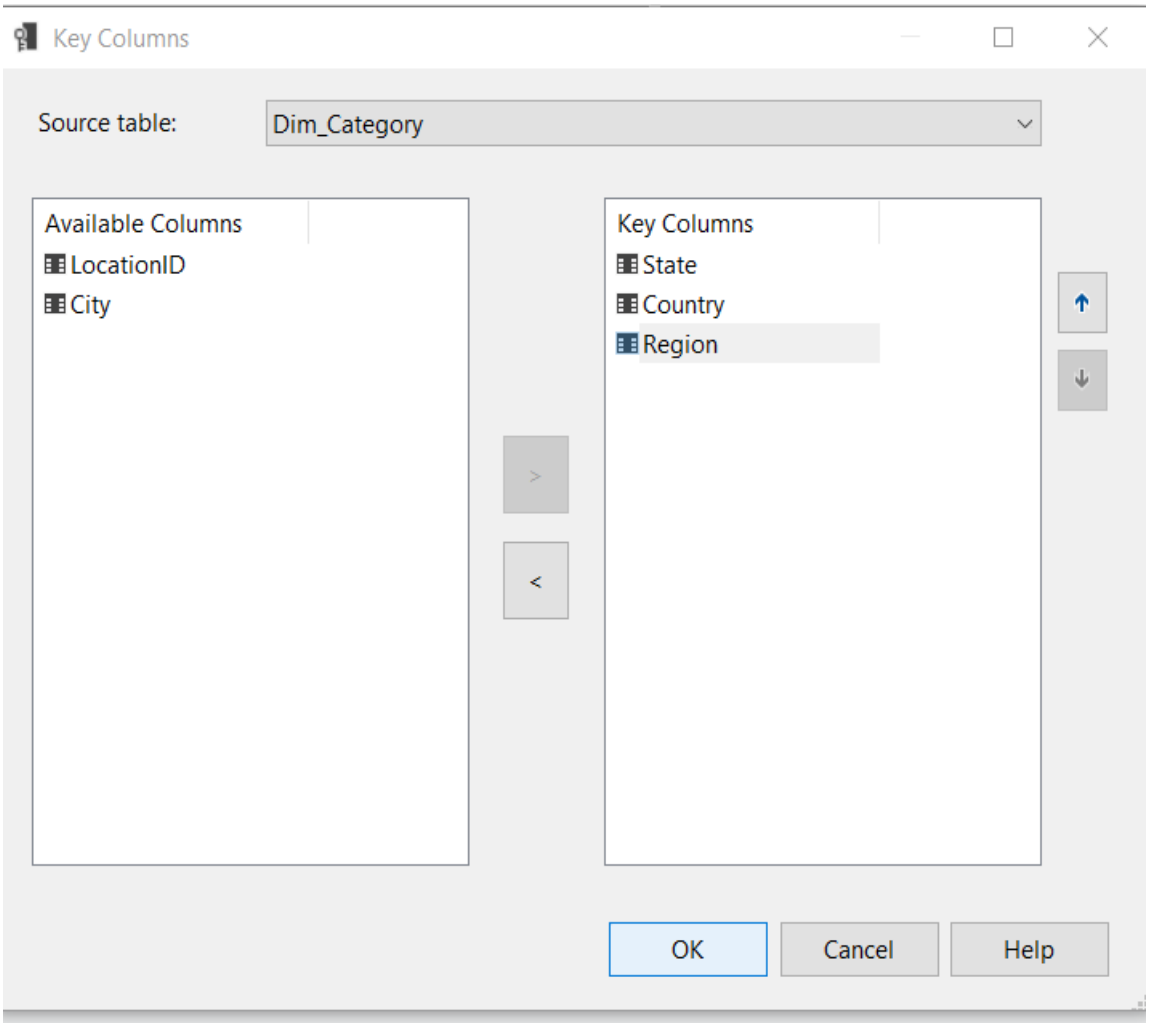
Hình 3.43 Định nghĩa Key Columns cho thuộc tính phân cấp City

Bước 4b: Chọn **Name Column** cho City là "City" và nhấn "Ok".



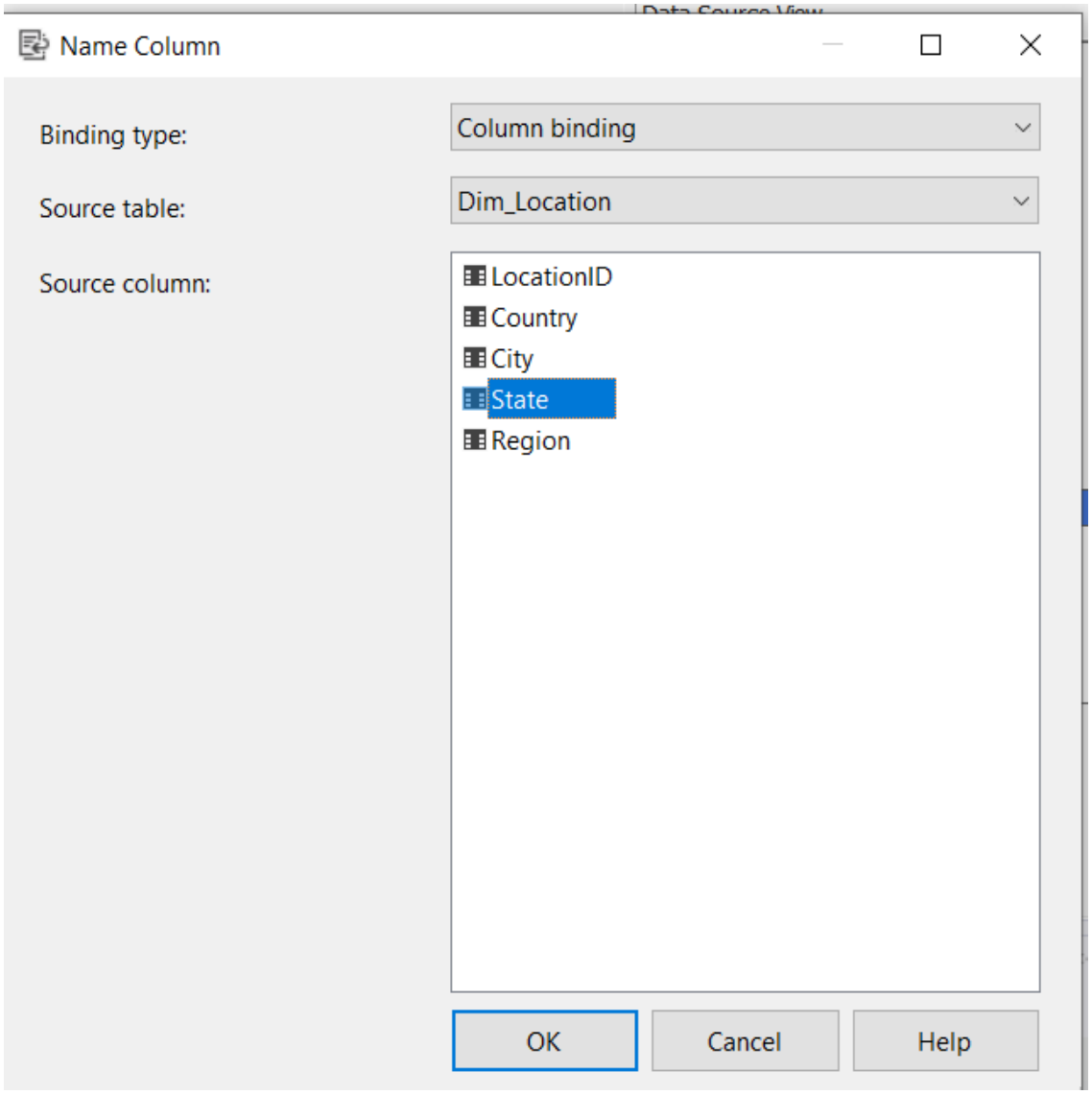
Hình 3.44 Định nghĩa Name Column cho thuộc tính phân cấp City

Bước 4c: Ở thuộc tính "State", vì nó là thuộc tính cấp nhỏ hơn "Country", "Region" nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là "Country", "Region", sau đó nhấn "Ok".



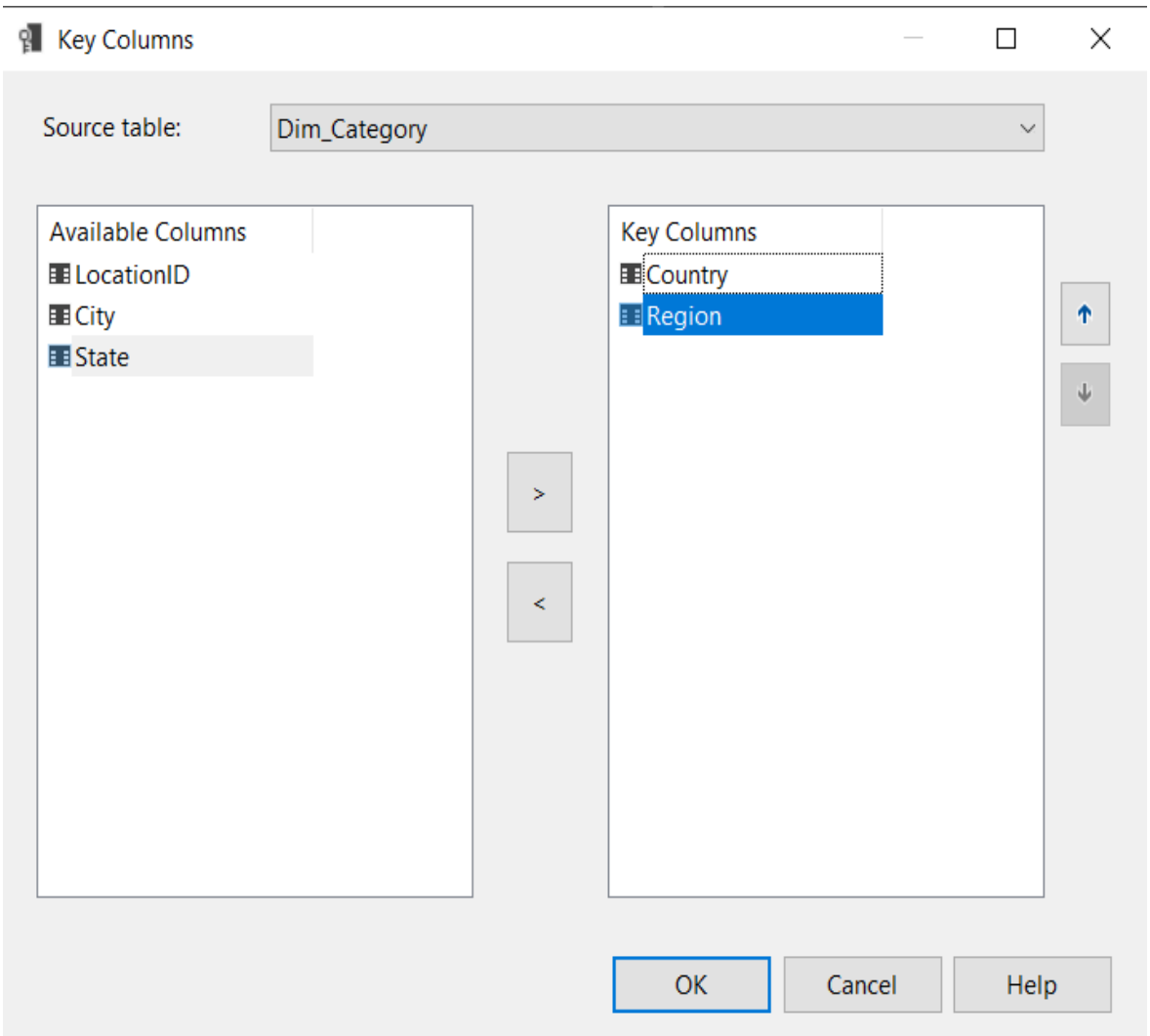
Hình 3.45 Định nghĩa Key Columns cho thuộc tính phân cấp State

Bước 4d: Chọn **Name Column** cho State là “State” và nhấn “Ok”.



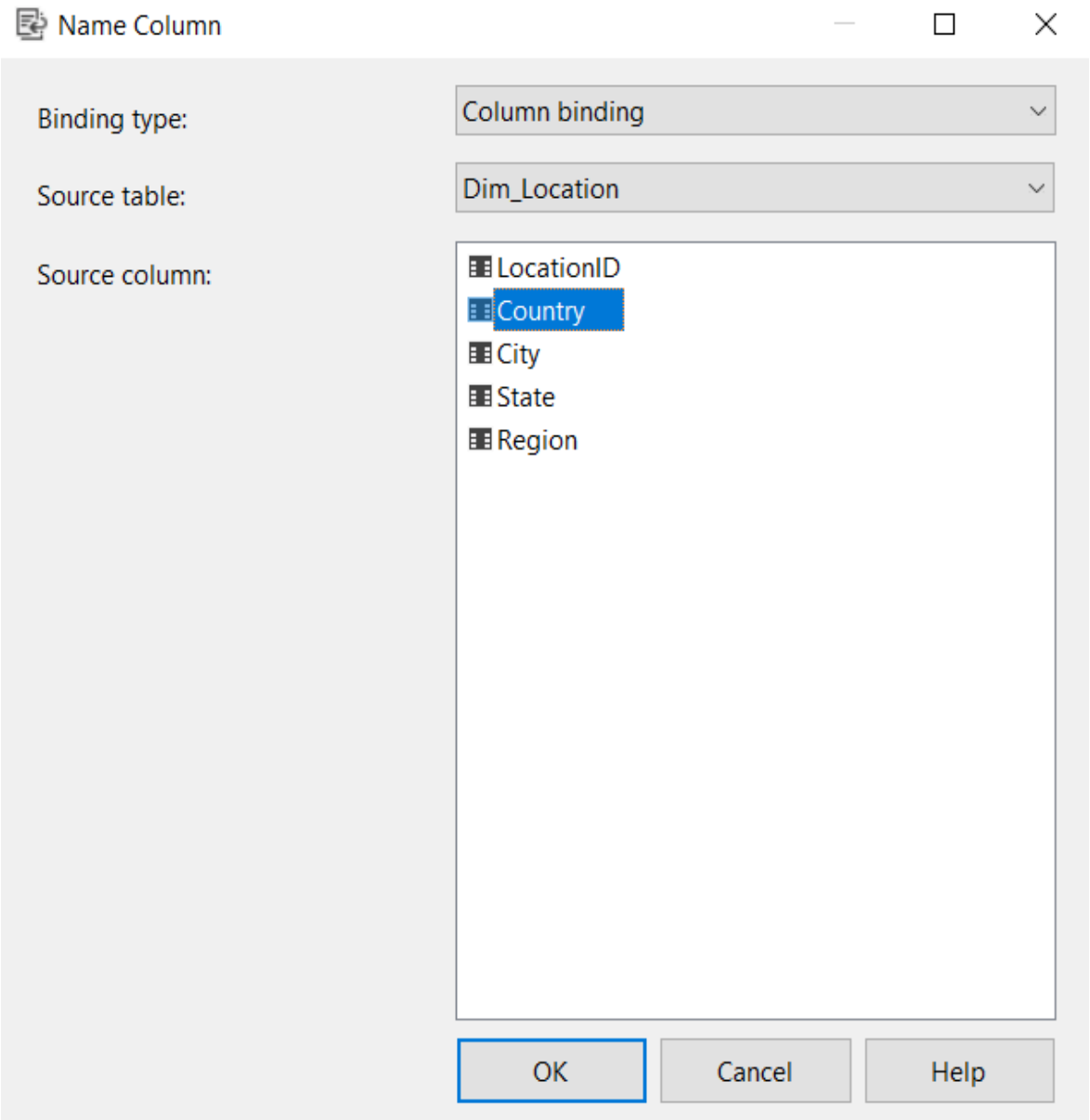
Hình 3.46 Định nghĩa Name Column cho thuộc tính phân cấp State

Bước 4e: Ở thuộc tính "Country", vì nó là thuộc tính cấp nhỏ hơn "Region" nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là "Region", sau đó nhấn "Ok".

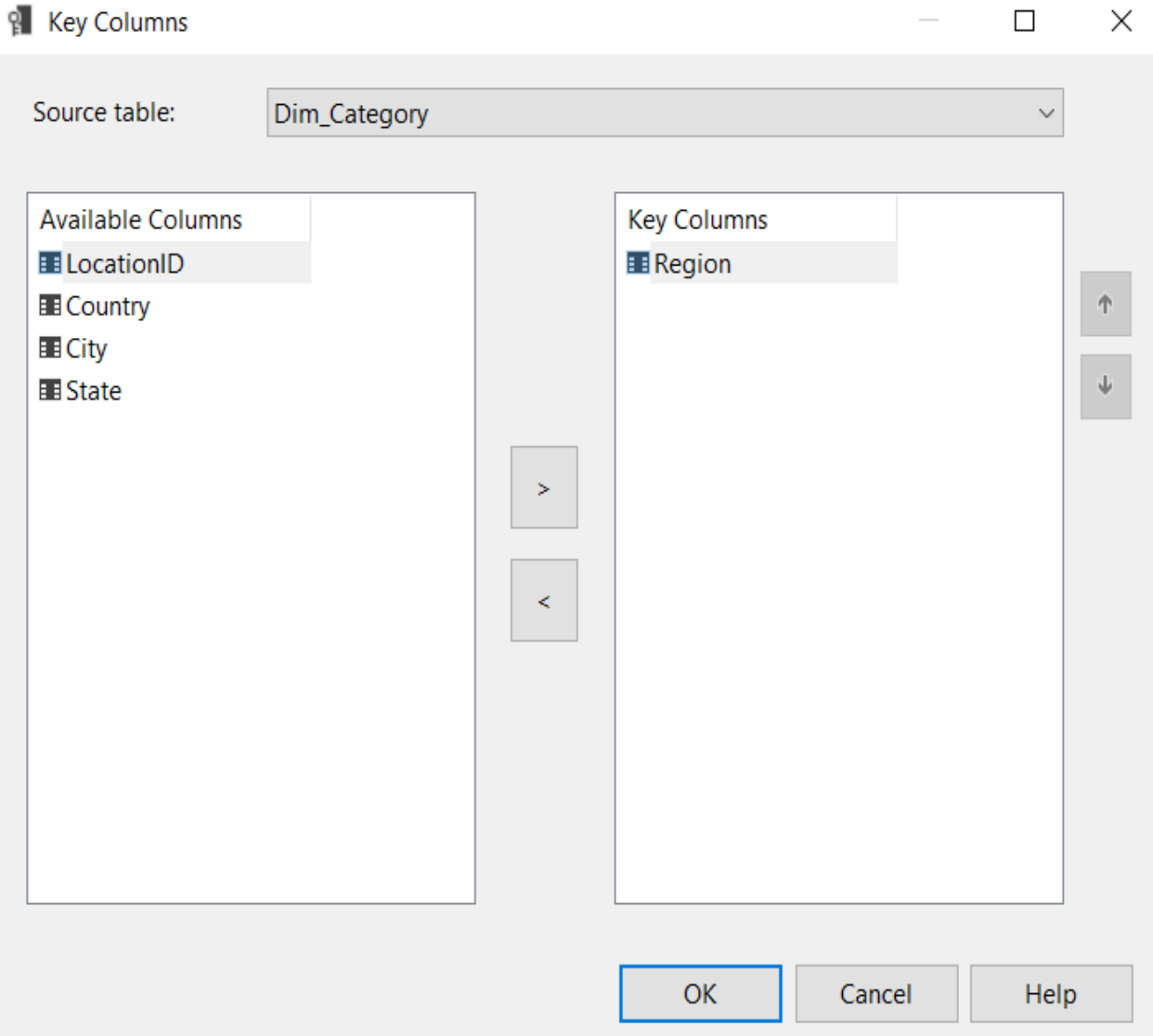


Hình 3.47 Định nghĩa Key Columns cho thuộc tính phân cấp Country

Bước 4f: Chọn **Name Column** cho Country là “Country” và nhấn “Ok”.

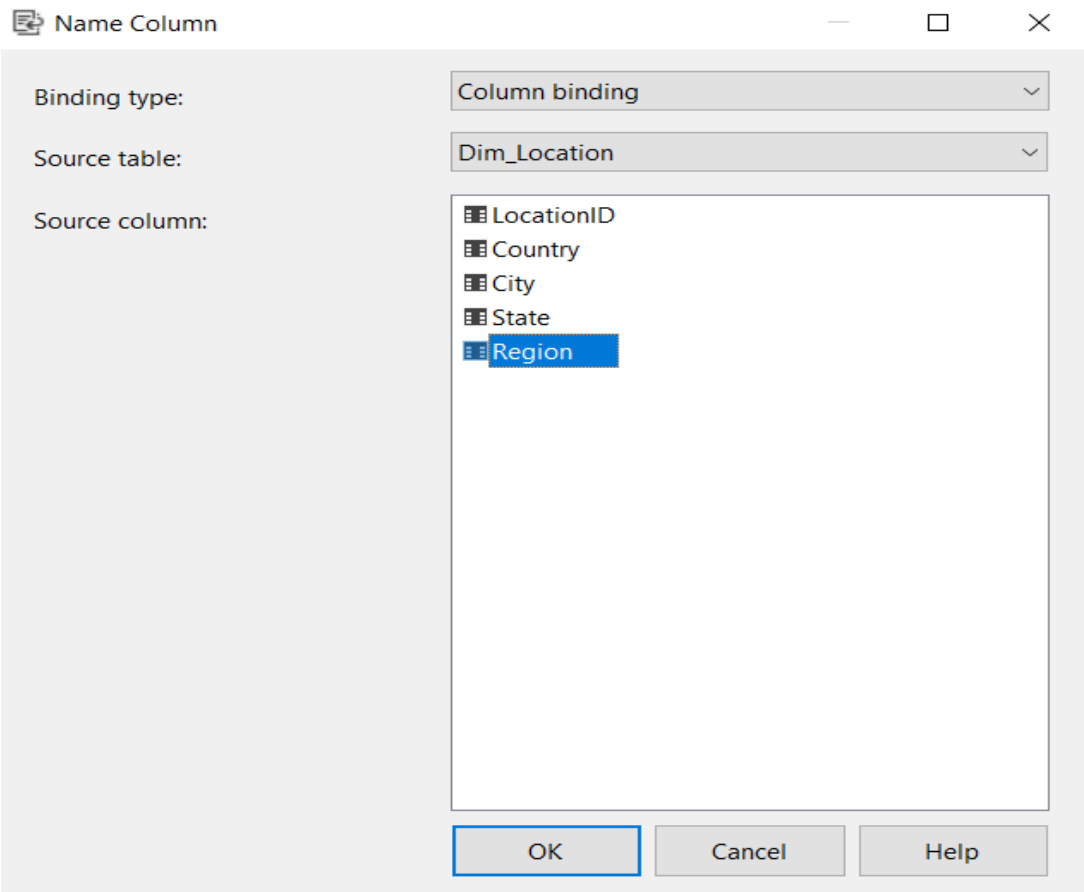


Bước 4g: Ở thuộc tính "Region", vì nó là thuộc tính cấp cao nhất nên sẽ lấy khóa dòng gồm chính nó, sau đó nhấn "Ok".



Hình 3.48 Định nghĩa Key Columns cho thuộc tính phân cấp Region

Bước 4h: Chọn Name Column cho Region là "Region" và nhấn "Ok".



Hình 3.49 Định nghĩa Name Column cho thuộc tính phân cấp Region

1.6.2. Dim Date

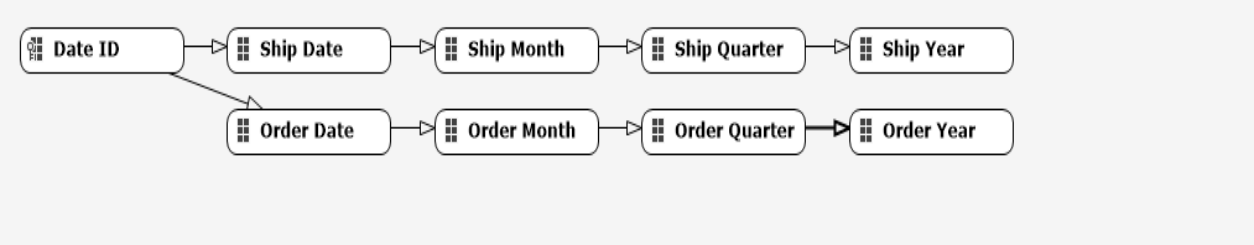
Bước 1: Ở "Dimension Structure", kéo các thuộc tính cần thiết để phân cấp từ **Dim Date** sang mục "Hierachies".

Hình 3.50 Kéo thả các thuộc tính phân cấp của Dim Date vào Hierachies

Bước 2: Để phân cấp cho ngày đặt hàng, kéo lần lượt theo thứ tự **Order Year → Order Month → Order Quarter → Order Year** sau đó để phân cấp cho ngày đặt hàng, kéo lần lượt theo thứ tự **Ship Year → Ship Quarter → Ship Month → Ship Date**.

Hình 3.51 Phân cấp cho đặt hàng

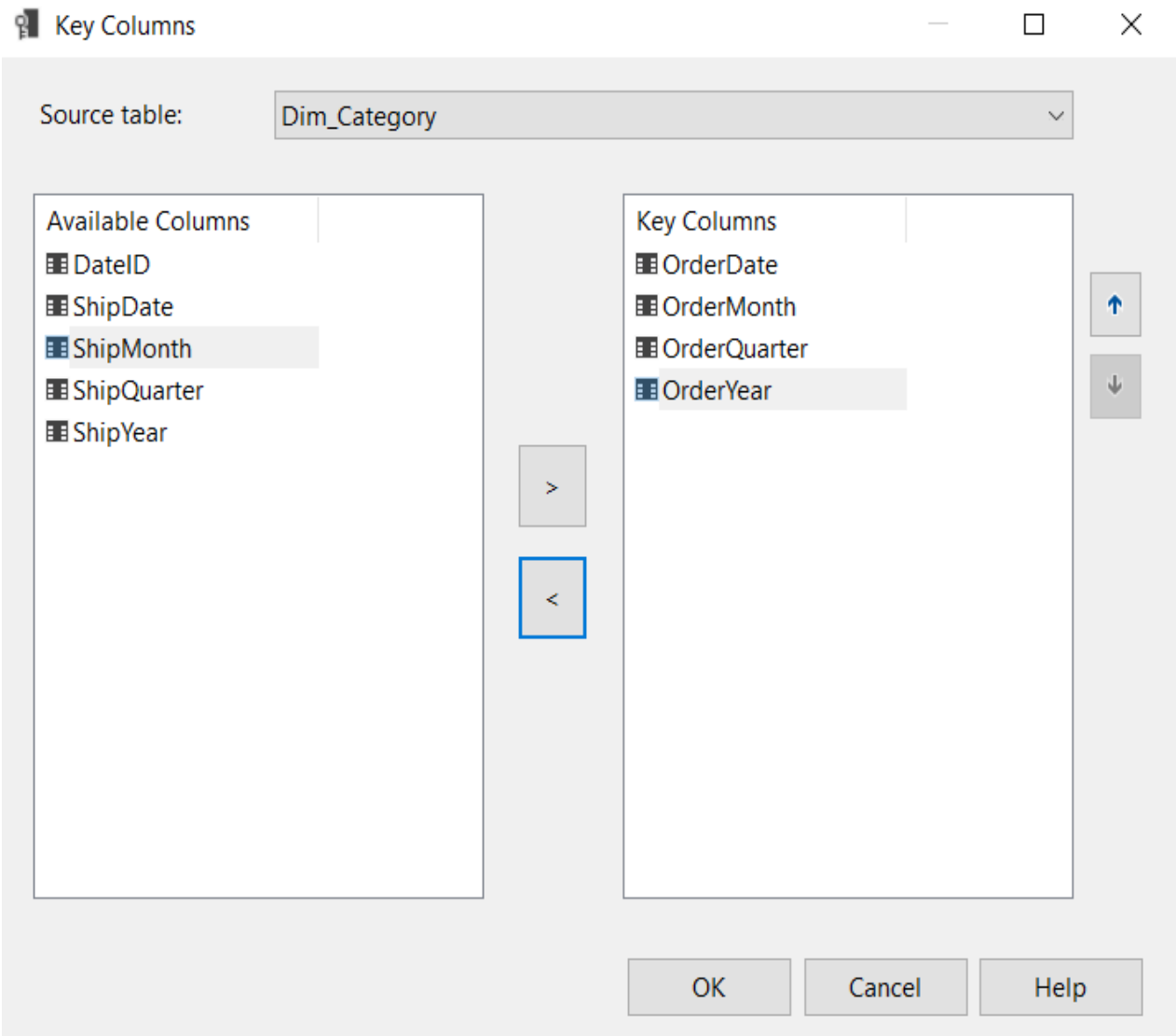
Bước 3: Ở mục Attribute Relationship, tiến hành định nghĩa lại mối quan hệ giữa các thuộc tính như hình



Hình 3.52 Định nghĩa lại mối quan hệ giữa các thuộc tính

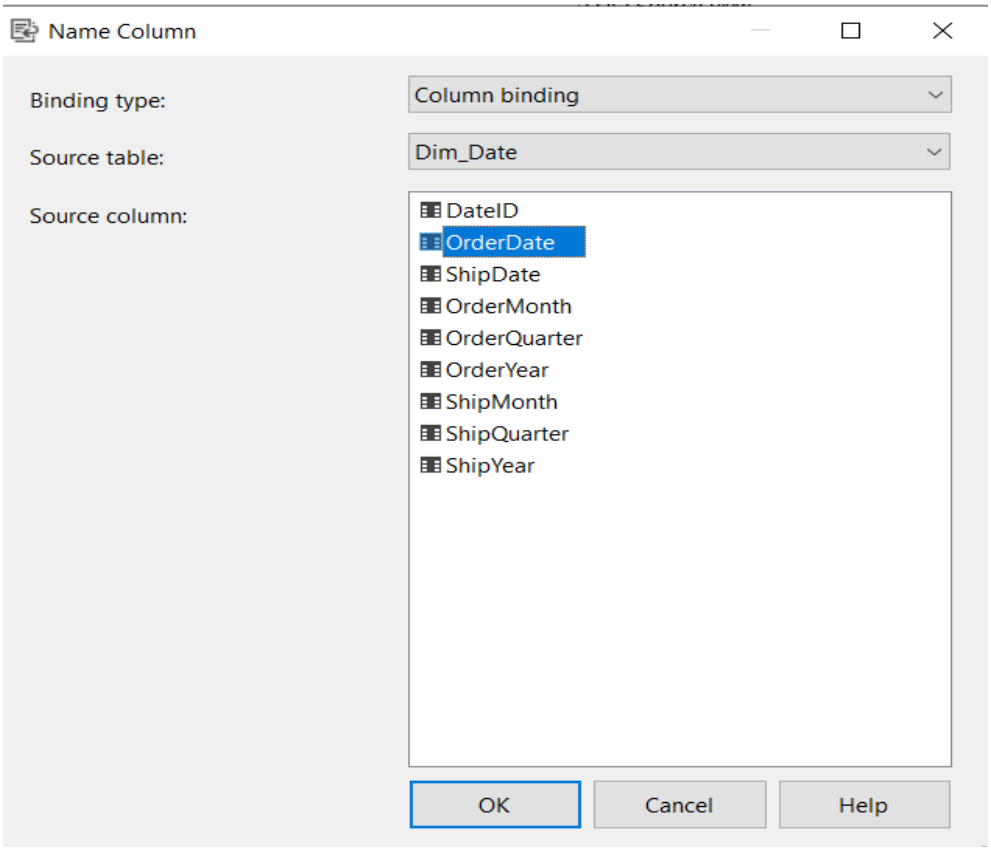
Bước 4: Định nghĩa lại Key Columns và Name Column cho Attribute cho phân cấp đặt hàng:

Bước 4a: Ở thuộc tính **"OrderDate"**, vì nó là thuộc tính cấp nhỏ nhất nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là **"OrderMonth"**, **"OrderQuarter", "OrderYear",** sau đó nhấn **"Ok"**



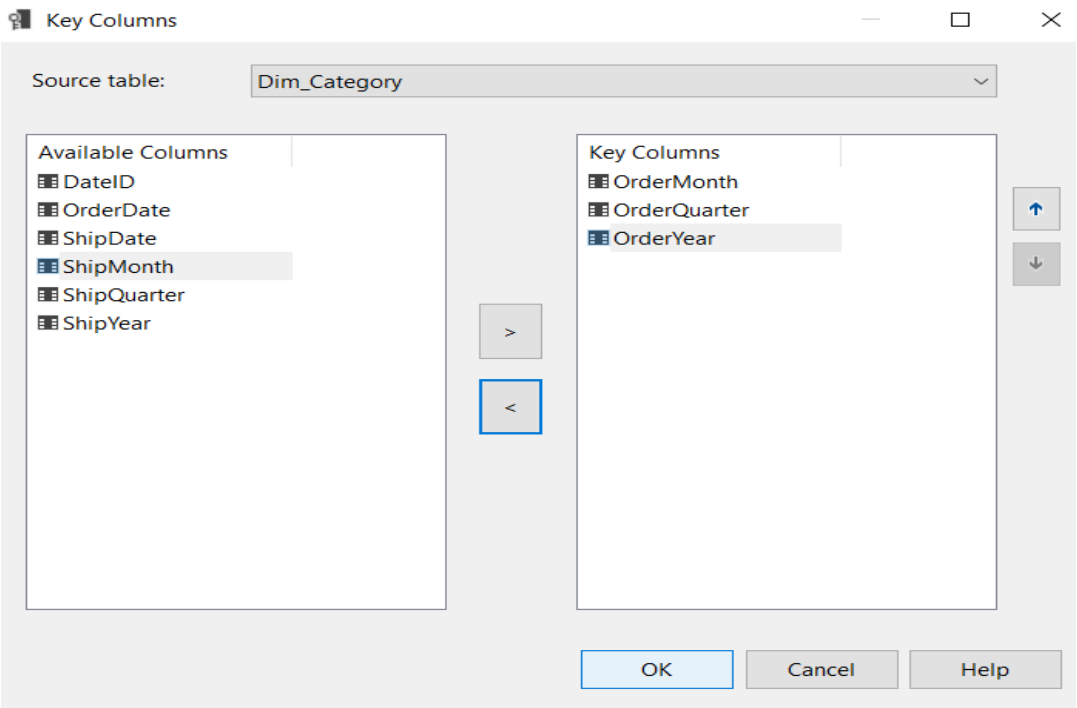
Hình 3.53 Định nghĩa lại Key Column cho OrderDate

Bước 4b: Chọn Name Column cho OrderDate là **"OrderDate"** và nhấn **"Ok"**



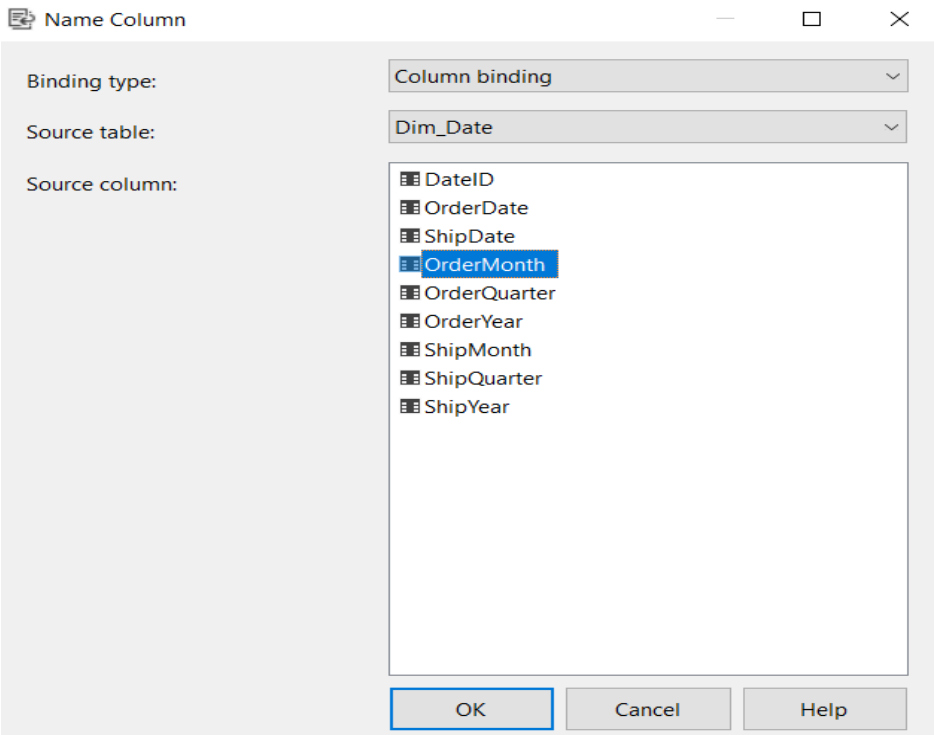
Hình 3.54 Chọn Name Column cho OrderDate

Bước 4c: Ở thuộc tính **"OrderMonth"**, vì nó là thuộc tính cấp nhỏ hơn **"OrderQuarter"**, **"OrderYear"** nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là **"OrderQuarter"**, **"OrderYear"**, sau đó nhấn **"Ok"**



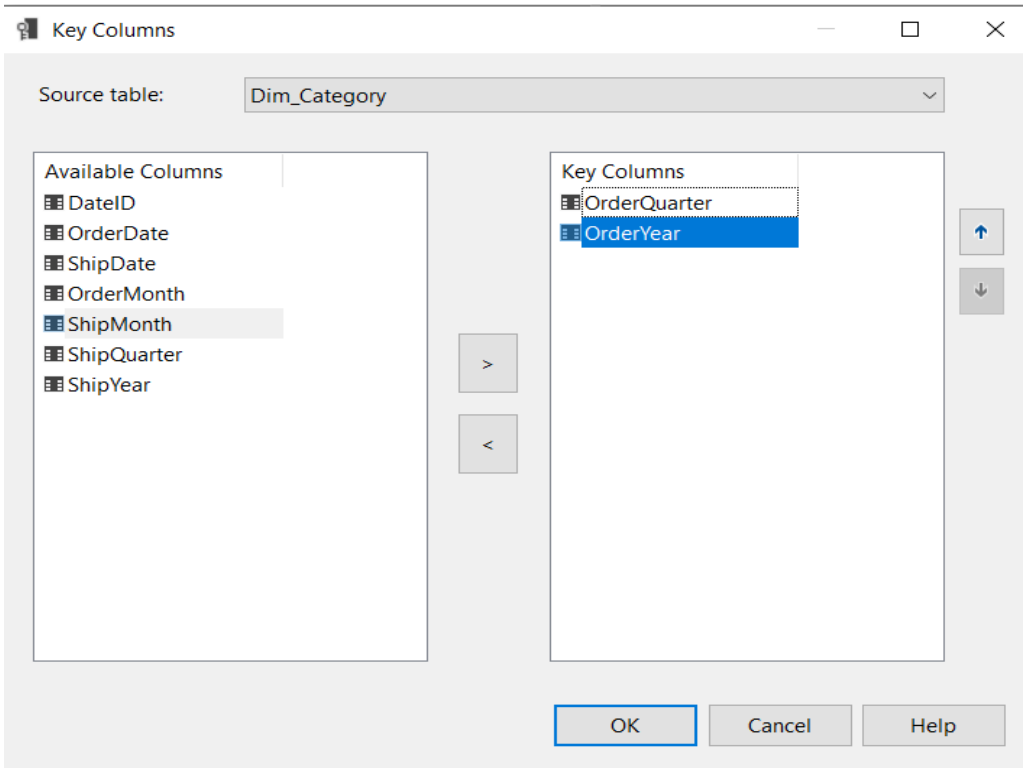
Hình 3.55 Định nghĩa lại Key Column cho OrderMonth

Bước 4d: Chọn Name Column cho OrderMonth là **"OrderMonth"** và nhấn **"Ok"**



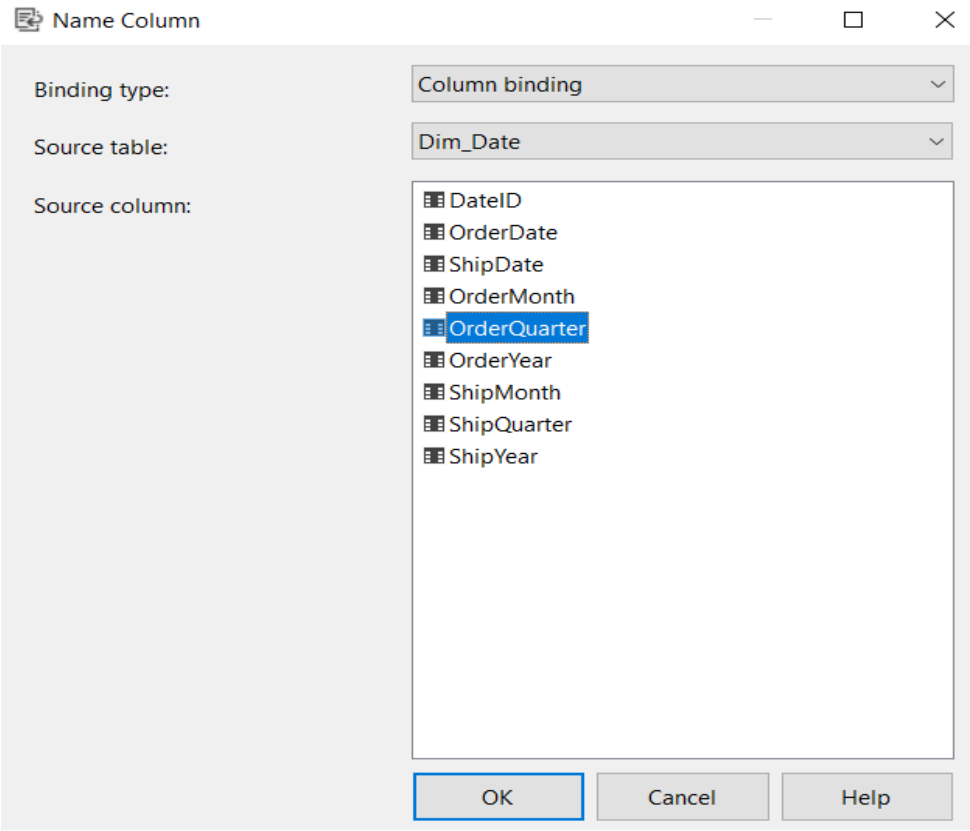
Hình 3.56 Chọn Name Column cho OrderMonth

Bước 4e: Ở thuộc tính **"OrderQuarter"**, vì nó là thuộc tính cấp nhỏ hơn **"OrderYear"**, **"OrderYear"** nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là **"OrderYear"**, sau đó nhấn **"Ok"**



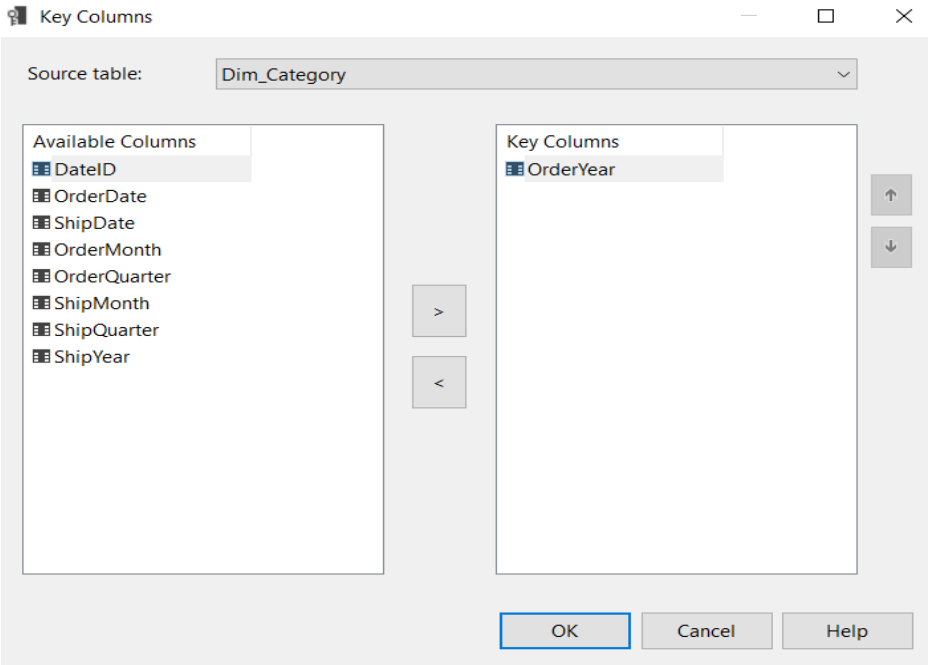
Hình 3. 57 Định nghĩa lại Key Column cho OrderQuarter

Bước 4f: Chọn Name Column cho OrderMonth là **"OrderMonth"** và nhấn **"Ok"**



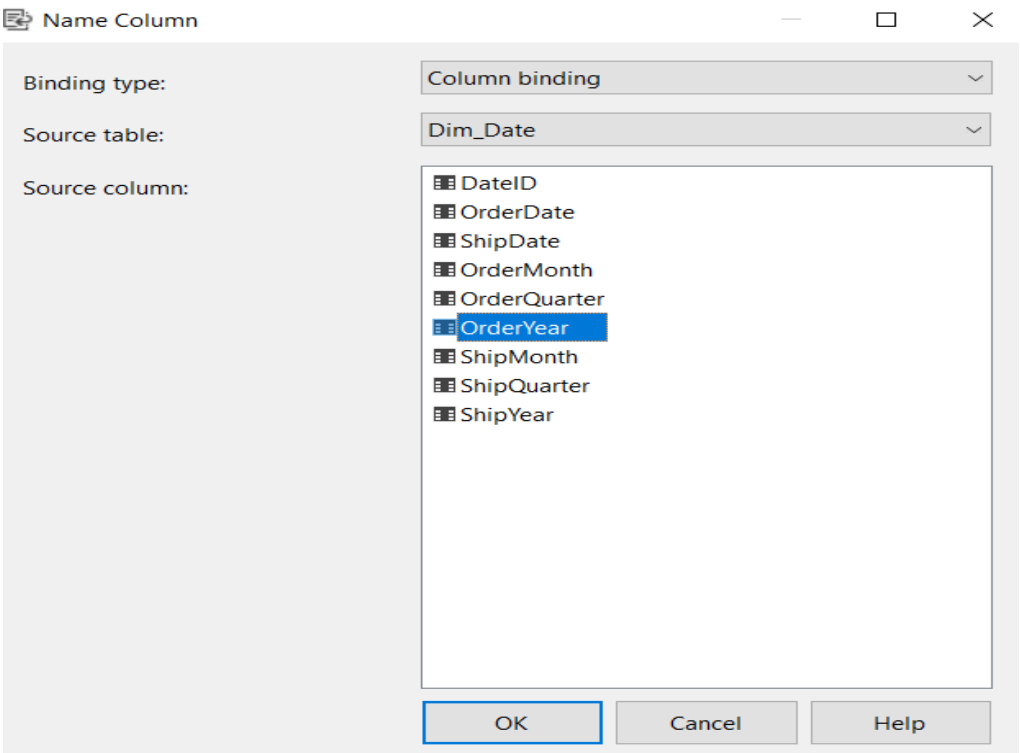
Hình 3. 58 Chọn Name Column cho OrderQuarter

Bước 4g: Ở thuộc tính **"OrderYear"**, vì nó là thuộc tính cấp cao nhất nên sẽ lấy khóa dòng gồm chính nó, sau đó nhấn **"Ok".**



Hình 3.59 Định nghĩa lại Key Column cho OrderYear

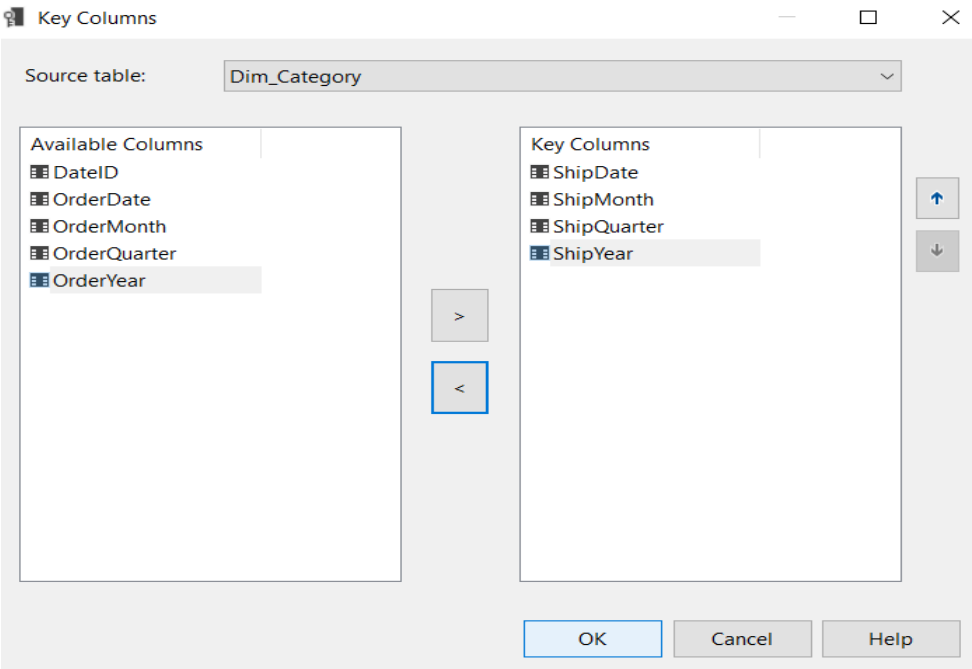
Bước 4h: Chọn Name Column cho OrderYear là **"OrderYear"** và nhấn **"Ok"**



Hình 3.60 Chọn Name Column cho OrderYear

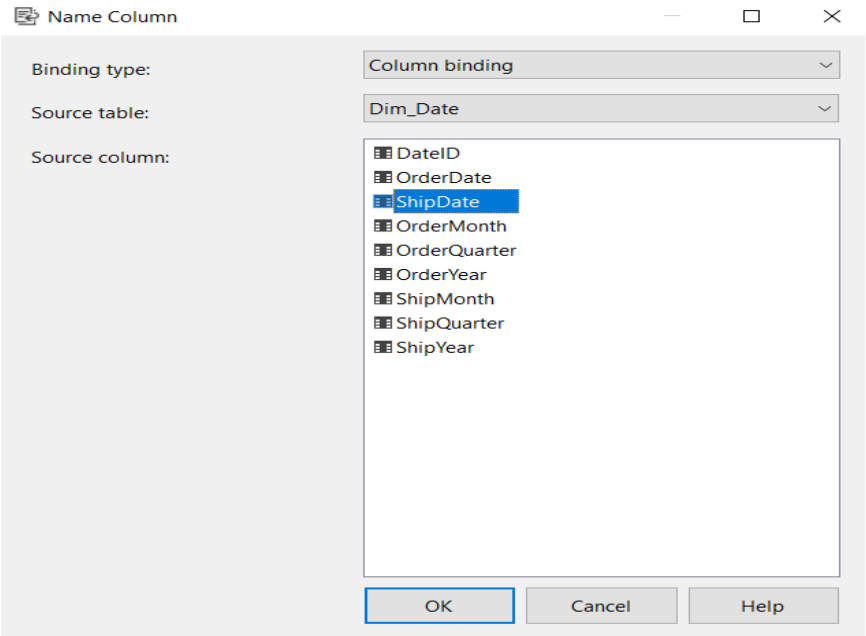
Bước 5: Định nghĩa lại Key Columns và Name Column cho Attribute cho phân cấp giao hàng:

Bước 5a: Ở thuộc tính **"ShipDate"**, vì nó là thuộc tính cấp nhỏ nhất nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là **"ShipMonth"**, **"ShipQuarter"**, **"ShipYear",** sau đó nhấn **"Ok".**



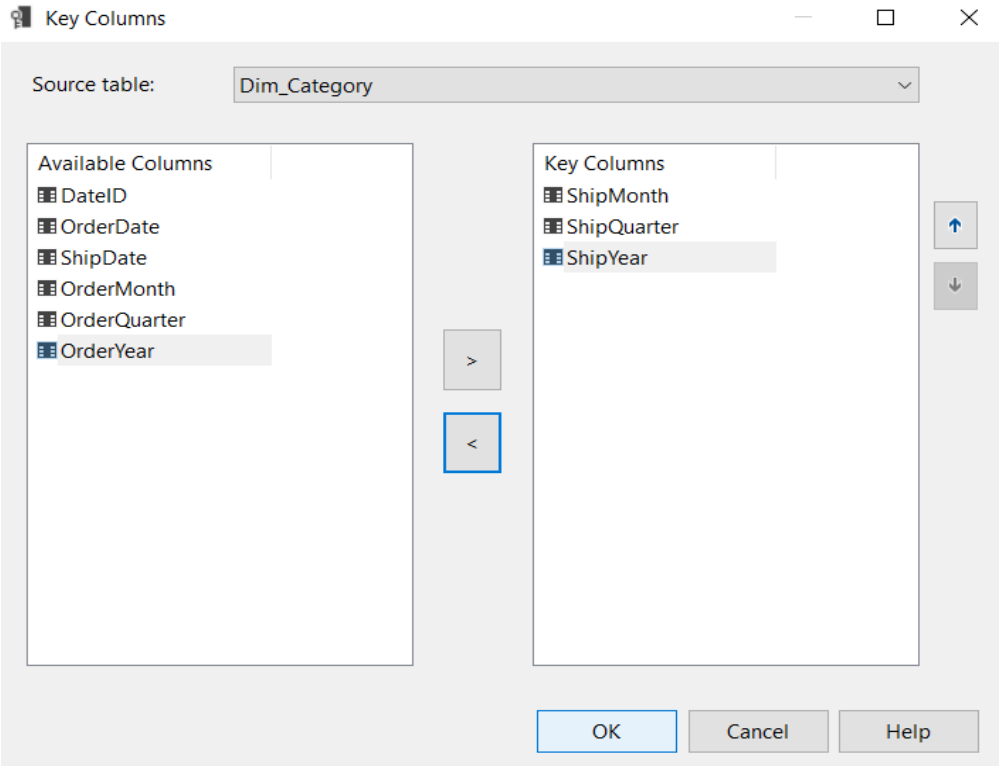
Hình 3 61 Định nghĩa Key Columns cho thuộc tính phân cấp ShipDat.

Bước 5b: Chọn **Name Column** cho ShipDate là **"ShipDate"** và nhấn **"Ok".**



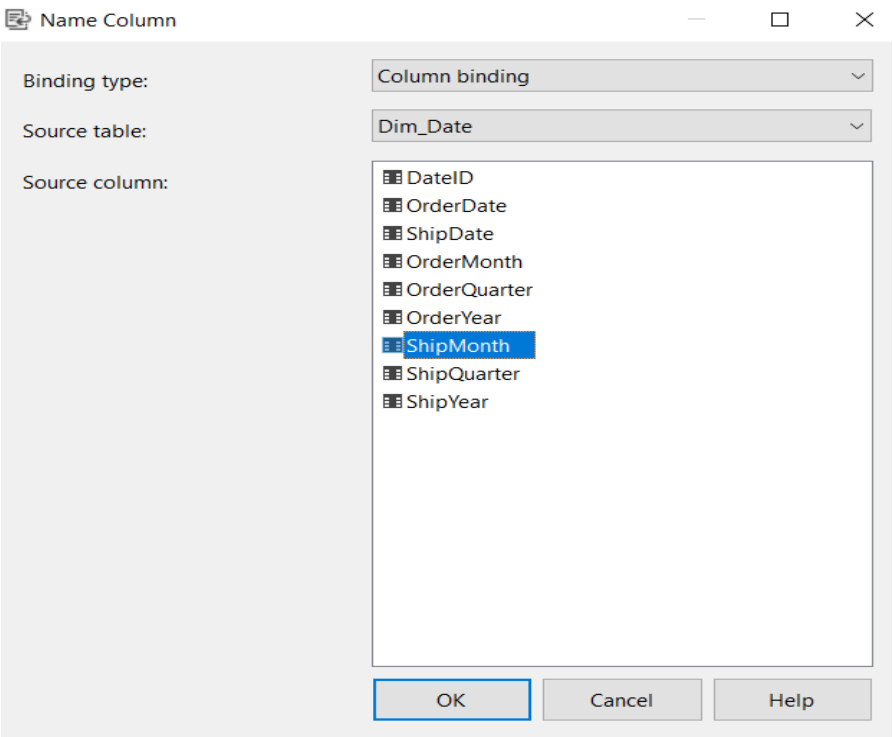
Hình 3.62 Định nghĩa Name Column cho thuộc tính phân cấp ShipDate

Bước 5c: Ở thuộc tính **"ShipMonth"**, vì nó là thuộc tính cấp nhỏ hơn **"ShipQuarter", "ShipYear"** nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là **"ShipQuarter", "ShipYear",** sau đó nhấn "Ok".



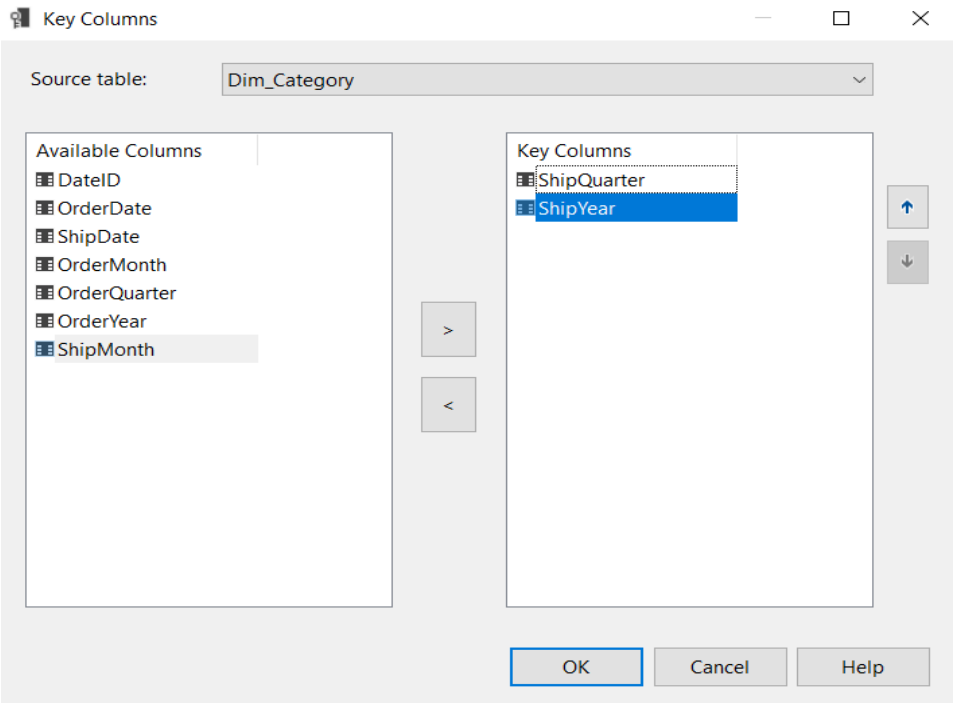
Hình 3.63 Định nghĩa lại Key Columns cho thuộc tính phân cấp ShipMonth

Bước 5d: Chọn **Name Column** cho **ShipMonth** là **"ShipMonth"** và nhấn "Ok".

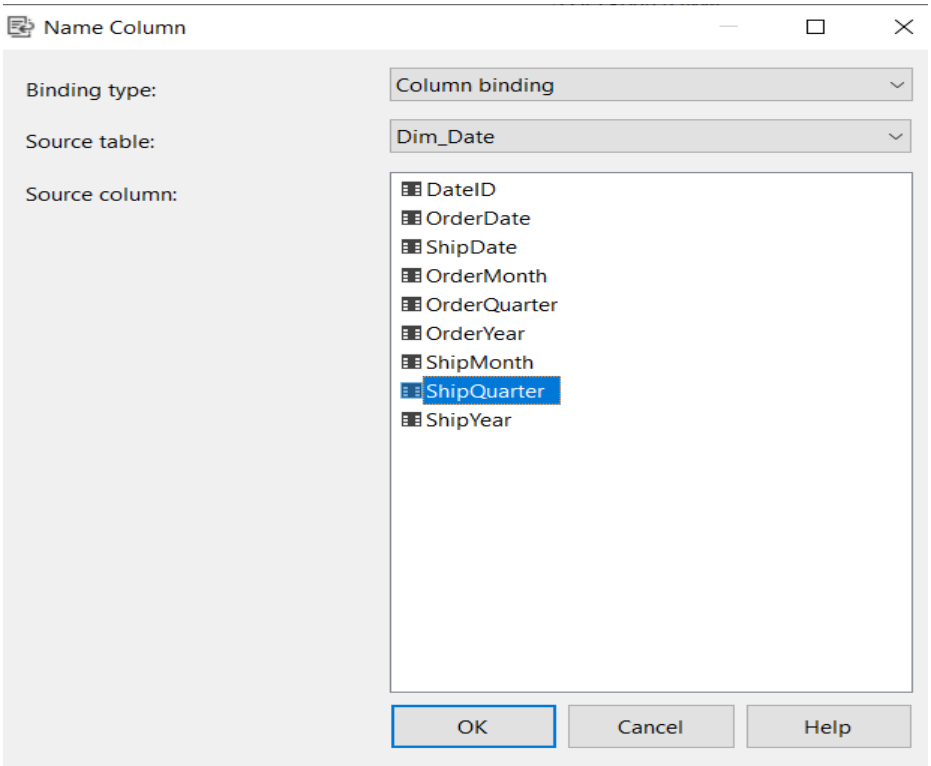


Hình 3.64 Định nghĩa Column Name cho thuộc tính phân cấp ShipMonth

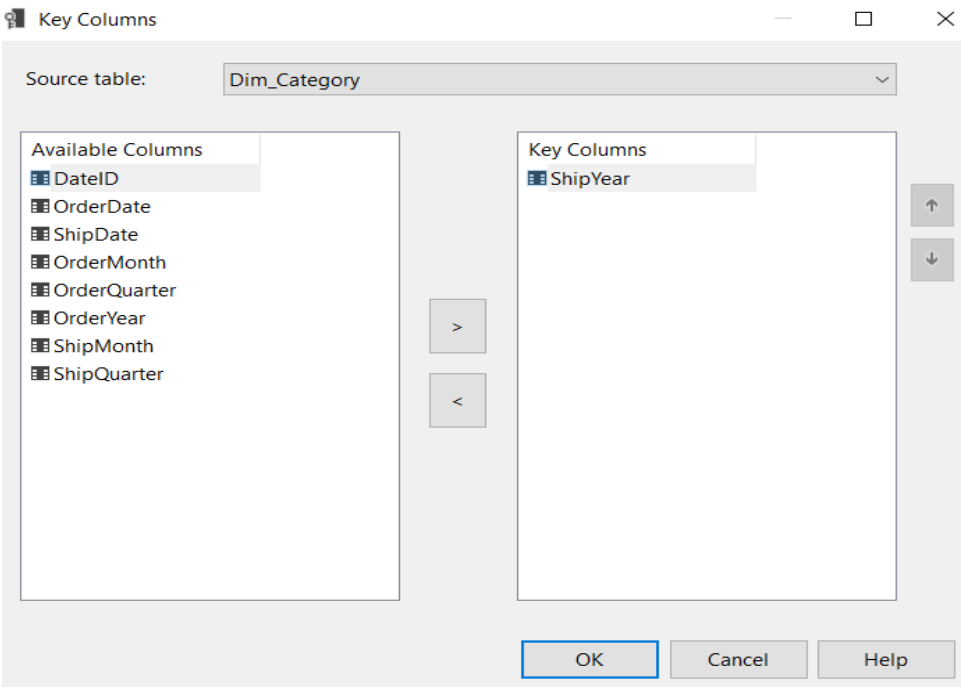
Bước 5e: Ở thuộc tính **"ShipQuarter"**, vì nó là thuộc tính cấp nhỏ hơn **"ShipYear"** nên sẽ lấy khóa dòng gồm chính nó và những thuộc tính cấp cao hơn là **"ShipYear",** sau đó nhấn "Ok".



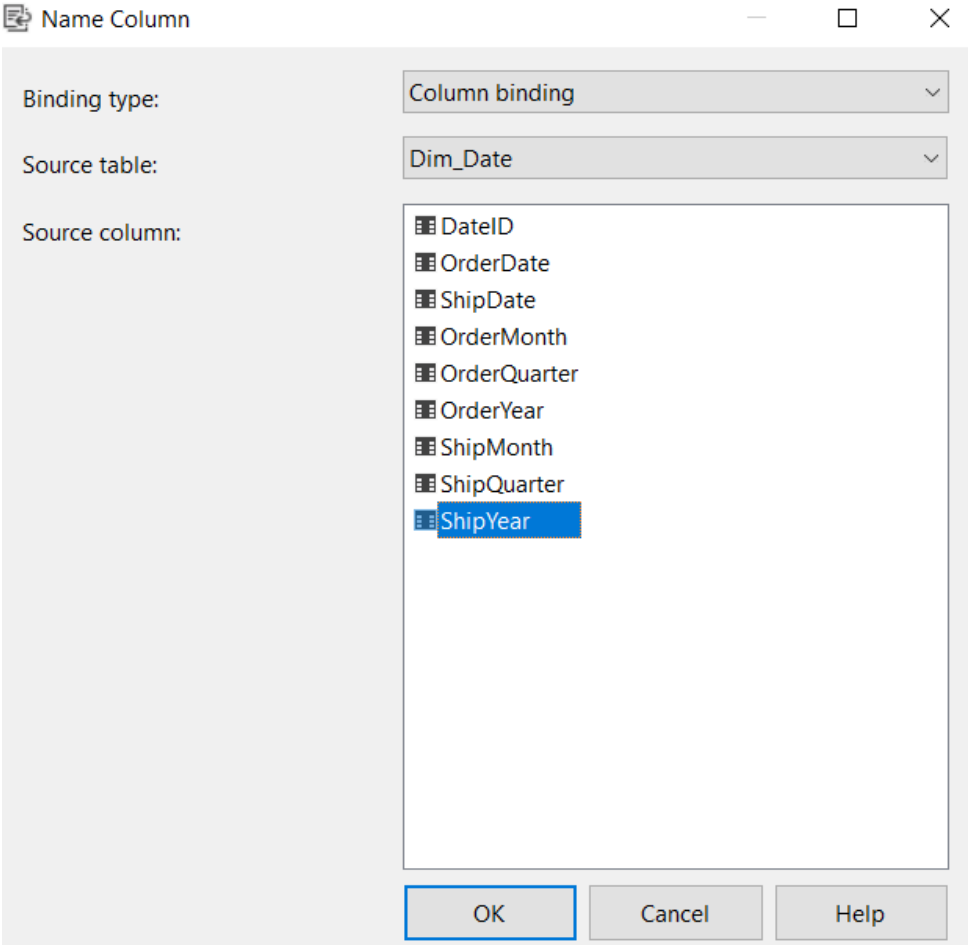
Bước 5f: Chọn Name Column cho ShipQuarter là “ShipQuarter” và nhấn “Ok”



Bước 5g: Ở thuộc tính “ShipYear”, vì nó là thuộc tính lớn nhất nên sẽ lấy khóa dòng gồm chính nó, sau đó nhấn “Ok”



Bước 5h: Chọn Name Column cho ShipYear là “ShipYear” và nhấn “Ok”

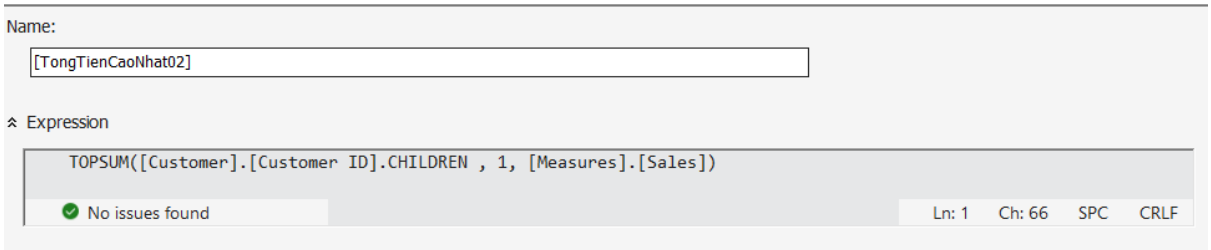


2. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ SSAS TRÊN CÁC KHỐI CUBE, NGÔN NGỮ MDX VÀ PIVOT TABLE

2.1. Cho biết khách hàng (mã khách hàng, tên khách hàng) có số tiền mua hàng cao nhất trong năm 2017?

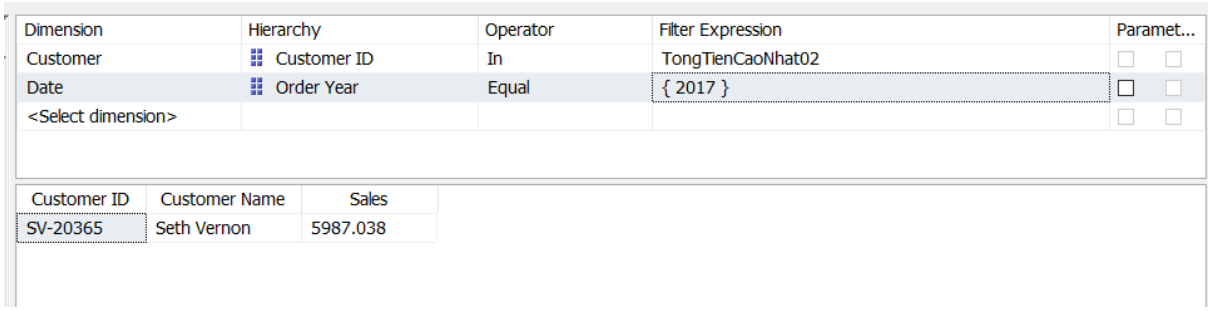
2.1.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo **Named Set** *[TongTienCaoNhat02]* như sau:



Hình 3.65 Tạo Named Set [TongTienCaoNhat02]

Bước 2: Tiến hành kéo các **Dimension** và chọn **Filter Expression** là 2017 và *[TongTienCaoNhat02]*; kết hợp với kéo độ đo Sales, Customer ID và Customer Name vào cửa sổ thực hiện truy vấn như ảnh để có kết quả bên dưới



Dimension	Hierarchy	Operator	Filter Expression	Paramet...
Customer	Customer ID	In	TongTienCaoNhat02	
Date	Order Year	Equal	{ 2017 }	
<Select dimension>				

Customer ID	Customer Name	Sales
SV-20365	Seth Vernon	5987.038

Hình 3.66 Thực hiện truy vấn trên Cube

2.1.2. Truy vấn bằng ngôn ngữ MDX:

```
--- cau1
select {[Measures].[Sales]
        as TongSoDonHang}
        on COLUMNS,
        {TOPSUM([Customer].[Customer ID].CHILDREN
        *[Customer].[Customer Name].CHILDREN, 1, [Measures].[Sales])}
        on ROWS
from [FurnitureStoreCube]
where [Date].[Order Year].&[2017]
```

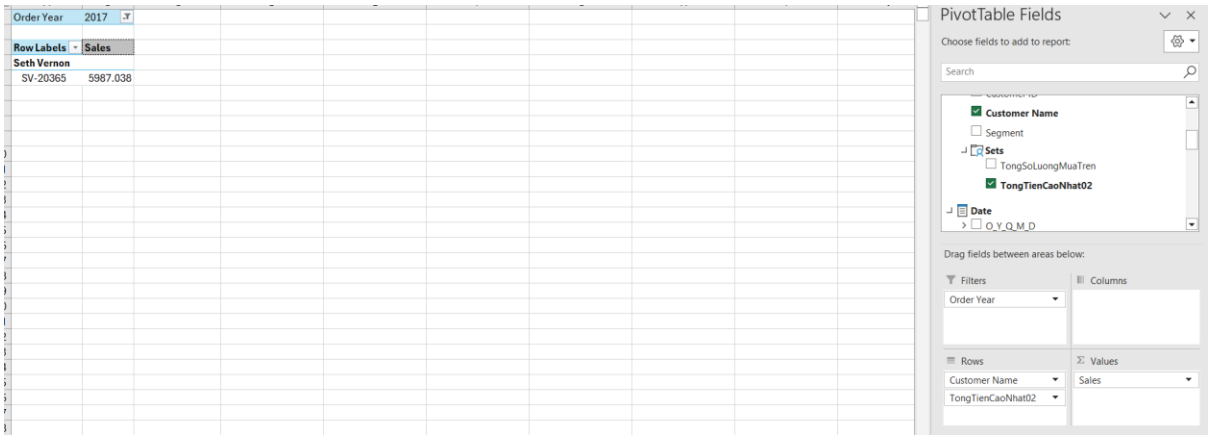
150 %

Messages | Results

		Sales
SV-20365	Seth Vernon	5987.038

Hình 3.67 Thực hiện truy vấn bằng ngôn ngữ MDX

2.1.3. Sử dụng Pivot Table



Hình 3. 68 Thực hiện truy vấn bằng PivotTable Excel

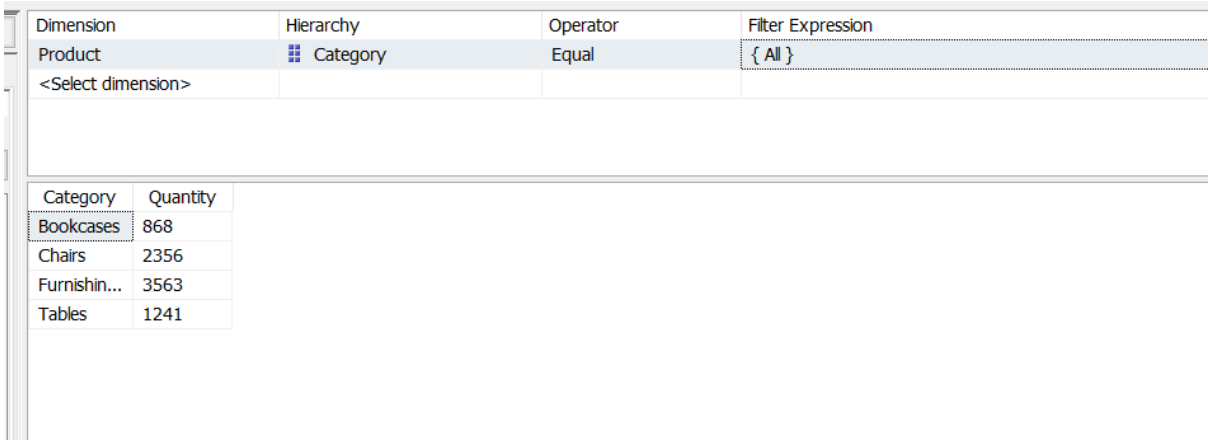
2.1.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau.

2.2. Đưa ra số lượng sản phẩm bán ra của từng loại nội thất?

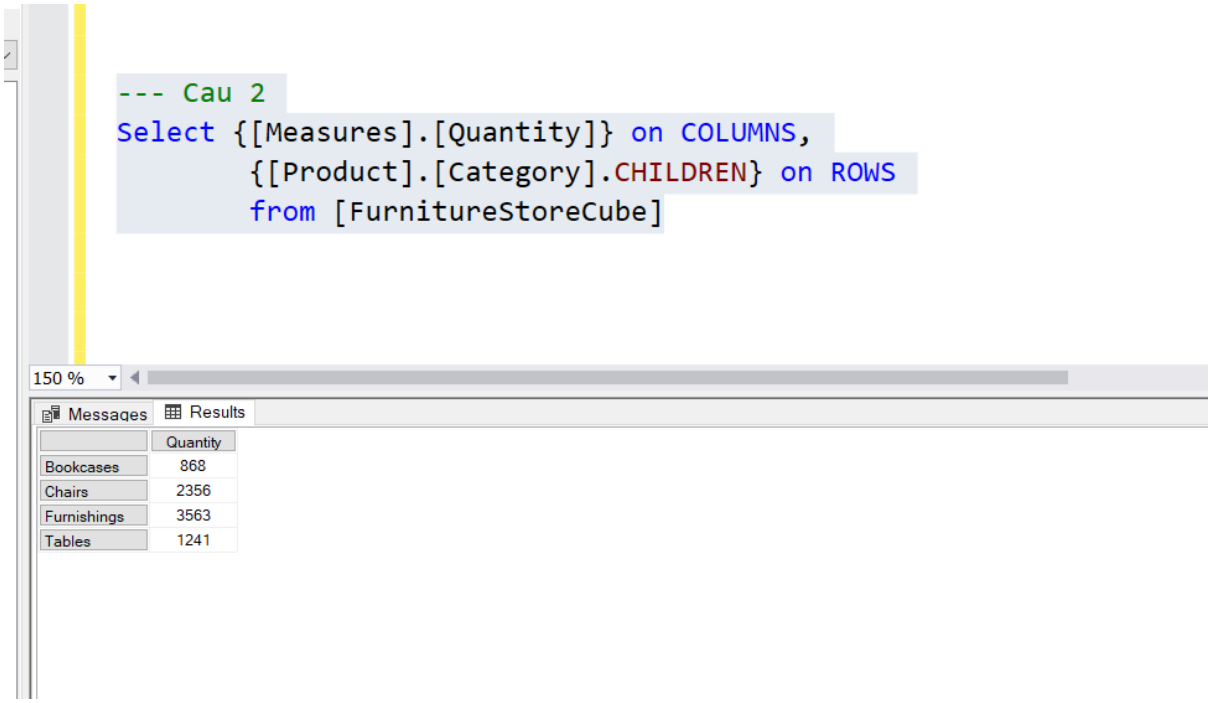
2.2.1. Phân tích dữ liệu trên Cube

Bước 1: Kéo độ đo Quantity và thuộc tính Category và độ đo Quantity vào cửa sổ thực hiện truy vấn để có kết quả như ảnh bên dưới



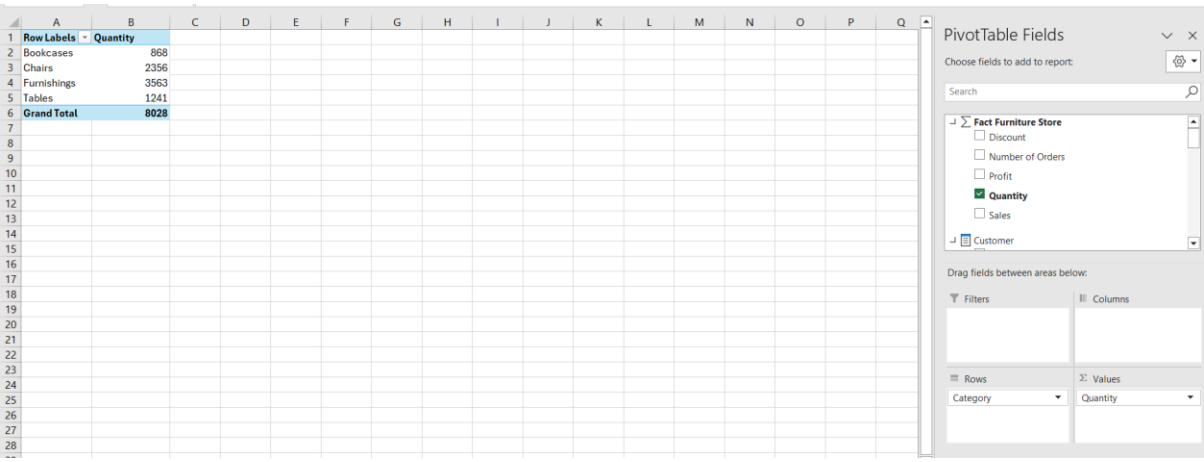
Hình 3.69 Thực hiện truy vấn trên Cube

2.2.2. Truy vấn bằng ngôn ngữ MDX



Hình 3.70 Thực hiện truy vấn bằng ngôn ngữ MDX

2.2.3. Sử dụng Pivot Table



Hình 3.71 Thực hiện truy vấn bằng PivotTable Excel

2.2.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.3. Cho biết thông tin khách hàng (Tên khách hàng) đặt hàng nhiều nhất trong bang Minnesota?

2.3.1. Phân tích dữ liệu trên Cube

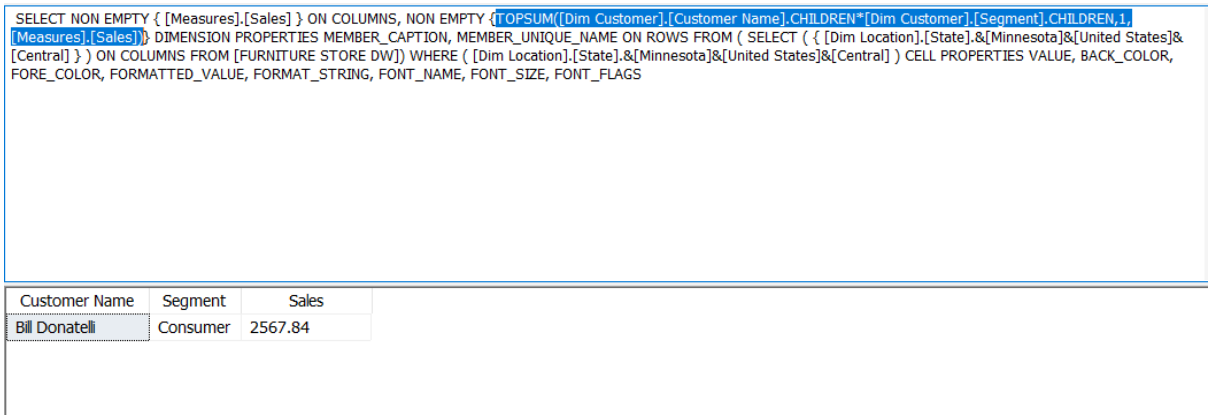
Bước 1: Chọn Dim Location, lọc theo điều kiện bang Minesota

Dimension	Hierarchy	Operator	Filter Expression
Dim Location	State	Equal	{ Minnesota }
<Select dimension>			

Customer Name	Sales
Benjamin Farhat	692.94
Bill Donatelli	2567.84
Carol Triggs	269.97
Charlotte Melton	607.52
Daniel Lacy	123.96
Dianna Wilson	2001.86
Gary Hansen	155.88
Justin Deggeller	18.84
Karen Bern	603.92
Karl Braun	53.34
Mark Packer	479.9
Philisse Overcash	29.22
Sung Pak	6.16

Hình 3.72 Chọn điều kiện lọc trên Cube

Chuyển sang chế độ ScriptMode và dán câu query vào vị trí như ảnh và thực hiện truy vấn để có kết quả



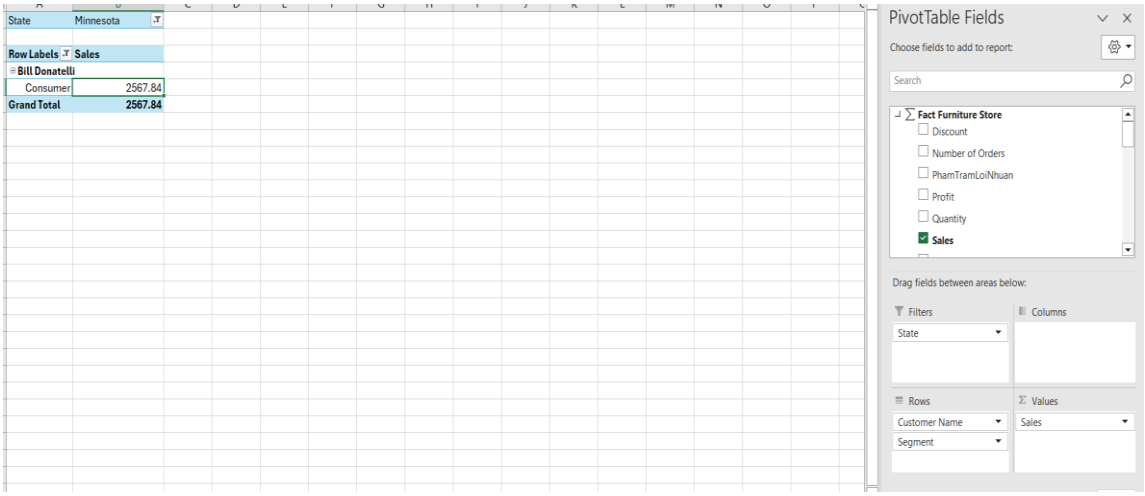
Hình 3.73 Chuyển sang chế độ Script Mode

2.3.2. Truy vấn bằng ngôn ngữ MDX

```
select {[Measures].[Sales]} on COLUMNS,
        {TOPSUM([Dim Customer].[Customer Name].CHILDREN
                *[Dim Customer].[Segment].CHILDREN, 1,
                [Measures].[Sales])
        } on ROWS
from [FURNITURE STORE DW]
where [Dim Location].[State].[Minnesota];
```

Hình 3.74 Thực hiện truy vấn bằng ngôn ngữ MDX

2.3.3. Sử dụng Pivot Table



Hình 3.75 Thực hiện truy vấn bằng PivotTable Excel

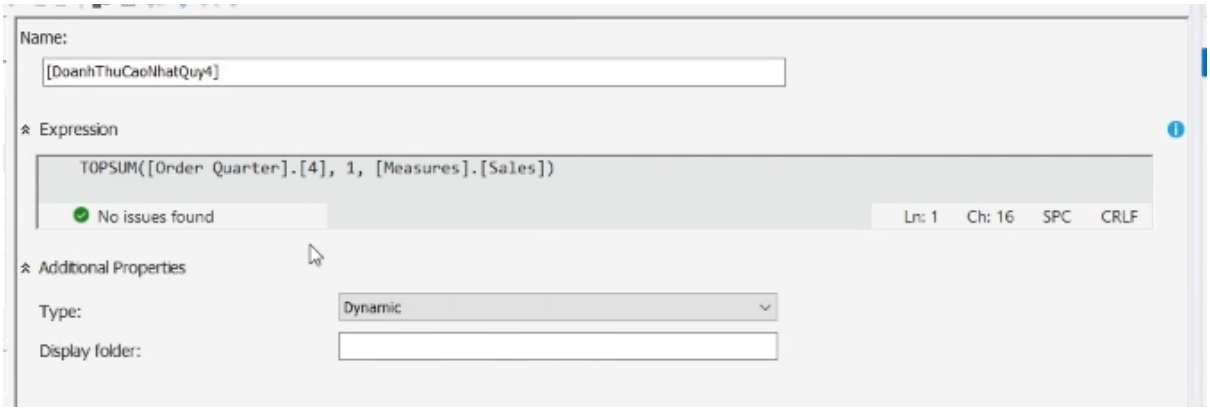
2.3.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.4. Quý 4 của năm nào có doanh thu bán hàng cao nhất?

2.4.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo **Named Set** *[DoanhThuCaoNhatQuy4]* như ảnh



Hình 3.76 Tạo NameSet [DoanhThuCaoNhatQuy4]

Bước 2: Chọn Dim Date, sử dụng **Filter Expression** là [DoanhThuCaoNhatQuy4] vừa tạo, kéo độ đo Sales và thuộc tính Order Year vào cửa sổ truy vấn, ta có kết quả

Hình 3. 77 Thực hiện truy vấn trên Cube

2.4.2. Truy vấn bằng ngôn ngữ MDX

```
----- Cau 4
select {[Measures].[Sales]}
        on COLUMNS,
        {TOPSUM([Date].[Order Year].CHILDREN*[Date].[Order Quarter].[4], 1, [Measures].[Sales])}
        on ROWS
from [FurnitureStoreCube]
```

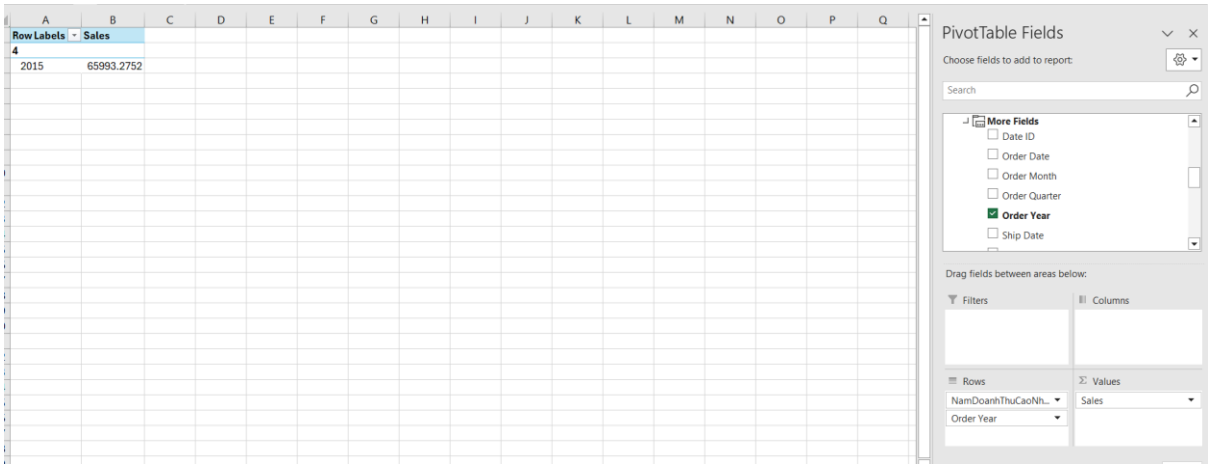
150 %

Messages | Results

		Sales
2015	4	65993.2752

Hình 3.78 Thực hiện truy vấn bằng ngôn ngữ MDX

2.4.3. Sử dụng Pivot Table



Hình 3. 79 Thực hiện truy vấn bằng PivotTable Excel

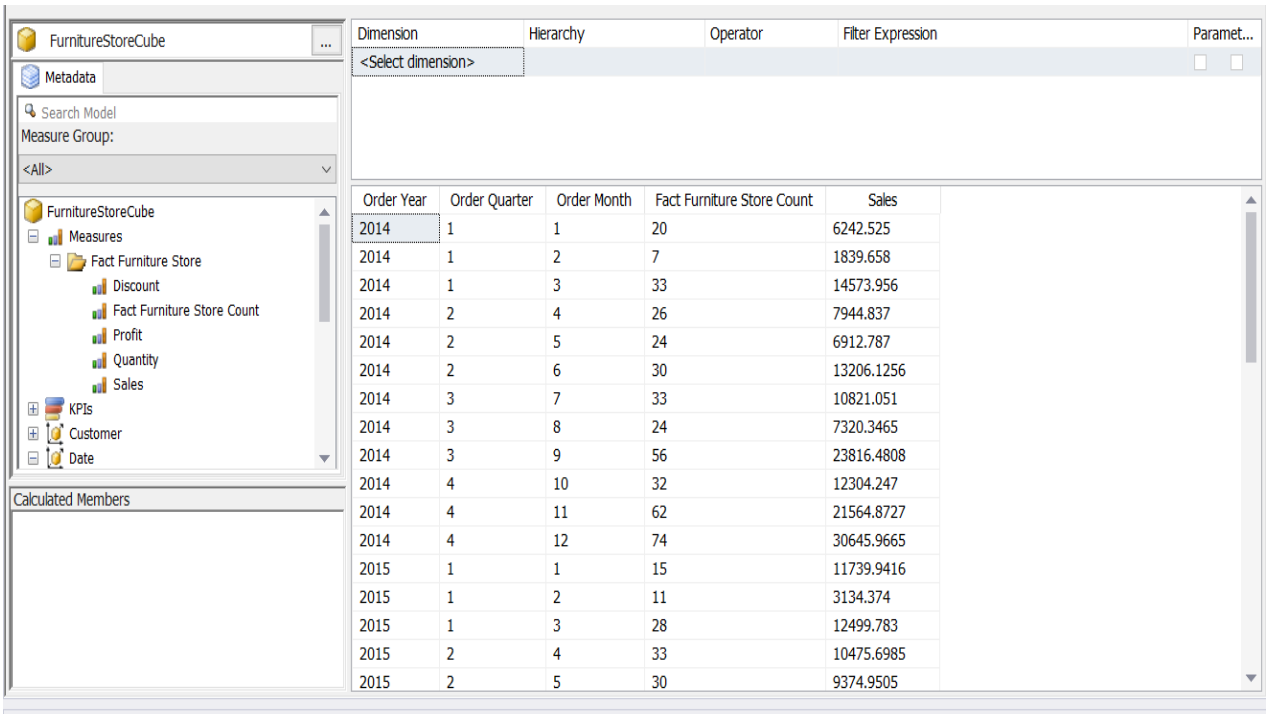
2.4.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.5. Theo từng tháng, quý, năm hãy liệt kê số đơn hàng đã được cửa hàng bán ra và tổng giá trị của các đơn hàng đó?

2.5.1. Phân tích dữ liệu trên Cube

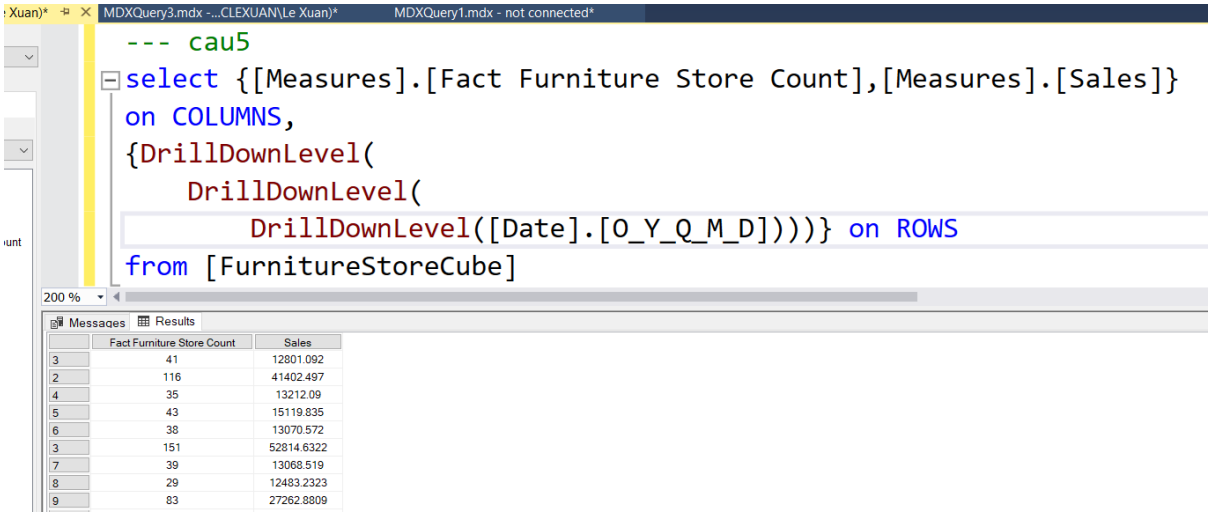
Bước 1: Kéo lần lượt thuộc tính Order Year, Order Quarter, Order Month và 2 độ đo Number of Orders và Sales để thực hiện truy vấn



Order Year	Order Quarter	Order Month	Fact Furniture Store Count	Sales
2014	1	1	20	6242.525
2014	1	2	7	1839.658
2014	1	3	33	14573.956
2014	2	4	26	7944.837
2014	2	5	24	6912.787
2014	2	6	30	13206.1256
2014	3	7	33	10821.051
2014	3	8	24	7320.3465
2014	3	9	56	23816.4808
2014	4	10	32	12304.247
2014	4	11	62	21564.8727
2014	4	12	74	30645.9665
2015	1	1	15	11739.9416
2015	1	2	11	3134.374
2015	1	3	28	12499.783
2015	2	4	33	10475.6985
2015	2	5	30	9374.9505

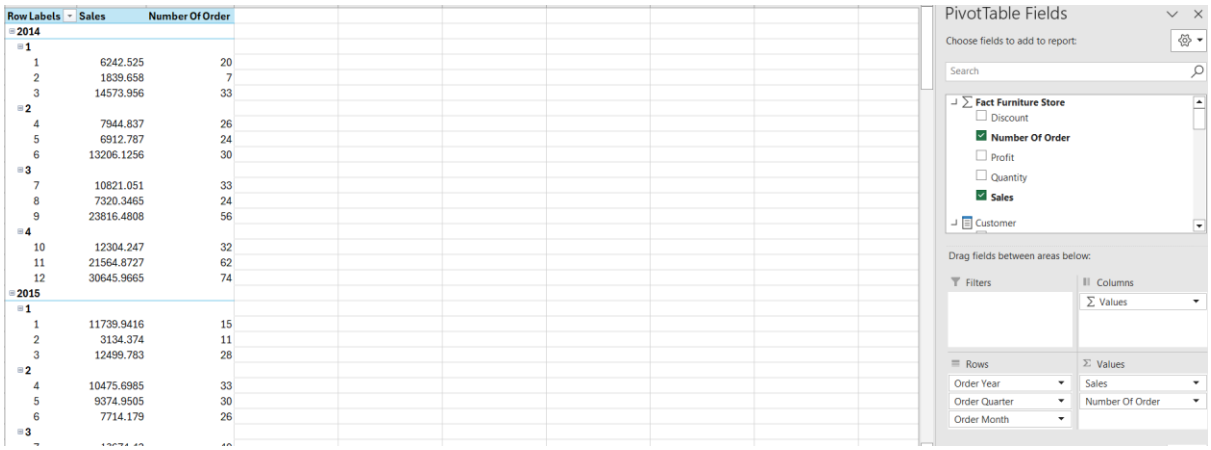
Hình 3.80 Thực hiện truy vấn trên Cube

2.5.2. Truy vấn bằng ngôn ngữ MDX



Hình 3.81 Thực hiện truy vấn bằng ngôn ngữ MDX

2.5.3. Sử dụng Pivot Table



Hình 3.82 Thực hiện truy vấn bằng PivotTable Excel

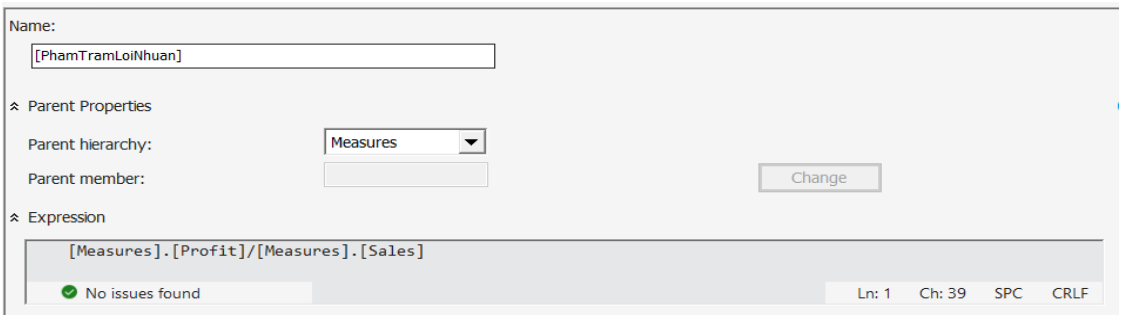
2.5.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.6. Liệt kê những thành phố có phần trăm lợi nhuận của đơn hàng trên 30%?

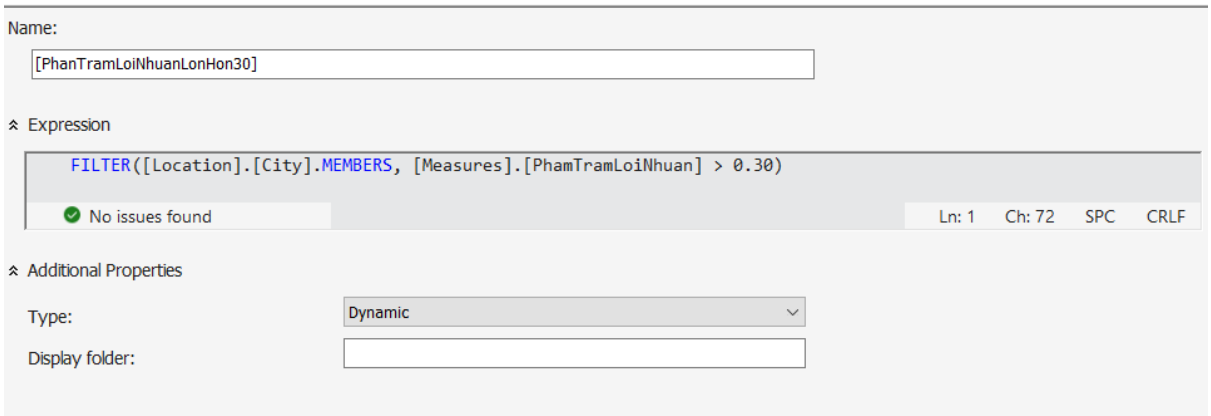
2.6.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo một **Calculated Member** *[PhamTramLoiNhuan]* như ảnh



Hình 3.83 Tạo Calculated Member [PhanTramLoiNhuan]

Bước 2: Tạo một **Named Set** *[PhanTramLoiNhuanLonHon30]*



Hình 3.84 Tạo NameSet [PhanTramLoiNhuanLonHon30]

Bước 3: Chọn Dim Location, sử dụng **Filter Expression** là *[PhanTramLoiNhuanLonHon30]* vừa tạo, kéo **Calculated Member** *[PhanTramLoiNhuan]* và thuộc tính City vào cửa sổ truy vấn, ta có kết quả:

Dimension	Hierarchy	Operator	Filter Expression
Location	City	In	PhanTramLoiNhuanLonHon30
<Select dimension>			

City	PhamTramLoiNhuan
Alexandria	0.390533911612952
Auburn	0.36
Bangor	0.31
Bellevue	0.39
Brentwood	0.44
Cambridge	0.44
Canton	0.39
Cincinnati	0.375
Citrus Heights	0.42
Danbury	0.36
Danville	0.42
Dearborn	0.305092613009923
Draper	0.42
East Point	0.34
Fairfield	0.353072439633639
Fayetteville	0.401011423873912
Fremont	0.44

Hình 3.85 Thực hiện truy vấn trên Cube

2.6.2. Truy vấn bằng ngôn ngữ MDX

```
--- cau6
select {[Measures].[PhamTramLoiNhuan]} on COLUMNS,
        {FILTER([Location].[City].MEMBERS,
         [Measures].[PhamTramLoiNhuan] > 0.30)} on ROWS
from [FurnitureStoreCube]
```

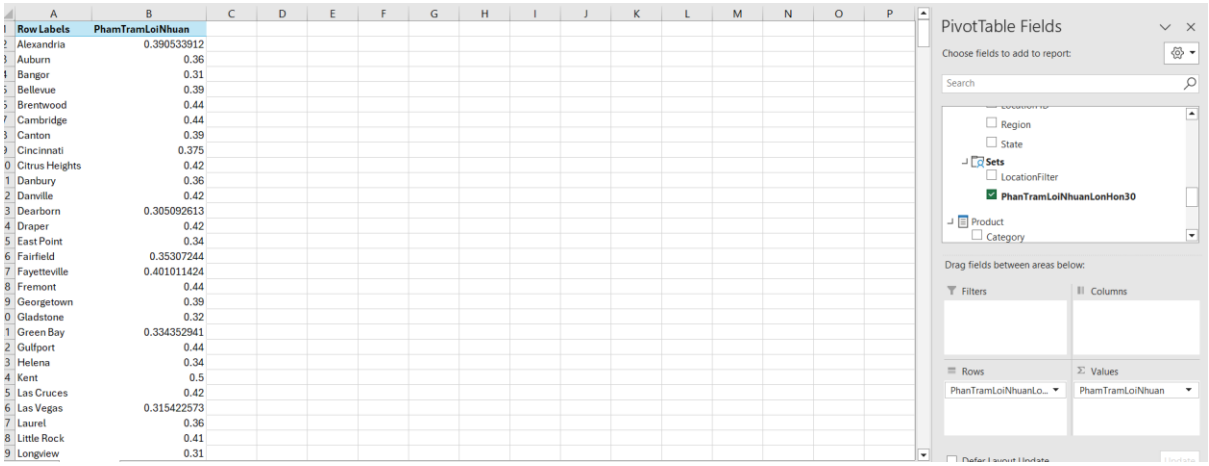
150 %

Messages | Results

	PhamTramLoiNhuan
Alexandria	0.390533911612952
Auburn	0.36
Bangor	0.31
Bellevue	0.39
Brentwood	0.44
Cambridge	0.44
Canton	0.39
Cincinnati	0.375
Citrus Heights	0.42
Danbury	0.36
Danville	0.42
Dearborn	0.305092613009923
Draper	0.42
East Point	0.34
Fairfield	0.353072439633639
Fayetteville	0.401011423873912

Hình 3.86 Thực hiện truy vấn bằng ngôn ngữ MDX

2.6.3. Sử dụng Pivot Table



Row Labels	PhamTramLoiNhuan
Alexandria	0.390533912
Auburn	0.36
Bangor	0.31
Bellevue	0.39
Brentwood	0.44
Cambridge	0.44
Canton	0.39
Cincinnati	0.375
Citrus Heights	0.42
Danbury	0.36
Danville	0.42
Dearborn	0.305092613
Draper	0.42
East Point	0.34
Fairfield	0.35307244
Fayetteville	0.401011424
Fremont	0.44
Georgetown	0.39
Gladstone	0.32
Green Bay	0.334352941
Gulfport	0.44
Helena	0.34
Kent	0.5
Las Cruces	0.42
Las Vegas	0.315422573
Laurel	0.36
Little Rock	0.41
Longview	0.31

Hình 3.87 Thực hiện truy vấn bằng PivotTable Excel

2.6.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.7. Hãy đưa ra tỉ lệ lợi nhuận thu được của top 3 sản phẩm có số lượng mua cao nhất của từng Bang?

2.7.1. Phân tích dữ liệu trên Cube

Bước 1: Copy đoạn query sau vào bộ nhớ tạm

```
Generate([Location].[State].CHILDREN,TOPCOUNT([Location].[State].CURRENTMEMBER*[Product].[Product Name].MEMBERS,3, [Measures].[PhamTramLoiNhuan]))
```

Hình 3.88 Copy đoạn query vào bộ nhớ tạm

Bước 2: Kéo các độ đo doanh thu, chiết khấu, **Calculated Member** *[PhanTramLoiNhuan]* và thực hiện truy vấn, sau đó, chuyển qua **Script Mode** và dán đoạn query vào vị trí bôi xanh như ảnh

```
SELECT NON EMPTY { [Measures].[PhamTramLoiNhuan] } ON COLUMNS, NON EMPTY { ([Location].[State].[State].ALLMEMBERS * [Product].[Product Name].[Product Name].ALLMEMBERS ) }
DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [FurnitureStoreCube] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE,
FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Hình 3.89 Chuyển sang ScriptMode

Bước 3: Bấm **Execute Query**, ta có kết quả:

```
SELECT NON EMPTY { [Measures].[PhamTramLoiNhuan] } ON COLUMNS, NON EMPTY {Generate([Location].[State].CHILDREN,TOPCOUNT([Location].[State].CURRENTMEMBER*[Product].[Product
Name].MEMBERS,3, [Measures].[PhamTramLoiNhuan])) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [FurnitureStoreCube] CELL PROPERTIES VALUE,
BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

State	Product Name	PhamTramLoiNhuan
Alabama	Acrylic Self-St...	0.38
Alabama	Master Caster...	0.34
Alabama	Stacking Tray...	0.31
Arizona	GE 48" Fluore...	0.375
Arizona	Eldon Wave ...	0.3
Arizona	Flat Face Post...	0.275
Arkansas	Eldon Pizzaz ...	0.44
Arkansas	3M Hangers ...	0.41
Arkansas	Howard Miller ...	0.4
California	GE 48" Fluore...	0.5

2.7.2. Truy vấn bằng ngôn ngữ MDX

```
--- cau7
select {[Measures].[PhamTramLoiNhuan]} on COLUMNS,
        {Generate([Location].[State].MEMBERS,
        TOPCOUNT([Location].[State].CURRENTMEMBER*[Product].[Product Name].MEMBERS,
                3, [Measures].[PhamTramLoiNhuan]))} on ROWS
from [FurnitureStoreCube]
```

50 %

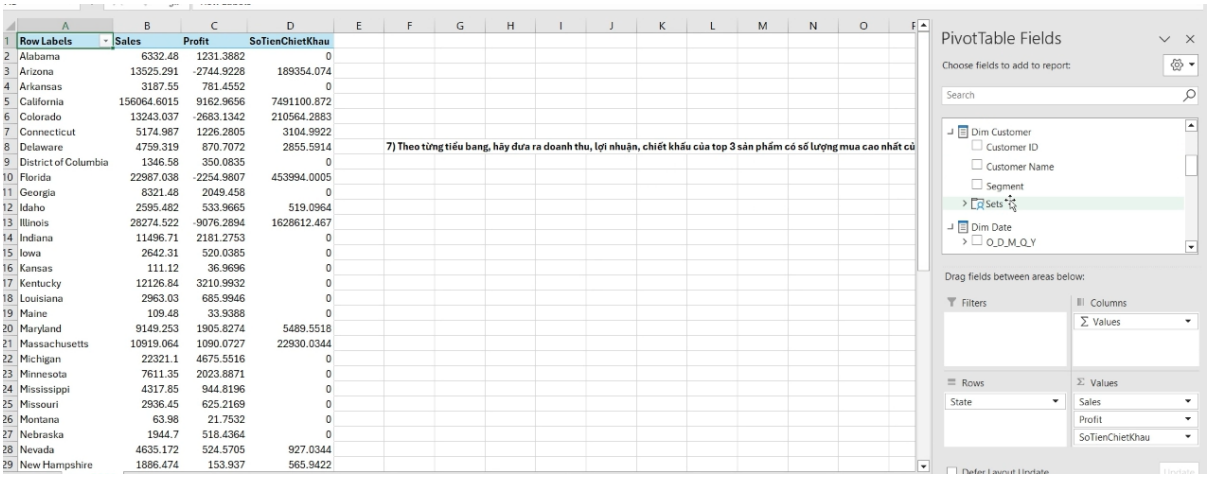
Messages | Results

		PhamTramLoiNhuan
Alabama	Acrylic Self-Standing Desk Frames	0.38
Alabama	Master Caster Door Stop, Gray	0.34
Alabama	Stacking Tray, Side-Loading, Legal, Smoke	0.31
Arizona	GE 48" Fluorescent Tube, Cool White Energy Saver, 34 Watts, 30/Box	0.375
Arizona	Eldon Wave Desk Accessories	0.3
Arizona	Flat Face Poster Frame	0.275
Arkansas	Eldon Pizzaz Desk Accessories	0.44
Arkansas	3M Hangers With Command Adhesive	0.41
Arkansas	Howard Miller 13-1/2" Diameter Rosebrook Wall Clock	0.4
California	GE 48" Fluorescent Tube, Cool White Energy Saver, 34 Watts, 30/Box	0.5
California	Floodlight Indoor Halogen Bulbs, 1 Bulb per Pack, 60 Watts	0.49
California	Longer-Life Soft White Bulbs	0.48
Colorado	Electrix 20W Halogen Replacement Bulb for Zoom-In Desk Lamp	0.35
Colorado	DAX Charcoal/Nickel-Tone Document Frame, 5 x 7	0.3125
Colorado	OIC Stacking Trays	0.3
Connecticut	3M Hangers With Command Adhesive	0.41

Hình 3.90 Thực hiện truy vấn bằng ngôn ngữ MDX

2.7.3. Sử dụng Pivot Table

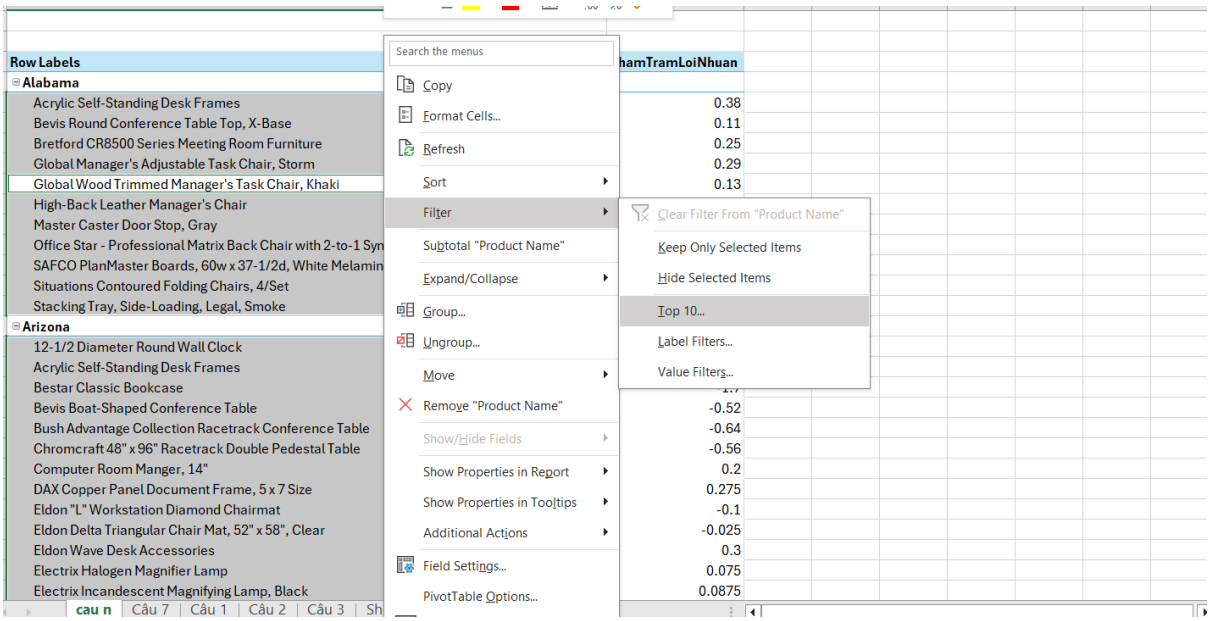
Bước 1: Kéo các thuộc tính và độ đo vào vị trí thích hợp như ảnh



Row Labels	Sales	Profit	SoTienChietKhau
Alabama	6332.48	1231.3882	0
Arizona	13525.291	-2744.9228	189354.074
Arkansas	3187.55	781.4552	0
California	156064.6015	9162.9656	7491100.872
Colorado	13243.037	-2683.1342	210564.2883
Connecticut	5174.987	1226.2805	3104.9922
Delaware	4759.319	870.7072	2855.5914
District of Columbia	1346.58	350.0835	0
Florida	22987.038	-2254.9807	453994.0005
Georgia	8321.48	2049.458	0
Idaho	2595.482	533.9665	519.0964
Illinois	28274.522	-9076.2894	1628612.467
Indiana	11496.71	2181.2753	0
Iowa	2642.31	520.0385	0
Kansas	111.12	36.9696	0
Kentucky	12126.84	3210.9932	0
Louisiana	2963.03	685.9946	0
Maine	109.48	33.9388	0
Maryland	9149.253	1905.8274	5489.5518
Massachusetts	10919.064	1090.0727	22930.0344
Michigan	22321.1	4675.5516	0
Minnesota	7611.35	2023.8871	0
Mississippi	4317.85	944.8196	0
Missouri	2936.45	625.2169	0
Montana	63.98	21.7532	0
Nebraska	1944.7	518.4364	0
Nevada	4635.172	524.5705	927.0344
New Hampshire	1886.474	153.937	565.9422

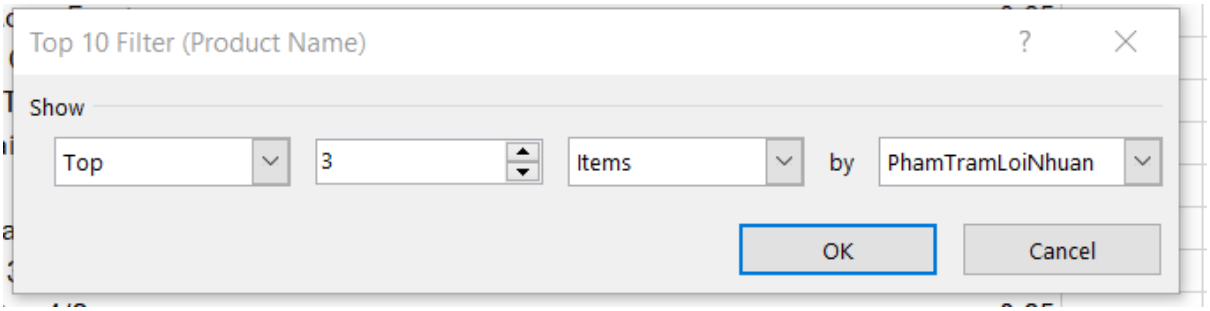
Hình 3. 91 Kéo thuộc tính và độ đo vào vị trí thích hợp

Bước 2: Sử dụng chức năng **Filter Top 10** sau khi đã chọn được tất cả các sản phẩm trong từng tiểu bang



Hình 3.92 Sử dụng Filter Top 10

Bước 3: Chọn **Filter Top 3 Items** và độ đo là PhanTramLoiNhuan



Hình 3. 93 Chọn Filter Top 3 và độ đo

Bước 4: Nhấn **"Ok"** và kết quả sẽ như ảnh sau

	A	B	C	D	E
1	Row Labels	Sales	Profit	SoTienChietKhau	
2	Alabama				
3	Acrylic Self-Standing Desk Frames	21.36	8.1168	0	
4	High-Back Leather Manager's Chair	1819.86	163.7874	0	
5	SAFCO PlanMaster Boards, 60w x 37-1/2d, White Melamine	1215.92	316.1392	0	
6	Arizona				
7	Flat Face Poster Frame	195.936	53.8824	78.3744	
8	Howard Miller 13" Diameter Pewter Finish Round Wall Clock	274.816	65.2688	109.9264	
9	KI Adjustable-Height Table	386.91	-185.7168	193.455	
10	Arkansas				
11	3M Hangers With Command Adhesive	22.2	9.102	0	
12	Hon Metal Bookcases, Gray	638.82	172.4814	0	
13	Office Star Flex Back Scooter Chair with Aluminum Finish Frame	605.34	145.2816	0	
14	California				
15	Chromcraft Round Conference Tables	3346.368	41.8296	2677.0944	
16	Dana Halogen Swing-Arm Architect Lamp	983.28	275.3184	0	
17	Lesro Sheffield Collection Coffee Table, End Table, Center Table, Corner Table	1427.4	-53.5275	1427.4	
18	Colorado				
19	Aluminum Document Frame	97.76	12.22	39.104	
20	Master Caster Door Stop, Large Neon Orange	52.416	15.0696	10.4832	
21	SAFCO Arco Folding Chair	1988.64	223.722	795.456	
22	Connecticut				
23	3M Hangers With Command Adhesive	22.2	9.102	0	
24	Hon GuestStacker Chair	1133.35	294.671	0	
25	Hon Metal Bookcases, Putty	638.82	185.2578	0	
26	Delaware				
27	Iceberg Nesting Folding Chair, 19w x 6d x 43h	291.1	75.686	0	
28	Office Star - Contemporary Swivel Chair with Padded Adjustable Arms and Flex Back	1268.82	266.4522	0	
29	Sauder Inglewood Library Bookcases	1025.88	235.9524	0	

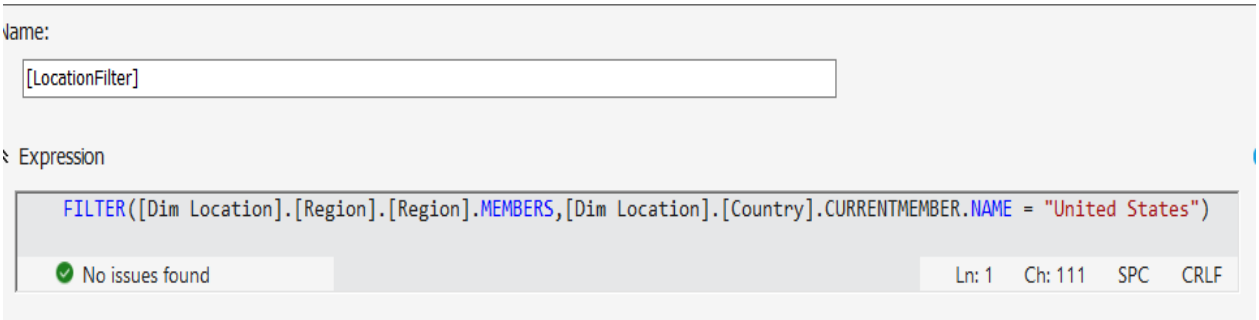
2.7.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.8. Cho biết tổng số tiền cửa hàng đã chiết khấu cho khách hàng khi mua loại sản phẩm 'Bookcases' theo từng khu vực của nước Mỹ?

2.8.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo mới một **Named Set** *[LocationFilter]* như ảnh



Hình 3.94 Tạo NamedSet [LocationFilter]

Bước 2: Chọn Dim Location, Product, sau đó chọn **Filter Expression** là *[LocationFilter]* và lọc các sản phẩm có loại là "Bookcases" và chọn **Calcalated Member** là *[SoTienChietKhau]*, thuộc tính là Region để có kết quả

Dimension	Hierarchy	Operator	Filter Expression
Location	Region	In	LocationFilter
Product	Category	Equal	{ Bookcases }
<Select dimension>			

Region	SoTienChietKhau
Central	281189.537952
East	674817.7436
South	30518.2136
West	658875.46005

Hình 3.95 Thực hiện truy vấn trên Cube

2.8.2. Truy vấn bằng ngôn ngữ MDX

```
--- cau 8
select {[Measures].[SoTienChietKhau]} on COLUMNS,
        {FILTER(
        [Location].[Region].[Region].MEMBERS,
        [Location].[Country].CURRENTMEMBER.NAME = "United States"
        )} on ROWS
from [FurnitureStoreCube]
where [Product].[Category].&[Bookcases]
```

150 %

Messages | Results

	SoTienChietKhau
Central	281189.537952
East	674817.7436
South	30518.2136
West	658875.46005

Hình 3.96 Thực hiện truy vấn bằng ngôn ngữ MDX

2.8.3. Sử dụng Pivot Table

Category	Bookcases
Row Labels	**Sales**
Central	24157.1768
East	43819.334
South	10899.362
West	36004.1235

Choose fields to add to report:

Search

Sets
- LocationFilter

Product
- Category
- Category ID
- Product ID
- Product Name

Drag fields between areas below:

Filters: Category

Columns

Rows: LocationFilter

Values: Sales

Hình 3.97 Thực hiện truy vấn bằng PivotTable Excel

2.8.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.9. Với từng loại khách hàng, hãy đưa ra tổng số lượng các đơn hàng được bán (quantity) theo từng loại hình vận chuyển?

2.9.1. Phân tích dữ liệu trên Cube

Bước 1: Kéo các thuộc tính Segment, ShipMode và độ đo Number of Orders sau đó thực hiện truy vấn để có kết quả

Dimension	Hierarchy	Operator	Filter Expression
<Select dimension>			

Segment	Ship Mode	Number of Orders
Consumer	First Class	164
Consumer	Same Day	66
Consumer	Second C...	236
Consumer	Standard ...	647
Corporate	First Class	104
Corporate	Same Day	28
Corporate	Second C...	121
Corporate	Standard ...	393
Home Office	First Class	59
Home Office	Same Day	25
Home Office	Second C...	70
Home Office	Standard ...	208

Hình 3.98 Thực hiện truy vấn trên Cube

2.9.2. Truy vấn bằng ngôn ngữ MDX

```
---- cau 9
select {[Measures].[Number of Orders]} on COLUMNS,
       {Generate([Customer].[Segment].[Segment].MEMBERS,
       [Customer].[Segment].CURRENTMEMBER *
       [Ship Mode].[Ship Mode].[Ship Mode].MEMBERS)} on ROWS
from [FurnitureStoreCube]
```

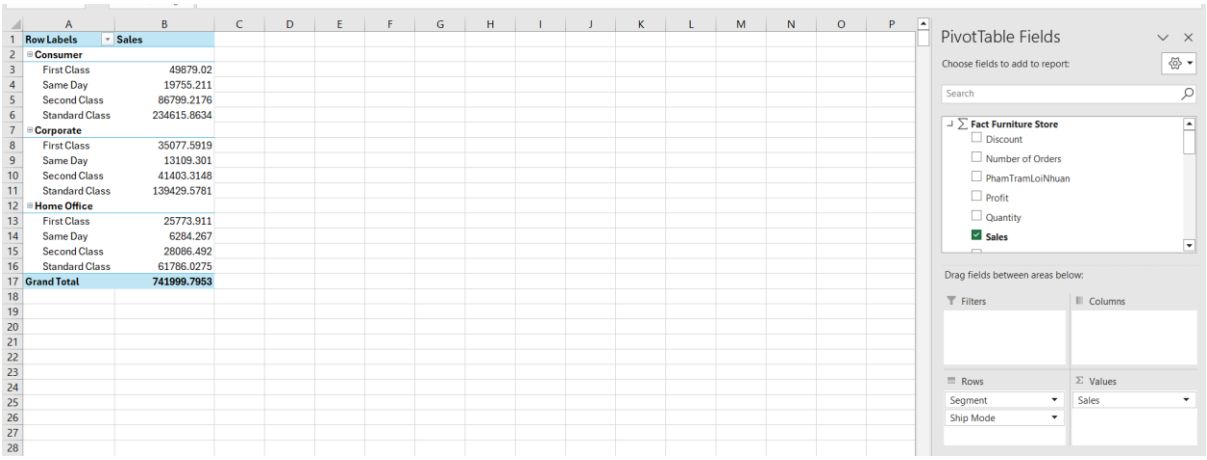
150 %

Messages | Results

		Number of Orders
Consumer	First Class	164
Consumer	Same Day	66
Consumer	Second Class	236
Consumer	Standard Class	647
Corporate	First Class	104
Corporate	Same Day	28
Corporate	Second Class	121
Corporate	Standard Class	393
Home Office	First Class	59
Home Office	Same Day	25
Home Office	Second Class	70
Home Office	Standard Class	208

Hình 3.99 Thực hiện truy vấn bằng ngôn ngữ MDX

2.9.3. Sử dụng Pivot Table



Row Labels	Sales
Consumer	
First Class	49879.02
Same Day	19755.211
Second Class	86799.2176
Standard Class	234615.8634
Corporate	
First Class	35077.5919
Same Day	13109.301
Second Class	41403.3148
Standard Class	139429.5781
Home Office	
First Class	25773.911
Same Day	6284.267
Second Class	28086.492
Standard Class	61786.0275
Grand Total	741999.7953

Hình 3.100 Thực hiện truy vấn bằng PivotTable Excel

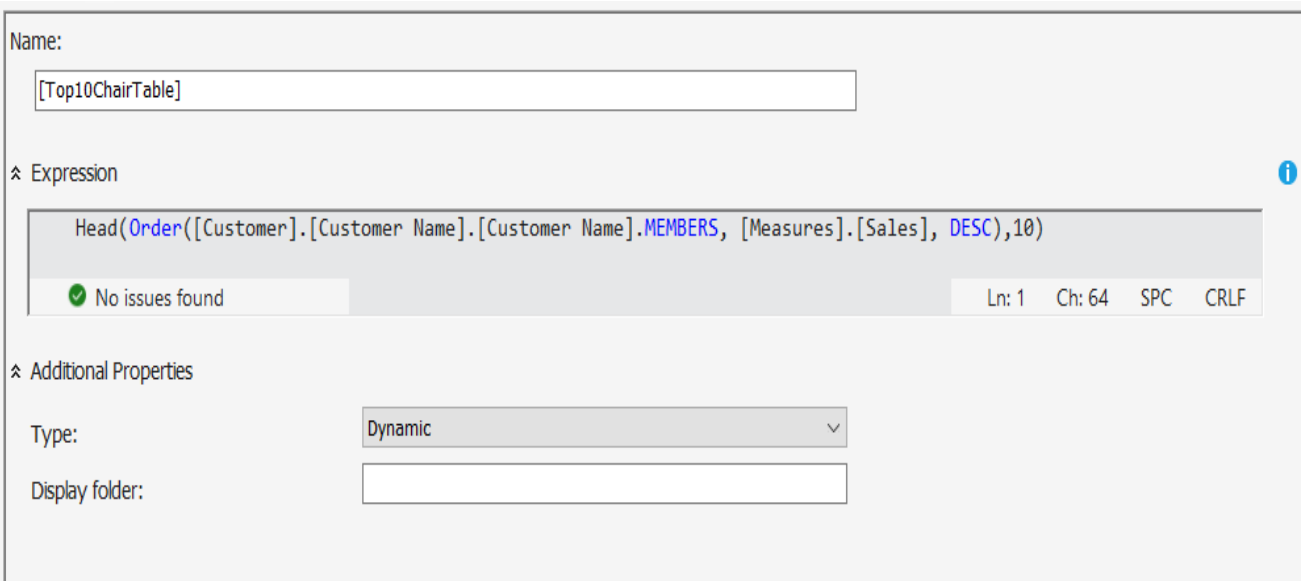
2.9.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.10. Hãy liệt kê thông tin khách hàng và loại khách hàng chi tiền mua sản phẩm thuộc phân loại tables hoặc chairs nhiều nhất?

2.10.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo một **Named Set** *[Top10ChairTable]* như ảnh



Hình 3.101 Tạo NamedSet [Top10ChairTable]

Bước 2: Chọn Dim Customer, Product, sau đó chọn Filter Expression là Top10ChairTable và lọc các sản phẩm có loại là “Chair” và “Table”, tiến hành kéo thuộc tính Customer Name và độ đo Sales vào cửa sổ truy vấn để có kết quả

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Customer	Customer Name	In	Top10ChairTable	
Product	Category	Equal	{ Chairs, Tables }	
<Select dimension>				

Customer Name	Sales
Bill Donatelli	4182.243
Brenda Bowman	4768.498
Caroline Jumper	4991.75
Greg Tran	2166.96
Joe Elijah	6842.408
Laura Armstrong	4921.424
Nora Preis	4584.684
Peter Fuller	199.764
Seth Vernon	8024.678
Tom Prescott	4416.174

Hình 3.102 Thực hiện truy vấn trên Cube

2.10.2. Sử dụng ngôn ngữ MDX

```
--- câu 10
SELECT {[Measures].[Sales]} ON COLUMNS,
NON EMPTY
    Head(
        Order(
            [Customer].[Customer Name].[Customer Name].MEMBERS,
            [Measures].[Sales],
            DESC
        ),
        10
    )
    ON ROWS
FROM [FurnitureStoreCube]
WHERE ({
    [Product].[Category].[Chairs],
    [Product].[Category].[Tables]
})
```

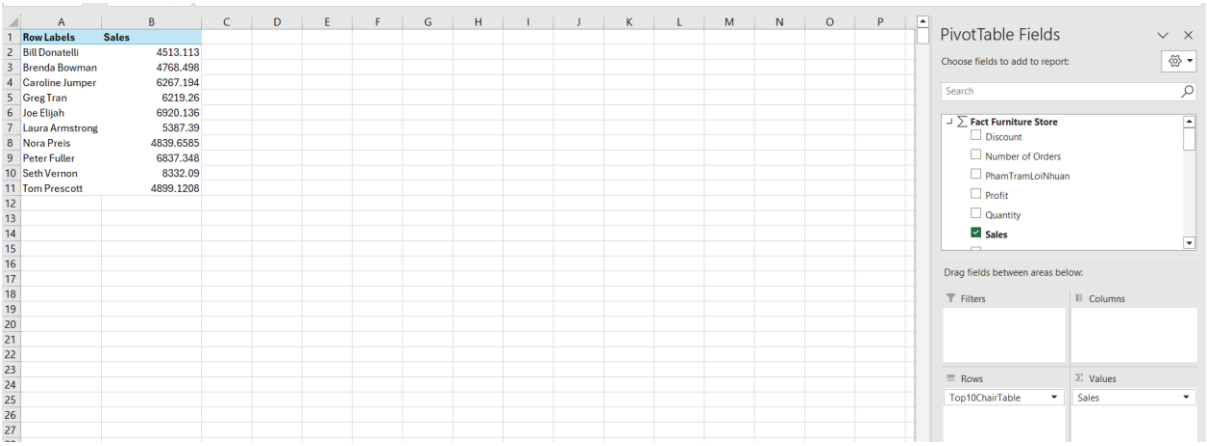
113 %

Messages | Results

	Sales
Seth Vernon	8024.678
Joe Elijah	6842.408
Caroline Jumper	4991.75
Laura Armstrong	4921.424
Brenda Bowman	4768.498
Nora Preis	4584.684
Tom Prescott	4416.174
Joseph Holt	4297.644
Tamara Willingham	4208.456
Bill Donatelli	4182.243

Hình 3.103 Thực hiện truy vấn bằng ngôn ngữ MDX

2.10.3. Sử dụng Pivot Table



Hình 3.104 Thực hiện truy vấn bằng PivotTable Excel

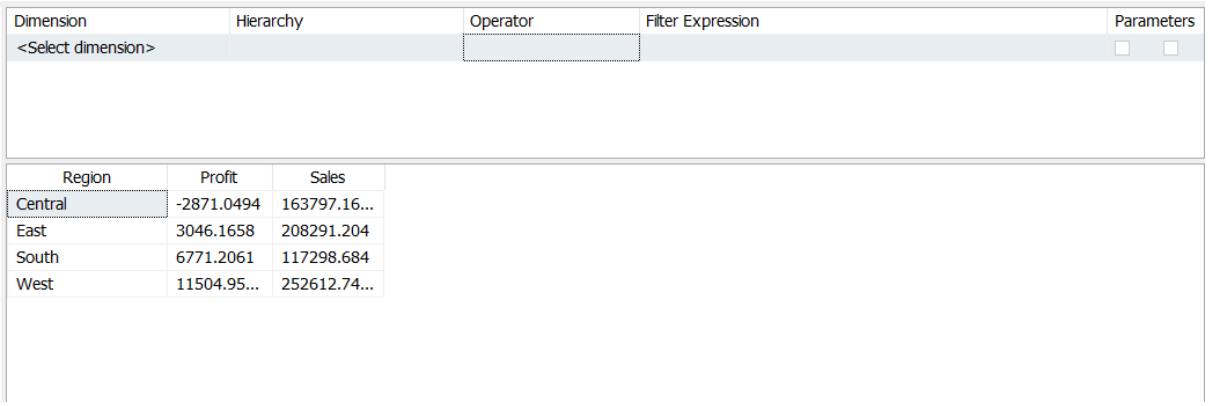
2.10.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.11. Cho biết tổng doanh số của từng khu vực bán hàng?

2.11.1. Phân tích dữ liệu trên Cube

Bước 1: Kéo các thuộc tính Region, Profit, Sales vào cửa sổ truy vấn và nhấn **Execute** để nhận kết quả



Dimension	Hierarchy	Operator	Filter Expression	Parameters
<Select dimension>				

Region	Profit	Sales
Central	-2871.0494	163797.16...
East	3046.1658	208291.204
South	6771.2061	117298.684
West	11504.95...	252612.74...

Hình 3.105 Thực hiện truy vấn trên Cube

2.11.2. Sử dụng ngôn ngữ MDX

```
select {[Measures].[Sales], [Measures].[Profit]} on COLUMNS,
        {[Location].[Region].CHILDREN} on ROWS
from [FurnitureStoreCube];
```

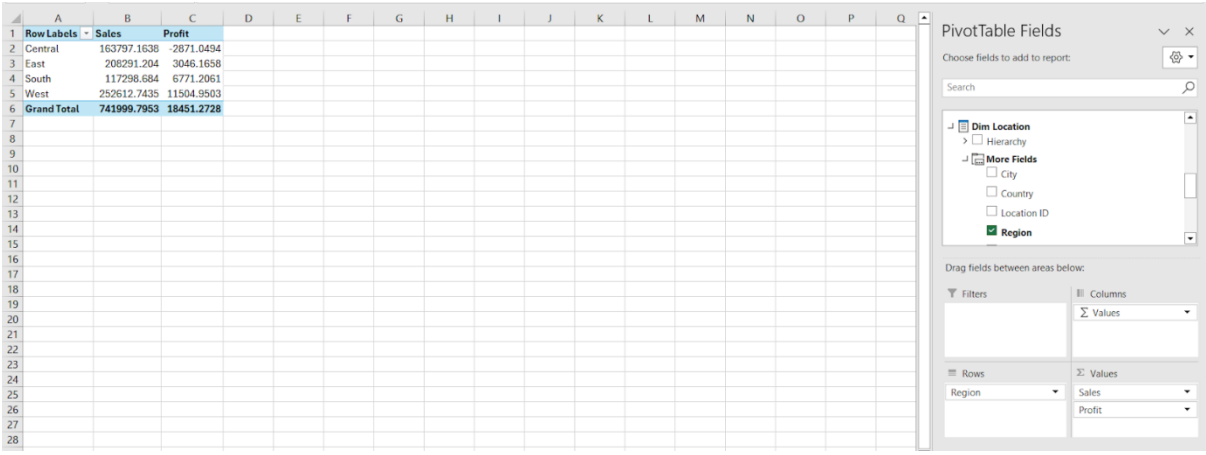
101 %

Messages | Results

	Sales	Profit
Central	163797.1638	-2871.0494
East	208291.204	3046.1658
South	117298.684	6771.2061
West	252612.7435	11504.9503

Hình 3.106 Thực hiện truy vấn bằng ngôn ngữ MDX

2.11.3. Sử dụng Pivot Table



Hình 3.107 Thực hiện truy vấn bằng PivotTable Excel

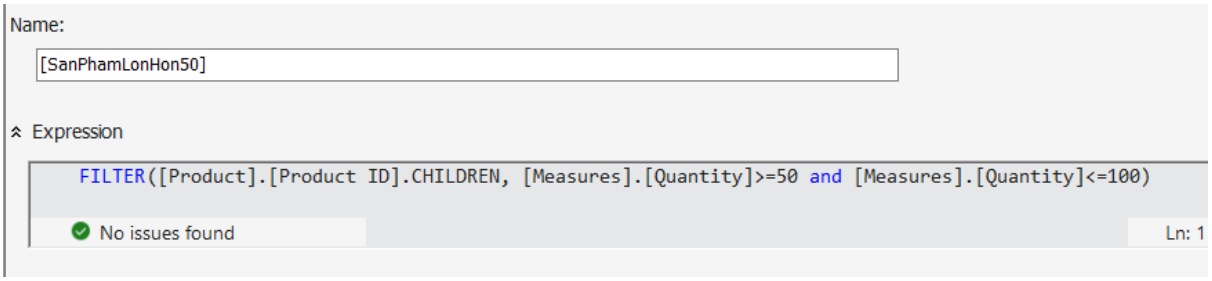
2.11.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.12. Liệt kê các sản phẩm có số lượng mua từ 50 đến 100?

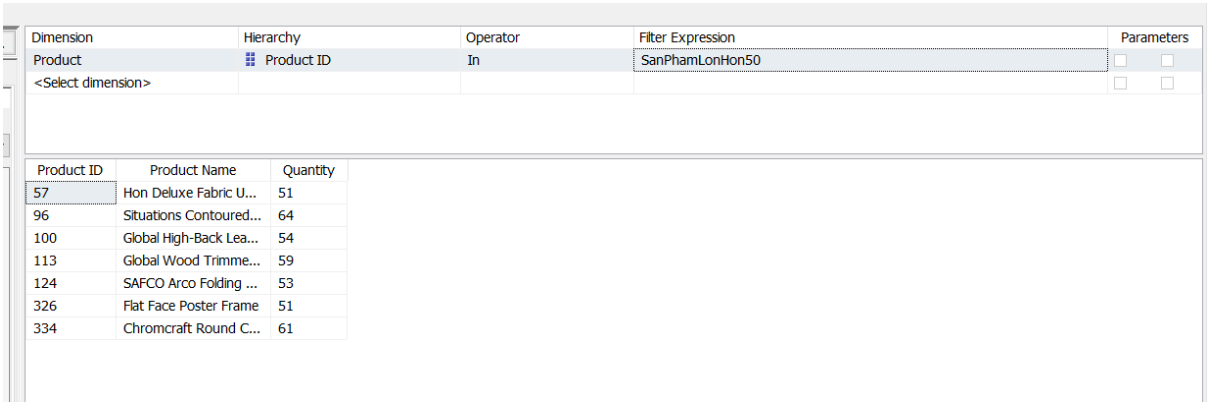
2.12.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo một **Named Set** *[SanPhamLonHon50]* như hình



Hình 3.108 Tạo NamedSet [SanPhamLonHon50]

Bước 2: Chọn Dim Product, sau đó chọn **Filter Expression** là *[SanPhamLonHon50]* kéo thuộc tính Product ID, Product Name và độ đo Quantity vào cửa sổ truy vấn để có kết quả



Hình 3.109 Thực hiện truy vấn trên Cube

2.12.2. Sử dụng ngôn ngữ MDX

```
select [Measures].[Quantity] on COLUMNS,
    {FILTER([Product].[Product ID].CHILDREN*[Product].[Product Name].CHILDREN,
    [Measures].[Quantity]>=50 and [Measures].[Quantity]<=100)} on ROWS
from [FurnitureStoreCube]
```

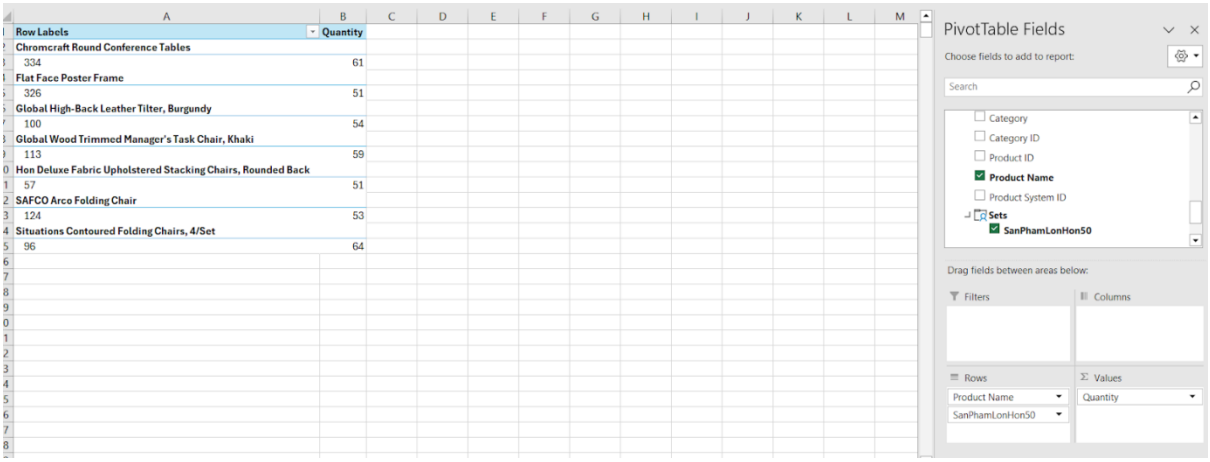
112 %

Messages | Results

		Quantity
57	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	51
96	Situations Contoured Folding Chairs, 4/Set	64
100	Global High-Back Leather Tilter, Burgundy	54
113	Global Wood Trimmed Manager's Task Chair, Khaki	59
124	SAFCO Arco Folding Chair	53
326	Flat Face Poster Frame	51
334	Chromcraft Round Conference Tables	61

Hình 3.110 Thực hiện truy vấn bằng ngôn ngữ MDX

2.12.3. Sử dụng Pivot Table



Hình 3.111 Thực hiện truy vấn bằng PivotTable Excel

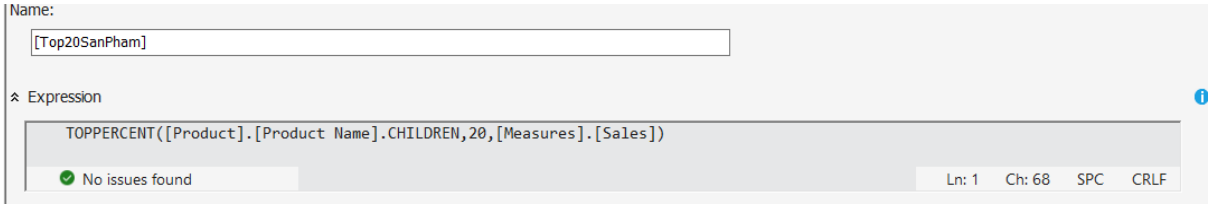
2.12.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau.

2.13. Đưa ra 20% phần trăm sản phẩm có doanh thu cao nhất, trong đó chỉ ra doanh thu theo từng loại hình vận chuyển được sử dụng (liệt kê nếu có)?

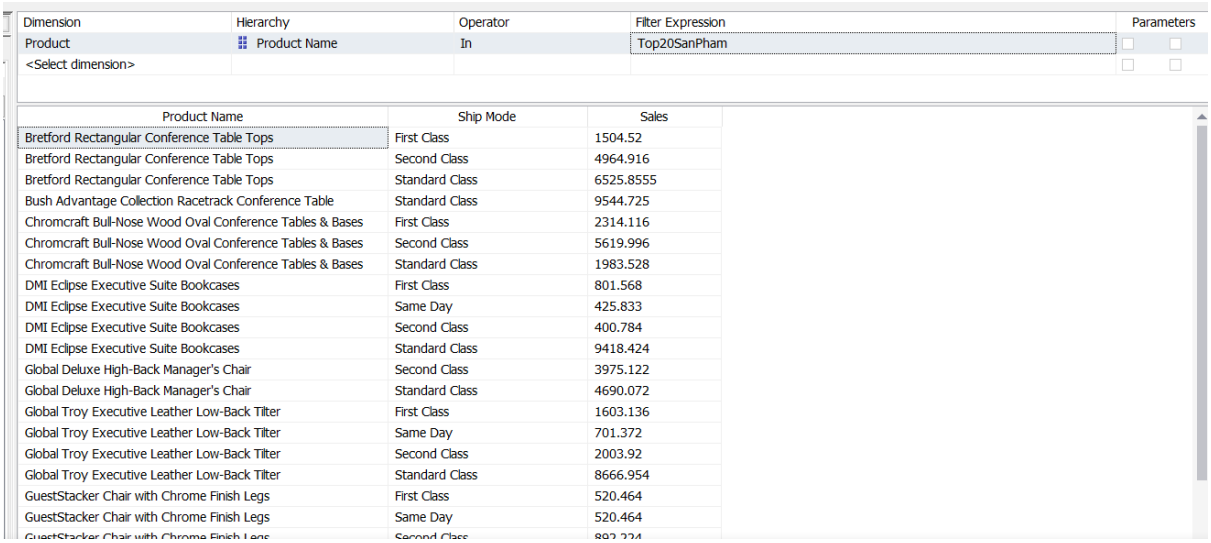
2.13.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo một **Named Set** *[Top20SanPham]*



Hình 3.112 Tạo NamedSet [Top20SanPham]

Bước 2: Chọn Filter Expression là *[Top20SanPham]*, kéo các thuộc tính Product Name, Ship Mode và độ đo Sales và nhấn Execute để ra kết quả



Dimension	Hierarchy	Operator	Filter Expression	Parameters
Product	Product Name	In	Top20SanPham	
<Select dimension>				

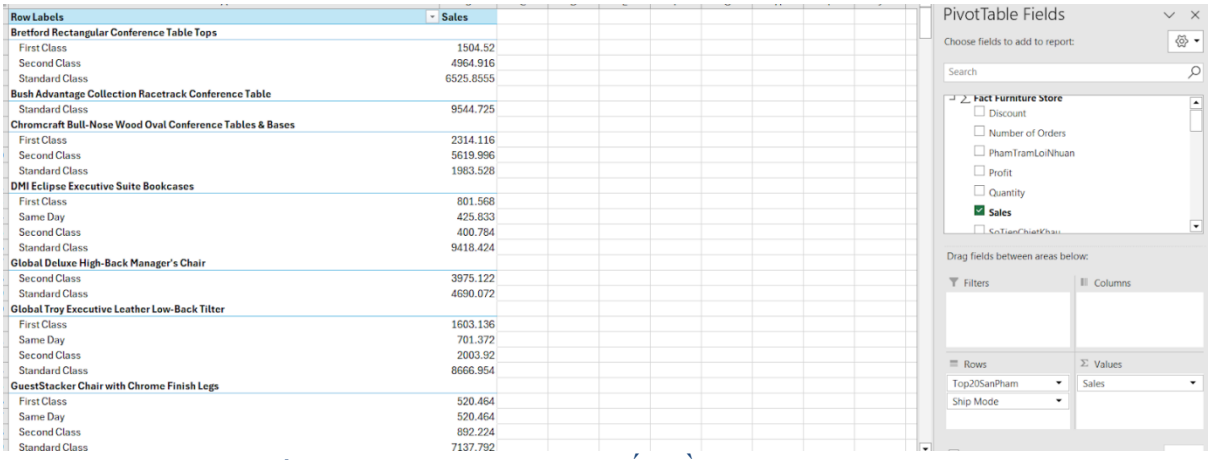
Product Name	Ship Mode	Sales
Bretford Rectangular Conference Table Tops	First Class	1504.52
Bretford Rectangular Conference Table Tops	Second Class	4964.916
Bretford Rectangular Conference Table Tops	Standard Class	6525.8555
Bush Advantage Collection Racetrack Conference Table	Standard Class	9544.725
Chromcraft Bull-Nose Wood Oval Conference Tables & Bases	First Class	2314.116
Chromcraft Bull-Nose Wood Oval Conference Tables & Bases	Second Class	5619.996
Chromcraft Bull-Nose Wood Oval Conference Tables & Bases	Standard Class	1983.528
DMI Eclipse Executive Suite Bookcases	First Class	801.568
DMI Eclipse Executive Suite Bookcases	Same Day	425.833
DMI Eclipse Executive Suite Bookcases	Second Class	400.784
DMI Eclipse Executive Suite Bookcases	Standard Class	9418.424
Global Deluxe High-Back Manager's Chair	Second Class	3975.122
Global Deluxe High-Back Manager's Chair	Standard Class	4690.072
Global Troy Executive Leather Low-Back Tilter	First Class	1603.136
Global Troy Executive Leather Low-Back Tilter	Same Day	701.372
Global Troy Executive Leather Low-Back Tilter	Second Class	2003.92
Global Troy Executive Leather Low-Back Tilter	Standard Class	8666.954
GuestStacker Chair with Chrome Finish Legs	First Class	520.464
GuestStacker Chair with Chrome Finish Legs	Same Day	520.464

2.13.2. Sử dụng ngôn ngữ MDX

```
SELECT
    {[Measures].[Sales]} ON COLUMNS,
    CROSSJOIN(
        TOPPERCENT(
            [Product].[Product Name].CHILDREN,
            20,
            [Measures].[Sales]
        ),
        [Ship Mode].[Ship Mode].MEMBERS
    ) ON ROWS
FROM [FurnitureStoreCube];
```

Hình 3.113 Thực hiện truy vấn bằng ngôn ngữ MDX

2.13.3. Sử dụng Pivot Table



Row Labels	Sales
Bretford Rectangular Conference Table Tops	
First Class	1504.52
Second Class	4964.916
Standard Class	6525.8555
Bush Advantage Collection Racetrack Conference Table	
Standard Class	9544.725
Chromcraft Bull-Nose Wood Oval Conference Tables & Bases	
First Class	2314.116
Second Class	5619.996
Standard Class	1983.528
DMI Eclipse Executive Suite Bookcases	
First Class	801.568
Same Day	425.833
Second Class	400.784
Standard Class	9418.424
Global Deluxe High-Back Manager's Chair	
Second Class	3975.122
Standard Class	4690.072
Global Troy Executive Leather Low-Back Tilter	
First Class	1603.136
Same Day	701.372
Second Class	2003.92
Standard Class	8666.954
GuestStacker Chair with Chrome Finish Legs	
First Class	520.464
Same Day	520.464
Second Class	892.224
Standard Class	7137.792

Hình 3.114 Thực hiện truy vấn bằng PivotTable Excel

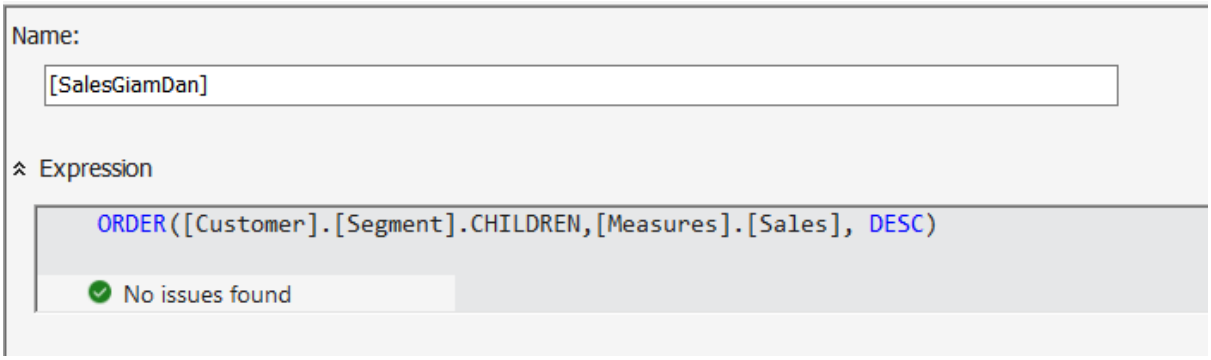
2.13.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.14. Thống kê lợi nhuận, phần trăm lợi nhuận từng loại khách hàng, sắp xếp theo phần trăm lợi nhuận giảm dần?

2.14.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo **Named Set** *[SalesGiamDan]*



Hình 3.115 Tạo NamedSet [SalesGiamDan]

2.14.2. Sử dụng ngôn ngữ MDX

```
select {[Measures].[Sales],[Measures].[Quantity]} on COLUMNS,
        {ORDER(
            [Customer].[Segment].CHILDREN,[Measures].[Sales], DESC
        )} on ROWS
FROM [FurnitureStoreCube]
```

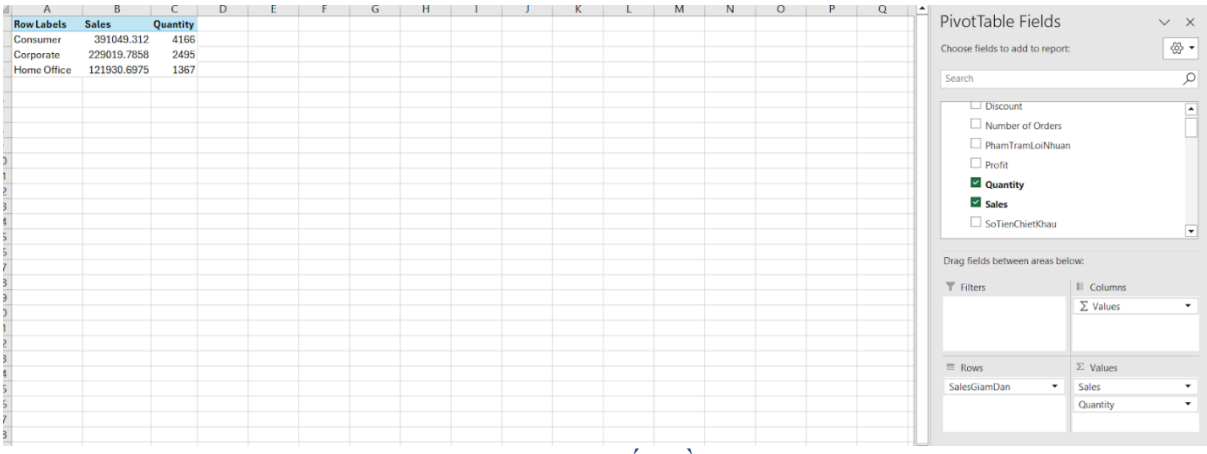
102 %

Messages | Results

	Sales	Quantity
Consumer	391049.312000001	4166
Corporate	229019.7858	2495
Home Office	121930.6975	1367

Hình 3.116 Thực hiện truy vấn bằng ngôn ngữ MDX

2.14.3. Sử dụng Pivot Table



Hình 3.117 Thực hiện truy vấn bằng PivotTable Excel

2.14.4. Kết luận:

- Ba phương pháp phân tích dữ liệu đưa ra kết quả giống nhau

2.15. Liệt kê Top 10 khách hàng có số đơn hàng ít nhất chọn sử dụng dịch vụ giao hàng First Class và danh sách tương tự nhưng sử dụng dịch vụ giao hàng Standard Class?

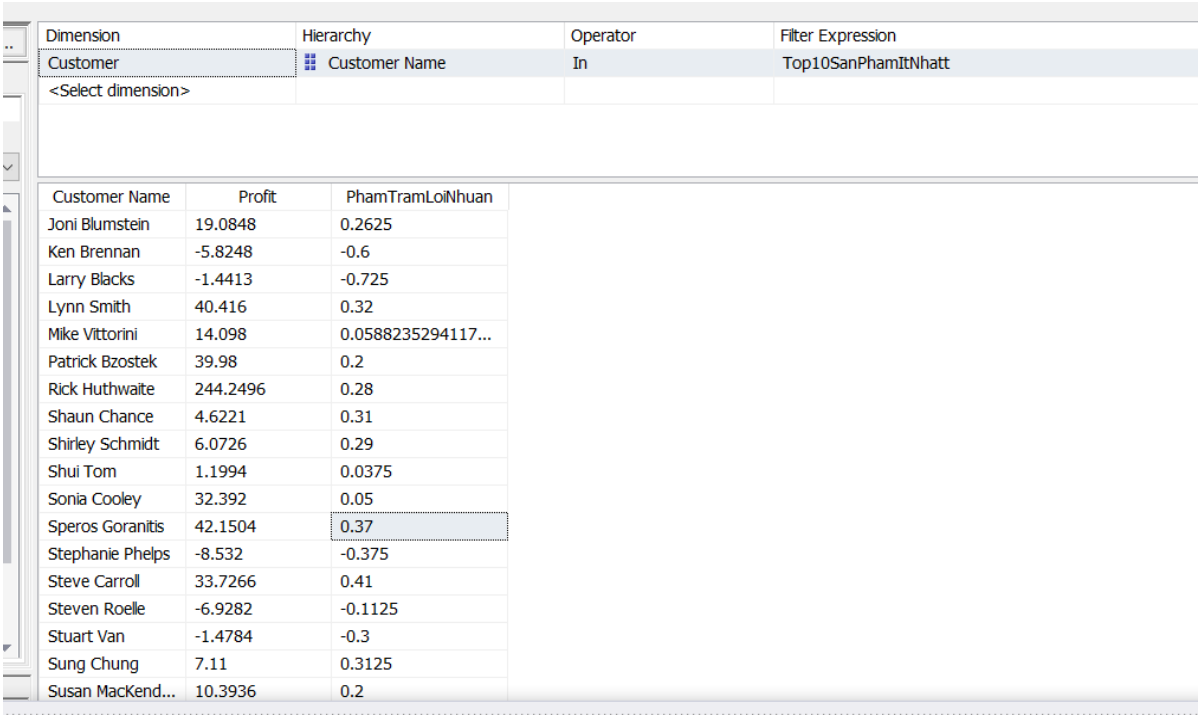
2.15.1. Phân tích dữ liệu trên Cube

Bước 1: Tạo **Named Set** *[Top10SanPhamItNhat]*

```
[Top10SanPhamItNhatt]

Expression

UNION(
        BottomCount(
        FILTER(
            [Customer].[Customer Name].MEMBERS,
            ([Ship Mode].[Ship Mode].&[First Class])
        ),10,
        [Measures].[Number of Orders]
    ),
    BottomCount(
        FILTER(
            [Customer].[Customer Name].MEMBERS,
            ([Ship Mode].[Ship Mode].&[Standard Class])
        ),10,
        [Measures].[Number of Orders]
    )

No issues found
```

Hình 3. 118 Tạo NamedSet Top10SanPhamItNhat

Bước 2: Chọn Dim Customer, sau đó chọn **Filter Expression** là *[Top10SanPhamItNhat]*, thuộc tính là Customer Name, độ đo Profit và **Calculated Member** là *[PhanTramLoiNhuan]* để có kết quả



Dimension	Hierarchy	Operator	Filter Expression
Customer	Customer Name	In	Top10SanPhamItNhatt
<Select dimension>			

Customer Name	Profit	PhamTramLoiNhuan
Joni Blumstein	19.0848	0.2625
Ken Brennan	-5.8248	-0.6
Larry Blacks	-1.4413	-0.725
Lynn Smith	40.416	0.32
Mike Vittorini	14.098	0.0588235294117...
Patrick Bzostek	39.98	0.2
Rick Huthwaite	244.2496	0.28
Shaun Chance	4.6221	0.31
Shirley Schmidt	6.0726	0.29
Shui Tom	1.1994	0.0375
Sonia Cooley	32.392	0.05
Speros Goranitis	42.1504	0.37
Stephanie Phelps	-8.532	-0.375
Steve Carroll	33.7266	0.41
Steven Roelle	-6.9282	-0.1125
Stuart Van	-1.4784	-0.3
Sung Chung	7.11	0.3125
Susan MacKend...	10.3936	0.2

Hình 3.119 Thực hiện truy vấn trên Cube

2.15.2. Sử dụng ngôn ngữ MDX

```
------
select {[Measures].[Profit],[Measures].[PhamTramLoiNhuan]} on COLUMNS,
        {
        UNION(
            BottomCount(
                FILTER(
                    [Customer].[Customer Name].MEMBERS,
                    ([Ship Mode].[Ship Mode].&[First Class])),10,[Measures].[Number of Orders]),
            BottomCount(
                FILTER(
                    [Customer].[Customer Name].MEMBERS,
                    ([Ship Mode].[Ship Mode].&[Standard Class])),10,[Measures].[Number of Orders]
    )
        )} on ROWS
FROM [FurnitureStoreCube];
-----------------------------------------------
```

102 %

Messages | Results

	Profit	PhamTramLoiNhuan
Susan MacKendrick	10.3936	0.2
Stuart Van	-1.4784	-0.3
Stephanie Phelps	-8.532	-0.375
Rick Huthwaite	244.2496	0.28
Patrick Bzostek	39.98	0.2
Mike Vittorini	14.098	5.88235294117647E-02
Lynn Smith	40.416	0.32
Larry Blacks	-1.4413	-0.725
Ken Brennan	-5.8248	-0.6
Joni Blumstein	19.0848	0.2625
Yoseph Carroll	14.9292	0.12
Toby Braunhardt	17.3754	0.42
Sung Chung	7.11	0.3125
Steven Roelle	-6.9282	-0.1125
Steve Carroll	33.7266	0.41
Speros Goranitis	42.1504	0.37

Hình 3.120 Thực hiện truy vấn bằng ngôn ngữ MDX

2.15.3. Sử dụng Pivot Table

Row Labels	PhamTramLoiNhuan	Profit
Joni Blumstein	0.2625	19.0848
Ken Brennan	-0.6	-5.8248
Larry Blacks	-0.725	-1.4413
Lynn Smith	0.32	40.416
Mike Vittorini	0.058823529	14.098
Patrick Bzostek	0.2	39.98
Rick Huthwaite	0.28	244.2496
Shaun Chance	0.31	4.6221
Shirley Schmidt	0.29	6.0726
Shui Tom	0.0375	1.1994
Sonia Cooley	0.05	32.392
Speros Goranitis	0.37	42.1504
Stephanie Phelps	-0.375	-8.532
Steve Carroll	0.41	33.7266
Steven Roelle	-0.1125	-6.9282
Stuart Van	-0.3	-1.4784
Sung Chung	0.3125	7.11
Susan MacKendrick	0.2	10.3936
Toby Braunhardt	0.42	17.3754
Yoseph Carroll	0.12	14.9292

Hình 3.121 Thực hiện truy vấn bằng PivotTable Excel

2.15.4. Kết luận:

- Ba phương pháp cho lại kết quả giống nhau.

CHƯƠNG 4. QUÁ TRÌNH LẬP BÁO BIỂU(SSRS)

1. QUÁ TRÌNH SSRS VỚI VISUAL STUDIO REPORTING SERVICE VÀ REPORT BUILDER

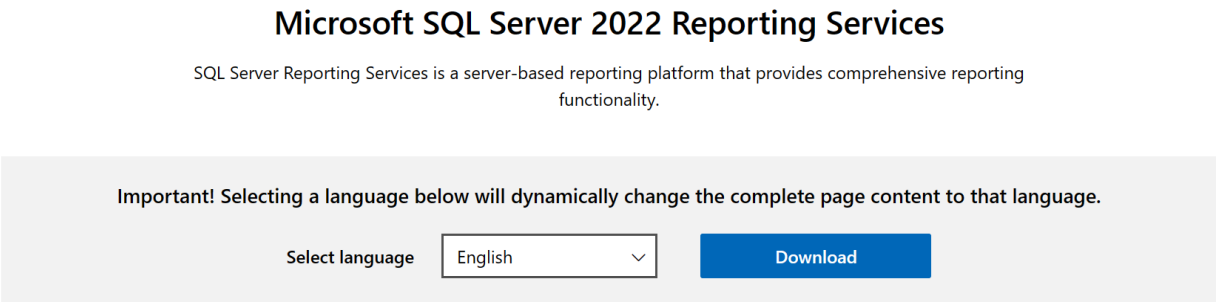
1.1. Cài đặt các công cụ cần thiết:

Những công cụ và phần mềm cần cài đặt để thực hiện quá trình SSRS:

- SQLServerReportingServices.exe: Đây là công cụ và dịch vụ giúp ích trong việc khởi tạo, deploy(triển khai) và quản lý các báo biểu.
- ReportBuilder.msi: Là công cụ tạo báo biểu cho SQL Server Reporting Services. Công cụ này cho phép khởi tạo, quản lý và công bố các báo biểu lên SQL Server Reporting Services, đồng thời tạo các bộ dữ liệu chia sẻ.

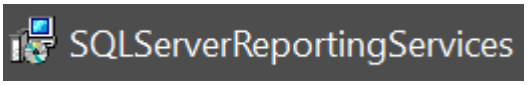
1.1.1. Cài đặt SQL Server Reporting Services:

Bước 1: Tiến hành tải công cụ Microsoft SQL Server 2022 Reporting Services trên trang web của Microsoft.



Hình 4.1 Tải công cụ Microsoft SQL Server 2022 Reporting Services

Bước 2: Sau khi đã tải về thành công cài đặt về thì ta tiến hành nhấn đúp vào file tải để tiến hành cài đặt công cụ Microsoft SQL Server 2022 Reporting Services

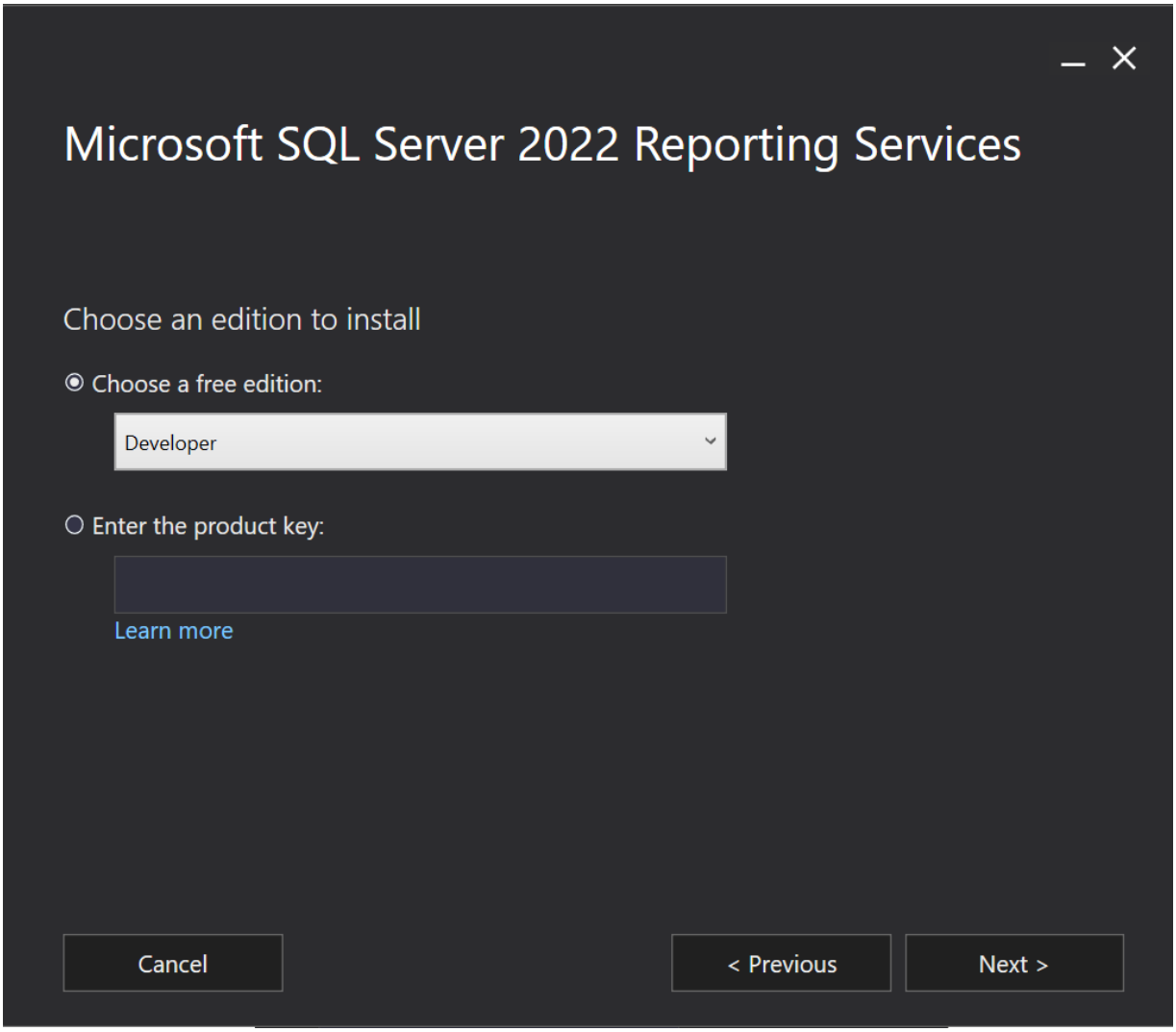


Hình 4.2 Tiến hành cài đặt công cụ

Bước 3: Khi giao diện cài đặt được bật lên ta nhấn chọn **Install Reporting Services** để đến bước tiếp theo.

Hình 4.3 Chọn Install Reporting Services

Bước 4: Tại mục chọn **Choose an edition to install** thì ta tích vào ô **Choose a free edition** và chọn phiên bản **Developer** để có thể sử dụng đầy đủ các tính năng cần thiết. Cuối cùng ta nhấn **Next**.

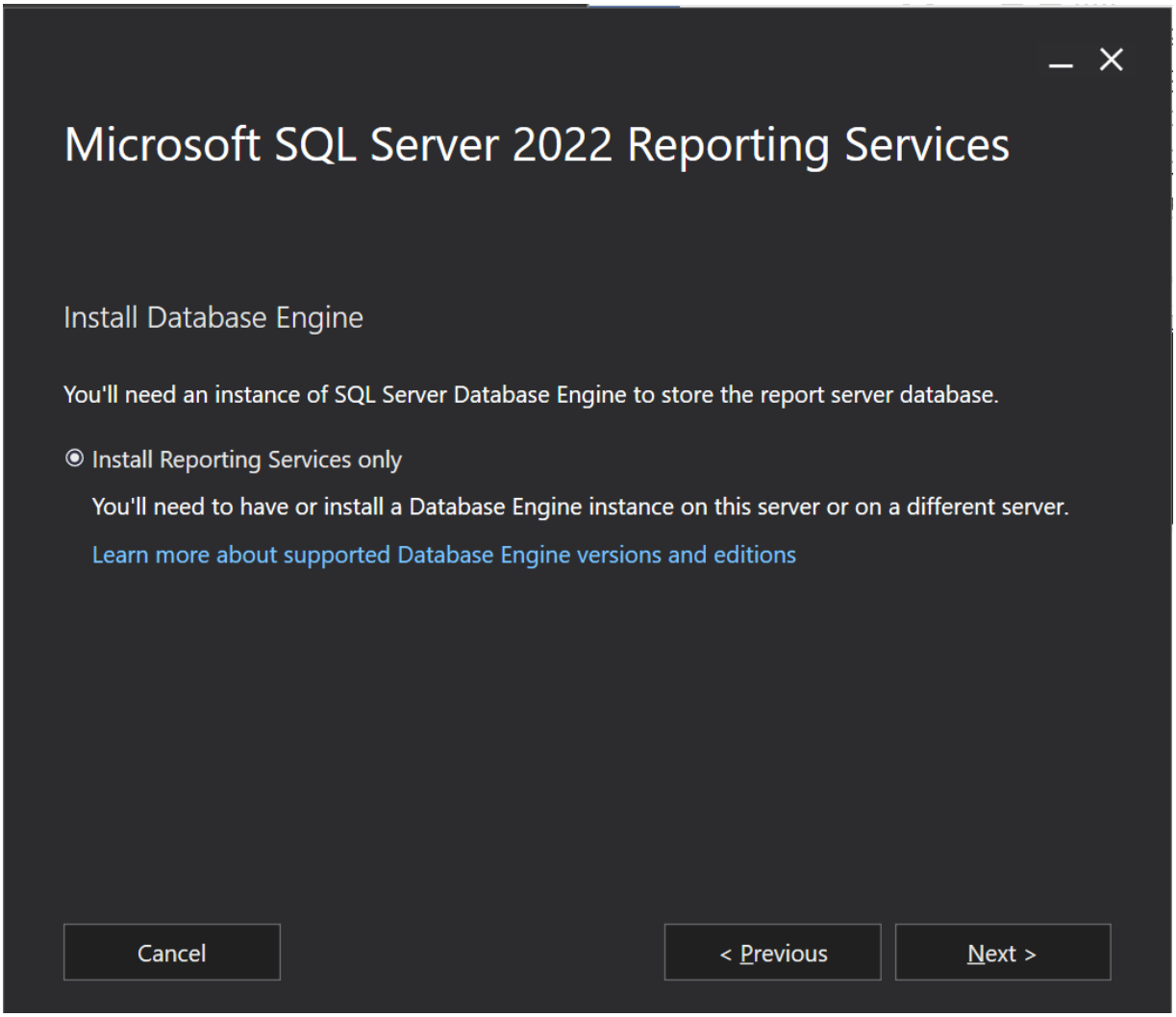


Hình 4.4 Chọn phiên bản cài đặt Reporting Services

Bước 5: Tại giao diện **Review the license terms**, chọn **I accept the license terms** và nhấn **Next**.

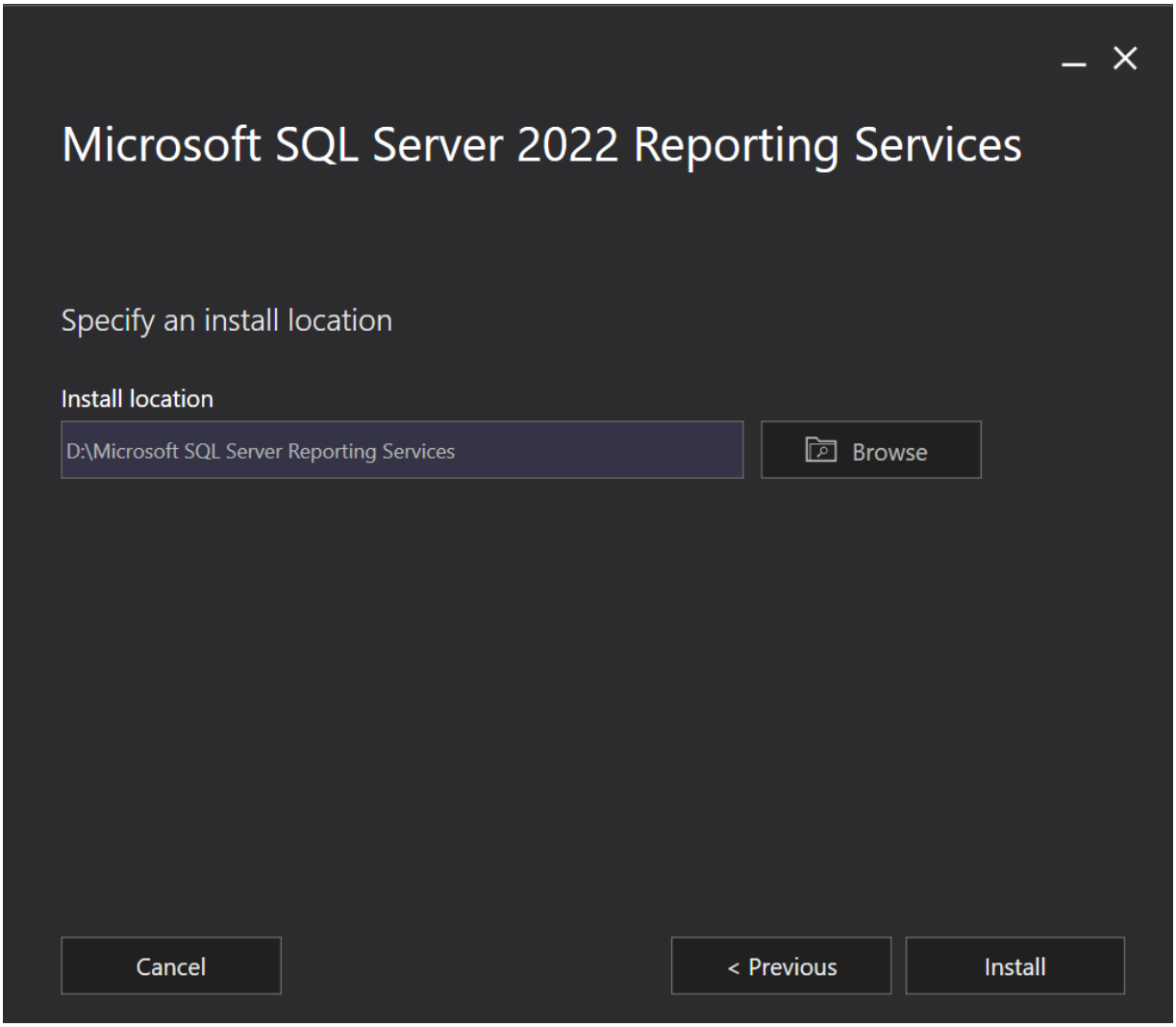
Hình 4.5 Chọn Chấp nhận quy định

Bước 6: Ta nhấn **Next**.



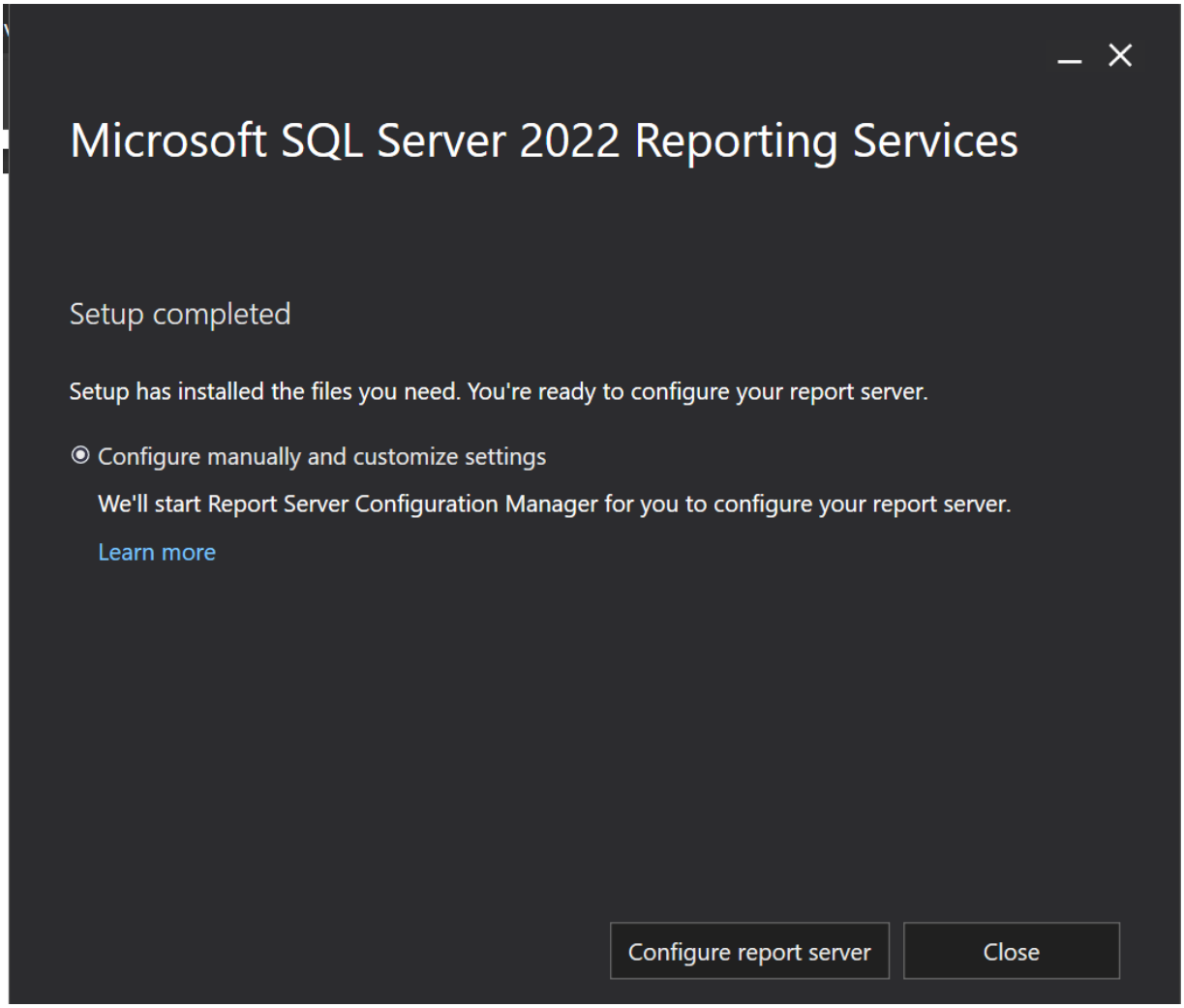
Hình 4.6 Chọn Next

Bước 7: Ở mục chọn đường dẫn lưu trữ cài đặt thì ta có thể để mặc định hoặc chọn một thư mục khác mà chúng ta muốn. Sau khi đã thiết lập thì nhấn **Install** để chính thức cài đặt ứng dụng.



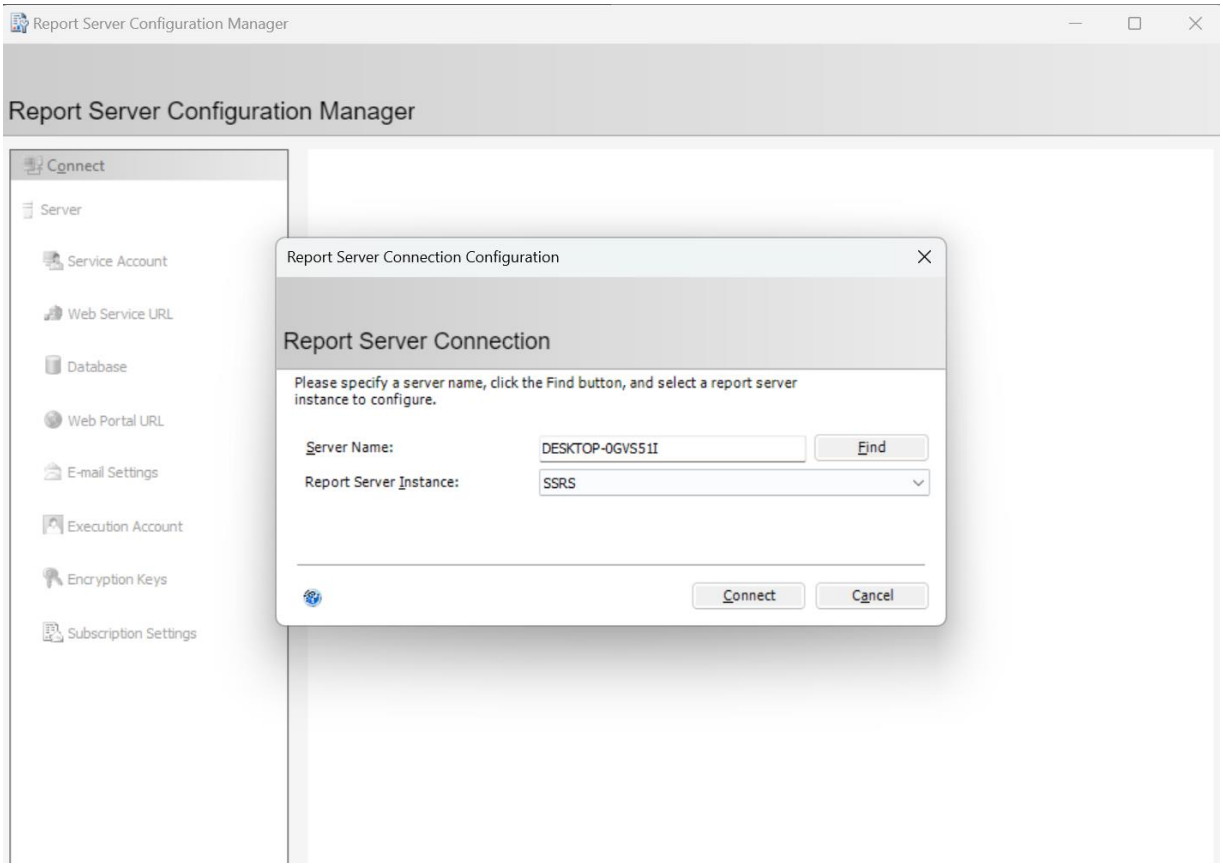
Hình 4.7 Chọn đường dẫn lưu trữ cài đặt

Bước 8: Sau khi cài đặt thành công thì ta chọn **Configure report server** để mở giao diện **Report Server Configuration Manager**.



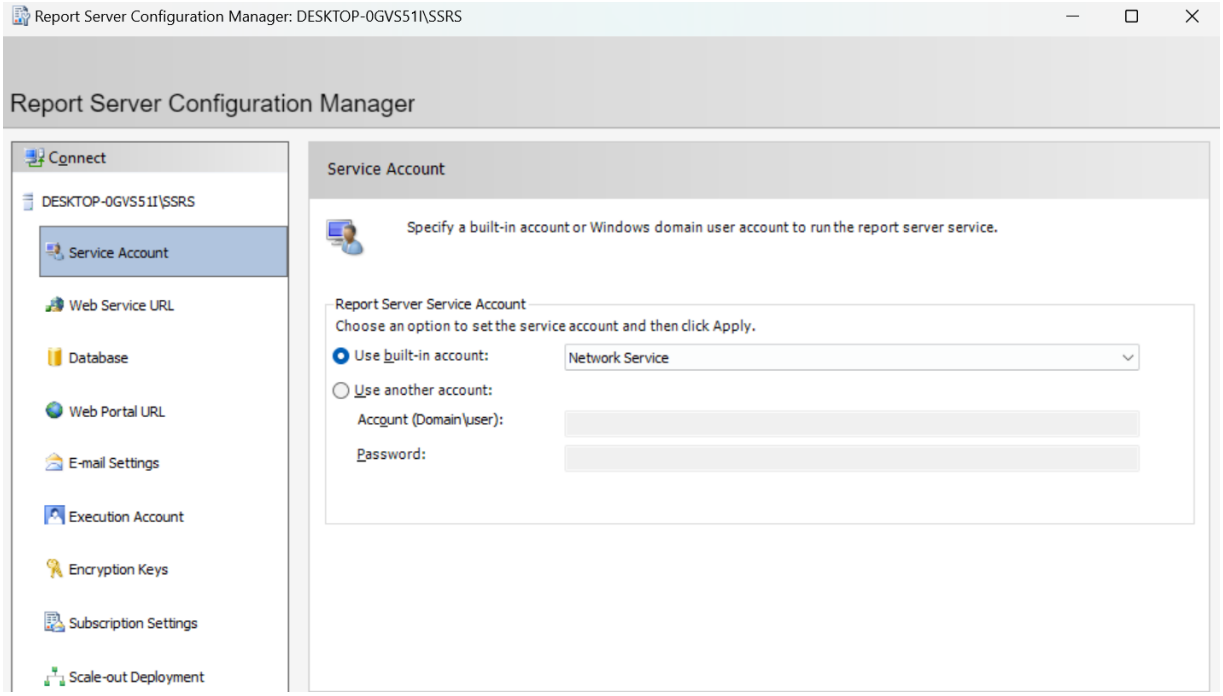
Hình 4.8 Configure Report Server

Bước 9: Khi giao diện thiết lập cài đặt cho Report Server hiện lên thì ta sẽ chọn tên server và chọn **Report Server Instance** là **SSRS**. Sau đó, nhấn connect để thiết lập kết nối.



Hình 4.9 Thiết lập cài đặt cho Report Server (1)

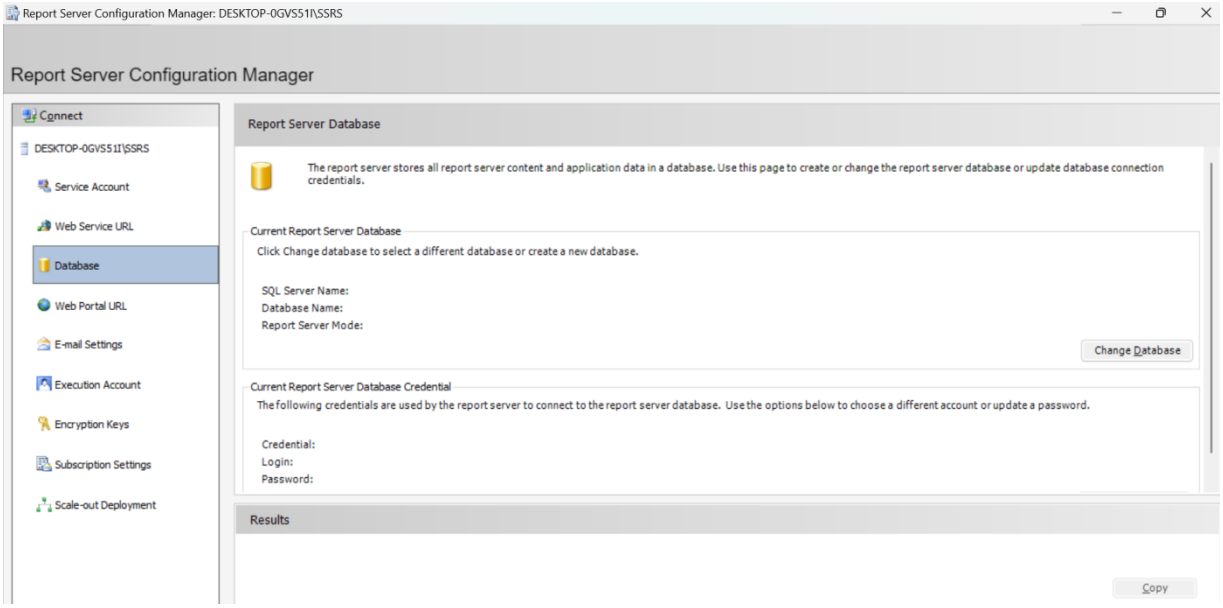
Bước 10: Tại giao diện tổng thể của Report Server Configuration Manager ta chuyển qua tab Service Account → Tích chọn **Use built-in account** → **Network service** và nhấn **Apply** để hoàn thành thiết lập.



Hình 4.10 Thiết lập cài đặt cho Report Server (2)

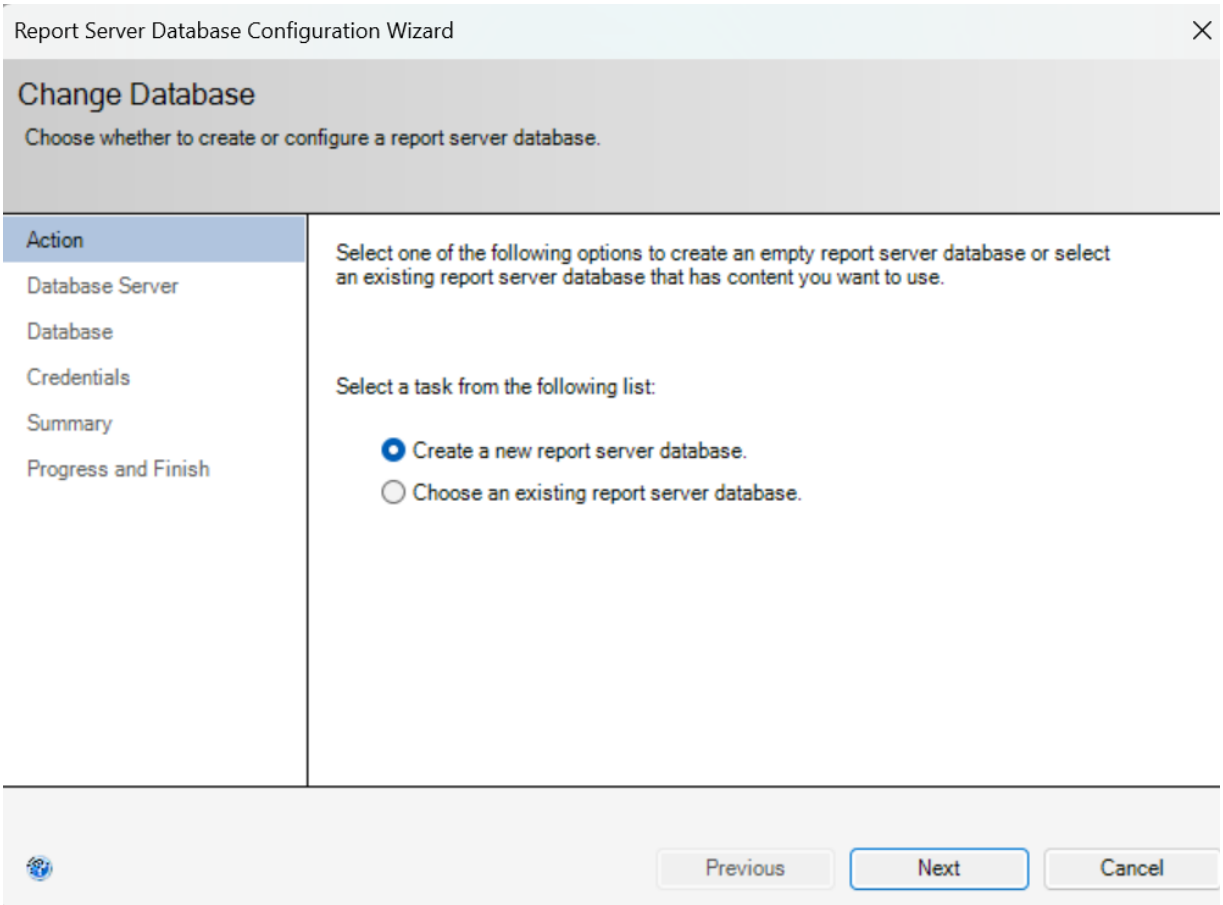
Hình 4.11 Hoàn thành thiết lập Report Server

Bước 11: Tại tab **Database,** để tạo mới một dataset ta sẽ chọn **Change database.**



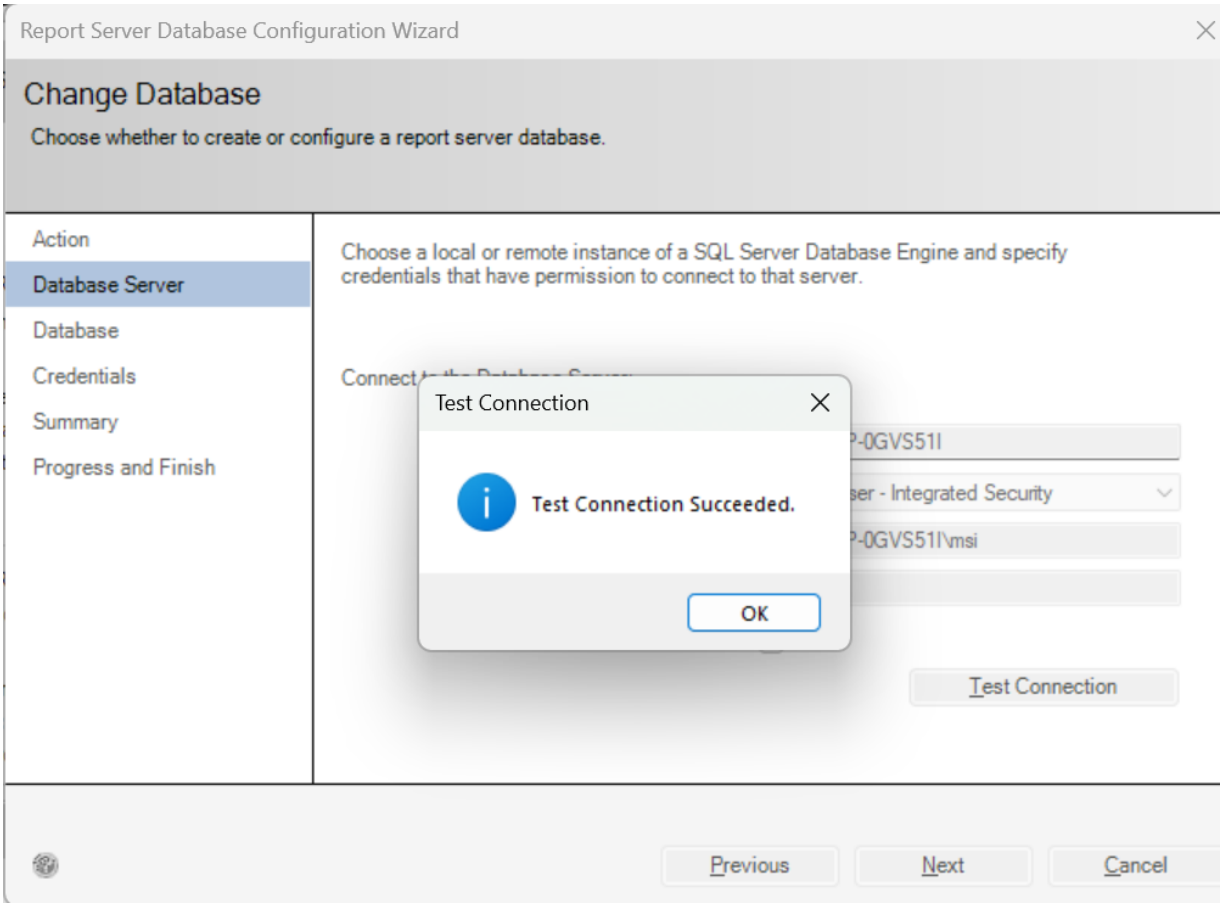
Hình 4.12 Tạo mới Dataset

Bước 12: Hộp thoại hiện lên thì ta chọn **Create a new report server database** và chọn Next.



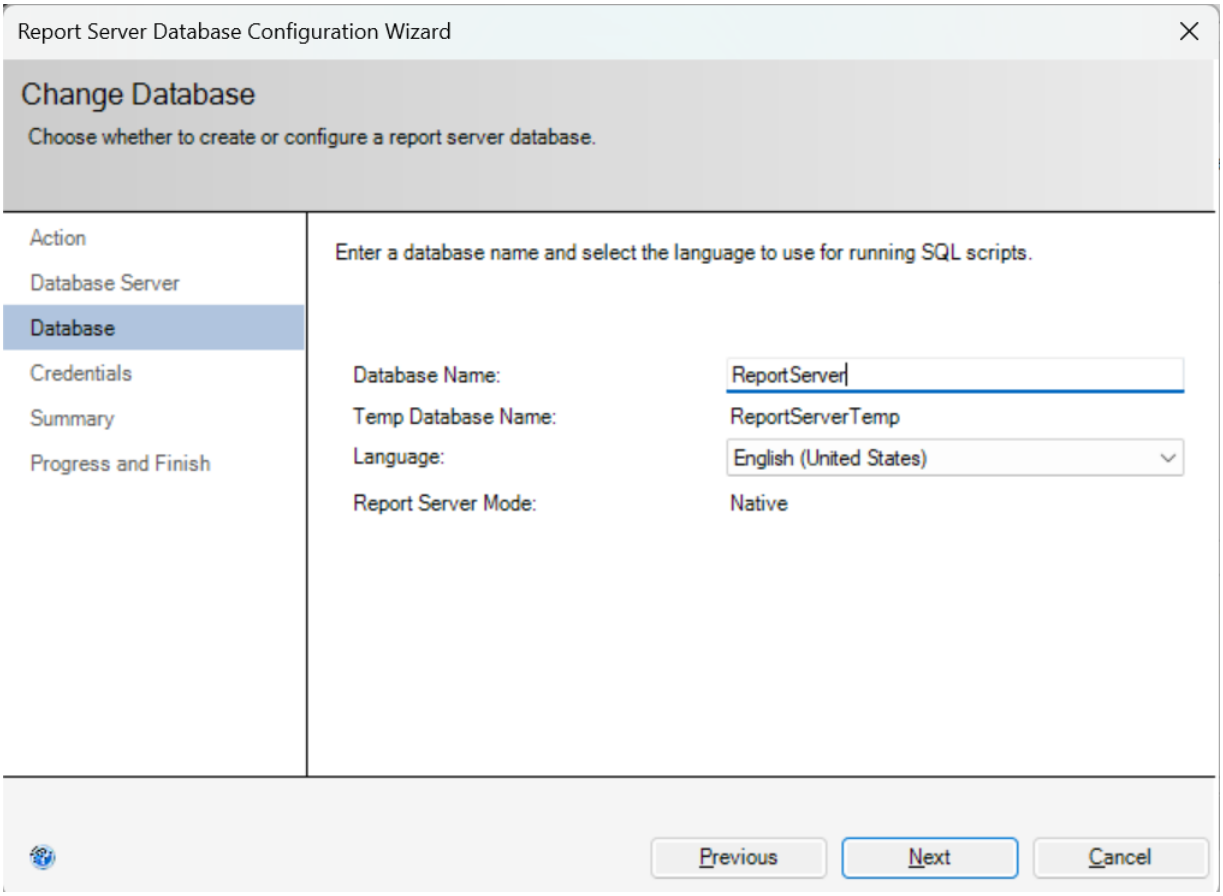
Hình 4.13 Tạo mới Report Server Database

Bước 13: Ở bước kết nối tới **Database Server** thì ta sẽ để mọi thông tin mặc định và nhấn Next để sang bước tiếp theo. Trước khi sang bước kế tiếp thì ta cũng có thể nhấn **Test Connection** để kiểm tra trạng thái kết nối.



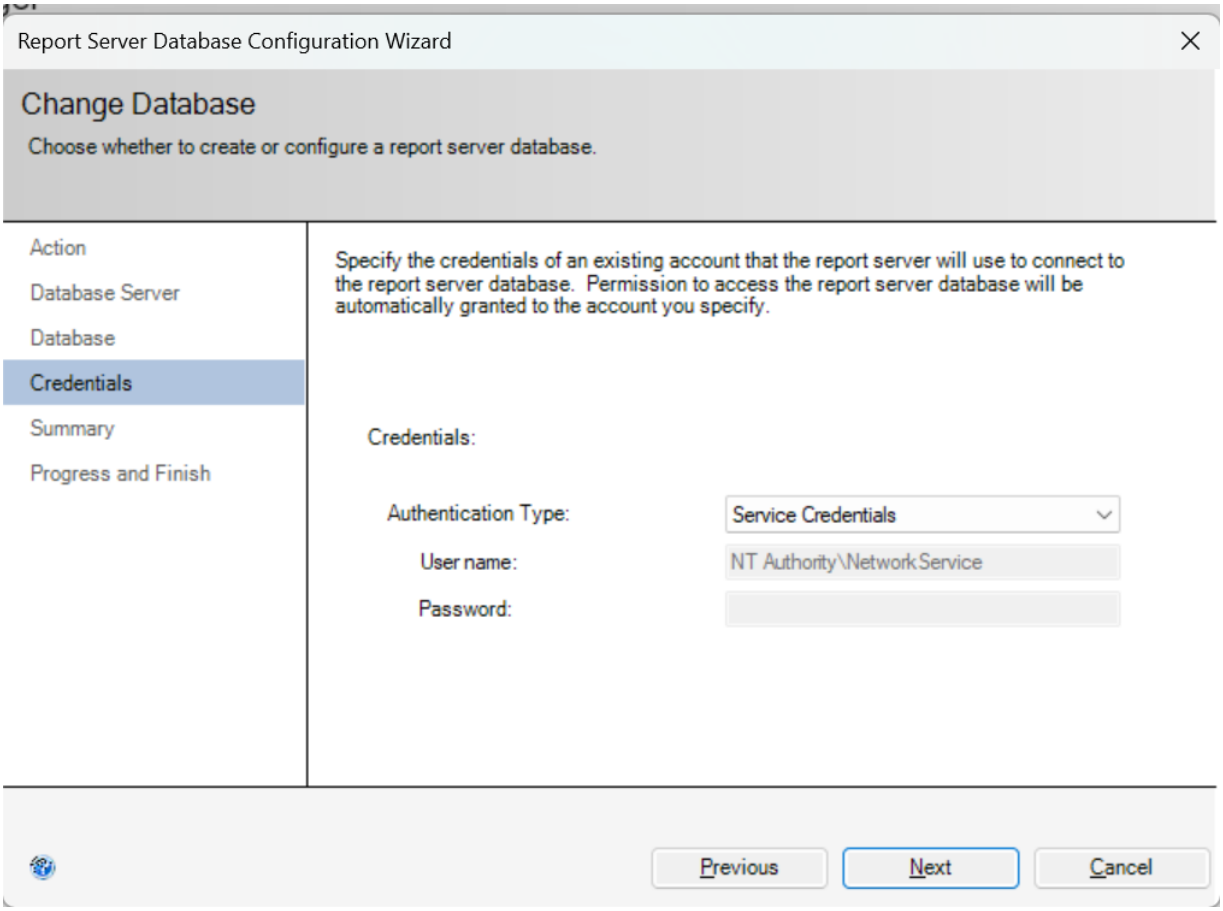
Hình 4.14 Tạo mới Database Server

Bước 14: Ở bước này ta đặt tên cho database và chọn Next.



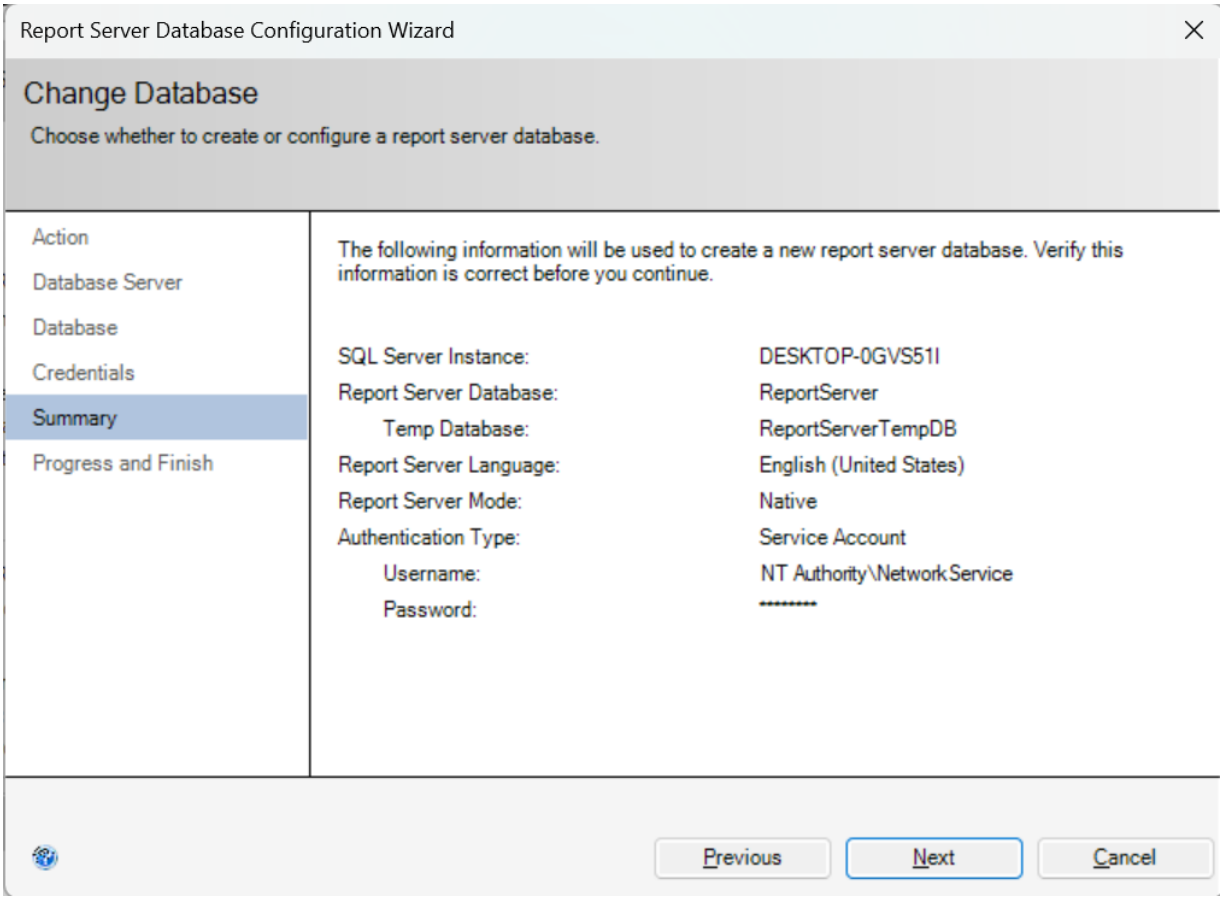
Hình 4.15 Đặt tên cho Database

Bước 15: Trong tab **Credentials** thì ta tiếp tục chọn Next.



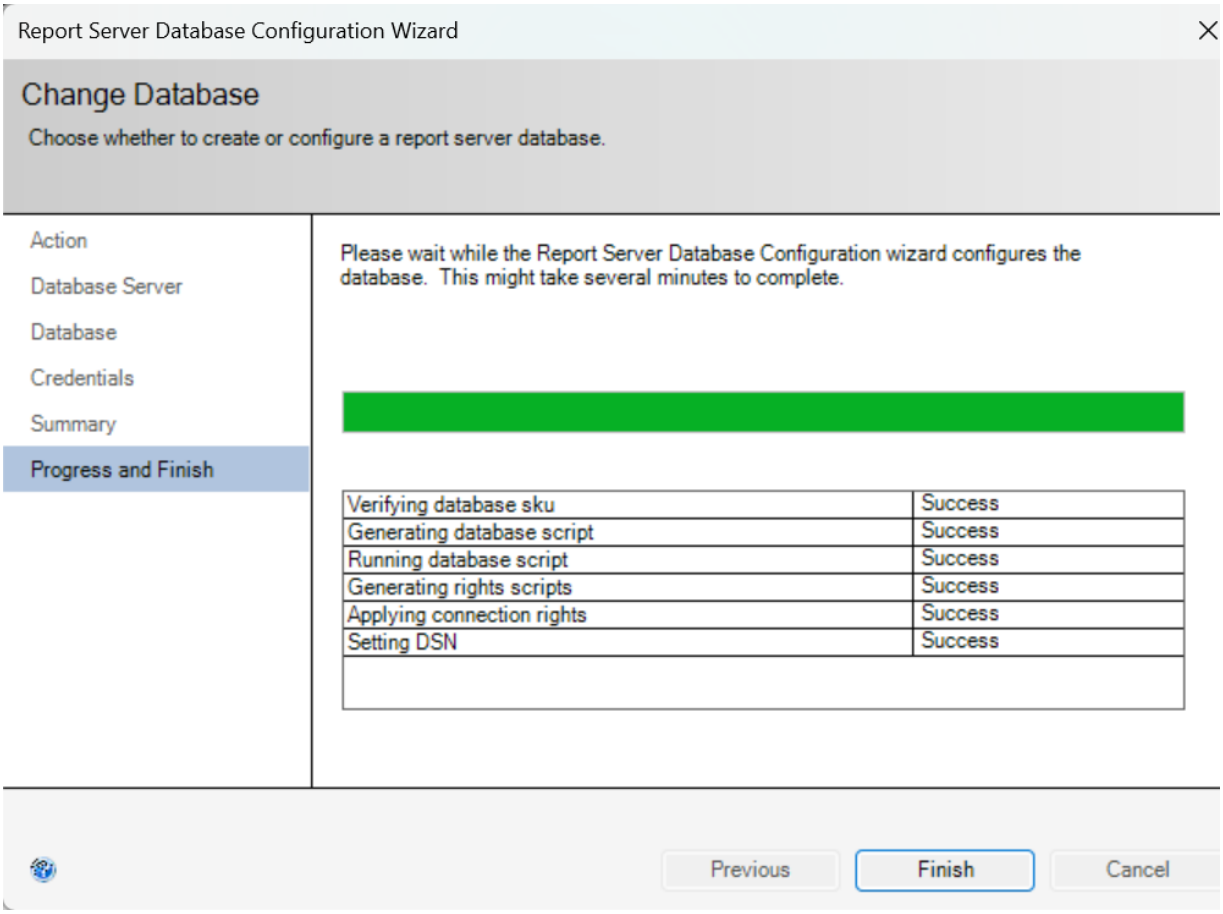
Hình 4.16 Cấu hình Credentials

Bước 16: Sau khi đã kiểm tra lại mọi thiết lập trong tab **Summary** và nhấn **Next** để hoàn thành tạo Database.



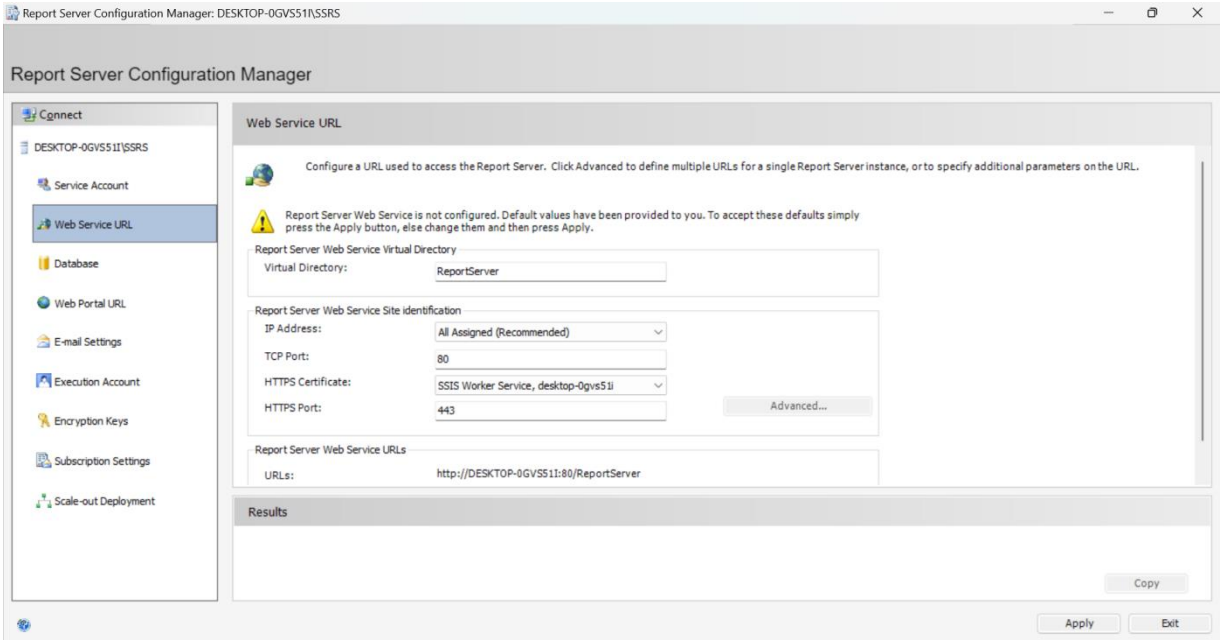
Hình 4.17 Xem lại các cài đặt cơ sở dữ liệu của Report Server

Bước 17: Sau khi hệ thống đã tạo mới database thì nhấn **Finish** để kết thúc quá trình khởi tạo.



Hình 4.18 Kết quả của quá trình tạo cơ sở dữ liệu của Report Server

Bước 18: Ta chuyển qua tab **Web Service URL** để cài đặt đường dẫn tới **Report Server** của chúng ta vừa tạo, với đường dẫn này ta có thể truy xuất và xem được cài báo biểu mà ta đã từng tạo và **deploy**. Sau khi chọn các thông tin cần thiết thì chọn **Apply** để áp dụng thiết lập.

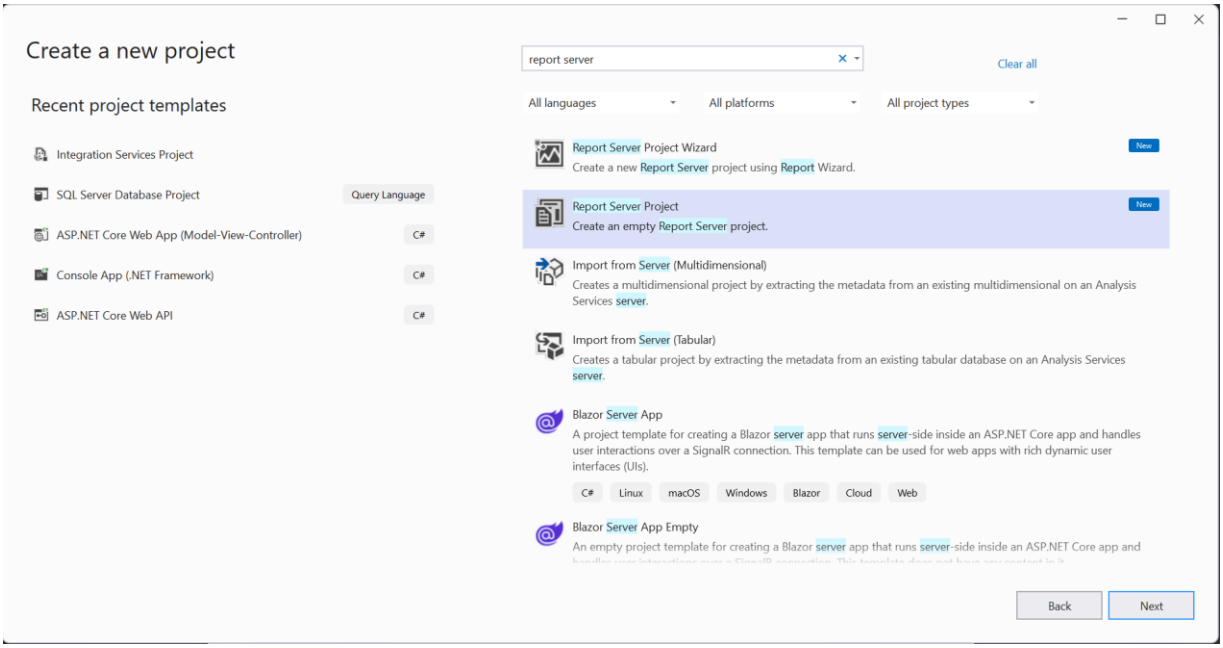


Hình 4.19 Khởi tạo đường dẫn truy cập tới Report Server

1.2. Quá trình tạo lập báo biểu bằng công cụ Microsoft SSRS(Report Builder)

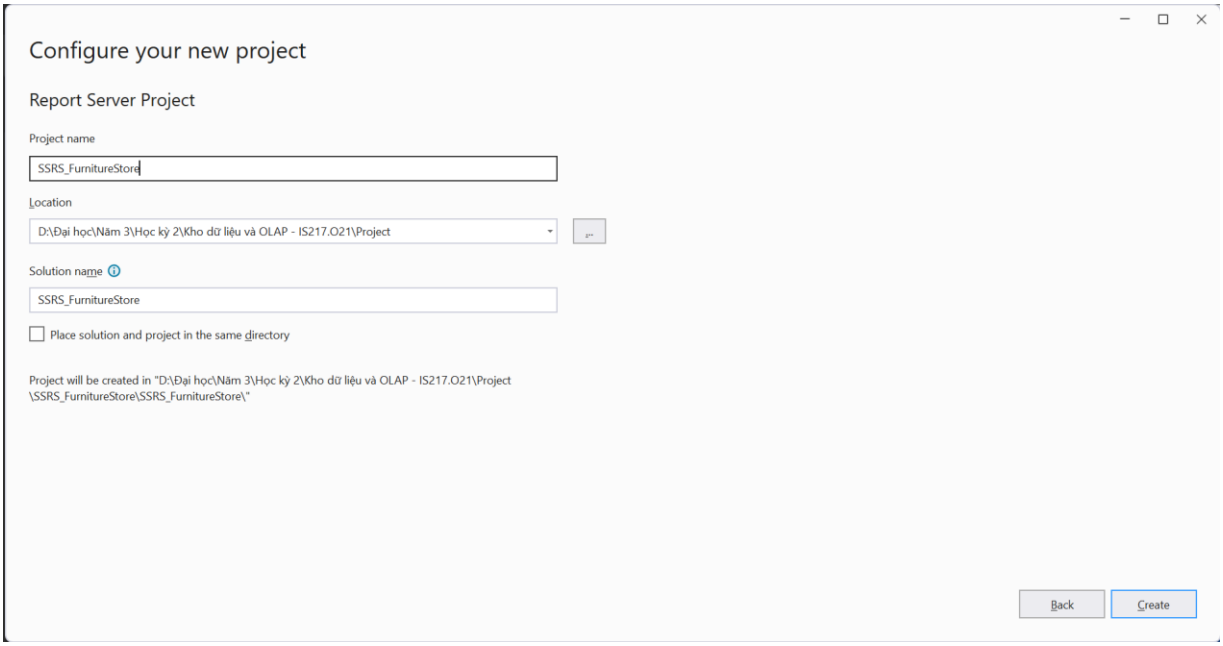
1.2.1. Tạo project SSRS

Bước 1: Mở **Visual Studio Community 2022 → Create a new project →** Chọn **Report Server Project** và nhấn **Next** để tới bước tiếp theo.



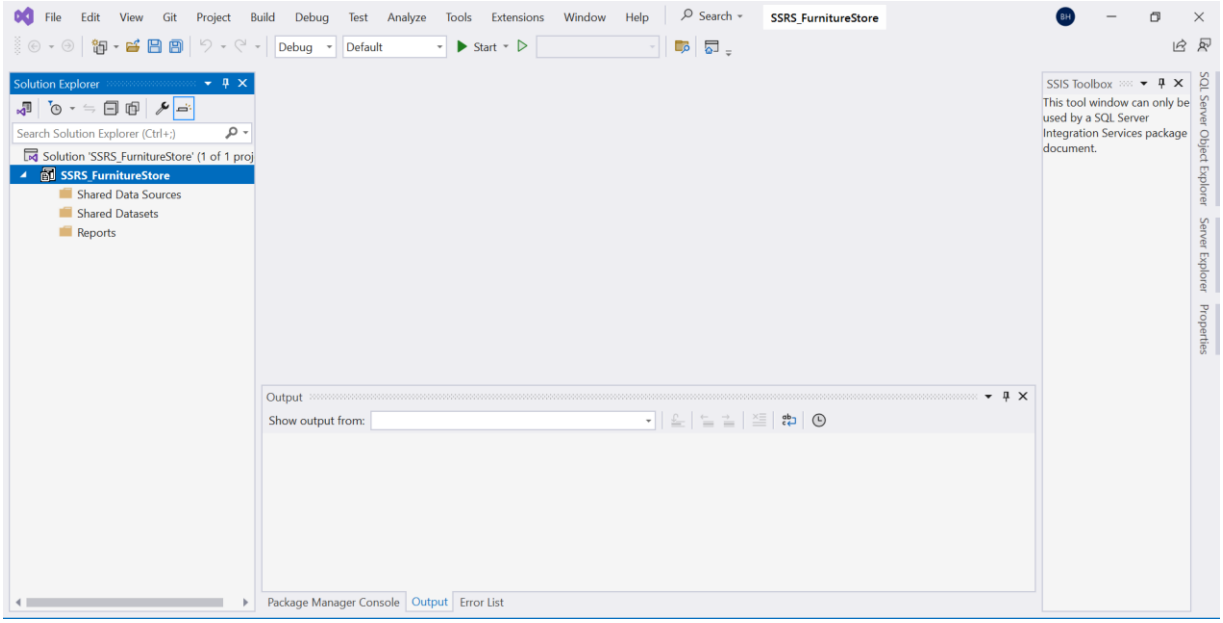
Hình 4.20 Tạo mới một dự án SSRS

Bước 2: Đặt tên cho dự án và định đường dẫn để lưu dự án. Cuối cùng nhấn **Create** để tạo dự án.



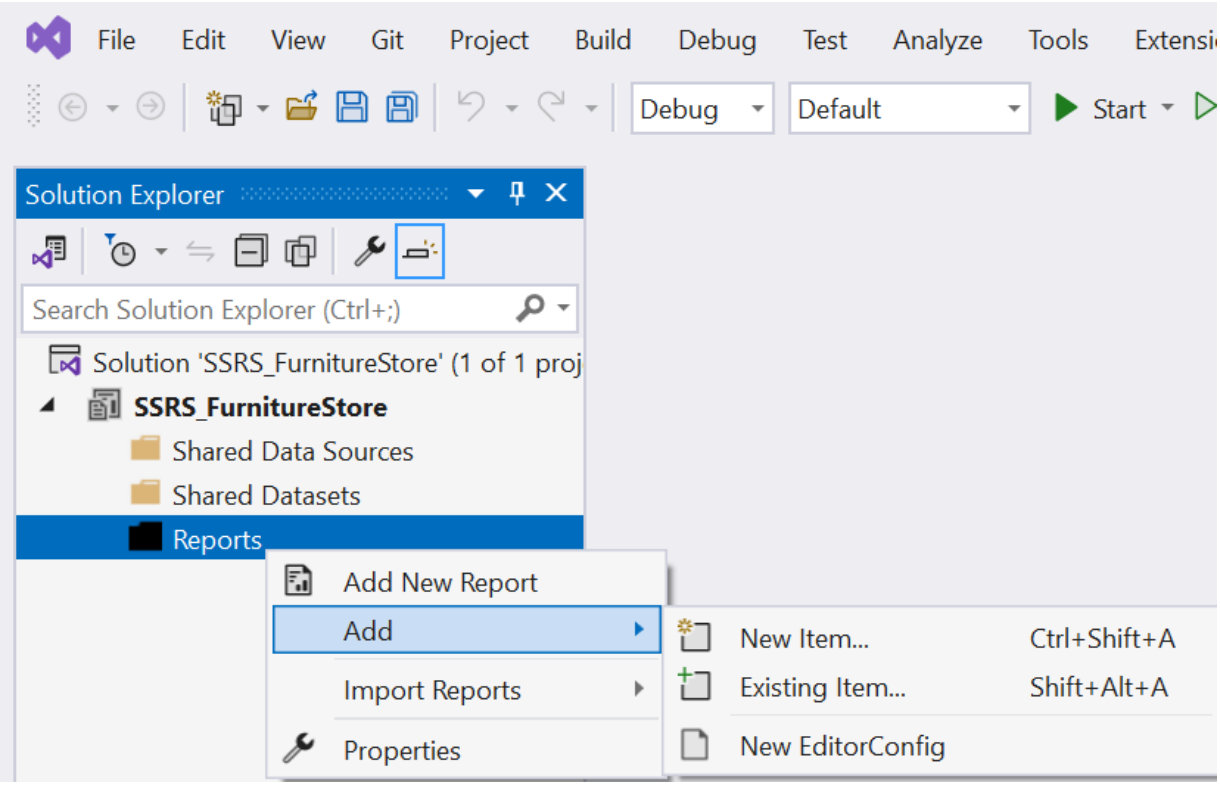
Hình 4.21 Đặt tên dự án và hoàn thành tạo dự án

Bước 3: Sau khi tạo dự án thành công thì đây là giao diện của ứng dụng.



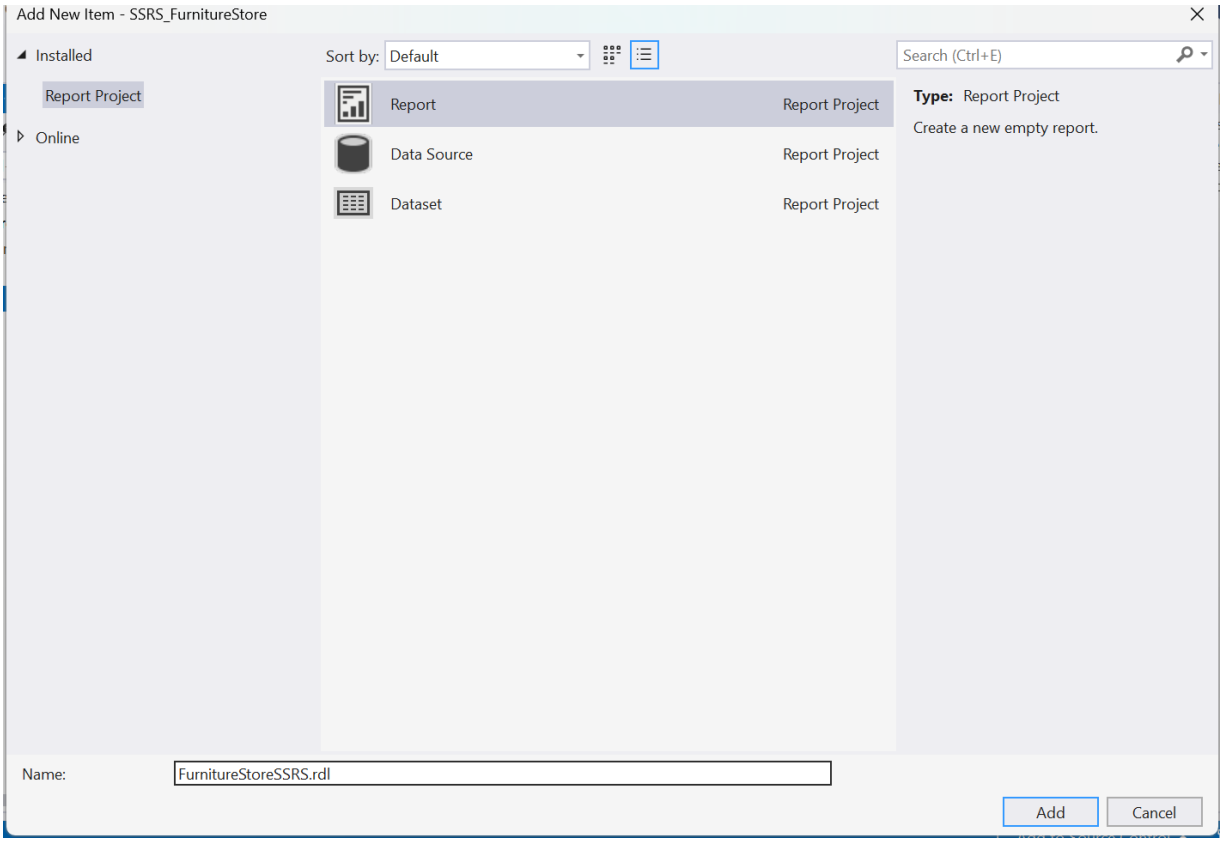
Hình 4.22 Giao diện dự án SSRS

Bước 4: Tại cửa sổ **Solution Explorer** ở phía bên trái, nhấn chuột phải vào **Reports → Add → Add new item**.

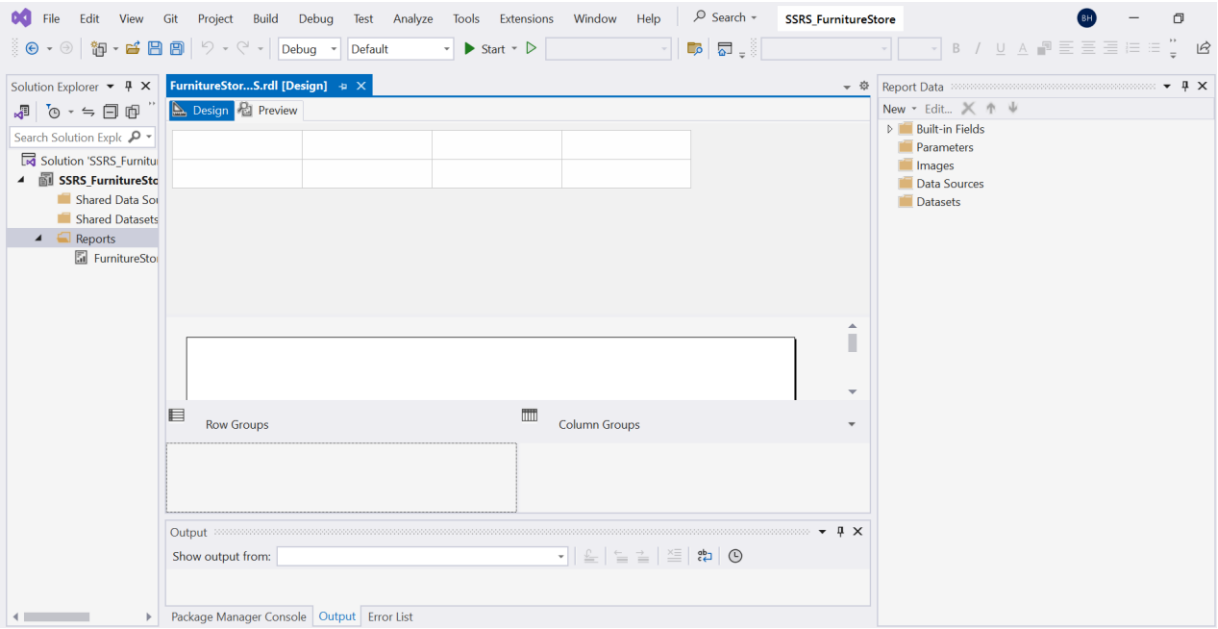


Hình 4.23 Tạo mới một Report(1)

Bước 5: Chọn **Report → Đặt tên → Add** để tiến hành tạo **Report**.



Hình 4.24 Tạo mới một Report(2)

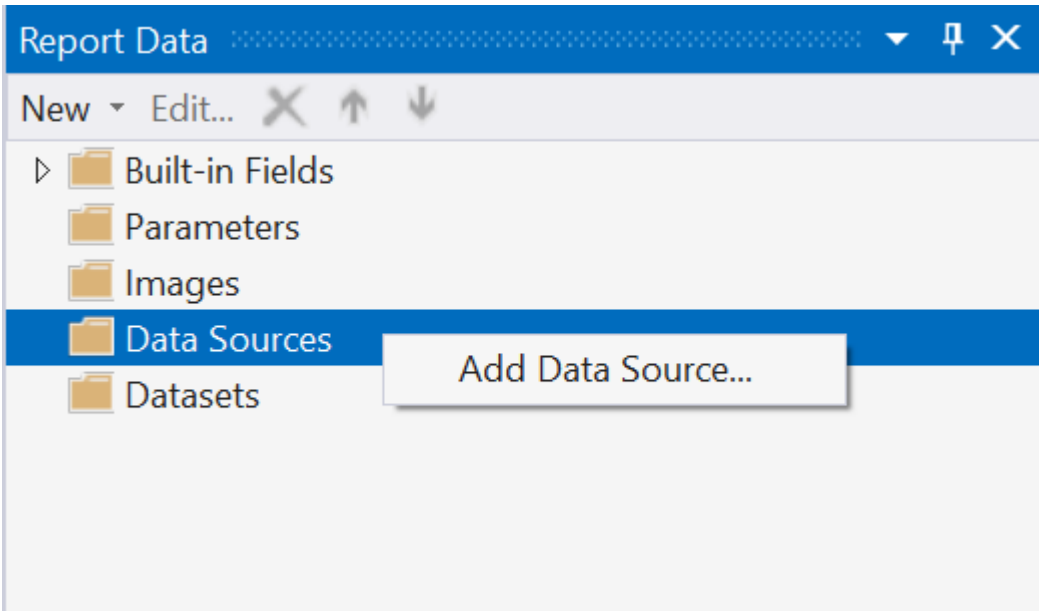


Hình 4.25 Giao diện của Report

- Trong cửa sổ làm việc của dịch vụ SSRS vừa mới được tạo ra thì ta có:
  - **Tab Design:** Để thiết kế Report
  - **Tab Preview**: Để xem trước kết quả Report.

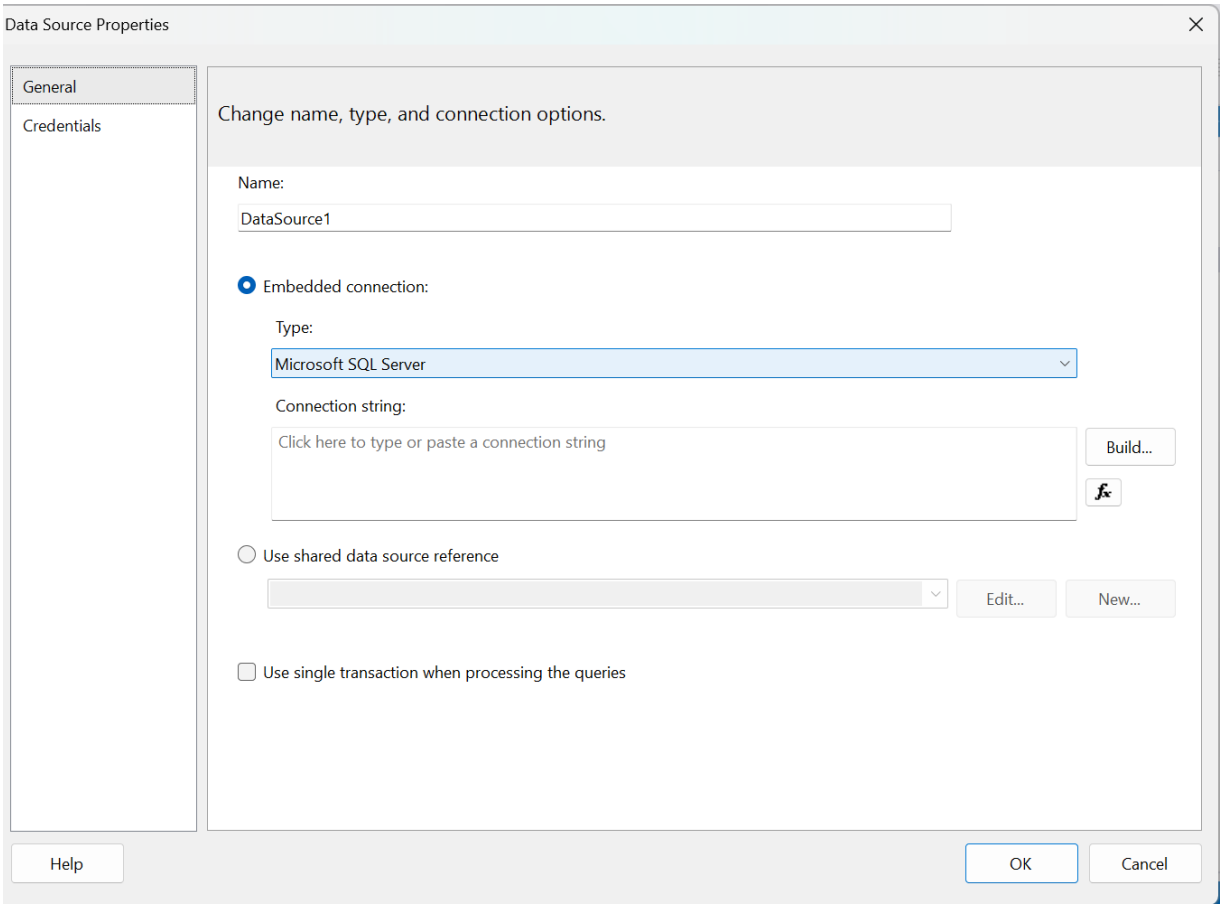
1.2.2. Tạo kết nối với kho dữ liệu:

Bước 1: Tại thanh **Report Data** → Chuột phải vào **Data Source** → chọn **Add Data Source**… như hình minh hoạ sau:



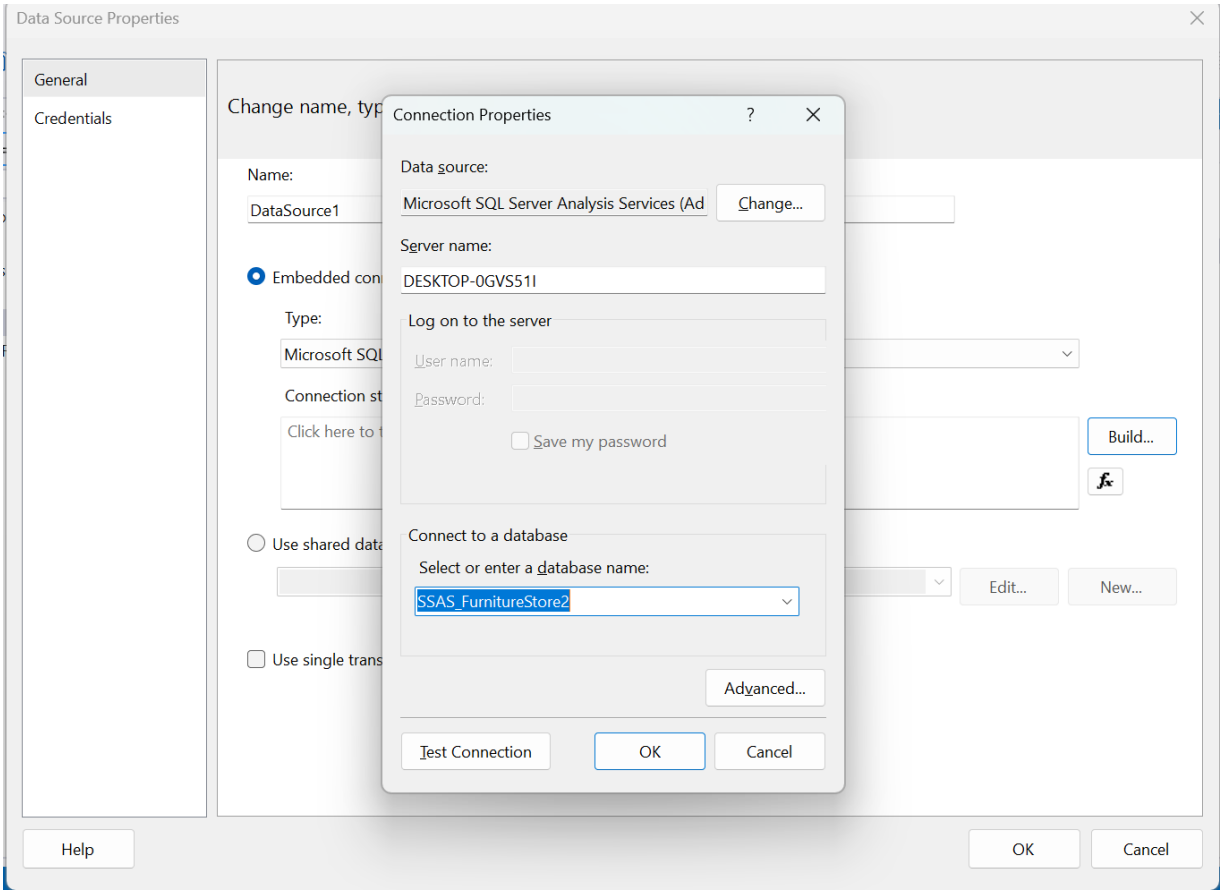
Hình 4.26 Tạo mới một Data Source

Bước 2: Khi giao diện **Data Source Properties** hiển thị lên thì ta chọn **Build** để tạo một chuỗi kết nối đến **Data Source**.

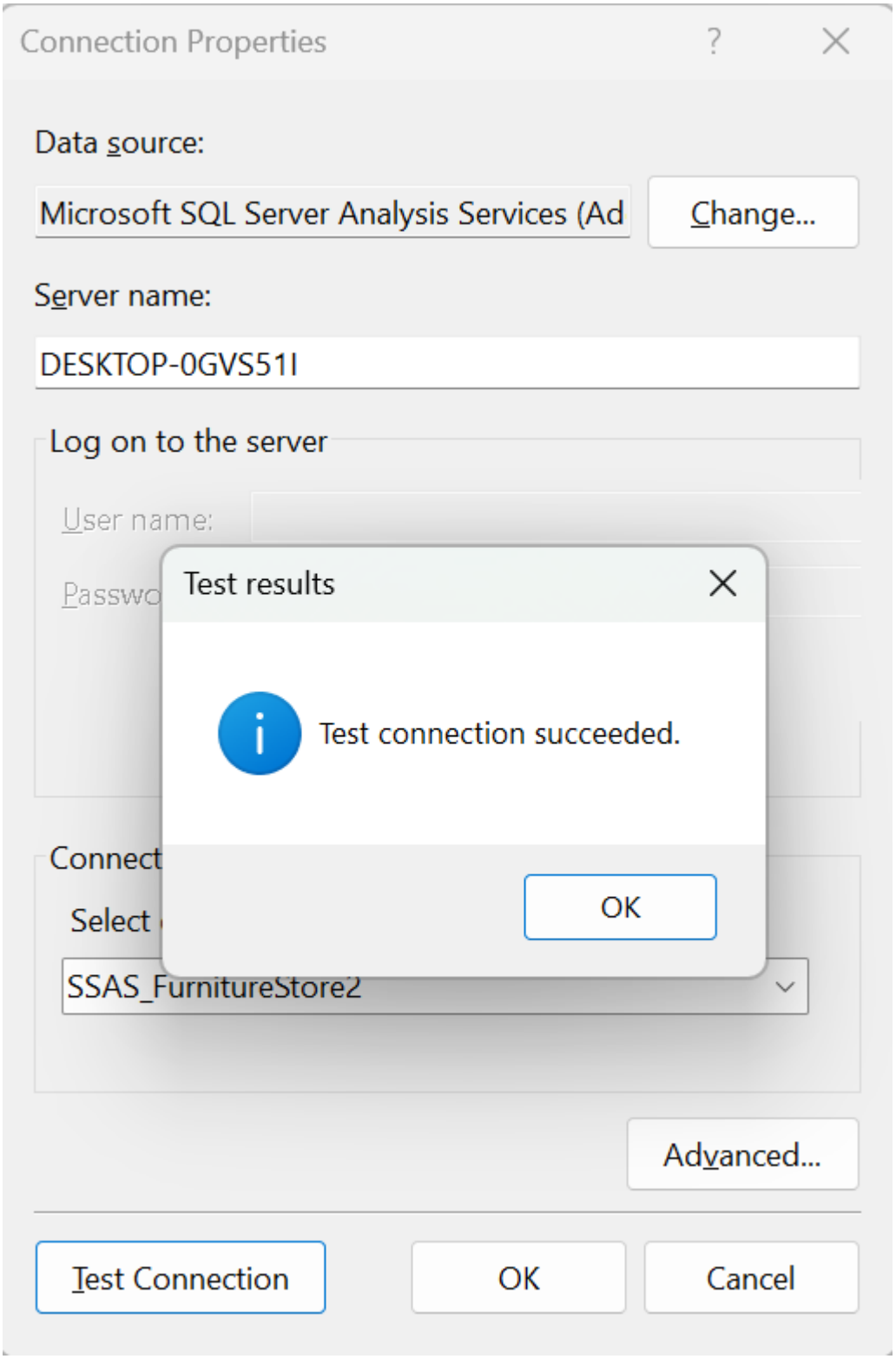


Hình 4.27 Tạo Connection String(1)

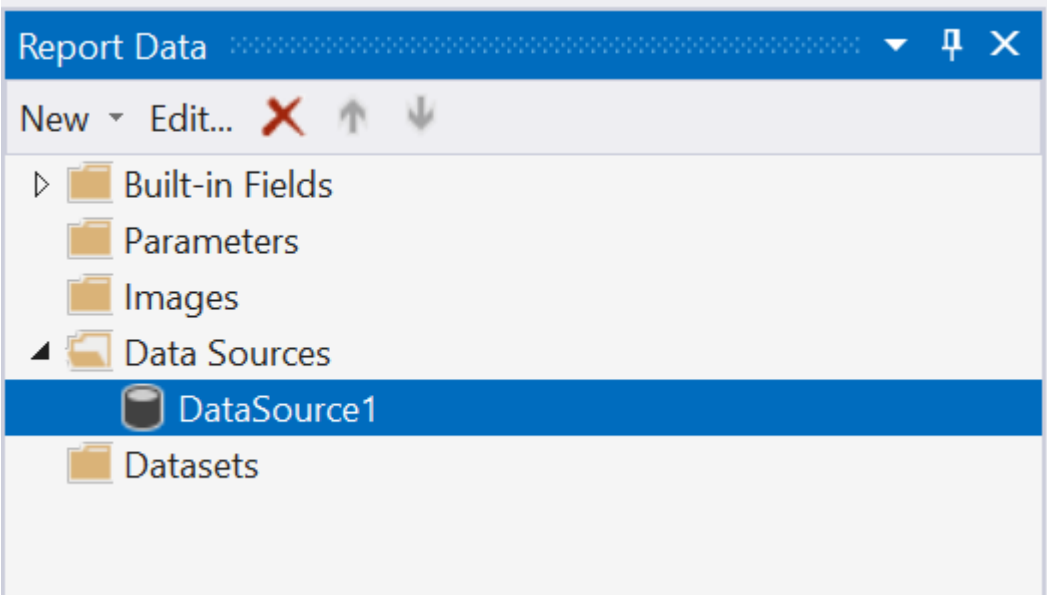
Bước 2: Trường **Type** ta chọn **Microsoft SQL Server Analysis Service** → Chọn **Type** và cấu hình như hình minh hoạ → Nhấn **Test Connection** để thử kết nối. Nếu như kết nối thành công thì nhấn **OK** → **OK.**



Hình 4.28 Tạo Connection String(2)



Hình 4.29 Kiểm tra kết nối thành công

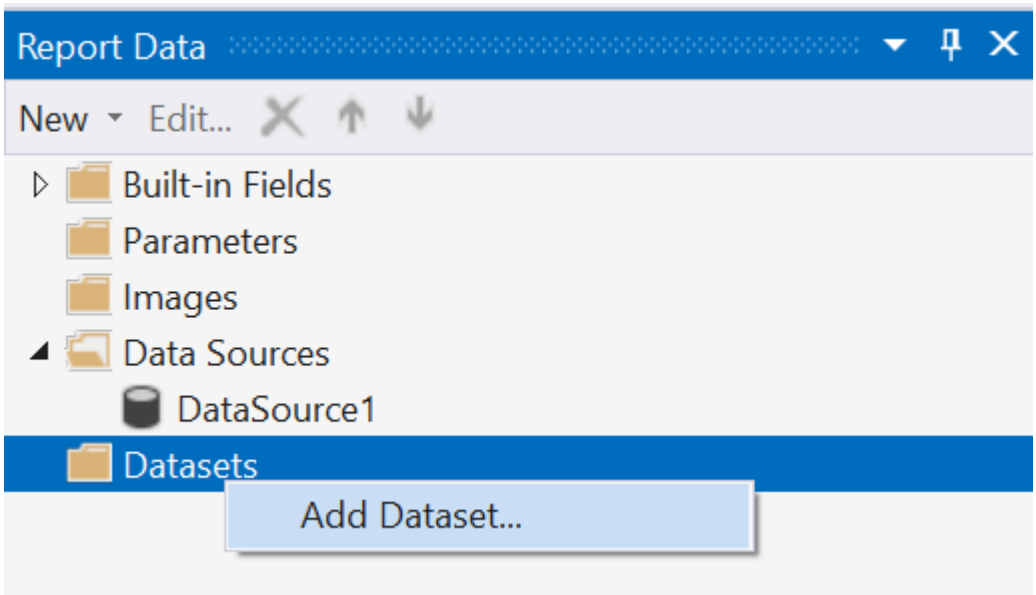


Hình 4.30 Kết quả

1.3. Thực hiện lập báo biểu trên Report Builder:

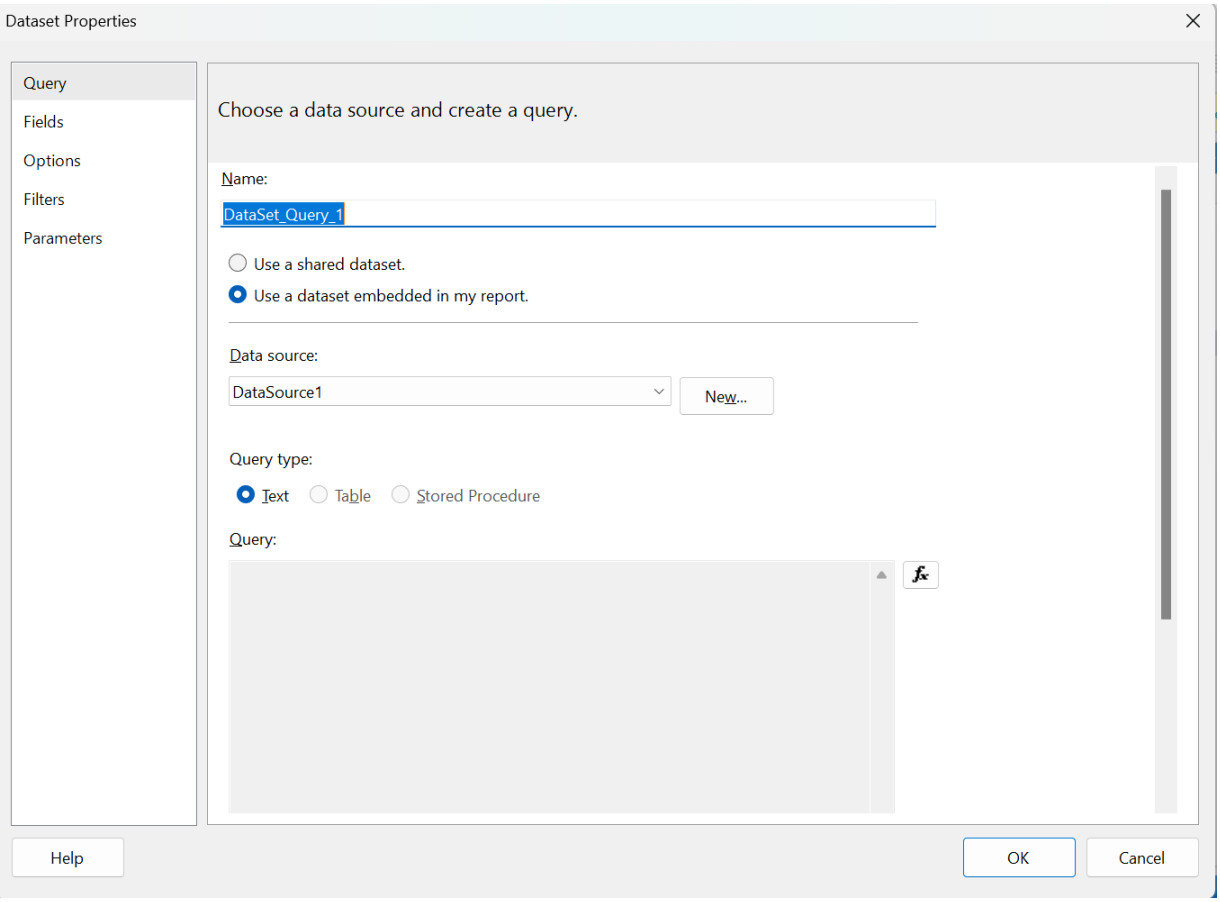
1.3.1. Báo biểu 1 (Report Grouping): Đưa doanh số, lợi nhuận, tổng số lượng mua của từng loại sản phẩm.

Bước 1: Thêm datasets cần thiết cho báo biểu 1 đặt tên là **DataSet_Query_1** bằng cách nhấn chuột phải vào **Datasets → Add new dataset…**



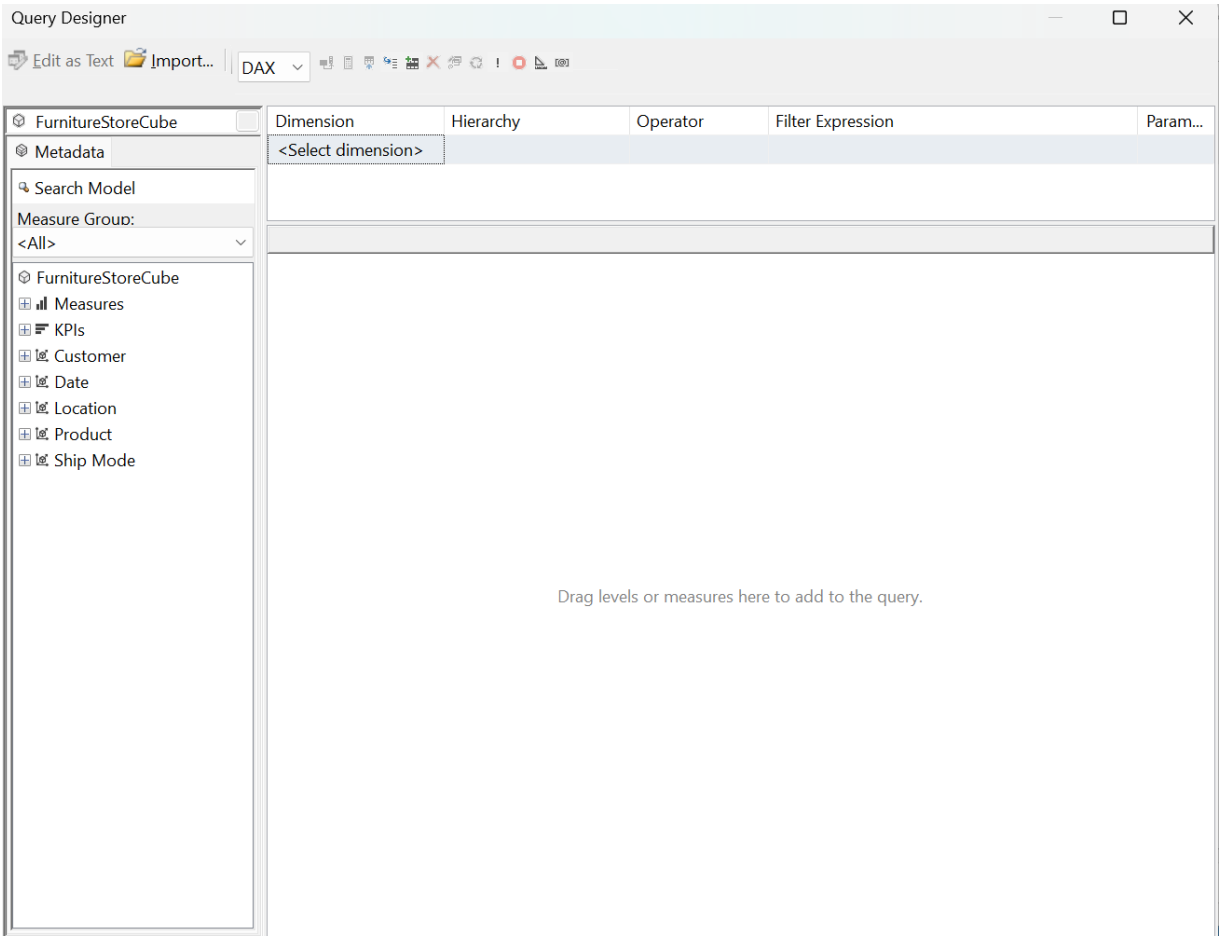
Hình 4.31 Tạo Dataset

Bước 2: Trong tab **Datasets properties** → Trường **Query**. Tại trường **Name** đặt tên là **Dataset_Query_1** chọn chế độ **Use a dataset embedded in my report**. Trường **Data Source** chọn data source tại phần tạo kết nối data source.



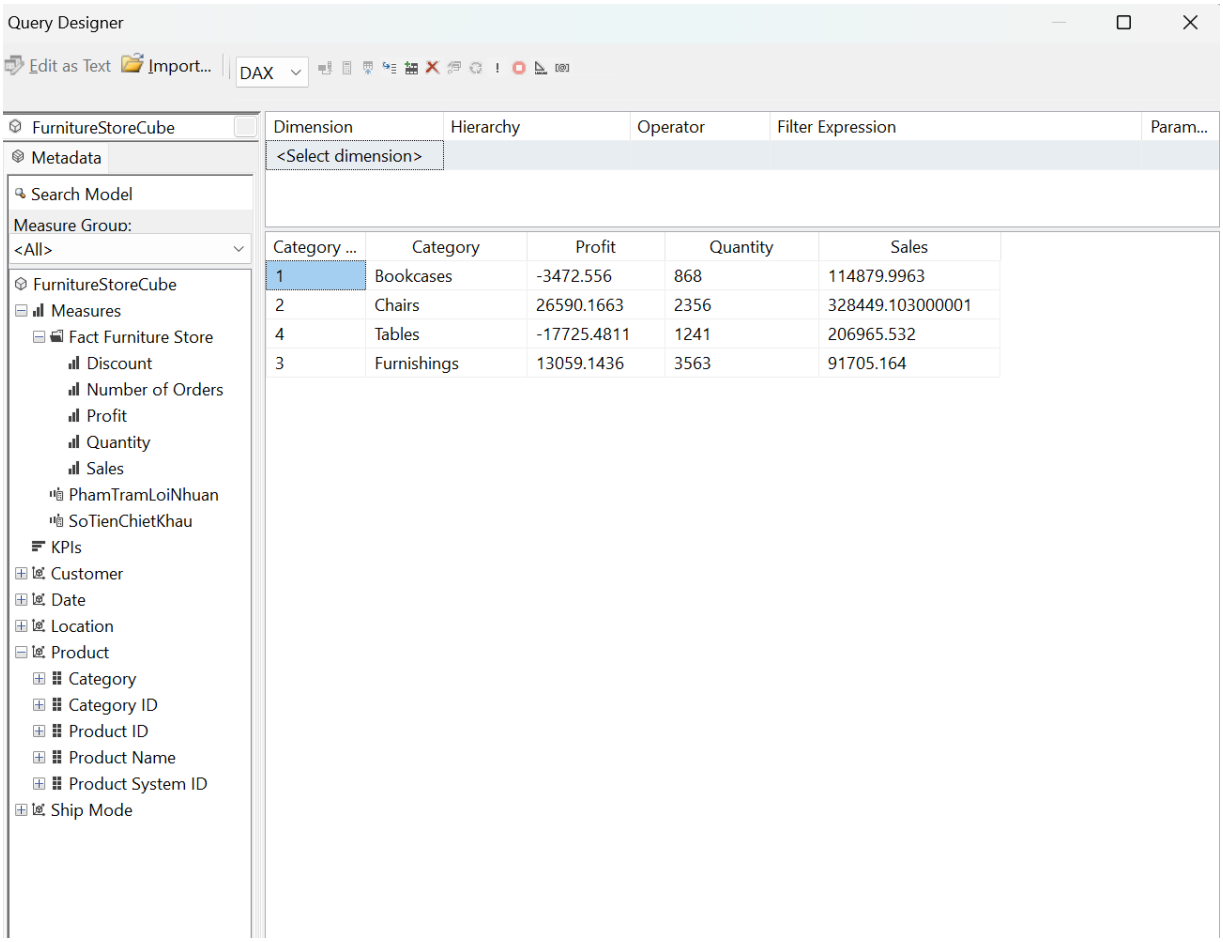
Hình 4.32 Chọn Data Source

Bước 3: Chọn **Query Designer** hộp thoại sẽ hiện ra:



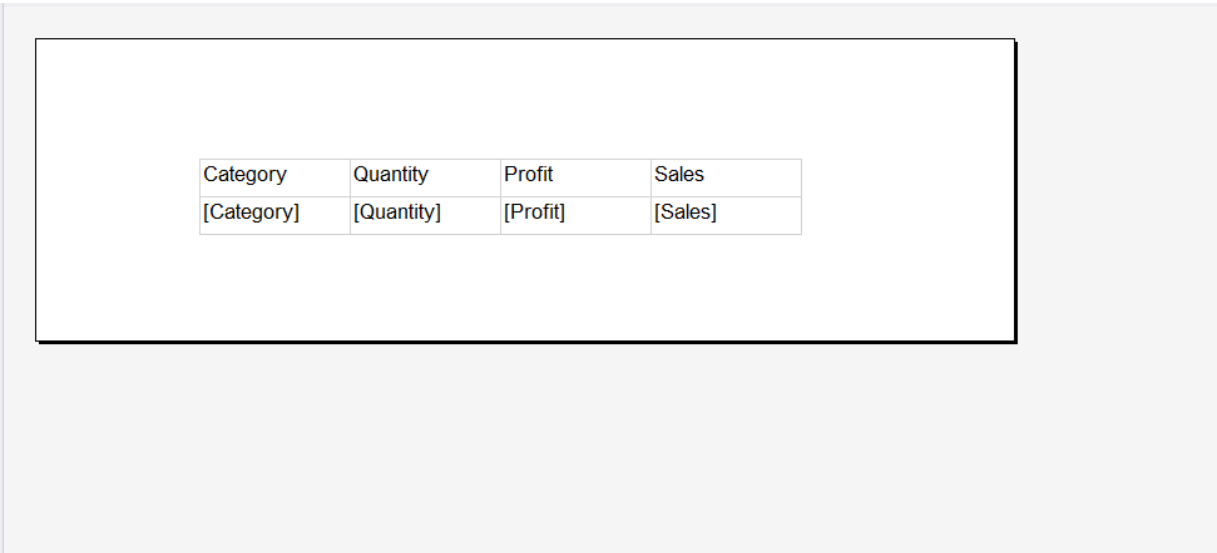
Hình 4.33 Giao diện Query Designer

Bước 4: Chọn những thuộc tính cần thiết cho báo biểu là: **CategoryID**, **Category, Profit, Quantity, Sales** như hình minh hoạ giống với cách chọn trong quá trình **SSAS** bằng công cụ **BI** sau đó nhấn **OK**.



Hình 4.34 Kéo thả các thuộc tính để lấy dữ liệu cho báo biểu

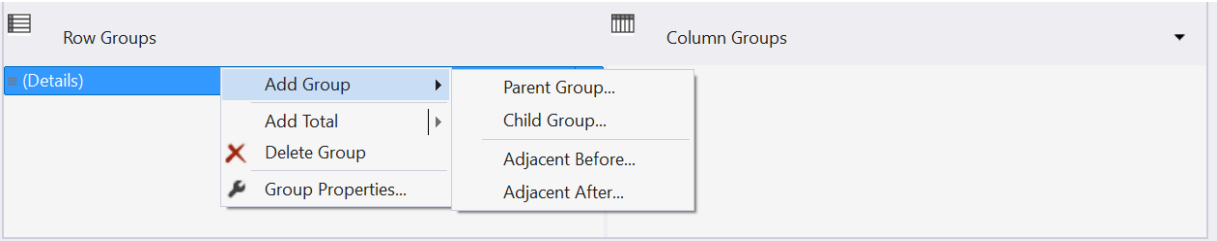
Bước 5: Qua trường **Toolbox** kéo table qua vùng **Design**, kéo những thuộc tính vào bảng **table** như hình minh hoạ:



Hình 4.35 Thiết kế của Report

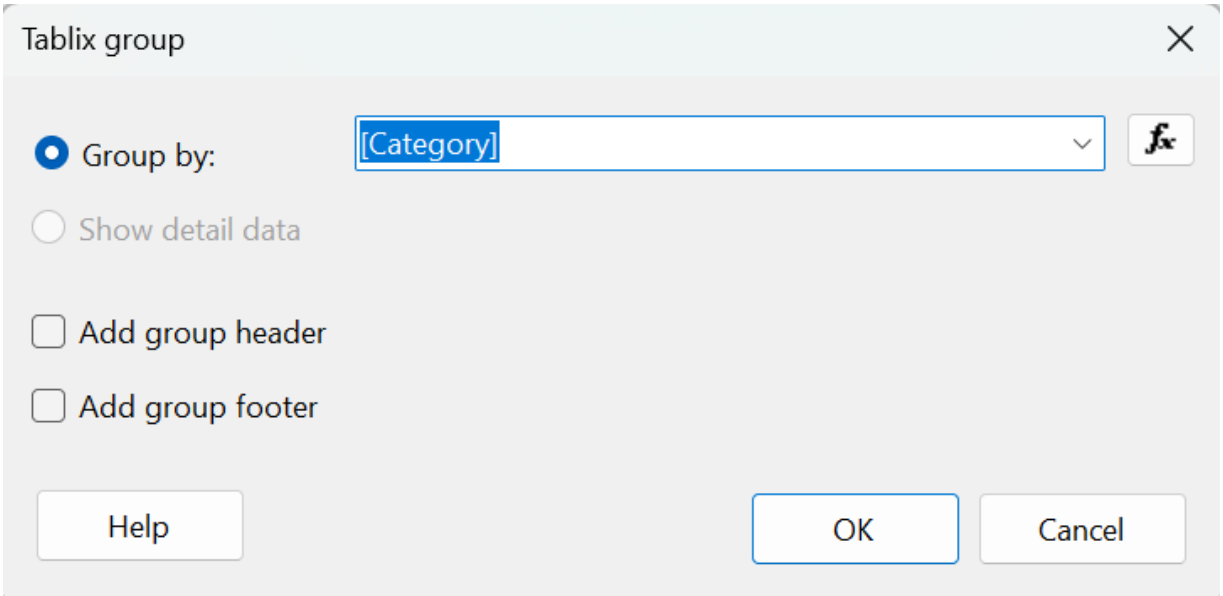
Bước 6: Định dạng cho báo biểu.

Bước 7: Tuỳ chọn **Grouping** bằng các bước sau: Nhấn chuột phải vào **Detail → Add Group → Parent Group.**



Hình 4.36 Tạo Grouping(1)

Bước 8: Chọn **Group by Category**



Hình 4.37 Tạo grouping(2)

Bước 9: Tiến hành format lại báo biểu và xem trước báo biểu bằng tab **Preview**:

BÁO CÁO KINH DOANH SẢN PHẨM

Category	Quantity	Profit	Sales
Bookcases	$868.00	-$3,472.56	114879.9963
Chairs	$2,356.00	$26,590.17	328449.103000001
Furnishings	$3,563.00	$13,059.14	91705.164
Tables	$1,241.00	-$17,725.48	206965.532

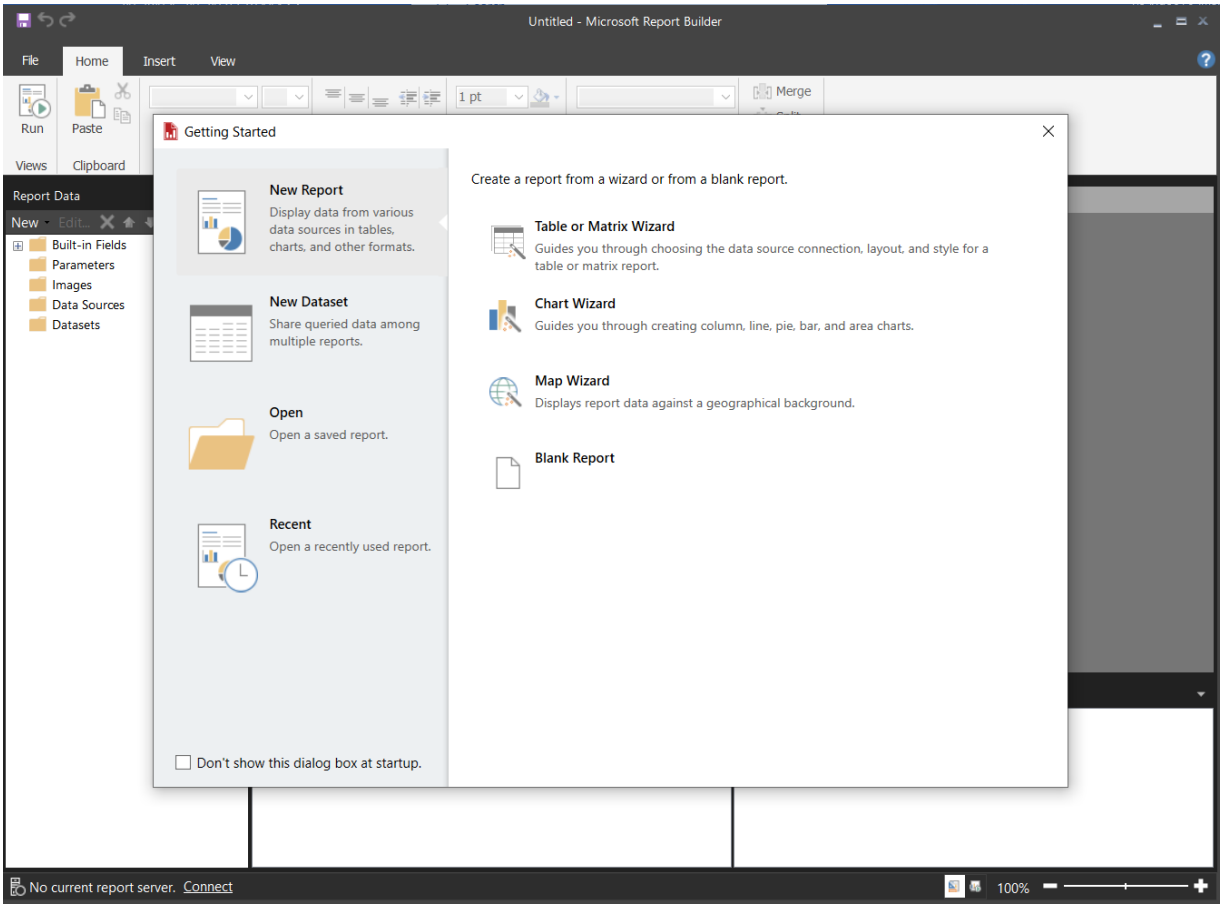
Người lập: Bùi Hữu Bằng 13/05/2024 20:54:34

Hình 4.38 Kết quả tạo báo biểu

Bước 10: Đăng Report lên server bằng chọn chức năng Connect Report Builder để kết nối tới Report Server mà ta đã tạo ở những bước đầu tiên.

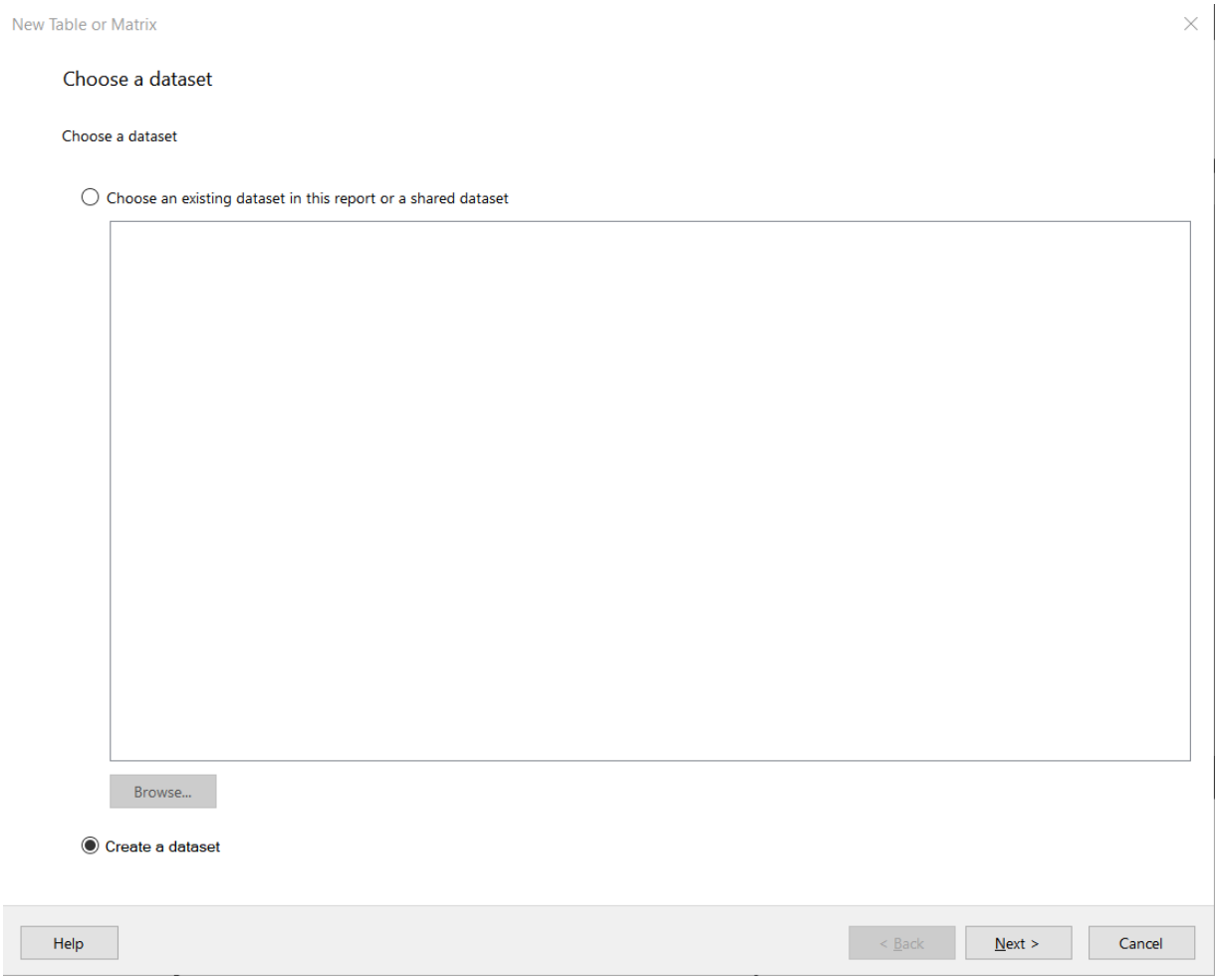
1.3.2. Báo biểu 2 (Report Grouping + Total + Matrix): Báo cáo về doanh thu của các thành phố thuộc nước Mỹ từ trong khoảng thời gian 2014 - 2017 theo các khoảng thời gian tháng, quí, năm.

Bước 1: Ở đây chúng ta sẽ sử dụng công cụ **Report Builder** để thực hiện tạo báo biểu 2, đầu tiên chúng ta vào công cụ **Report Builder** → Chọn **Table or Matrix Wizard** để tiến hành tạo một báo biểu có dạng **Matrix**.



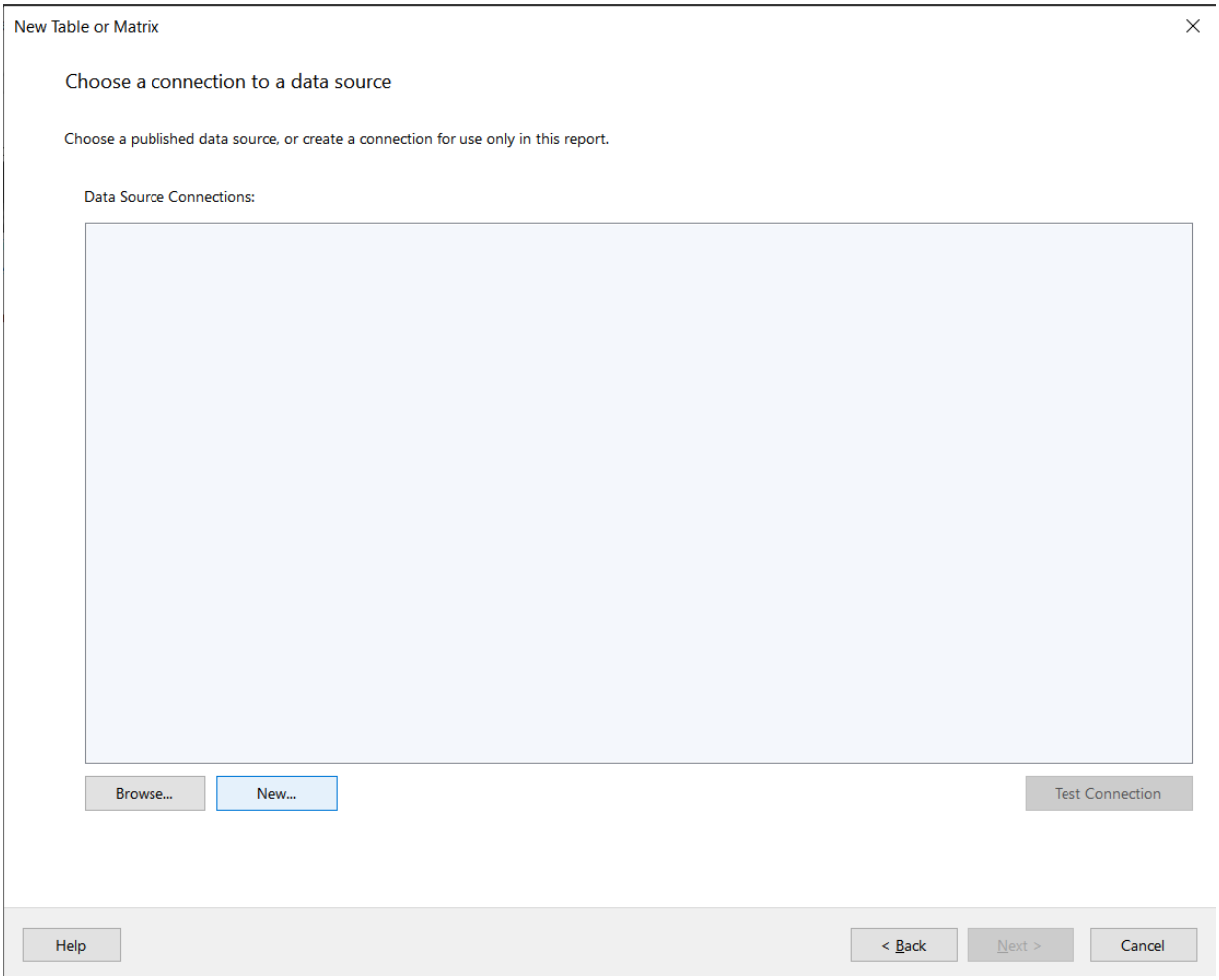
Hình 4.39 Tạo báo biểu mới có dạng Matrix

Bước 2: Tại bước chọn Dataset thì ta tích chọn ô **Create a dataset** để tiến hành tạo một dataset mới → **Next** để tiếp tục.



Hình 4.40 Thêm mới một Dataset

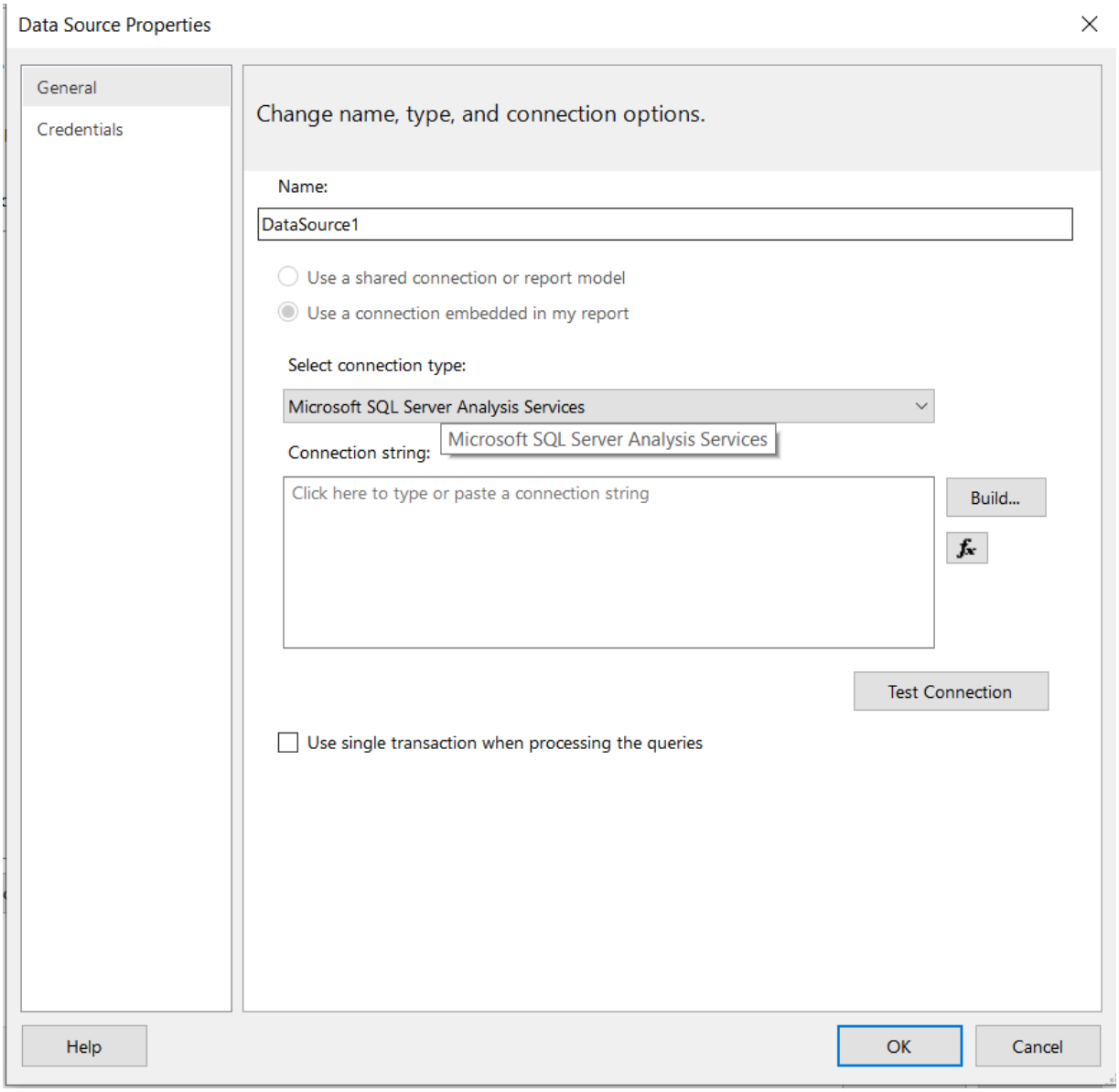
Bước 3: Để có thể tạo một dataset mới thì đầu tiên ta cần tạo một kết nối tới **Data Source** của quá tình SSAS, ở giao diện chọn kết nối tới một nguồn dữ liệu ta chọn **New** để tiến hành tạo một kết nối mới.



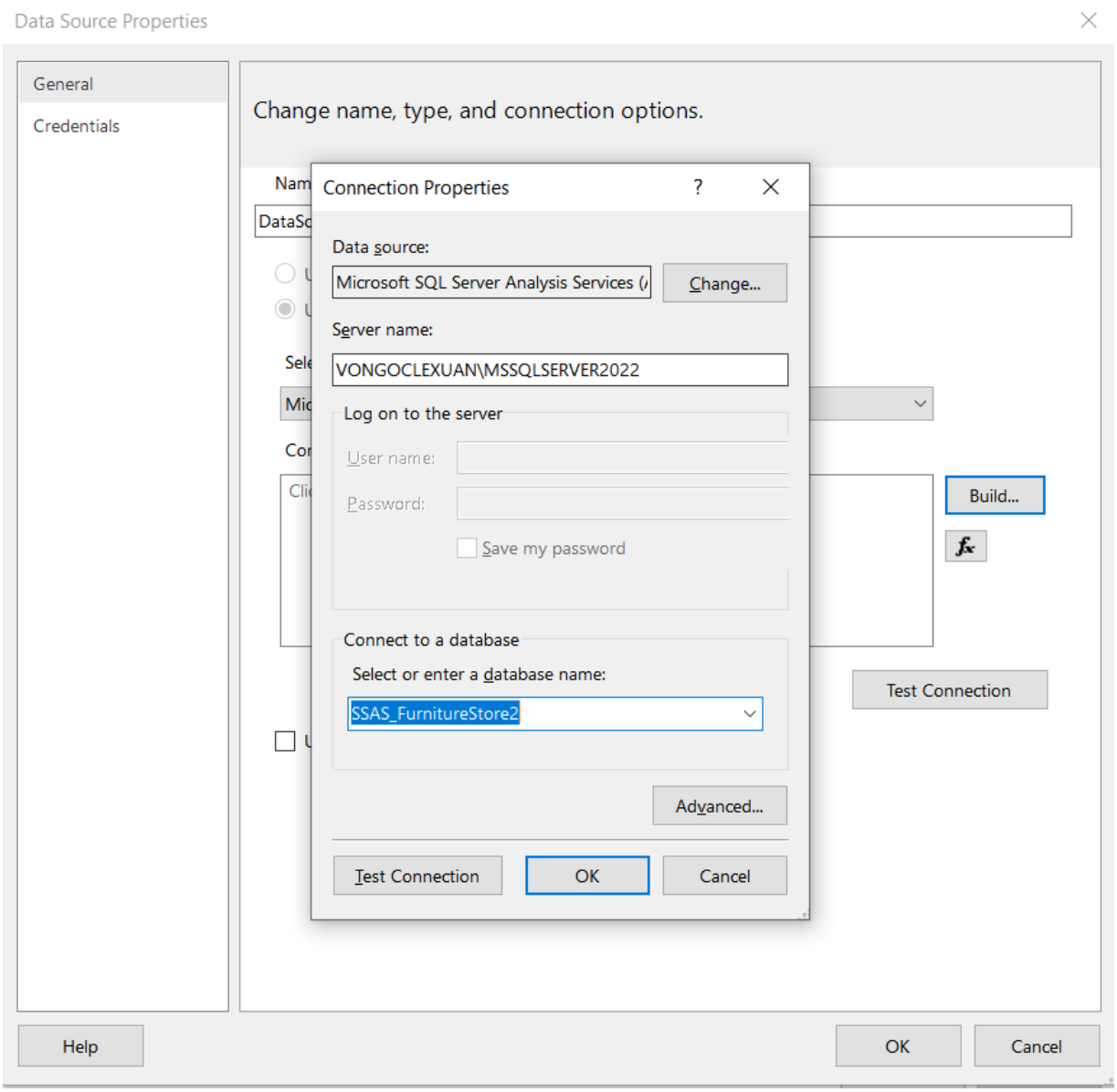
Hình 4.41 Tạo kết nối mới tới Data Source

Bước 4: Cũng tương tự như với dịch vụ **SSRS** của **Visual Studio 2022** thì ở đây chúng ta sẽ có giao diện tương tự như với **Data Source Properties** với những cài đặt tương tự, chúng ta sẽ tiến hành các bước tuần tự như sau: Đặt tên cho Data Source(Có thể để mặc định) → Chọn loại kết nối là **Microsoft SQL Server Analysis Services → Build →** Điền tên của server vào ô **Server name →** Chọn database trong ô **Select or enter a database name → OK .**

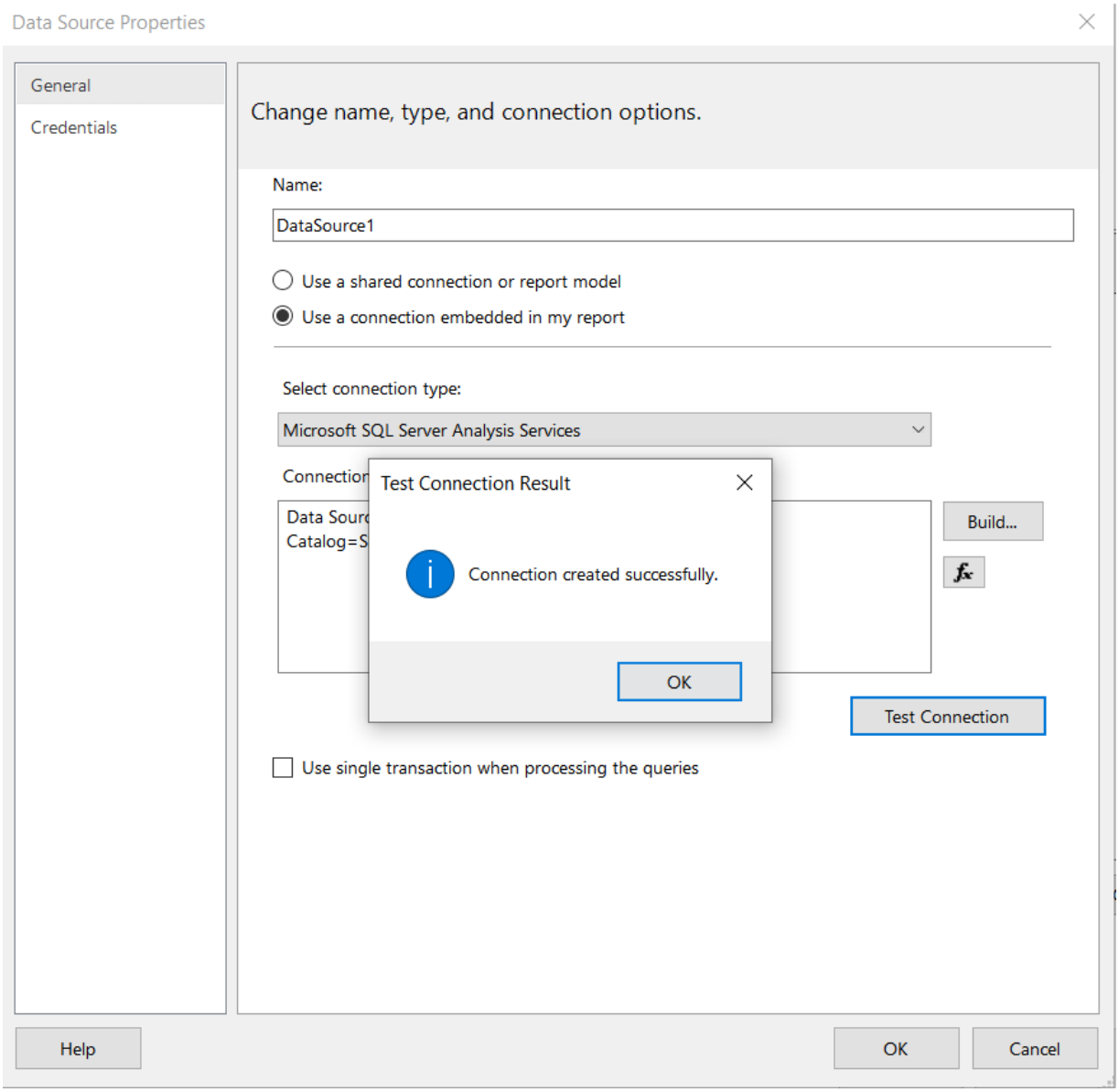
Sau khi đã kiểm kết nối đã được thiết lập thành công bằng chức năng **Test Connection** thì ta nhấn **OK** để hoàn thành việc tạo kết nối và nhấn **Next** để chuyển sang bước tiếp theo.



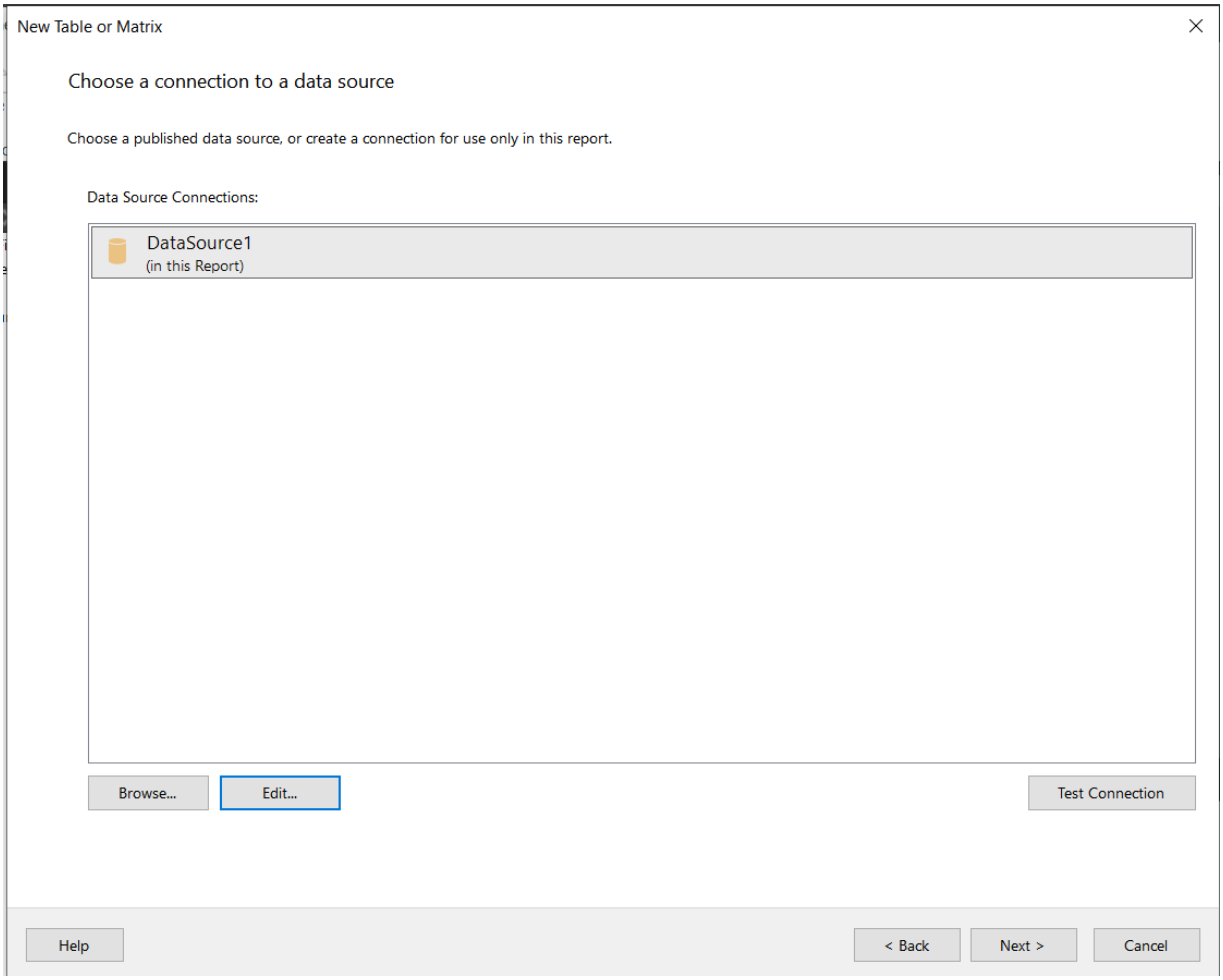
Hình 4.42 Tạo kết nối tới Data Source(1)



Hình 4.43 Tạo kết nối tới Data Source(2)

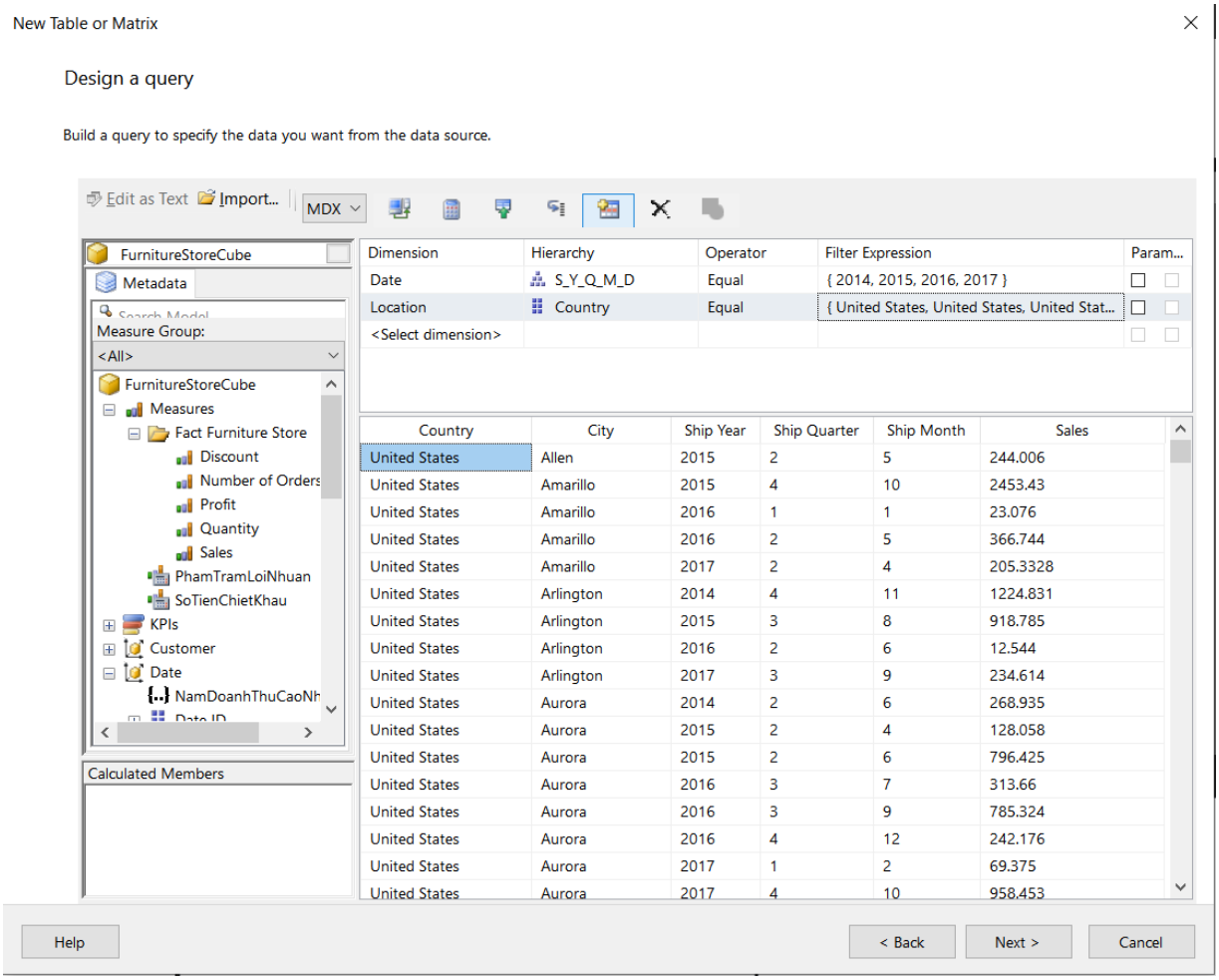


Hình 4.44 Kiểm thử kết nối tới Data Source



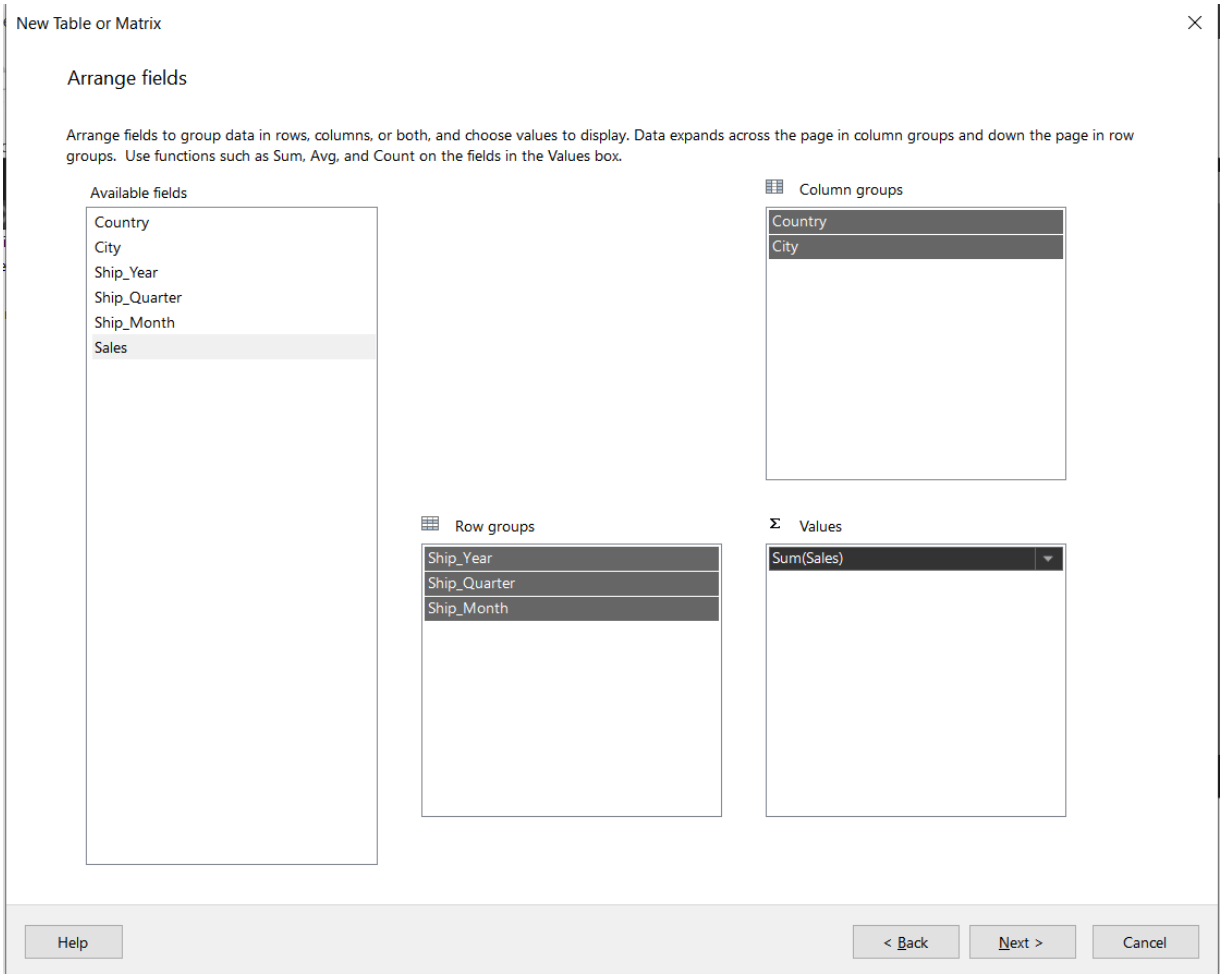
Hình 4.45 Chọn kết nối tới Data Source

Bước 5: Tại bước **Design a query**, chúng ta cần phải tạo câu truy vấn **MDX** để lấy dữ liệu cần thiết cho báo biểu 2. Ta lần lượt kéo các thuộc tính cần thiết và nhấn vào dòng chữ **Click to execute the query** để chạy thử câu truy vấn nhằm kiểm tra tính đúng đắn của câu lệnh truy vấn **MDX**. Sau khi đã xác nhận câu truy vấn **MDX** là đúng thì ta nhấn **Next** để chuyển sang bước thiết lập **Arrange Fields**.



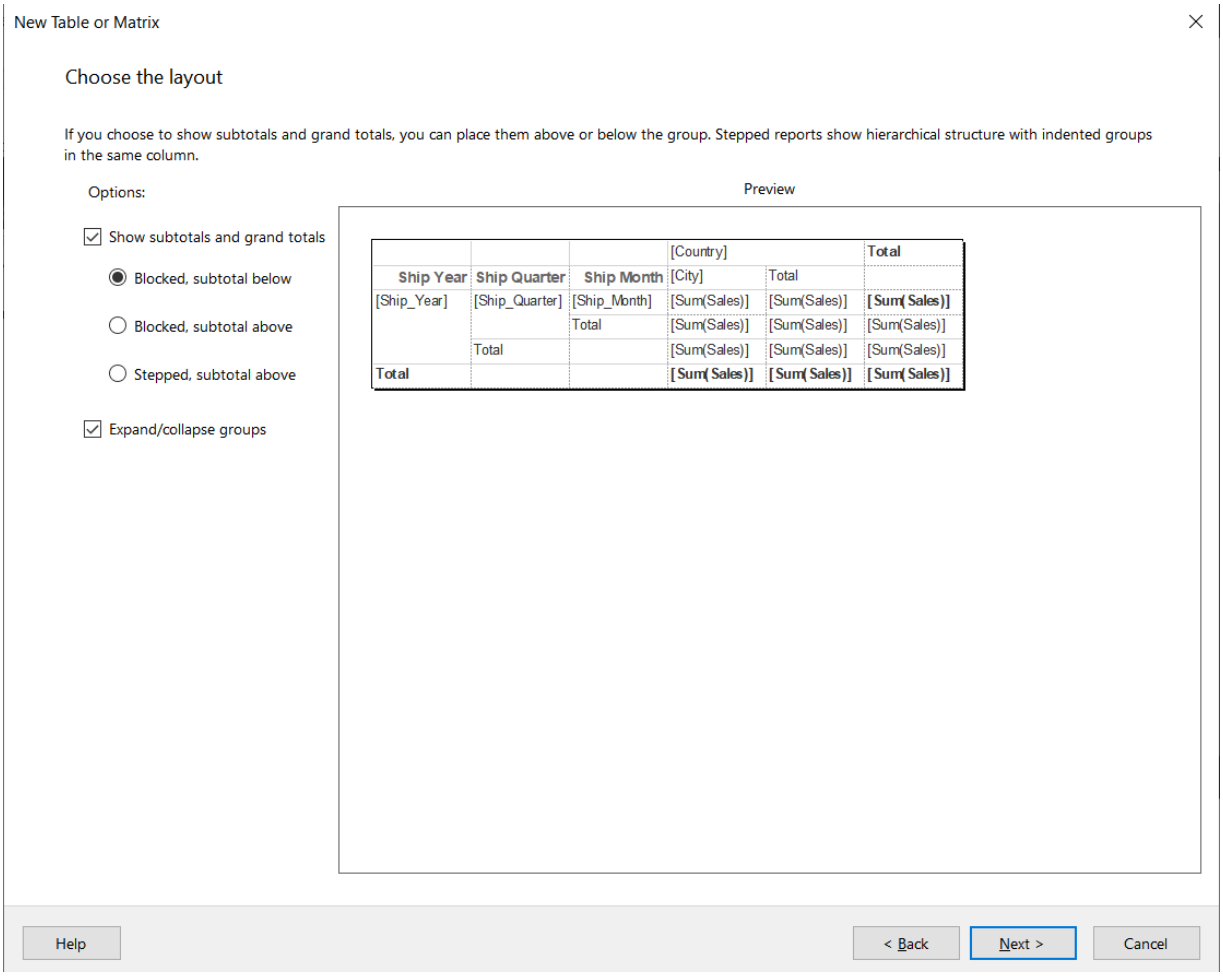
Hình 4.46 Kéo các thuộc tính để thực hiện câu lệnh MDX lấy dữ liệu cho báo biểu

Bước 6: Tại bước thiết lập **Arrange Fields**, ta lần lượt kéo các thuộc tính **Ship_Year, Ship_Quarter, Ship_Month** theo đúng thứ tự vào ô **Row groups** vì ta sẽ tiến hành nhóm các dòng dữ liệu theo các thuộc tính này. Tương tự, ta sẽ kéo các thuộc tính **Country, City** theo đúng thứ tự vào ô **Column groups** vì ta sẽ thể hiện doanh thu của các thành phố của nước Mỹ và cuối cùng đó chính là thuộc tính **Sales** sẽ kéo vào ô **Values**. Cuối cùng, nhấn **Next** để hoàn thành thiết lập bước này.

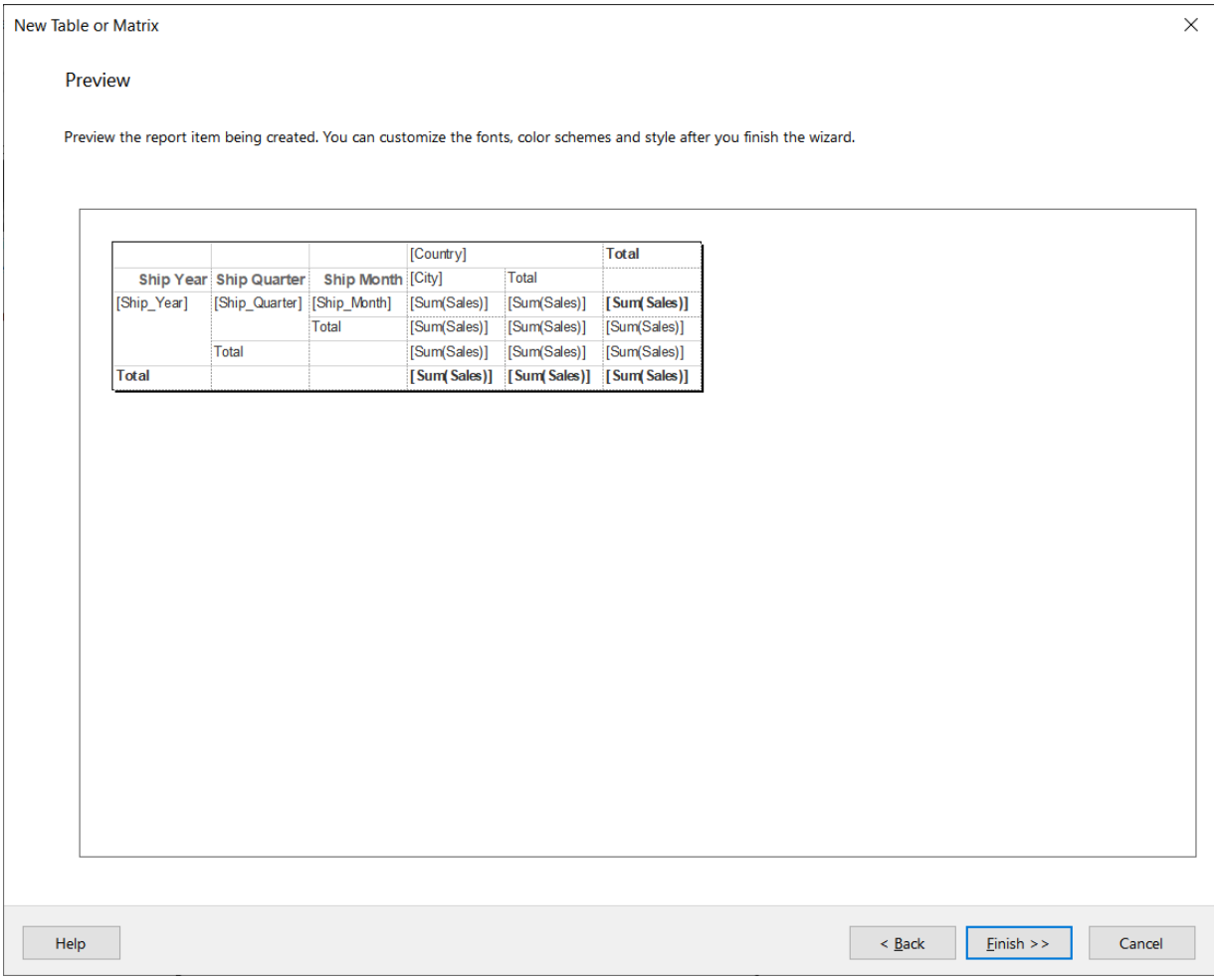


Hình 4.47 Sắp xếp các thuộc tính thành các cụm dữ liệu

Bước 7: Tại bước **Choose the layout**, ta chọn các tùy chọn cần thiết và tiếp tục nhấn **Next** để chuyển sang màn hình **Preview**(Xem trước). Khi đã xác nhận báo biểu có định dạng đúng như ta mong muốn thì ta nhấn **Finish** để hoàn thành việc tạo báo biểu số 2.

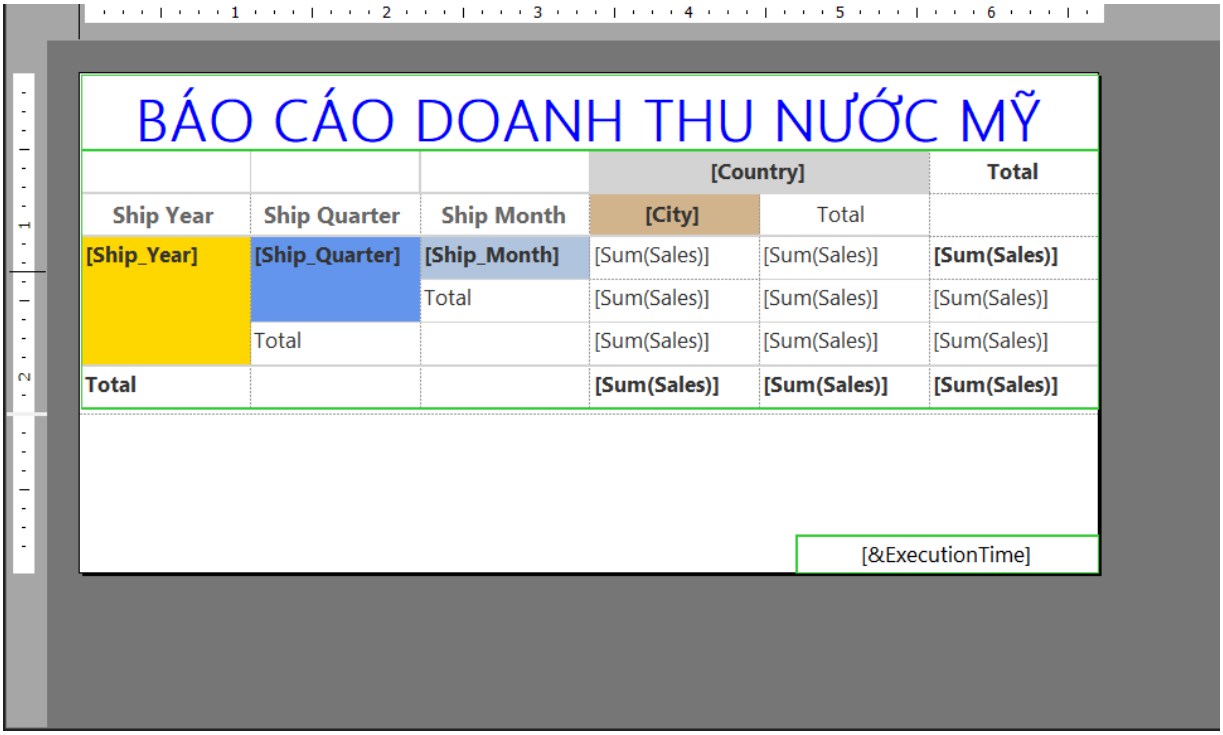


Hình 4.48 Chọn Layout cho báo biểu



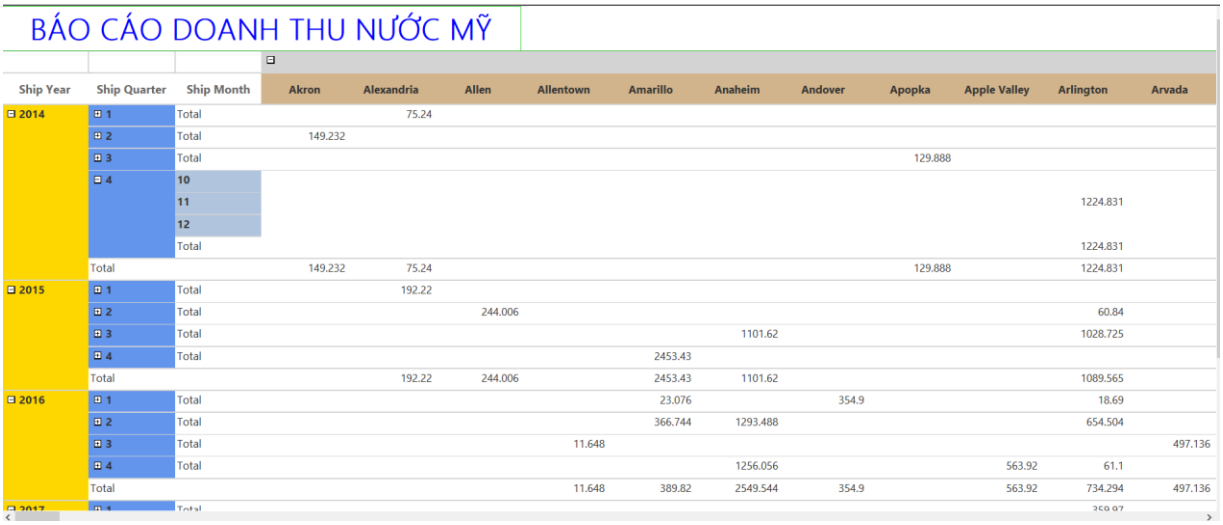
Hình 4.49 Xem lại báo biểu

Bước 8: Ta tiến hành định dạng lại báo biểu để dễ nhìn hơn.



Hình 4.50 Định dạng lại báo biểu

Bước 9: Ta chọn chức năng Run để xem thử kết quả của báo biểu thể hiện đúng theo như yêu cầu của báo biểu số 2.



BÁO CÁO DOANH THU NƯỚC MỸ

Ship Year	Ship Quarter	Ship Month	Akron	Alexandria	Allen	Allentown	Amarillo	Anaheim	Andover	Apopka	Apple Valley	Arlington	Arvada
2014	1	Total		75.24									
	2	Total	149.232										
	3	Total								129.888			
	4	10											
		11										1224.831	
		12											
		Total										1224.831	
	Total		149.232	75.24						129.888		1224.831	
2015	1	Total		192.22									
	2	Total			244.006							60.84	
	3	Total						1101.62				1028.725	
	4	Total					2453.43						
	Total			192.22	244.006		2453.43	1101.62				1089.565	
2016	1	Total					23.076		354.9			18.69	
	2	Total					366.744	1293.488				654.504	
	3	Total				11.648							497.136
	4	Total						1256.056			563.92	61.1	
	Total					11.648	389.82	2549.544	354.9		563.92	734.294	497.136

Hình 4.51 Kết quả tạo báo biểu

Bước 10: Xuất báo biểu ra file Pdf hoặc tiến hành đăng lên **Report Server**.

BÁO CÁO DOANH THU NƯỚC MỸ

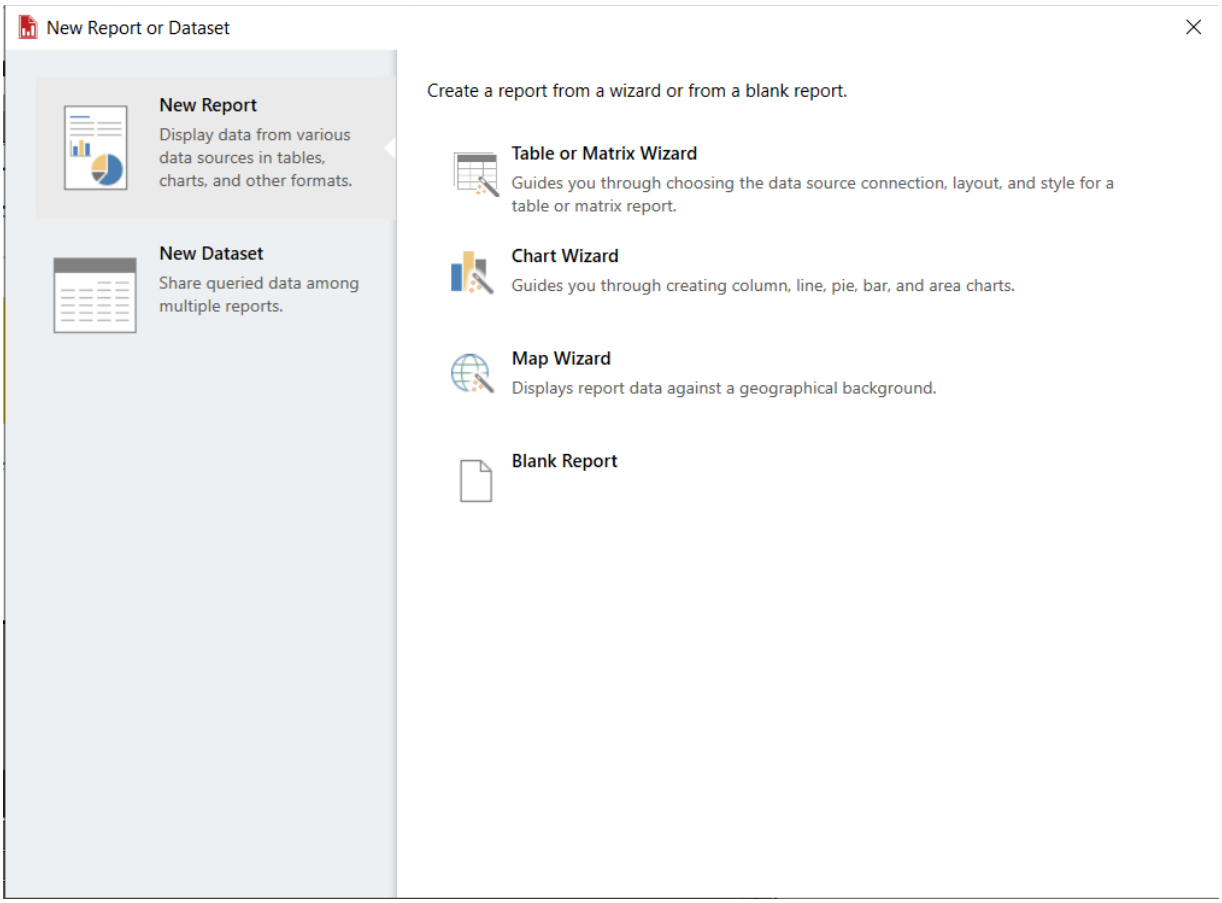
			United States	
Ship Year	Ship Quarter	Ship Month	Akron	Alexandria
2014	1	Total		75.24
	2	Total	149.232	
	3	Total		
	4	10		
		11		
		12		
		Total		
	Total		149.232	75.24
2015	1	Total		192.22
	2	Total		
	3	Total		
	4	Total		
	Total			192.22
2016	1	Total		
	2	Total		
	3	Total		
	4	Total		
	Total			
2017	1	Total		
	2	Total		
	3	Total		

5/14/2024 3:58:15 AM

Hình 4.52 Báo biểu dưới dạng PDF

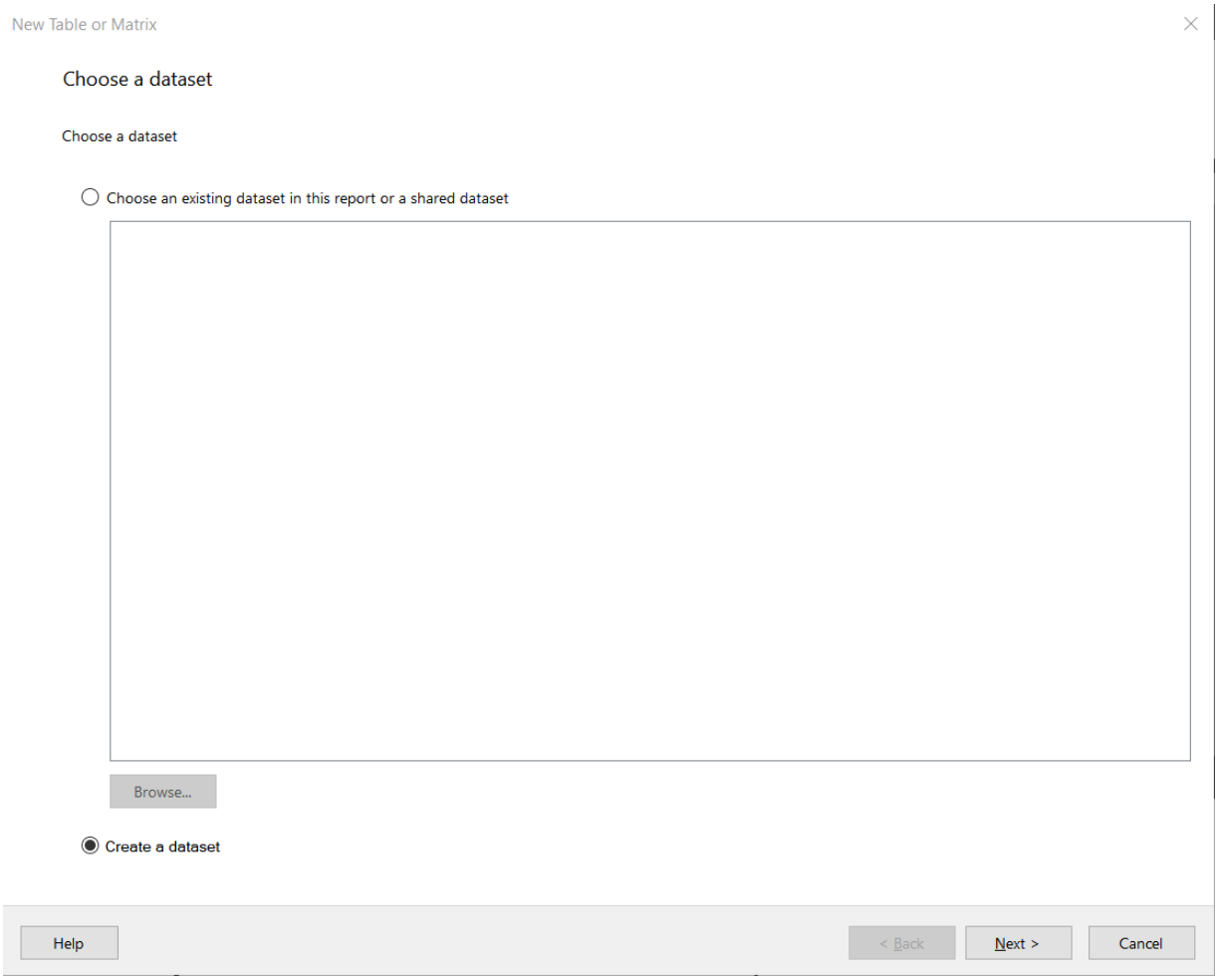
1.3.3. Báo biểu 3(Report Chart): Với từng loại khách hàng(Segment) hãy đưa ra doanh thu từ năm 2014 đến năm 2017. Kết quả của báo biểu là một biểu đồ dạng Column Chart.

Bước 1: Đầu tiên chúng ta vào công cụ **Report Builder** → Chọn **Chart Wizard** để tiến hành tạo một báo biểu có dạng là đồ thị.



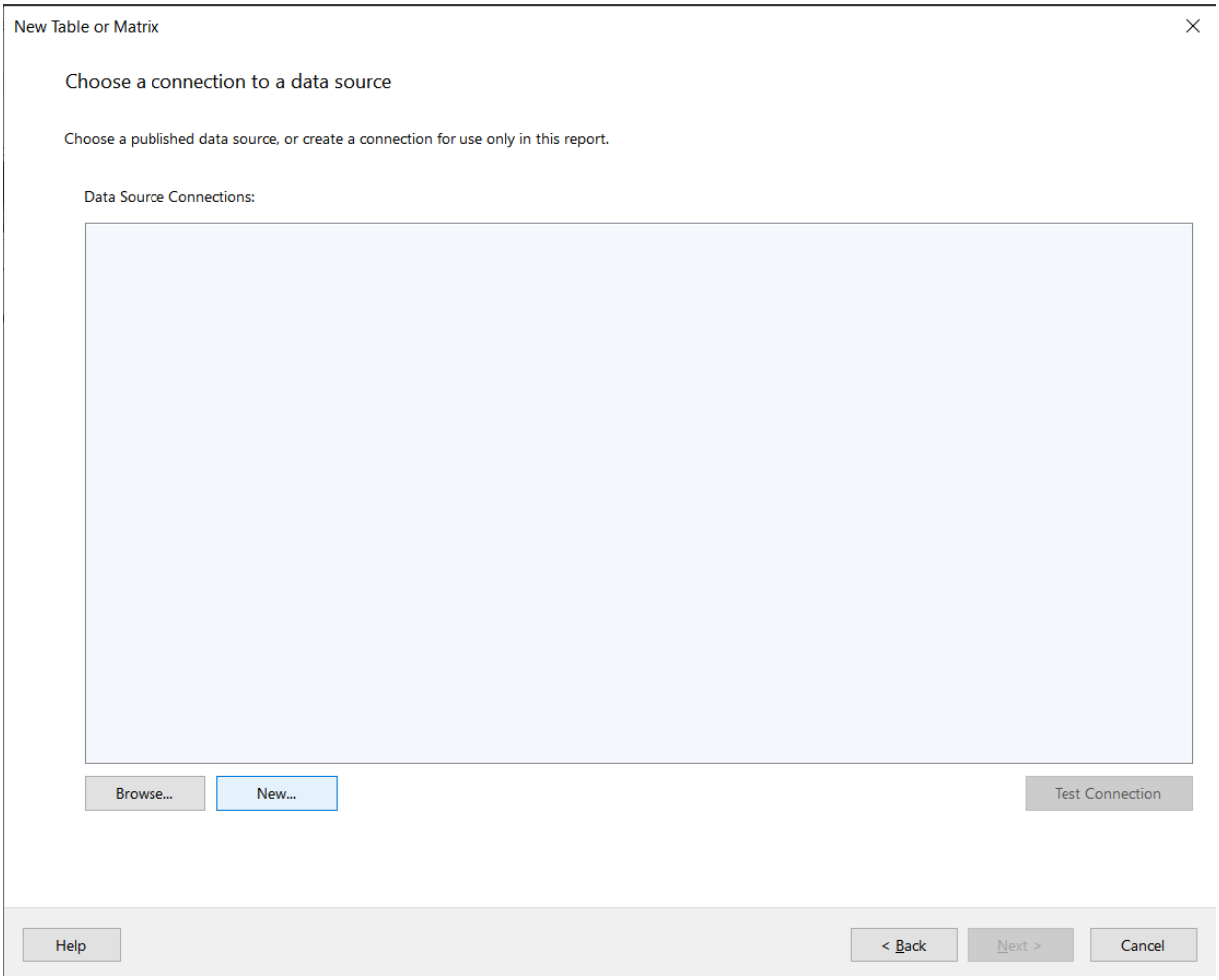
Hình 4.53 Tạo mới một báo biểu có dạng biểu đồ

Bước 2: Tại bước chọn **Dataset** thì ta tích chọn ô **Create a dataset** để tiến hành tạo một dataset mới → **Next** để tiếp tục.



Hình 4.54 Chọn tạo mới một Dataset

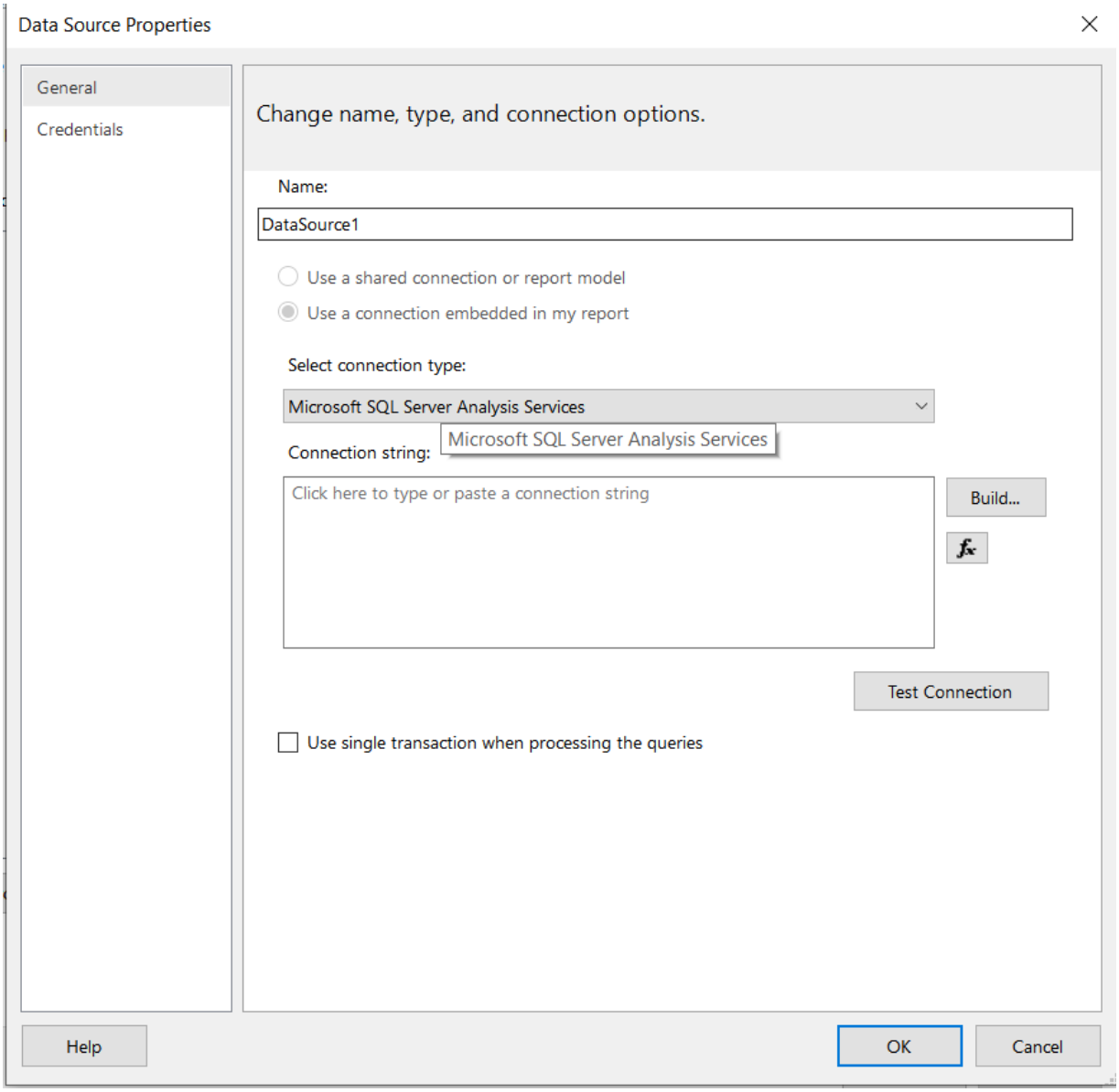
Bước 3: Để có thể tạo một dataset mới thì đầu tiên ta cần tạo một kết nối tới **Data Source** của quá tình SSAS, ở giao diện chọn kết nối tới một nguồn dữ liệu ta chọn **New** để tiến hành tạo một kết nối mới.



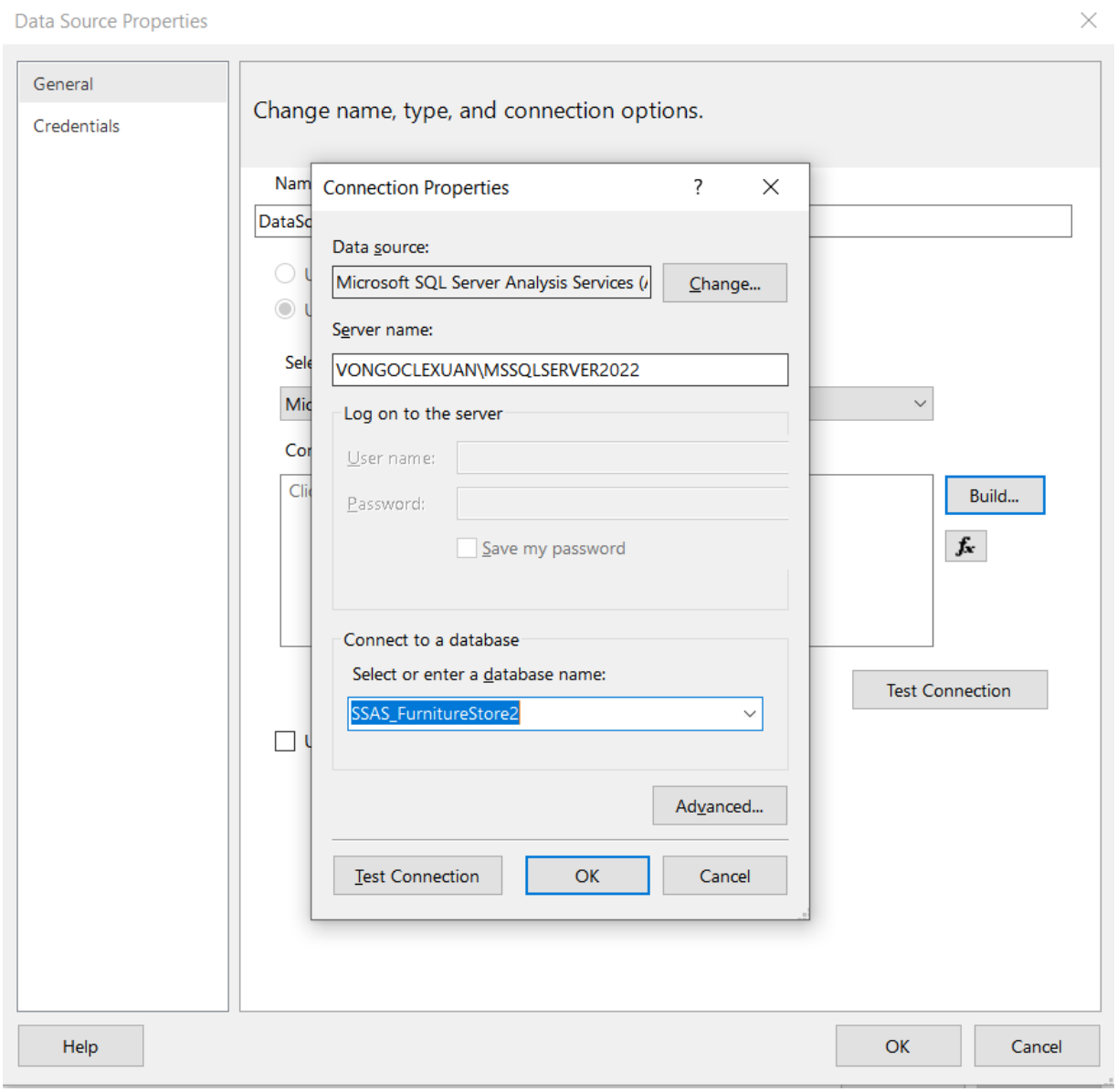
Hình 4.55 Tạo mới một kết nối tới Data Source

Bước 4: Tương tự với tạo báo biểu 2, ta thực hiện các bước như sau: Đặt tên cho Data Source(Có thể để mặc định) → Chọn loại kết nối là **Microsoft SQL Server Analysis Services → Build →** Điền tên của server vào ô **Server name →** Chọn database trong ô **Select or enter a database name → OK .**

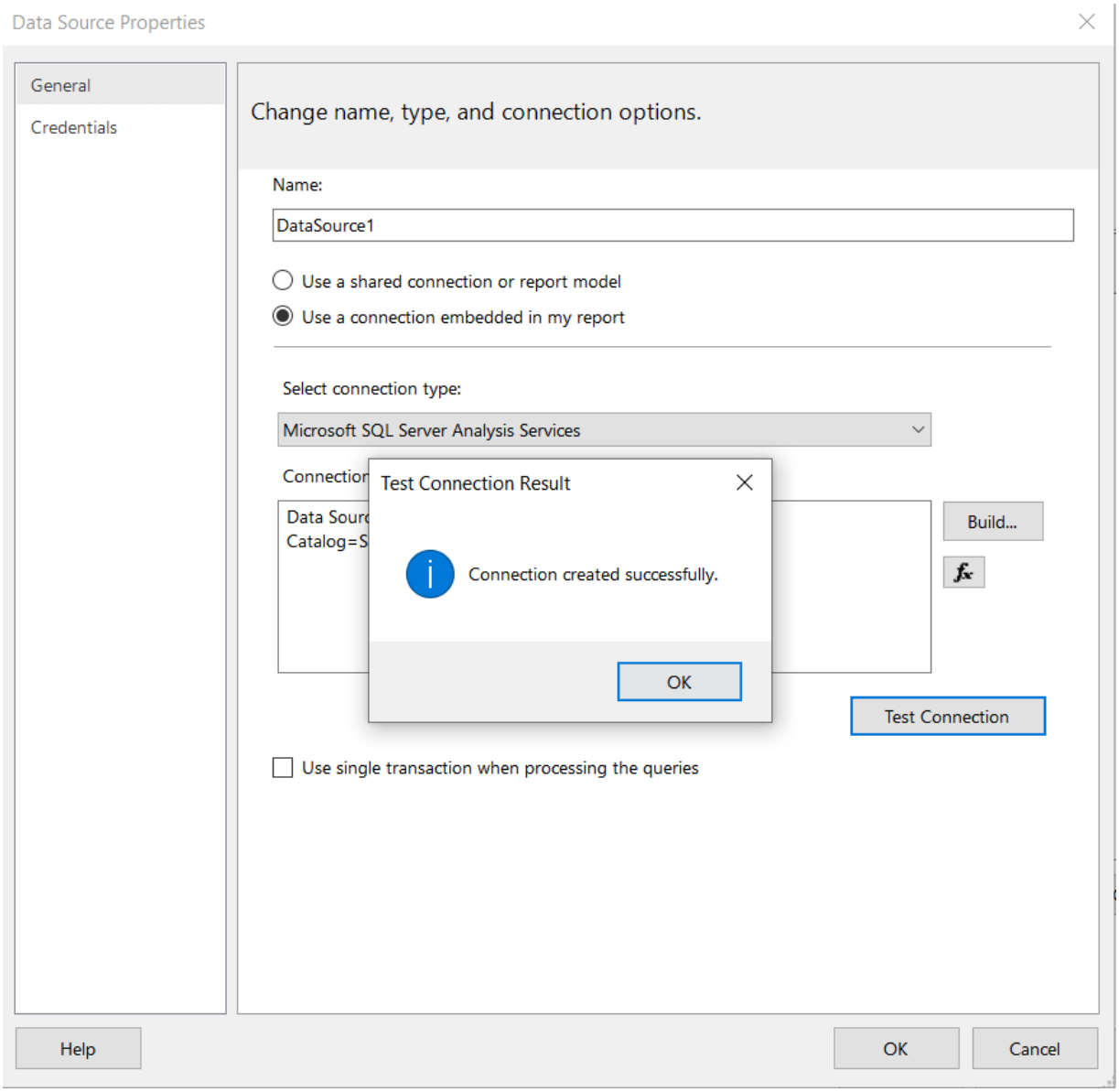
Sau khi đã kiểm kết nối đã được thiết lập thành công bằng chức năng **Test Connection** thì ta nhấn **OK** để hoàn thành việc tạo kết nối và nhấn **Next** để chuyển sang bước tiếp theo.



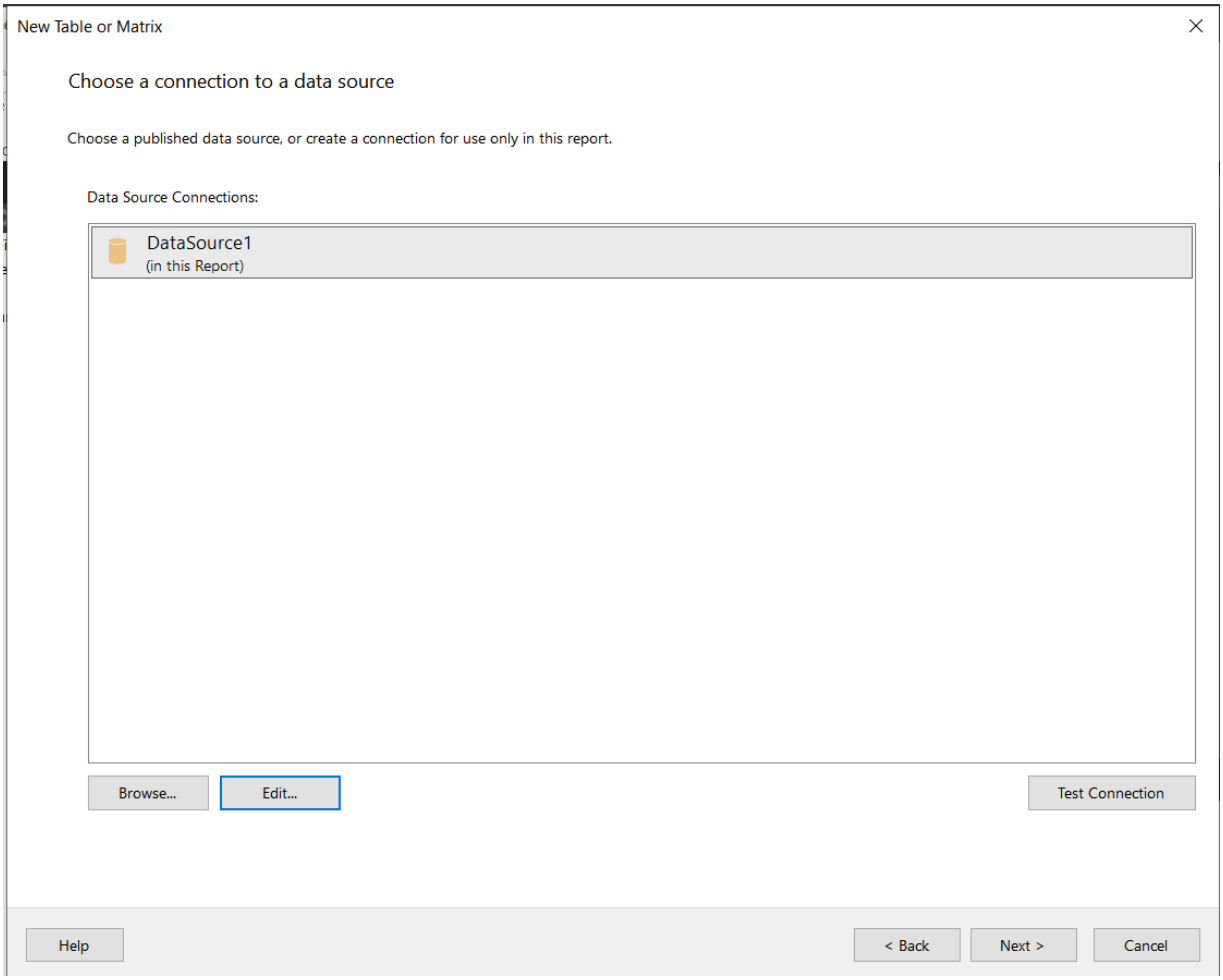
Hình 4.56 Tạo kết nối tới Data Source(1)



Hình 4.57 Tạo kết nối tới Data Source(2)

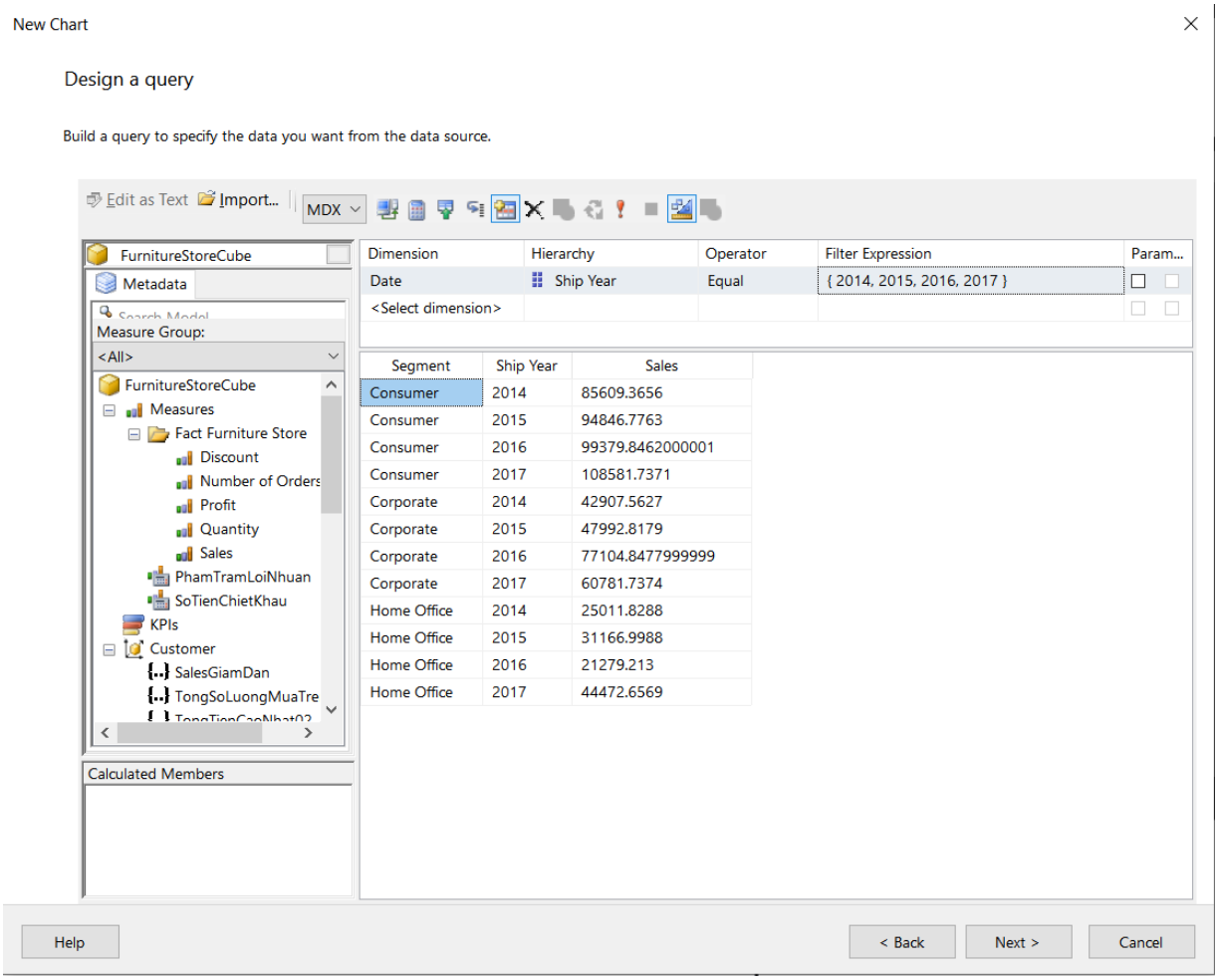


Hình 4.58 Kiểm tra kết nối tới Data Source thành công



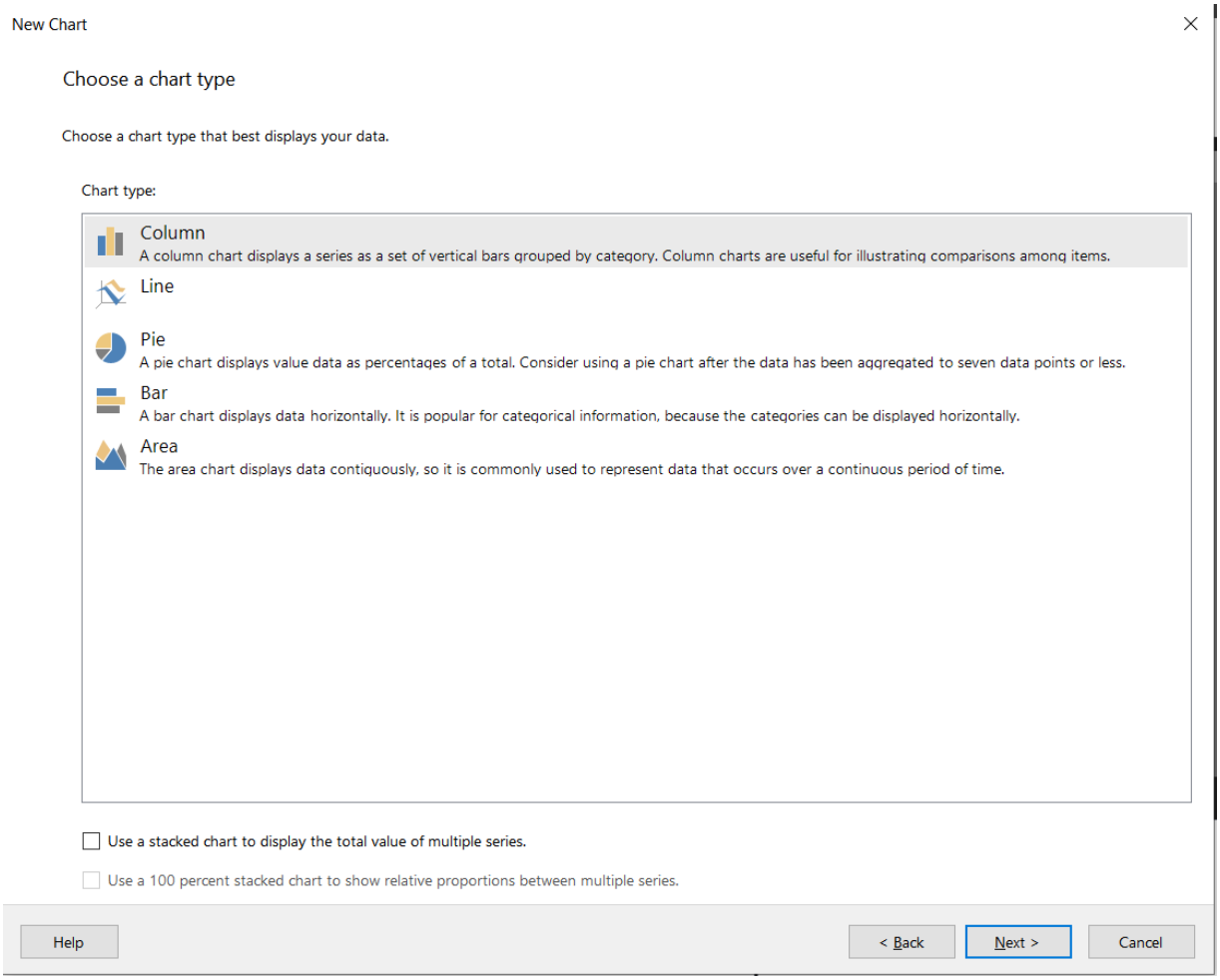
Hình 4.59 Chọn kết nối tới Data Source

Bước 5: Tại bước **Design a query**, chúng ta cần phải tạo câu truy vấn **MDX** để lấy dữ liệu cần thiết cho báo biểu 3. Ta lần lượt kéo các thuộc tính cần thiết(**Segment, Ship Year, Sales**) và nhấn vào dòng chữ **Click to execute the query** để chạy thử câu truy vấn nhằm kiểm tra tính đúng đắn của câu lệnh truy vấn **MDX**. Sau khi đã xác nhận câu truy vấn **MDX** là đúng thì ta nhấn **Next** để chuyển sang bước thiết lập **Arrange Fields**.



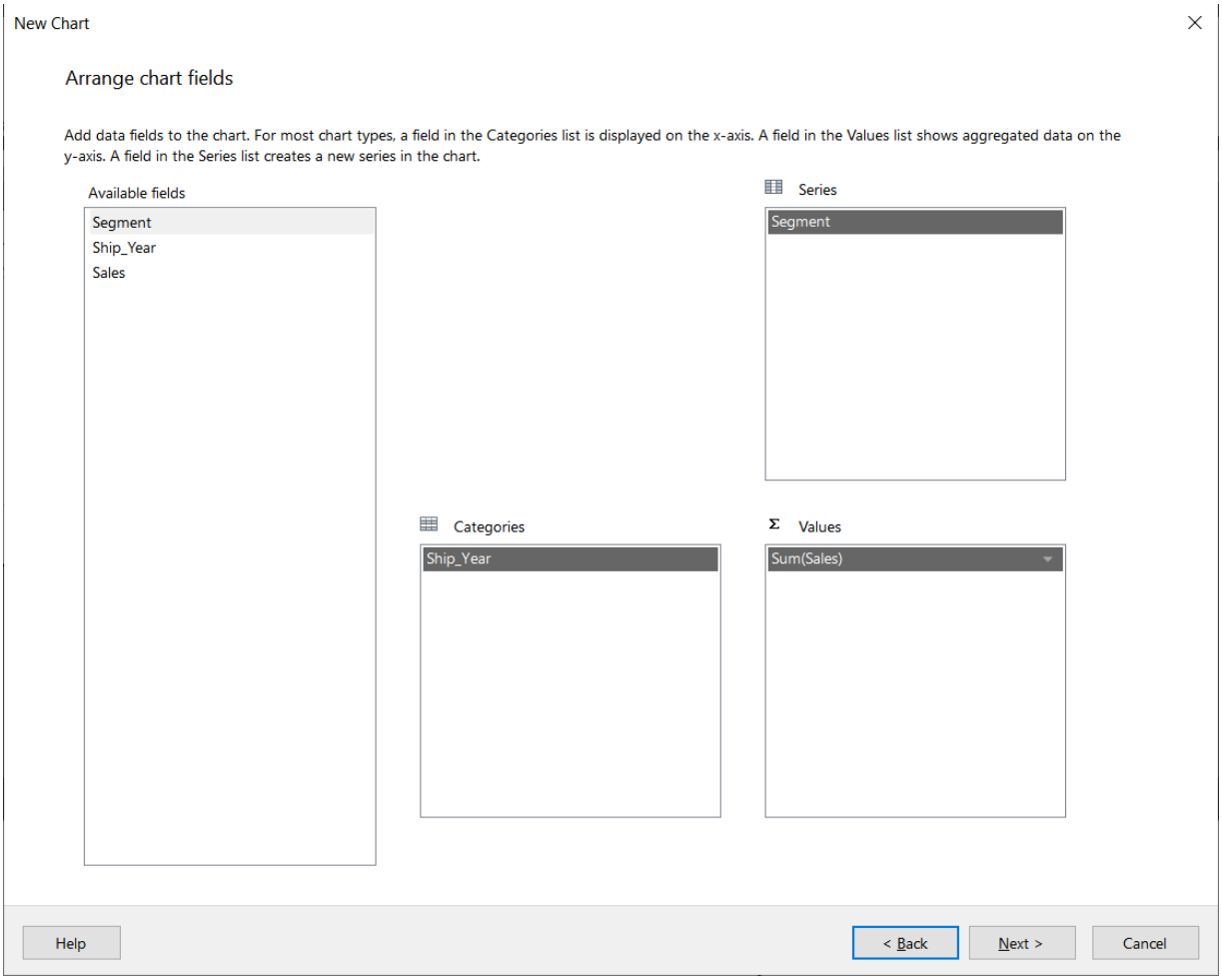
Hình 4.60 Kéo thả các thuộc tính để thực hiện câu lệnh MDX lấy dữ liệu

Bước 6: Tại bước chọn loại biểu đồ biểu diễn dữ liệu, vì ở đây dữ liệu được gộp theo các loại khách hàng khác nhau(**Consumer, Corporate, Home Office**) diễn ra trong khoảng thời gian từ năm **2014 đến năm 2017** vì vậy nhóm sẽ chọn biểu đồ dạng cột(**Column**) → Nhấn **Next** để chuyển sang bước tiếp theo.



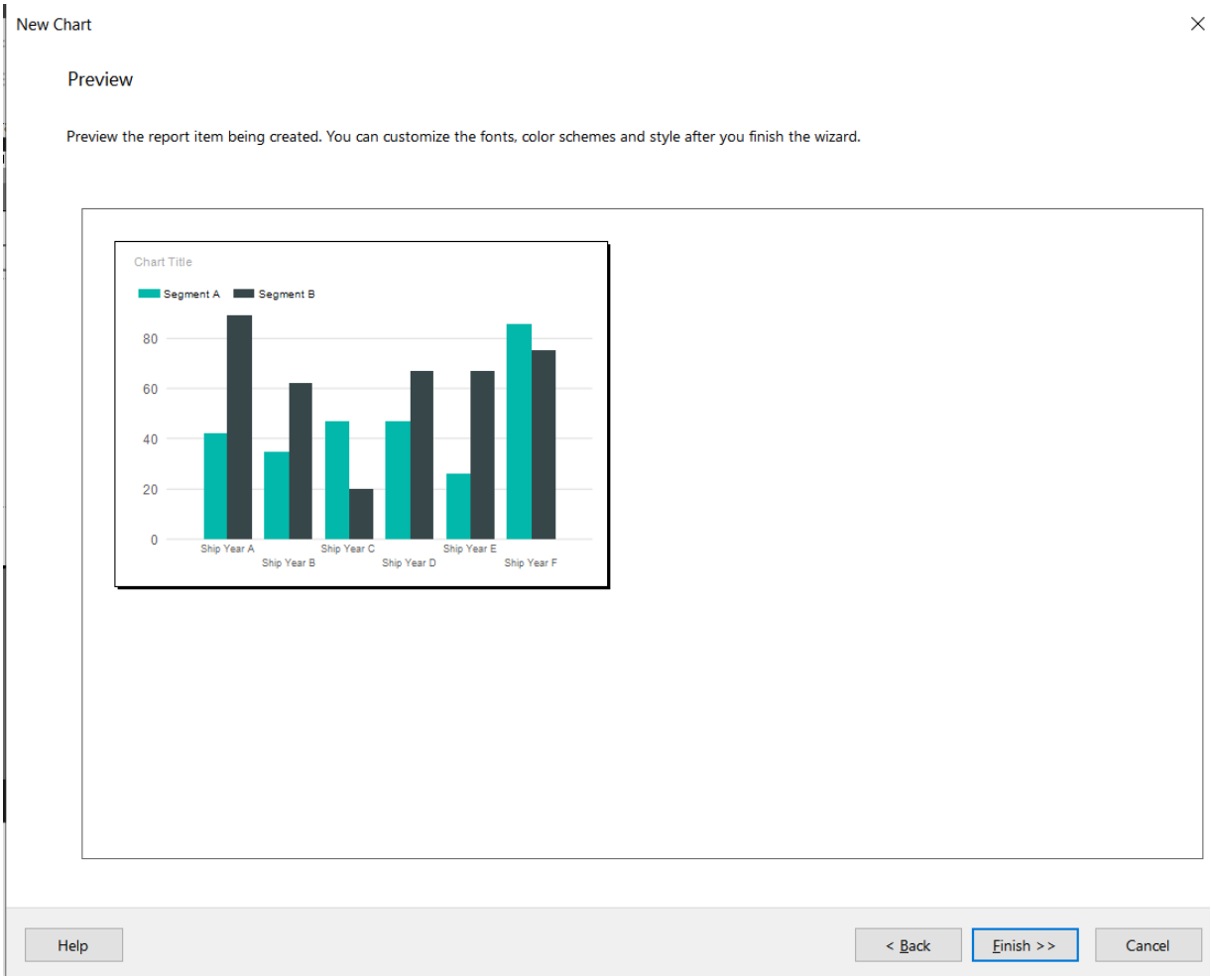
Hình 4.61 Chọn loại biểu đồ thể hiện báo biểu

Bước 7: Ở bước **Arrange chart Fields** thì ta lần lượt kéo thả các thuộc tính vào các mục như sau: **Segment** vào **Categories**, **Sales** vào **Values**, **Ship_Year** vào **Series**. Khi đã hoàn thành bước vừa nêu thì nhấn **Next** để đi đến bước thiết lập cuối cùng.



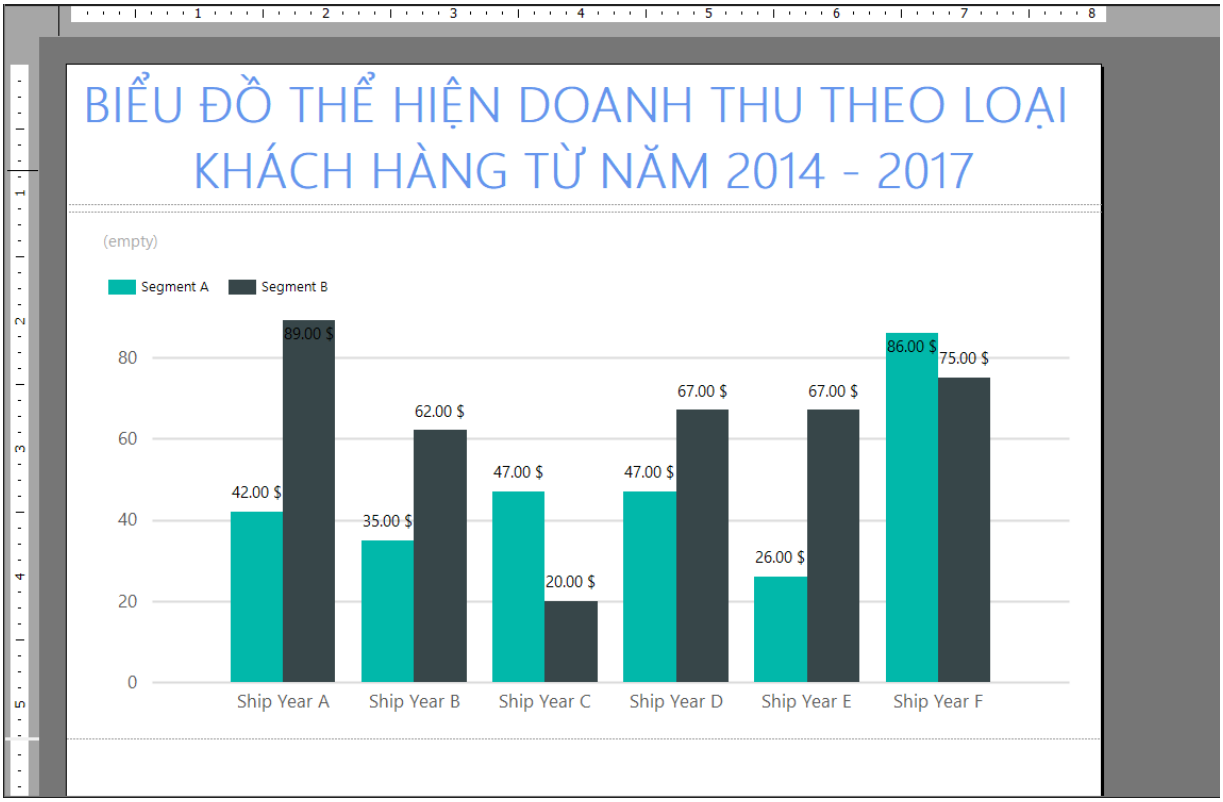
Hình 4.62 Sắp xếp các thuộc tính thành các cụm dữ liệu

Bước 8: Đến với bước khởi tạo báo biểu cuối cùng đó chính là bước xem trước biểu đồ. Ở đây, nếu như biểu đồ đã thể hiện đúng và đầy đủ những gì mà ta yêu cầu thì ta nhấn **Finish** >> để hoàn thành khởi tạo báo biểu dạng biểu đồ.



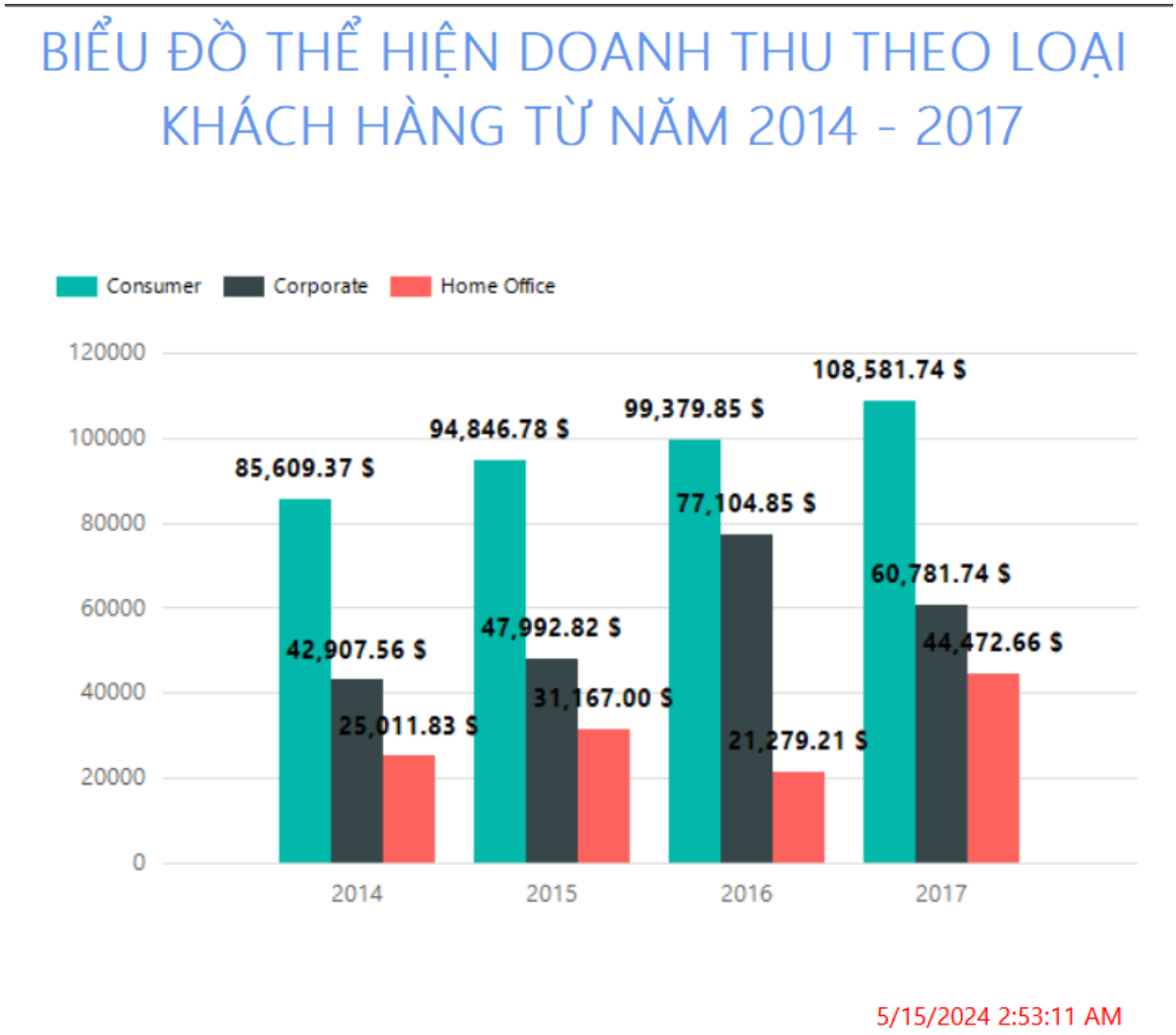
Hình 4.63 Xem trước báo biểu

Bước 9: Tiến hành định dạng lại báo biểu số 3.



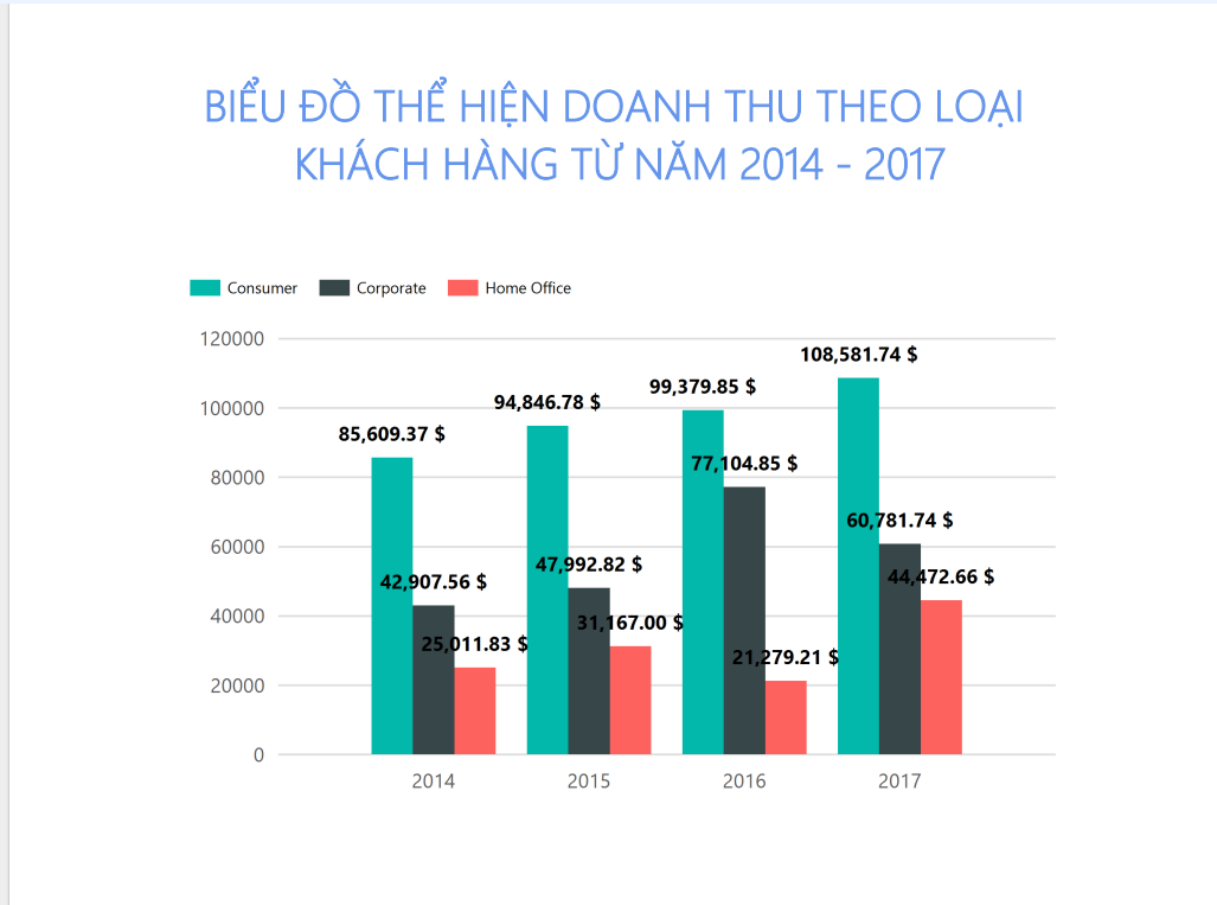
Hình 4.64 Định dạng lại báo biểu

Bước 10: Ta chọn chức năng **Run** để xem thử kết quả của báo biểu thể hiện đúng theo như yêu cầu của báo biểu số 3.



Hình 4.65 Kết quả tạo báo biểu

Bước 11: Xuất báo biểu ra file Pdf hoặc tiến hành đăng lên **Report Server**.



Hình 4.66 Xuất báo biểu dưới dạng PDF

2. QUÁ TRÌNH SSRS VỚI POWER BI

2.1. Cài đặt các công cụ cần thiết

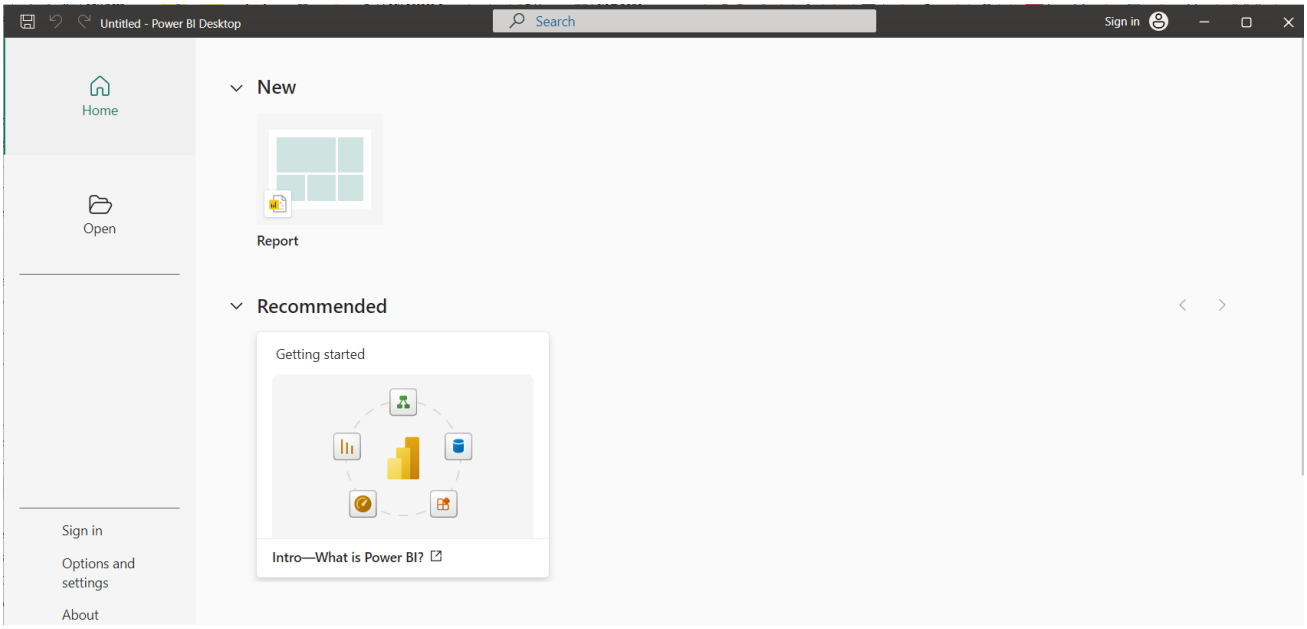
2.1.1. Cài đặt Power BI

- Truy cập vào đường link https://www.microsoft.com/en-us/power-platform/products/power-bi/ và dùng thử bằng một tài khoản Microsoft.
- Sau khi đã có tài khoản Power BI, tiến hành cài đặt Power BI về máy, truy cập https://www.microsoft.com/en-us/download/details.aspx?id=58494
- Cài đặt bình thường như các ứng dụng khác, sau khi cài đặt xong, đăng nhập vào Power BI bằng tài khoản vừa đăng ký dùng thử.

2.2. Quá trình tạo lập báo biểu bằng công cụ Power BI

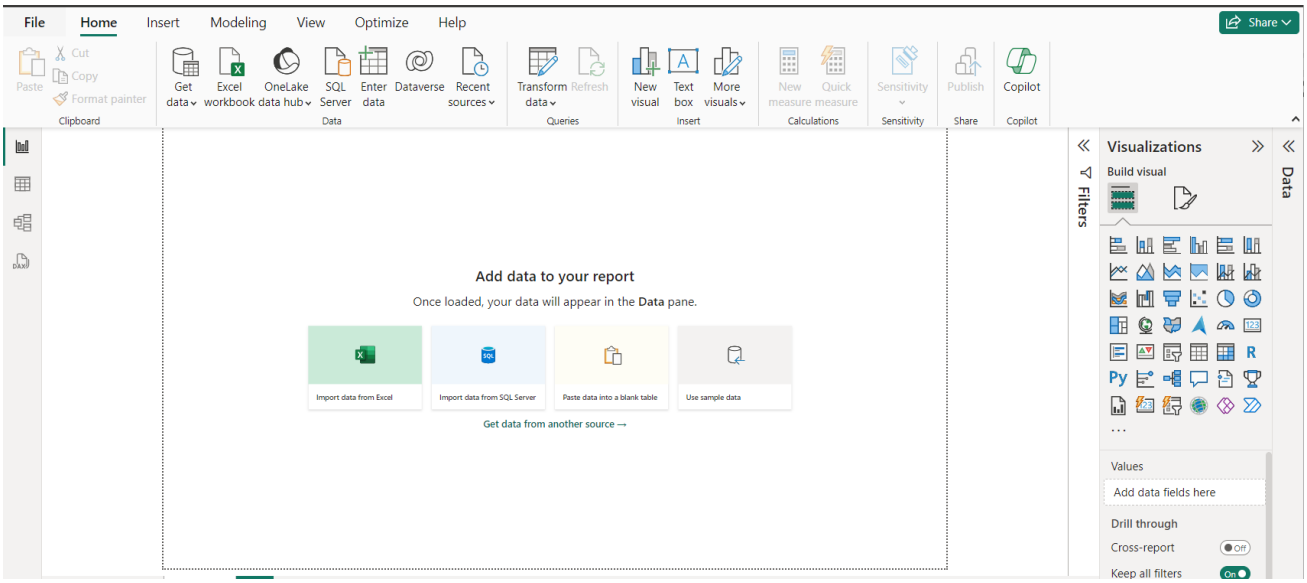
2.2.1. Tạo project SSRS trên Power BI

- Mở **Power BI**, đây là cửa sổ giới thiệu của phần mềm.



Hình 4.67 Giao diện mở đầu của Power BI

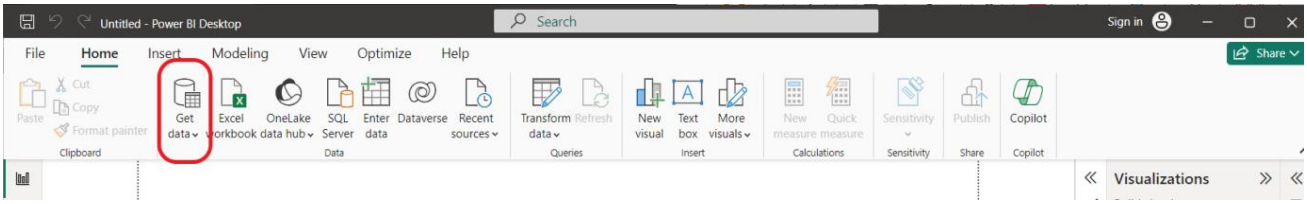
- Tiến hành tạo mới một **Project** bằng cách nhấn vào **Report** ở màn hình trên, giao diện hiện ra như hình dưới.



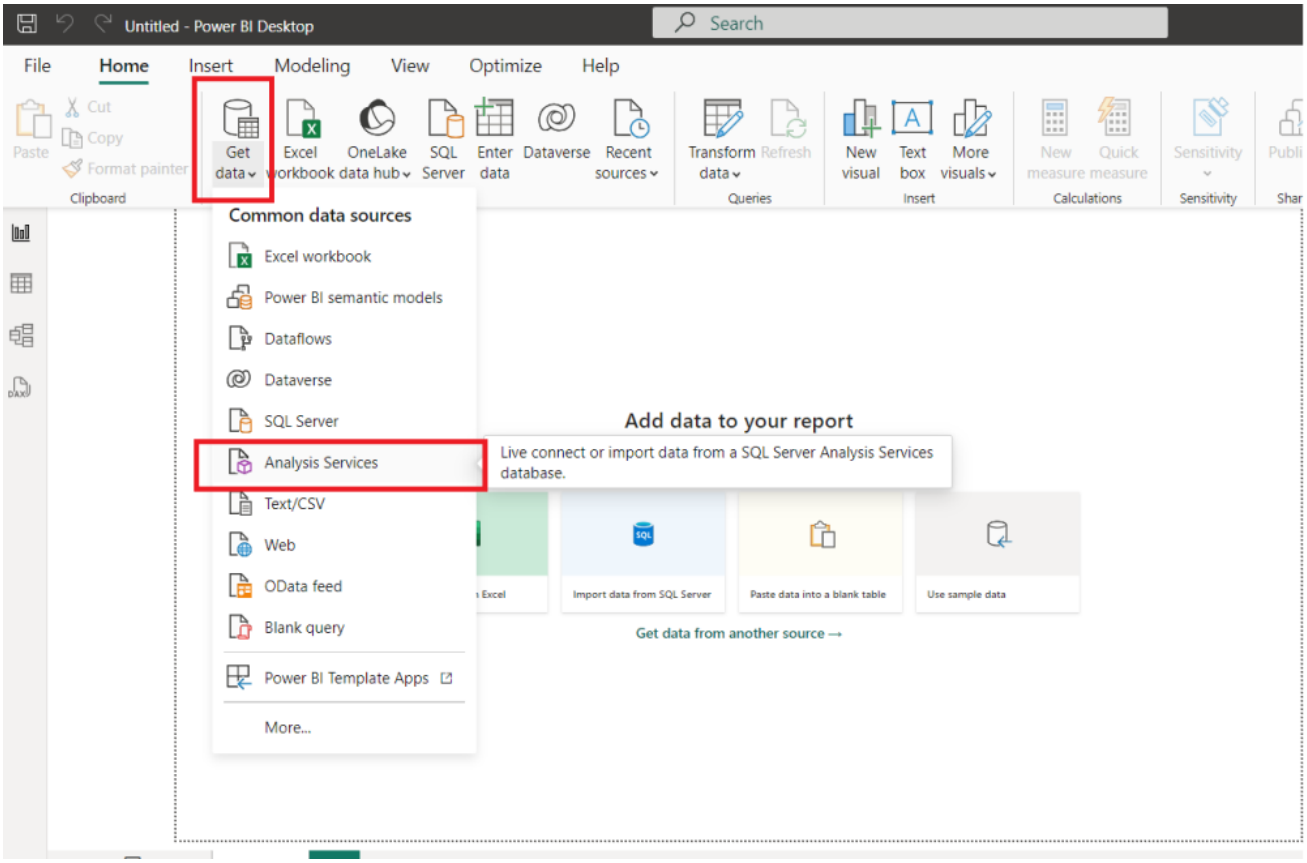
Hình 4.68 Tạo mới một Project

2.2.2. Tạo kết nối với kho dữ liệu

Bước 1: Nhấn vào nút **Get data** ở tab **Home**.



Bước 2: Nhấn vào nút **Analysis Service** để kết nối dữ liệu từ kho vào trong Power BI cho bước báo biểu.

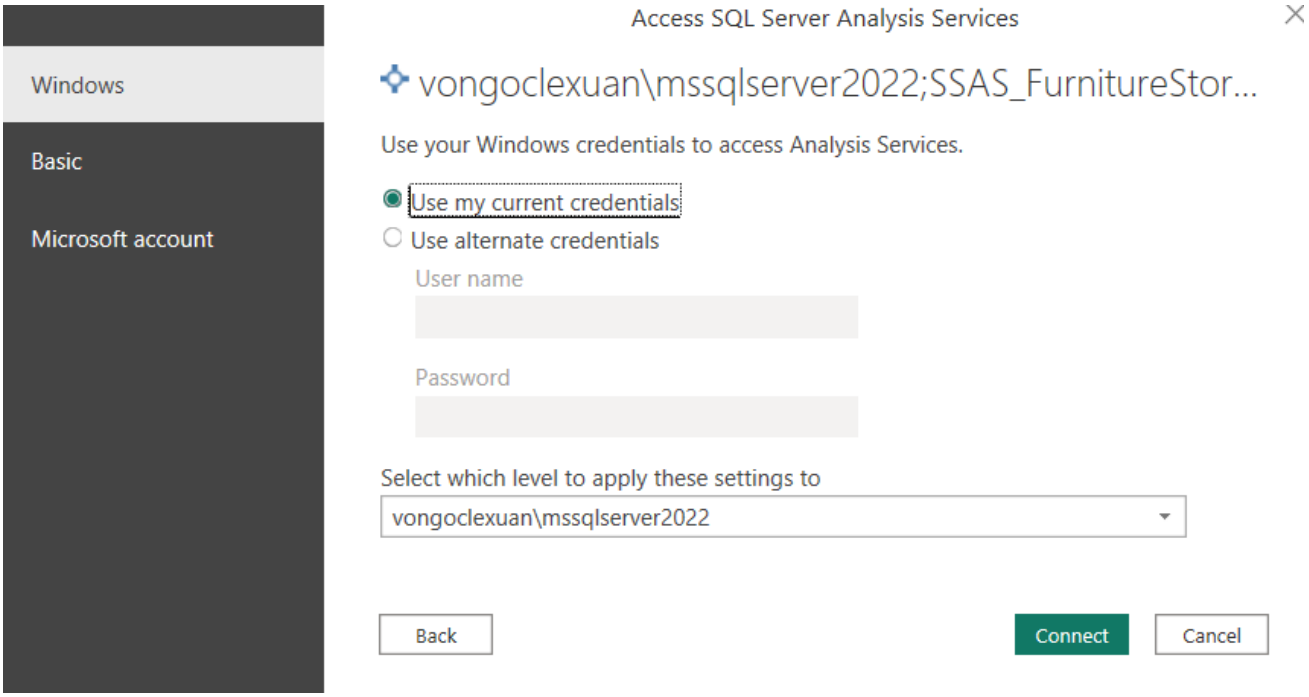


Hình 4.69 Chọn chức năng tạo kết nối tới kho dữ liệu thông qua Analysis Services

Bước 3: Nhập thông tin **Server** và **Database** vào khung, chọn **Connect Live** và nhấn **OK.**

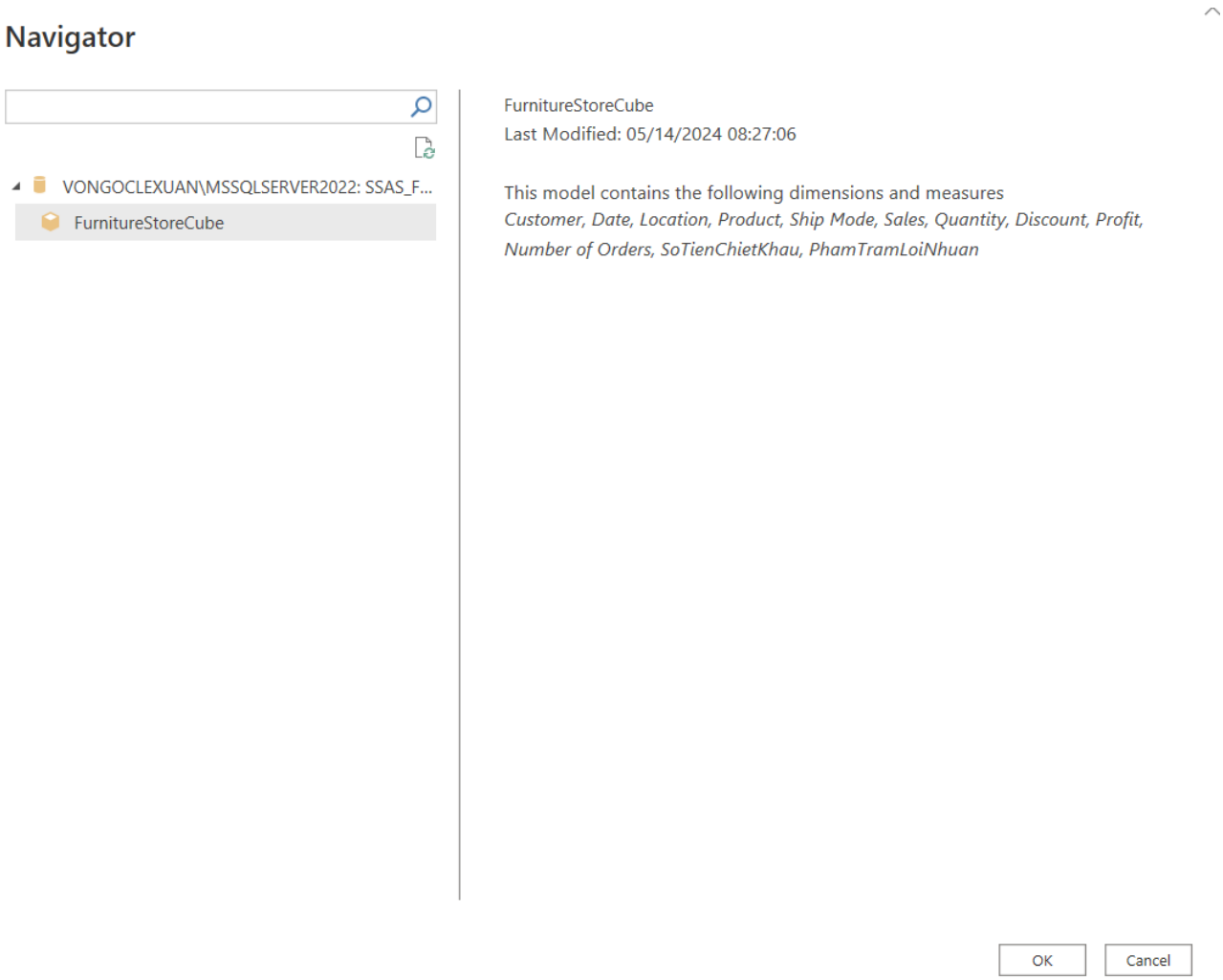
Hình 4.70 Nhập thông tin kết nối

Bước 4: Tiếp đó, chọn **Use my current credentials** để đăng nhập vào **SQL Analysis Service**



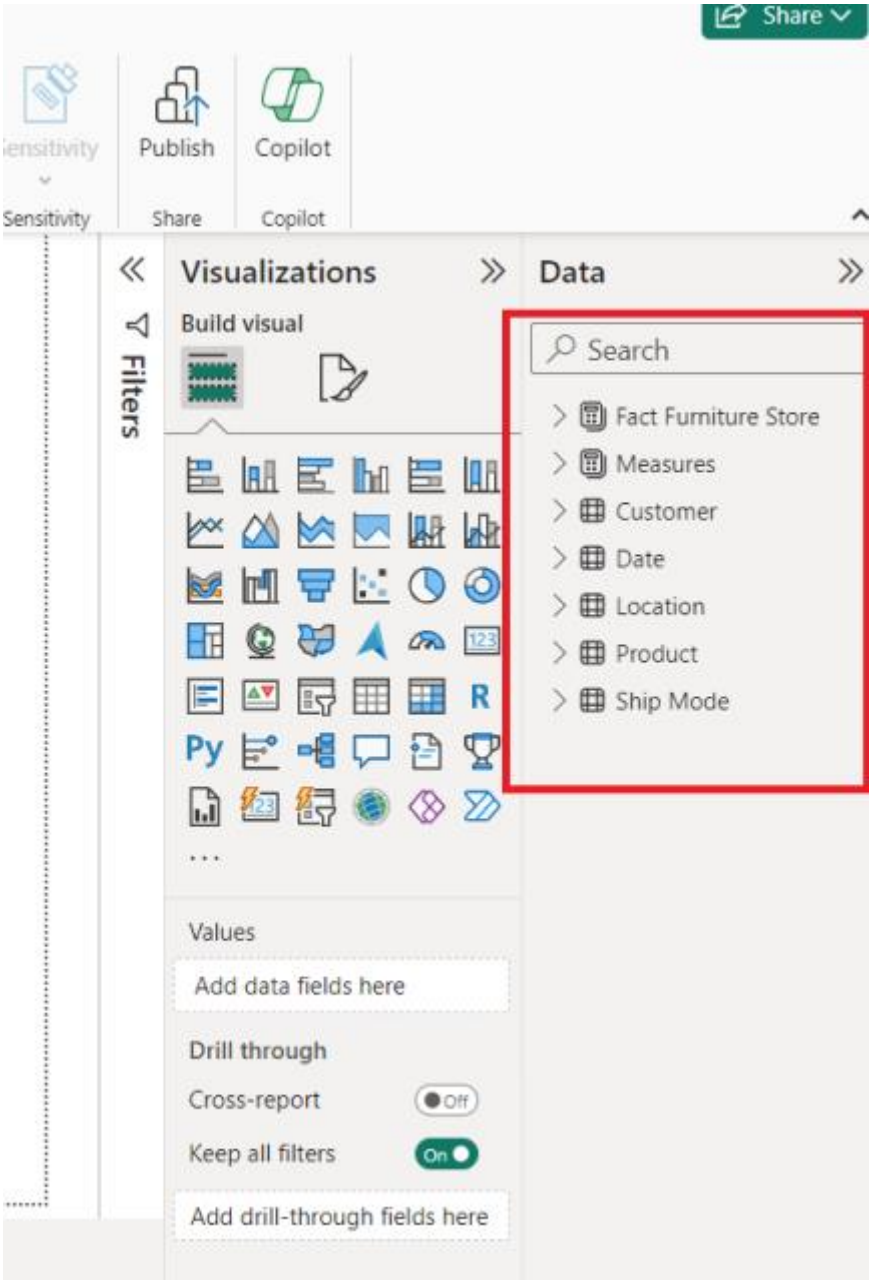
Hình 4.71 Chọn loại credentials dùng để đăng nhập vào SQL Analysis Service

Bước 5: Ở màn hình này, chúng ta xem lại thông tin kho dữ liệu và kiểm tra trước khi import vào Power BI.



Hình 4.72 Chọn Kho dữ liệu cần đưa vào dự án

Bước 6: Sau khi xác nhận thành công, thông tin về các bảng **Fact** và **Dimension** của kho dữ liệu xuất hiện ở góc phải màn hình như sau:

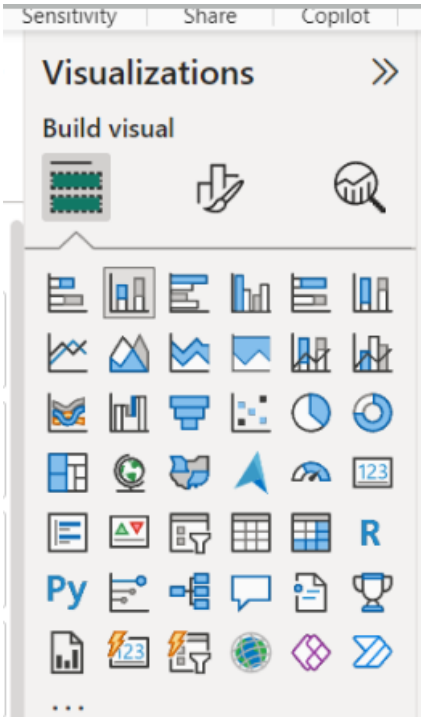


Hình 4.73 Thông tin về kho dữ liệu thể hiện ở cột Data

2.3. Thực hiện lập báo biểu trên Power BI

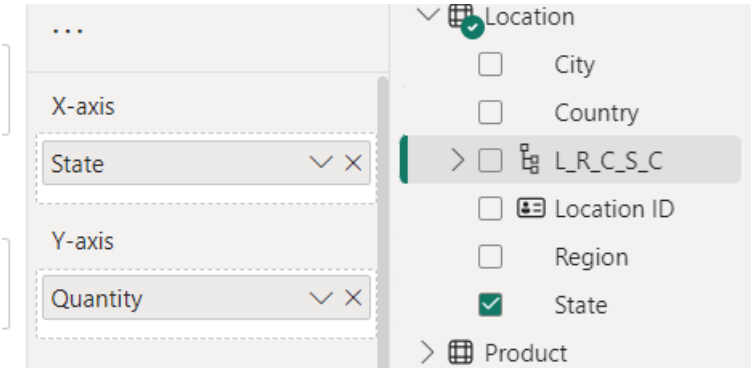
2.3.1. Báo biểu 4: Thống kê số sản phẩm được bán ra theo từng Bang (có thể lọc theo từng loại sản phẩm)

Bước 1: Chọn biểu đồ loại "**Stacked Column Chart**" cho báo biểu này.



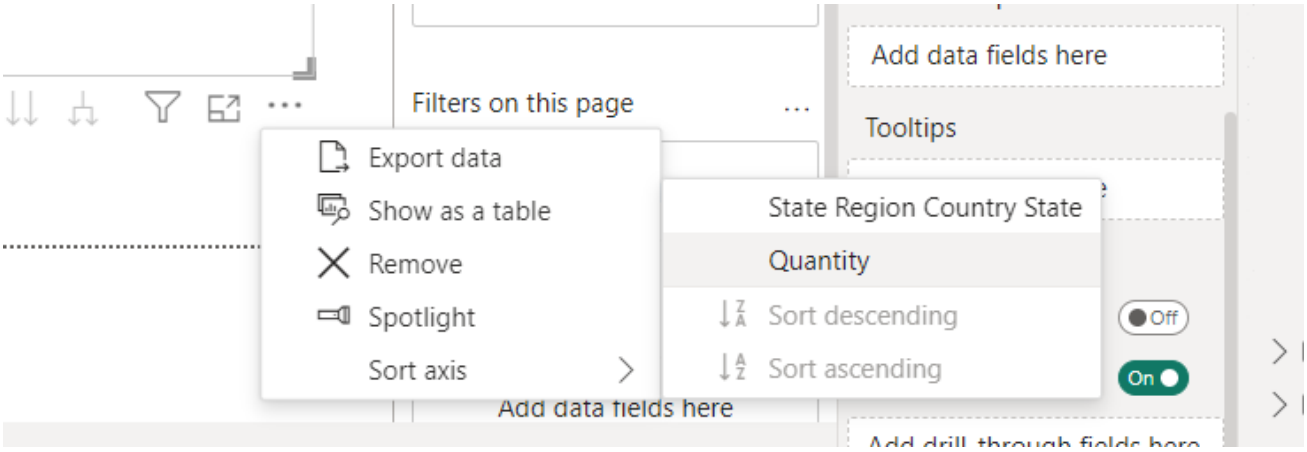
Hình 4.74 Chọn biểu đồ Stacked Column Chart

Bước 2: Ở **X-axis**: chọn thuộc tính **State** của **Dim Location** và **Y-axis**: chọn độ đo **Quantity**.



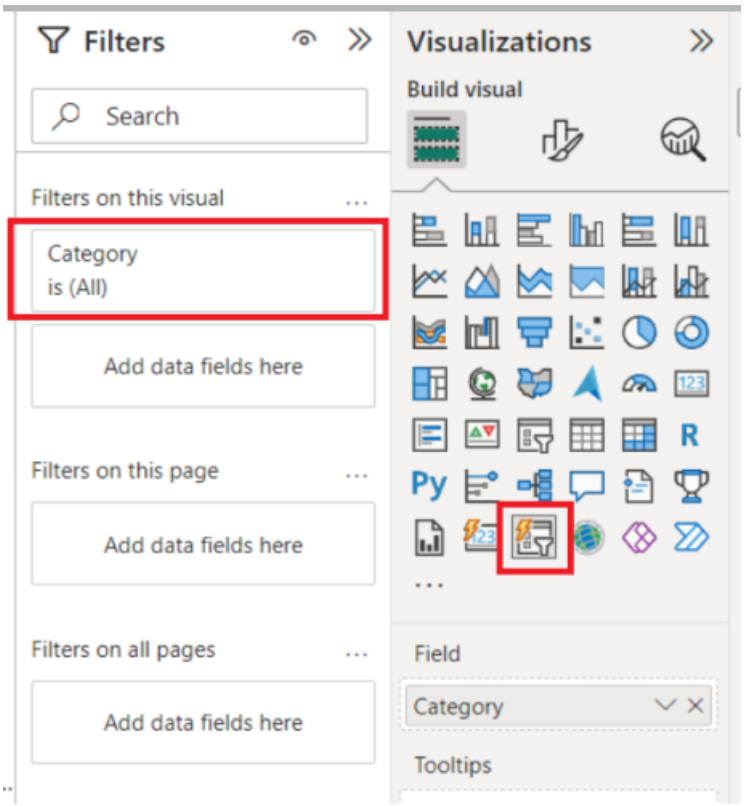
Hình 4.75 Cài đặt các thông số của biểu đồ

Bước 3: Chuột phải vào biểu đồ và chọn **Sort axis** là **Quantity** để biểu đồ hiển thị có sắp xếp theo giá trị.



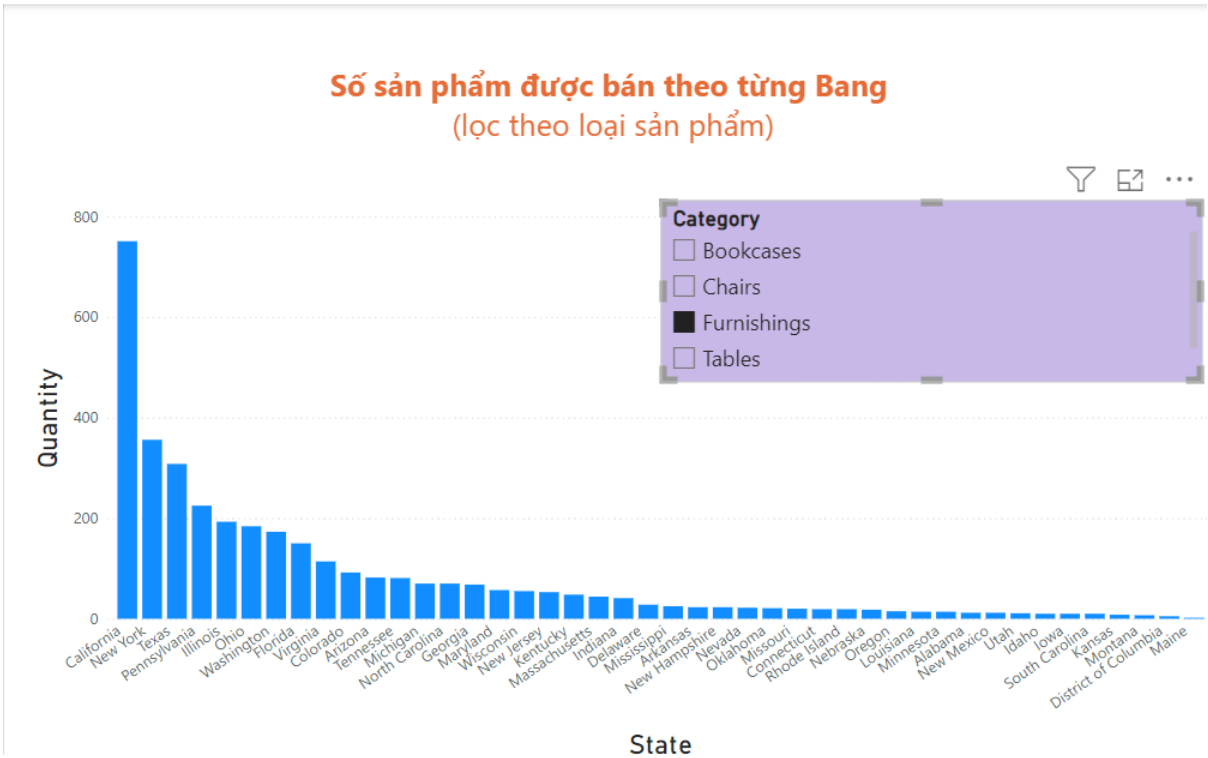
Hình 4.76 Chọn kiểu thể hiện của biểu đồ

Bước 4: Thêm **Slicer** bằng cách chọn **Slicer** và thêm thuộc tính filter là **Category**



Hình 4.77 Thêm slicer

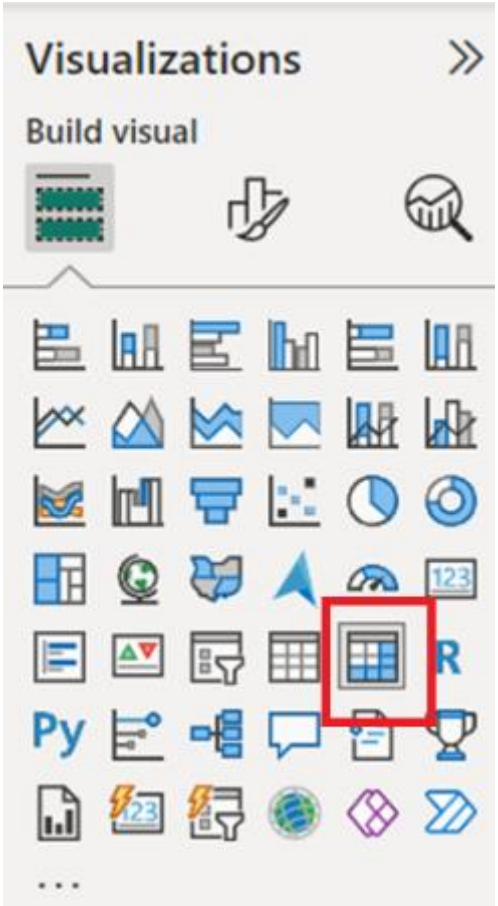
Bước 5: Đặt tên cho biểu đồ và thực hiện các bước format (màu biểu đồ, màu chữ, cỡ chữ,…) để được biểu đồ như ảnh .



Hình 4.78 Kết quả tạo báo biểu số 4

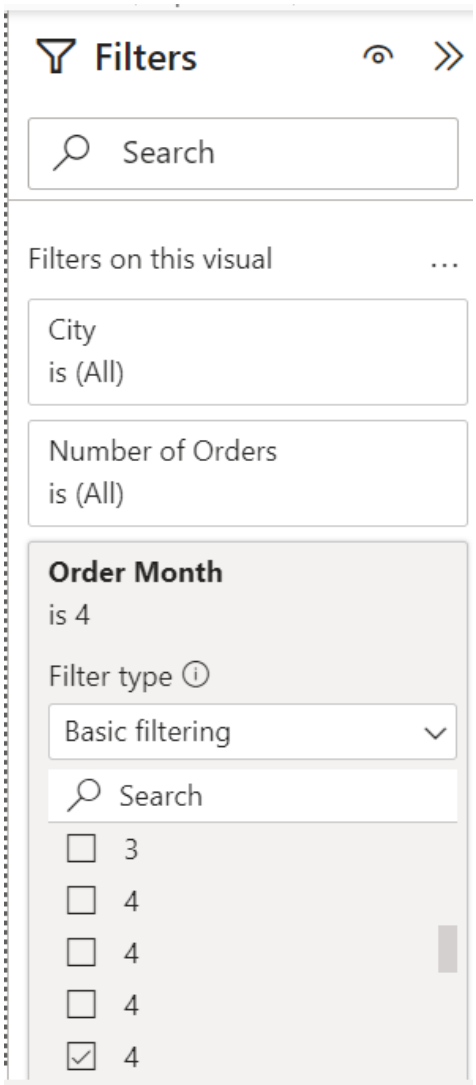
2.3.2. Báo biểu 5: Thống kê số đơn hang, số sản phẩm đặt mua theo từng sản phẩm, giá trị đơn hang và địa chỉ giao hàng (thành phố)

Bước 1: Chọn biểu đồ loại "**Matrix**" cho báo biểu



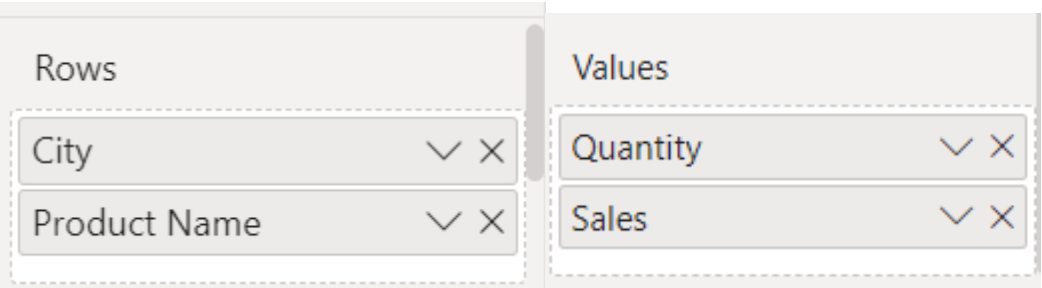
Hình 4.79 Chọn báo biểu dạng Matrix

Bước 2: Ở mục **Filter**, tích vào mục tháng **4/2017** để áp dụng bộ lọc.



Hình 4.80 Sử dụng bộ lọc

Bước 3: Ở mục **Rows:** chọn thuộc tính **City** của **Dim Location** và thuộc tính **Product Name** của **Dim Product** làm cột ngang và ở mục **Values**: chọn độ đo **Sales** và **Quantity**



Hình 4.81 Cài đặt các thông số cho biểu đồ

Bước 4: Đặt tên cho biểu đồ và thực hiện các bước format (màu biểu đồ, màu chữ, cỡ chữ,…) để được biểu đồ như ảnh

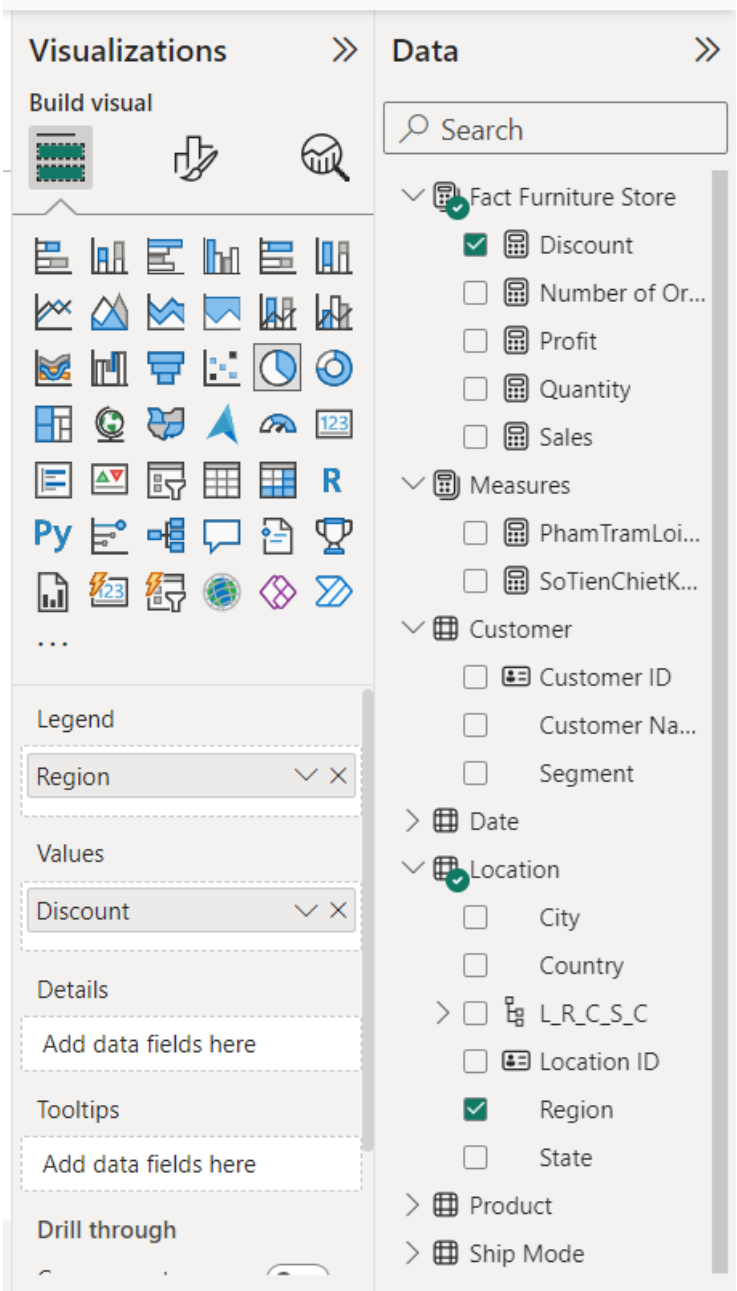
Thống kê số đơn hàng, số sản phẩm đặt mua theo từng sản phẩm, giá trị của các đơn hàng mà các thành phố đã đặt (T4-2017)

City	Number of Orders	Quantity	Sales
York	**1**	**3**	**422.06**
Global Leather Highback Executive Chair with Pneumatic Height Adjustment, Black	1	3	422.06
Yonkers	**2**	**9**	**851.48**
Tensor Computer Mounted Lamp	1	1	14.89
Chromcraft Round Conference Tables	1	8	836.59
Woonsocket	**1**	**2**	**52.96**
DAX Black Cherry Wood-Tone Poster Frame	1	2	52.96
Woodstock	**1**	**8**	**845.49**
Global Manager's Adjustable Task Chair, Storm	1	8	845.49
Wilson	**1**	**4**	**163.14**
Dana Fluorescent Magnifying Lamp, White, 36"	1	4	163.14
Wilmington	**6**	**29**	**3,135.94**
Sauder Inglewood Library Bookcases	1	6	1,025.88
Office Star - Contemporary Swivel Chair with Padded Adjustable Arms and Flex Back	1	9	1,268.82
Lesro Sheffield Collection Coffee Table, End Table, Center Table, Corner Table	1	4	199.84
Hon Metal Bookcases, Putty	1	4	283.92
Total	**2121**	**8028**	**741,999.80**

Hình 4.82 Kết quả tạo báo biểu 5

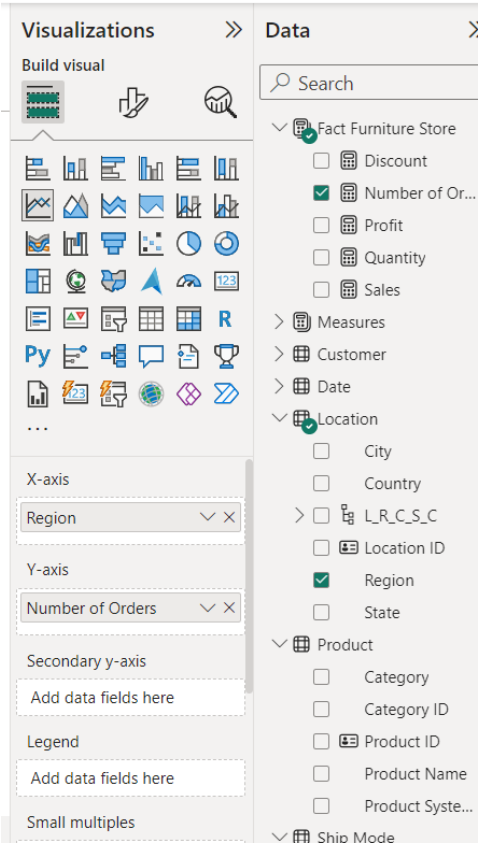
2.3.3. Báo biểu 6: Thống kê và so sánh số tiền chiết khấu và đơn hàng mua của các khu vực khảo sát

Bước 1: Chọn biểu đồ **Pie Chart** để thực hiện báo biểu này, sau đó ở mục **Legend** chọn thuộc tính **Region** ở **Dim Location** và mục **Values** chọn độ đo **Discount** để thống kê số tiền chiết khấu mà cửa hàng đã dành cho từng khu vực.



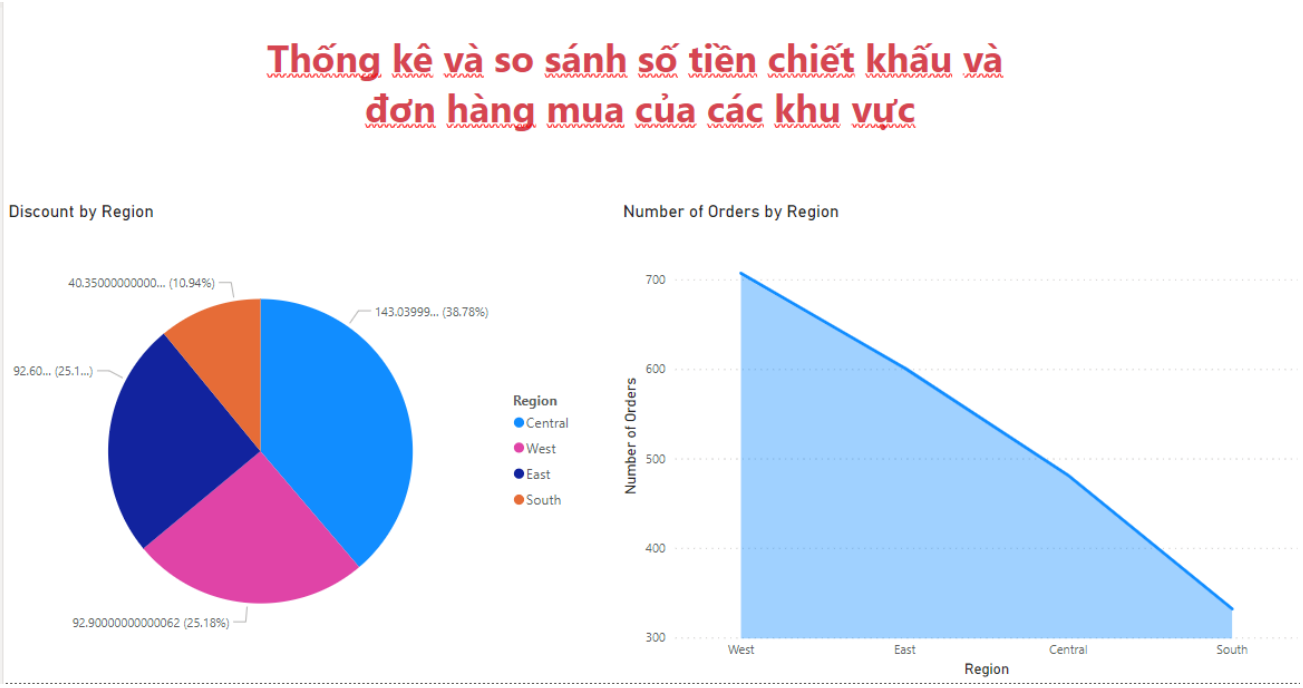
Hình 4.83 Chọn biểu đồ dạng Pie Chart

Bước 2: Chọn biểu đồ **Line chart** để thực hiện báo biểu này, sau đó mục **X-axis**: chọn thuộc tính **Region** ở **Dim Location** và mục **Y-axis**: chọn độ đo **Discount** để xem số đơn hàng của từng khu vực đặt được và so sánh để xác định ảnh hưởng của việc chiết khấu đến đặt hàng.



Hình 4.84 Chọn biểu đồ dạng Line Chart

Bước 3: Đặt tên cho biểu đồ và thực hiện các bước format (màu biểu đồ, màu chữ, cỡ chữ,…) để được biểu đồ như ảnh.



Hình 4.85 Kết quả tạo báo biểu số 6

CHƯƠNG 5. DATA MINING

1. Bộ dữ liệu dùng cho quá trình Data mining

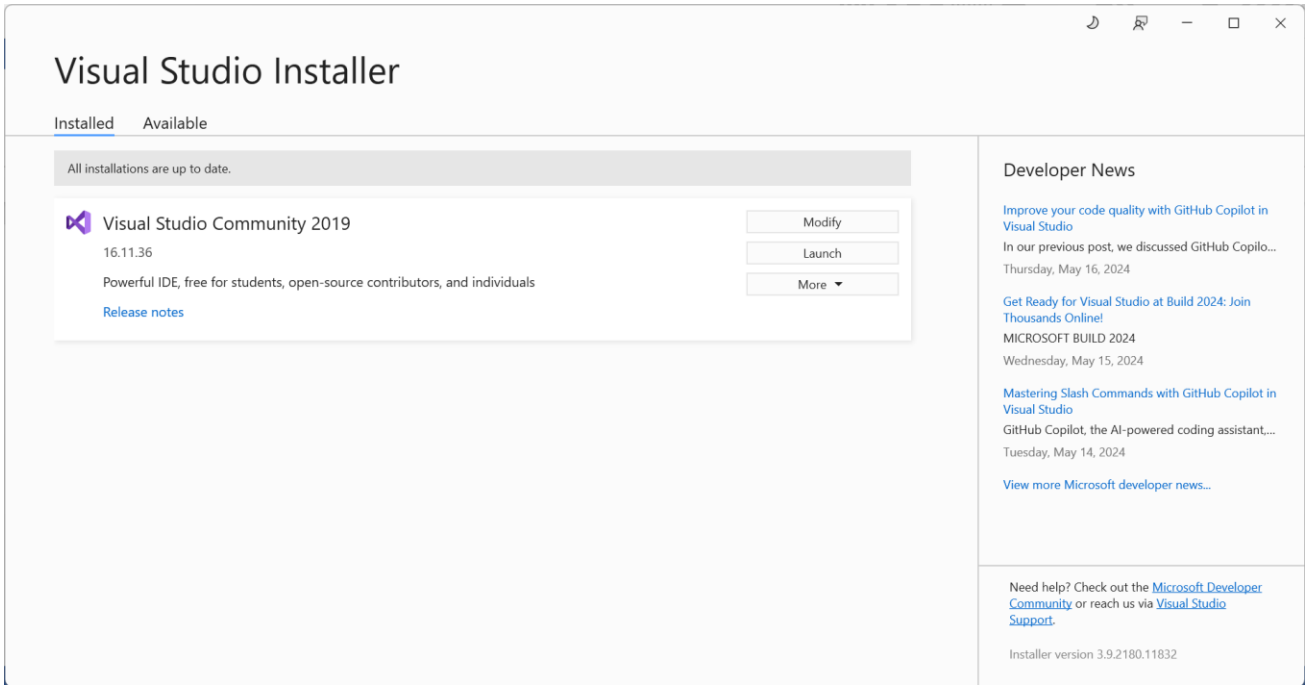
- Tên dataset: Indicators of Heart Disease.
- Đường dẫn đến bộ dataset: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
- Tệp dữ liệu sử dụng để phục vụ cho quá trình Khai thác dữ liệu có tên: *heart_2022_no_nans.csv*. Đậy là tệp dữ liệu đã được tiền xử lý các dữ liệu bị thiếu dữ liệu bởi tác giả của bộ dữ liệu, kích thước tệp là 89.98 MB với 40 cột dữ liệu và 246022 dòng dữ liệu.

2. Thiết lập các công cụ cần thiết để thực hiện quá trình Data Mining

- Các công cụ cần cài đặt để thực hiện quá trình Data mining:
    - Visual Studio 2019(hoặc 2017): Đây là một IDE phổ biến được sử dụng trong rất nhiều lĩnh vực khác nhau: Lập trình Web, lập trình ứng dụng Windows, Machine Learning… Lý do mà chúng ta cần cài bản 2019 trở về trước mà không phải bản mới nhất là vì chức năng Data Mining của Analysis Services đã ngừng hỗ trợ kể từ sau phiên bản Visual Studio 2019.
    - SSDT 2019: Là công cụ giúp tạo project SSAS trong Visual Studio 2019.

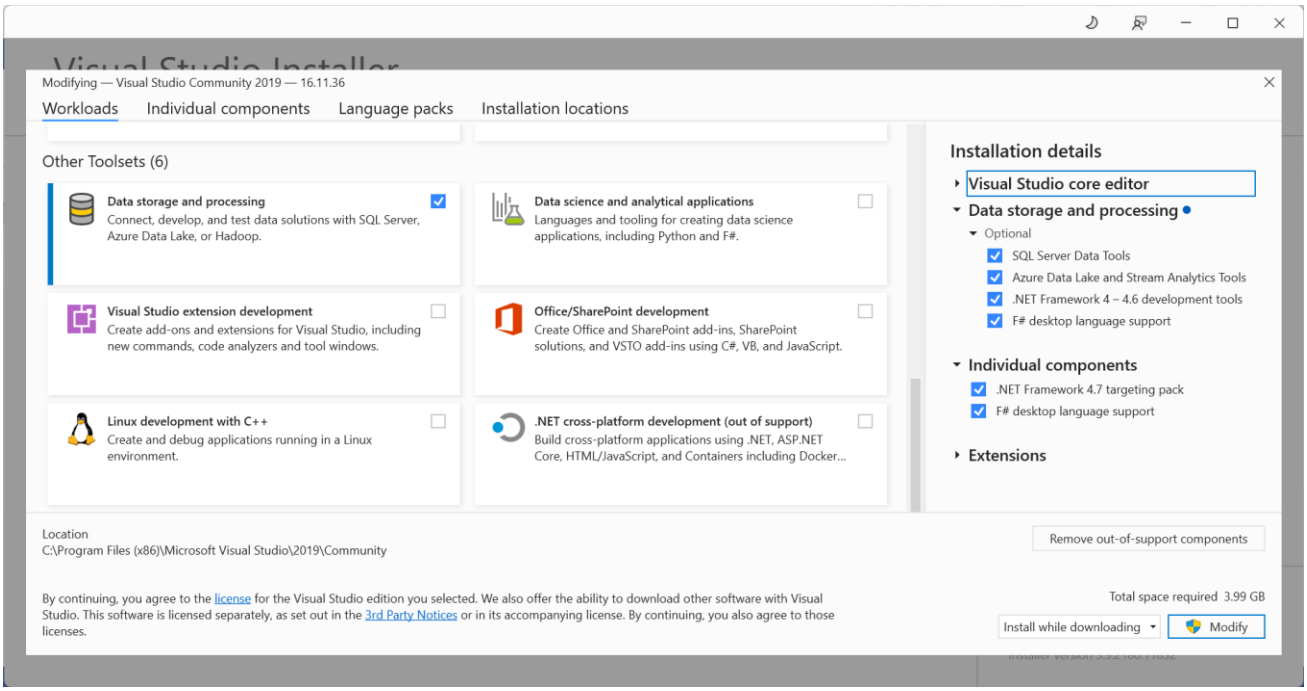
2.1. Cài đặt Visual Studio 2019

- Để cài đặt được Visual Studio 2019 thì tương tự với cách cài đặt các phiên bản khác, chúng ta sẽ truy cập vào đúng trang chủ của phần mềm và thực hiện việc cài đặt theo các hướng dẫn của nhà sản xuất.



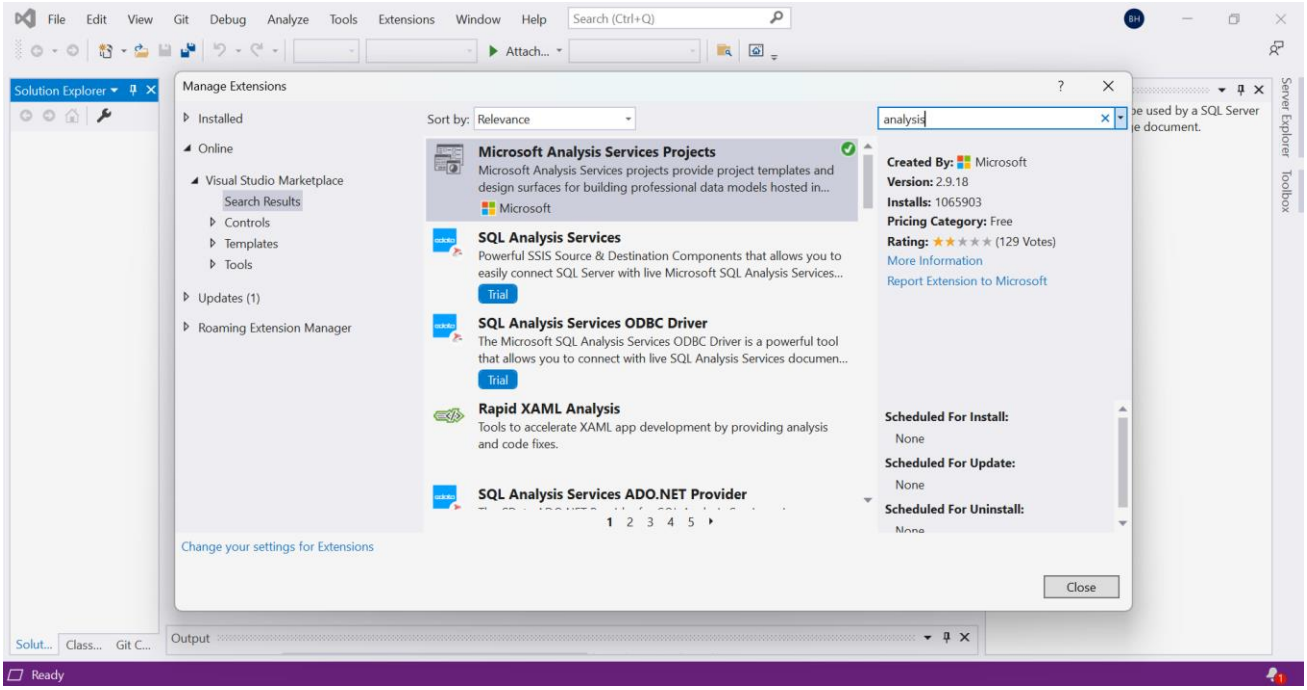
Hình 5.1 Kết quả cài đặt Visual Studio 2019

2.2. Cài đặt SSDT 2019



Bước 1: Để cài đặt được SSDT 2019, đầu tiên chúng ta mở **Visual Studio 2019 → Continue without code → Extensions → Manage Extensions →** Nhập từ khoá **Analysis Services** vào ô tìm kiếm → Chọn extension giống như hình và nhấn **Download** để tiến tới

trang tải extension.



Bước 2: Sau khi tải file cài đặt của extension về máy thì ta tiến nhấn đúp chuột vào file để mở hộp thoại cài đặt.

3. Thực hiện quá trình Data Mining

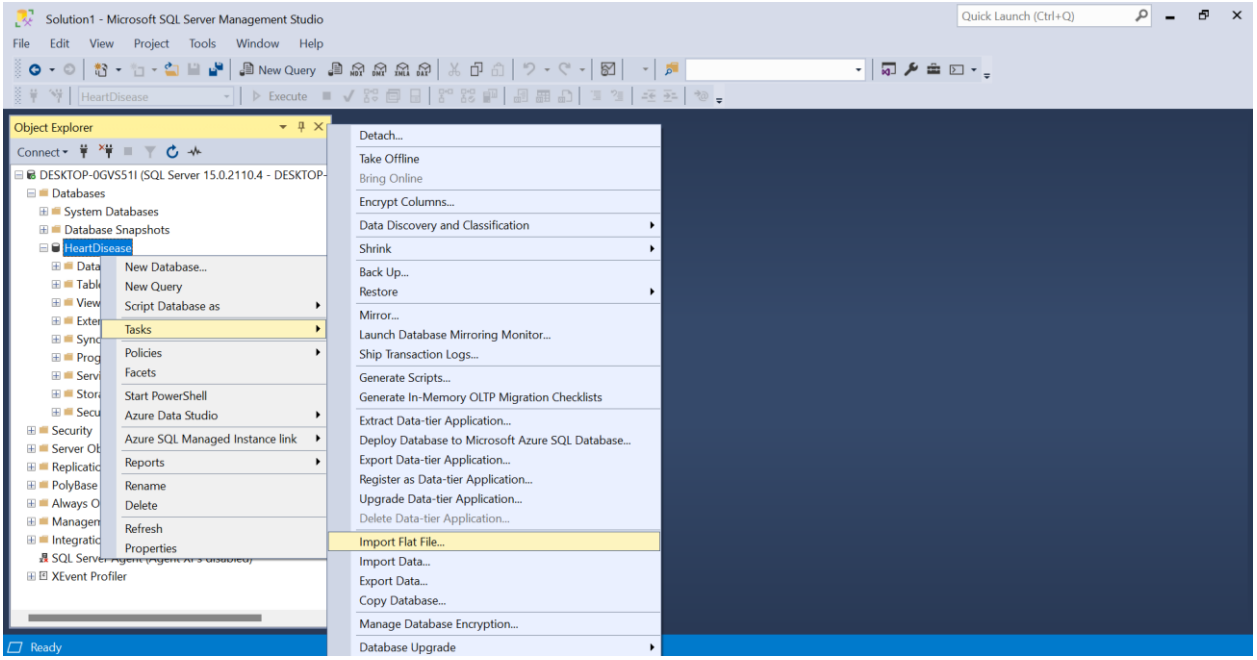
3.1. Quá trình đổ dữ liệu vào cơ sở dữ liệu

Bước 1: Tạo cơ sở dữ liệu để chứa bảng và sử dụng cơ sở dữ liệu đó để tiến hành đổ dữ liệu từ file csv

```
-- create database
create database HeartDisease

-- use database
use HeartDisease
```

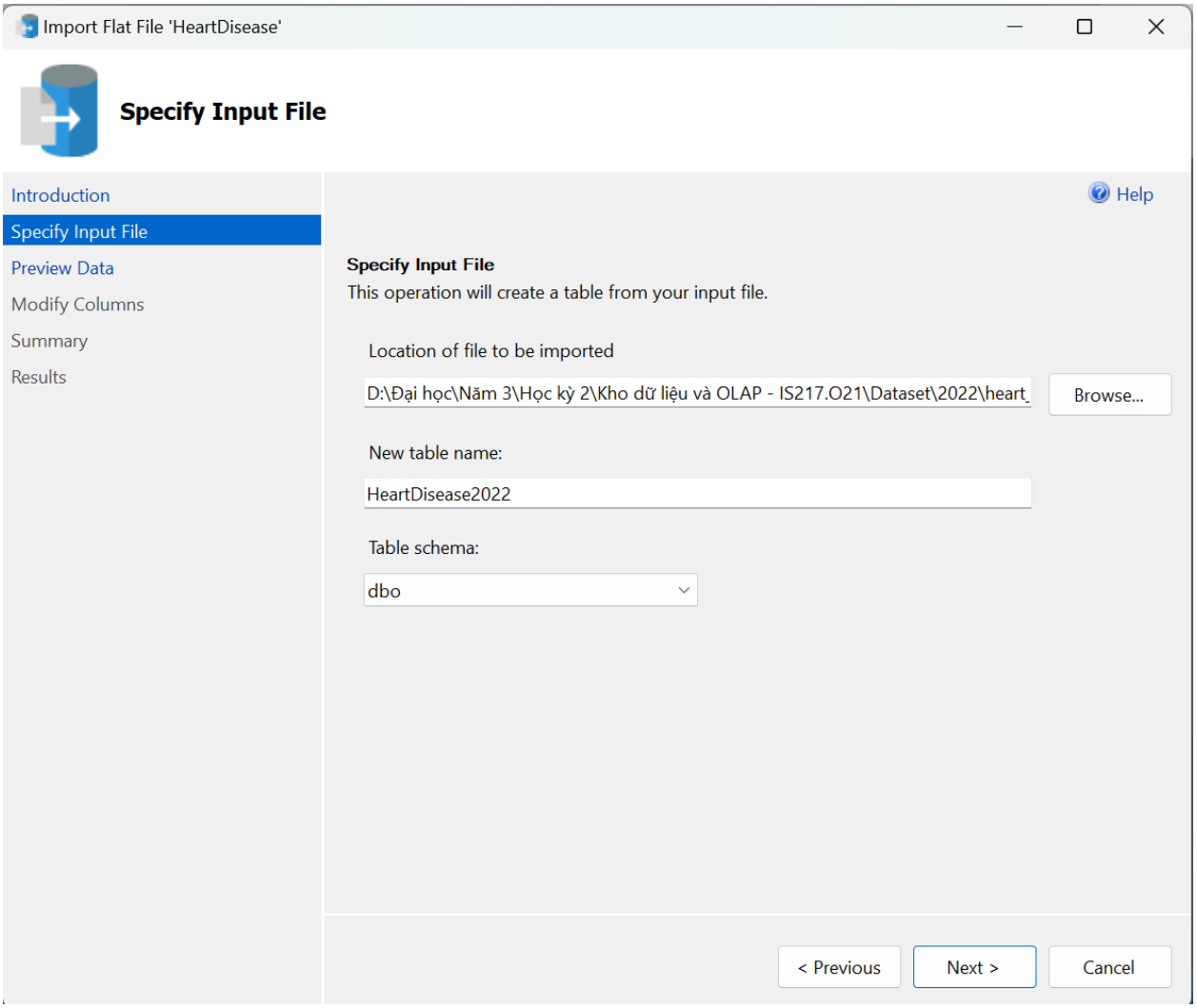
Hình 5.2 Tạo Database để chứa dữ liệu cần thiết cho quá trình Data Mining

Bước 2: Chuột phải vào **HeartDisease → Tasks → Import Flat File** để tiến hành đổ dữ liệu từ một file phẳng(Là một file mà dữ liệu được lưu trữ dưới dạng một bảng 2 chiều mà trong đó mỗi dòng là một bản ghi độc lập của bảng, không có bất kỳ mối quan hệ nào tới dữ liệu khác).



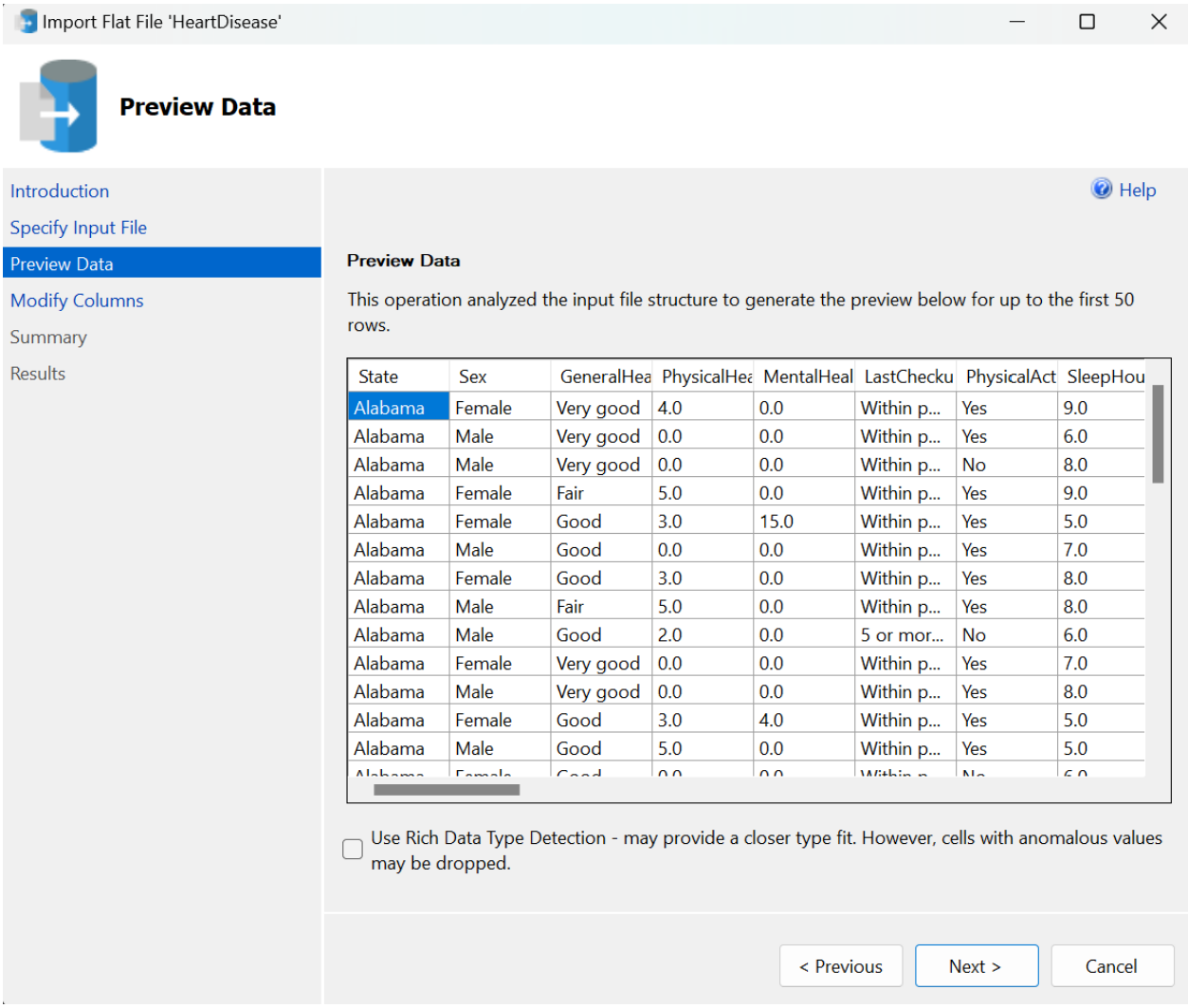
Hình 5.3 Chọn chức năng để import dữ liệu từ một file .csv

Bước 3: Chọn đường dẫn tới file csv chứa bộ dữ liệu, đặt tên cho bảng và chọn schema cho bảng(có mặc định).



Hình 5.4 Chọn file .csv làm đầu vào đổ dữ liệu

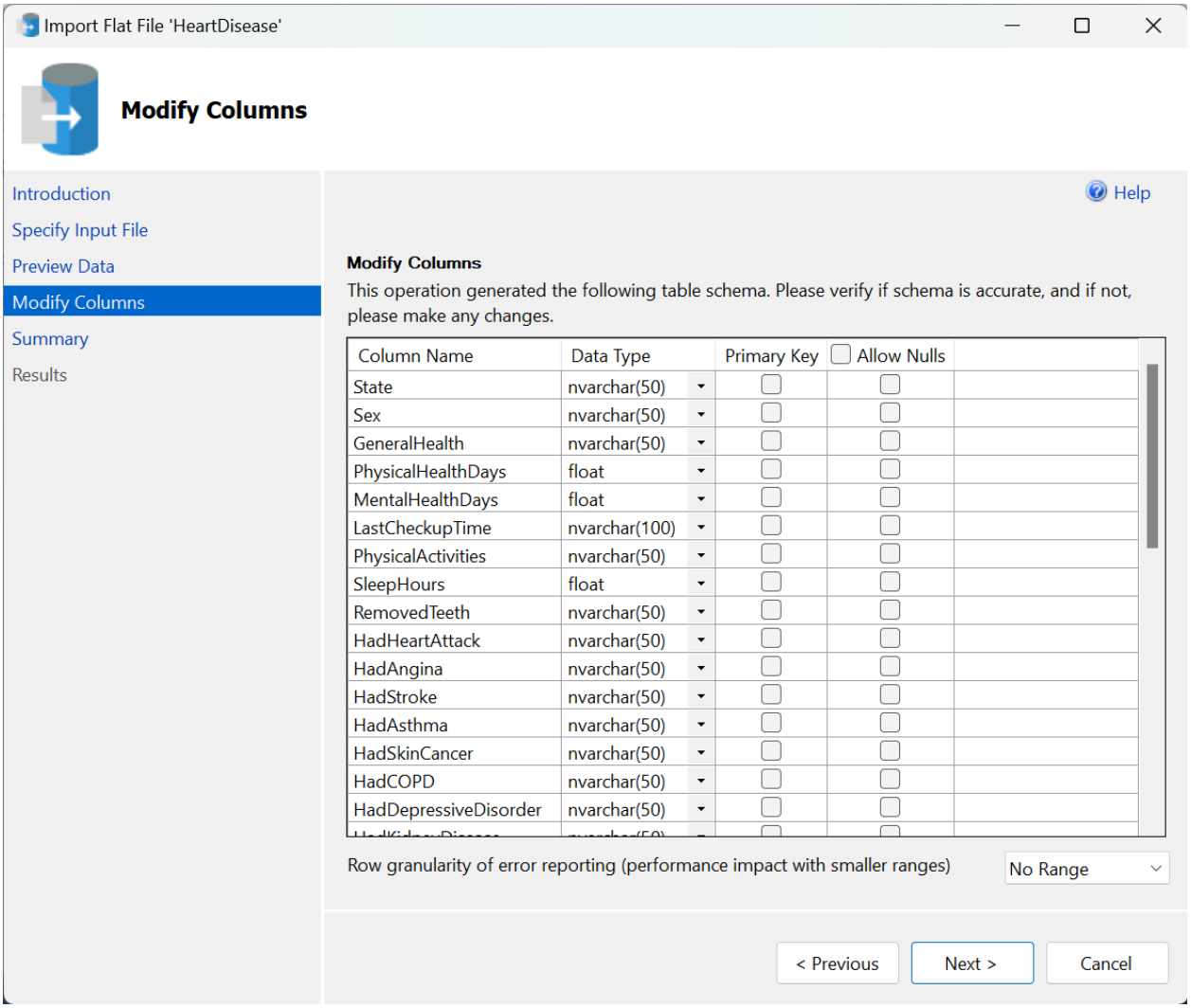
Bước 4: Ở bước Preview Data chúng ta có thể xem trước được dữ liệu trong file đầu vào của chúng ta sẽ được tổ chức như thế nào.



State	Sex	GeneralHea	PhysicalHea	MentalHeal	LastChecku	PhysicalAct	SleepHou
Alabama	Female	Very good	4.0	0.0	Within p...	Yes	9.0
Alabama	Male	Very good	0.0	0.0	Within p...	Yes	6.0
Alabama	Male	Very good	0.0	0.0	Within p...	No	8.0
Alabama	Female	Fair	5.0	0.0	Within p...	Yes	9.0
Alabama	Female	Good	3.0	15.0	Within p...	Yes	5.0
Alabama	Male	Good	0.0	0.0	Within p...	Yes	7.0
Alabama	Female	Good	3.0	0.0	Within p...	Yes	8.0
Alabama	Male	Fair	5.0	0.0	Within p...	Yes	8.0
Alabama	Male	Good	2.0	0.0	5 or mor...	No	6.0
Alabama	Female	Very good	0.0	0.0	Within p...	Yes	7.0
Alabama	Male	Very good	0.0	0.0	Within p...	Yes	8.0
Alabama	Female	Good	3.0	4.0	Within p...	Yes	5.0
Alabama	Male	Good	5.0	0.0	Within p...	Yes	5.0

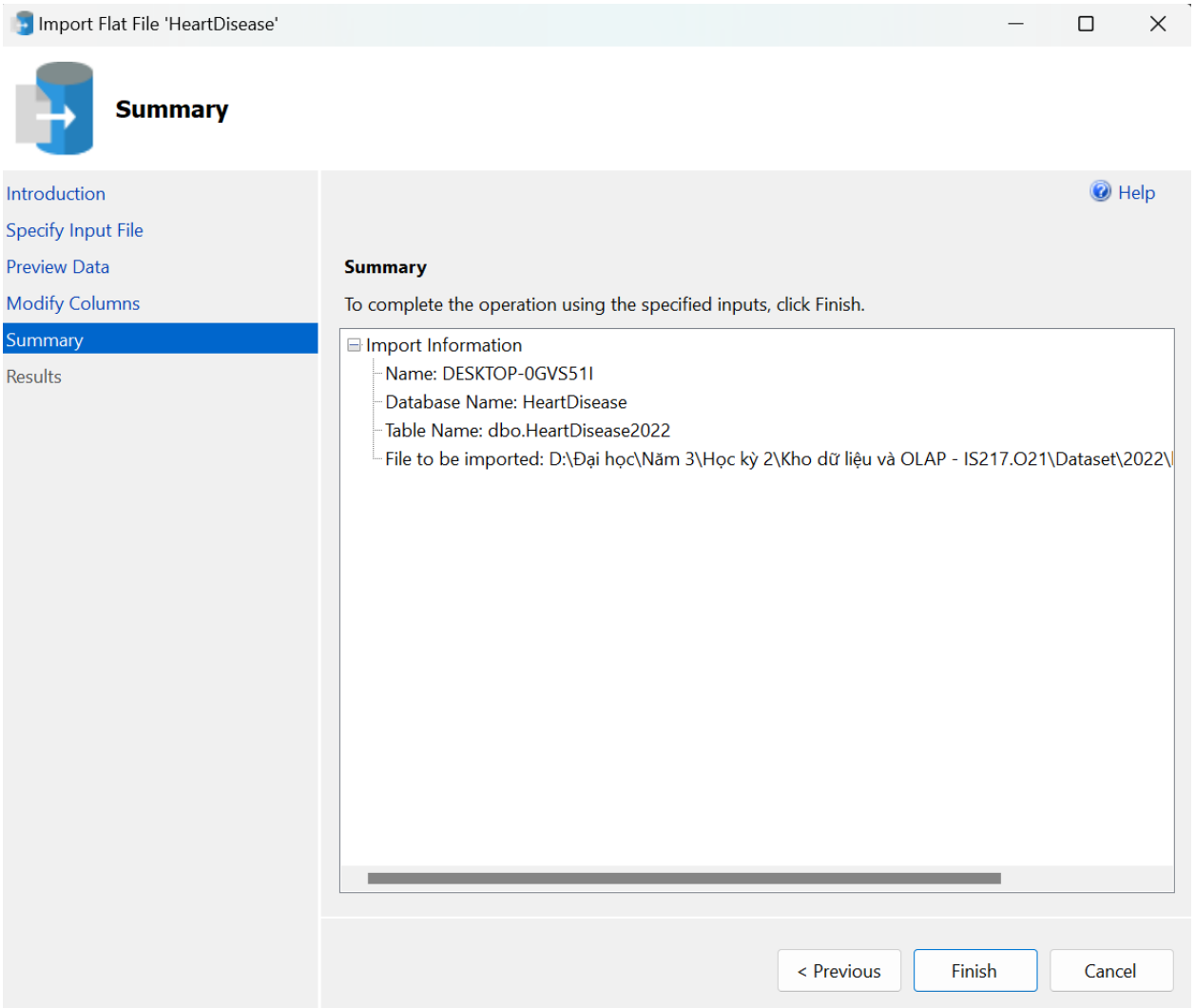
Hình 5.5 Dữ liệu trích xuất được từ file csv

Bước 5: Điều chỉnh lại kiểu dữ liệu và ràng buộc của các thuộc tính của bảng dữ liệu.



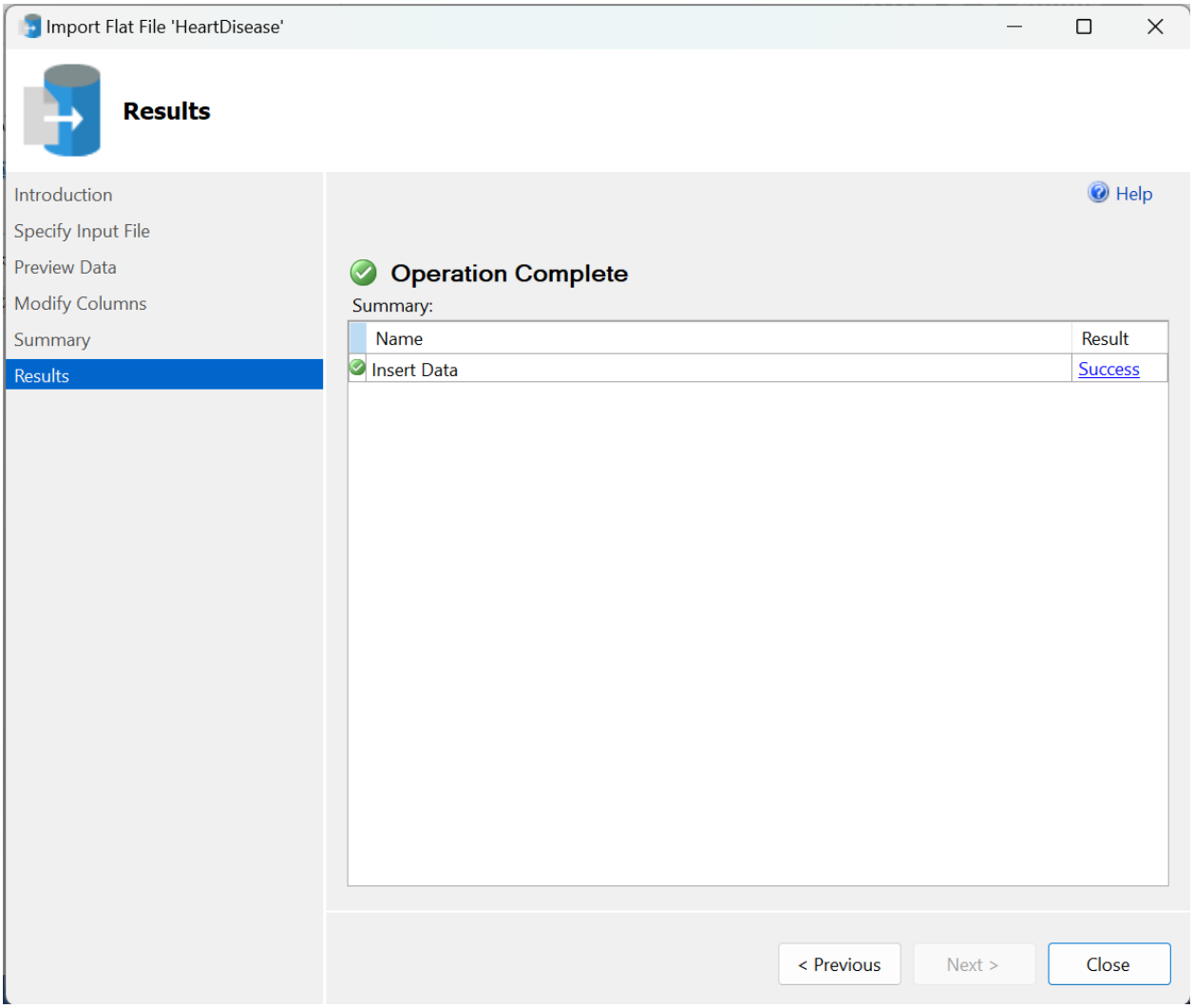
Hình 5.6 Chỉnh sửa thông tin các cột dữ liệu của bảng

Bước 6: Tại tab Summary sẽ tóm gọn lại các thông tin của dữ liệu được đổ vào.



Hình 5.7 Tóm tắt thông tin của quá trình đổ dữ liệu từ file csv

Bước 7: Kết quả của quá trình đổ dữ liệu file csv.



Hình 5.8 Thông báo quá trình đổ dữ liệu thành công

Bước 8: Vì trong file csv chứa bộ dữ liệu gốc ban đầu không có cột nào đóng vai trò làm khoá chính cho bảng được vì vậy ở đây nhóm quyết định tạo thêm một cột mới có tên là **PatientID** với giá trị tự động tăng(IDENTITY(1,1)).

```
-- add auto increment patient id
ALTER TABLE HeartDisease2022 ADD PatientID INT IDENTITY(1, 1) CONSTRAINT HeartDisease_PK PRIMARY KEY
```

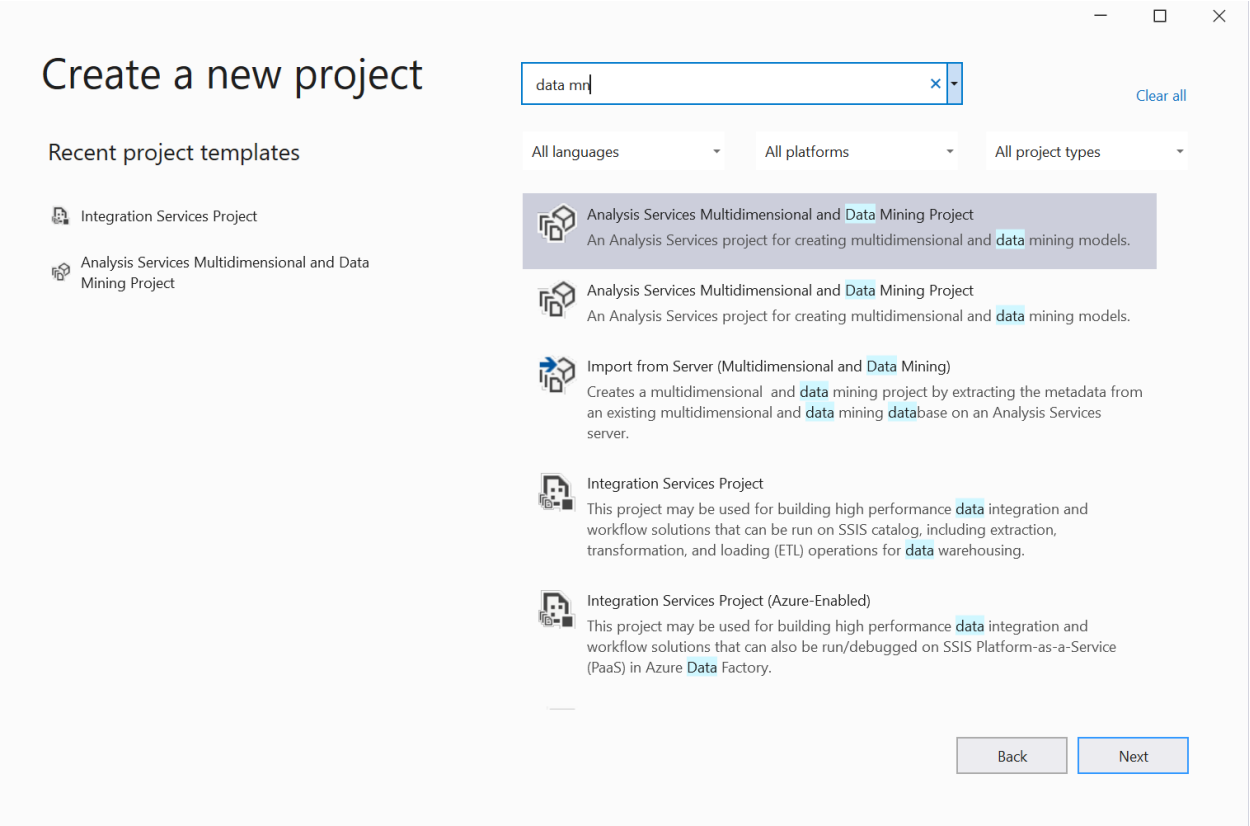
Column Name	Data Type	Allow Nulls
BMI	float	☐
AlcoholDrinkers	nvarchar(50)	☐
HIVTesting	nvarchar(50)	☐
FluVaxLast12	nvarchar(50)	☐
PneumoVaxEver	nvarchar(50)	☐
TetanusLast10Tdap	nvarchar(100)	☐
HighRiskLastYear	nvarchar(50)	☐
CovidPos	nvarchar(100)	☑
PatientID	int	☐
		☐

Hình 5.9 Tạo cột mới đóng vai trò làm khoá chính của bảng

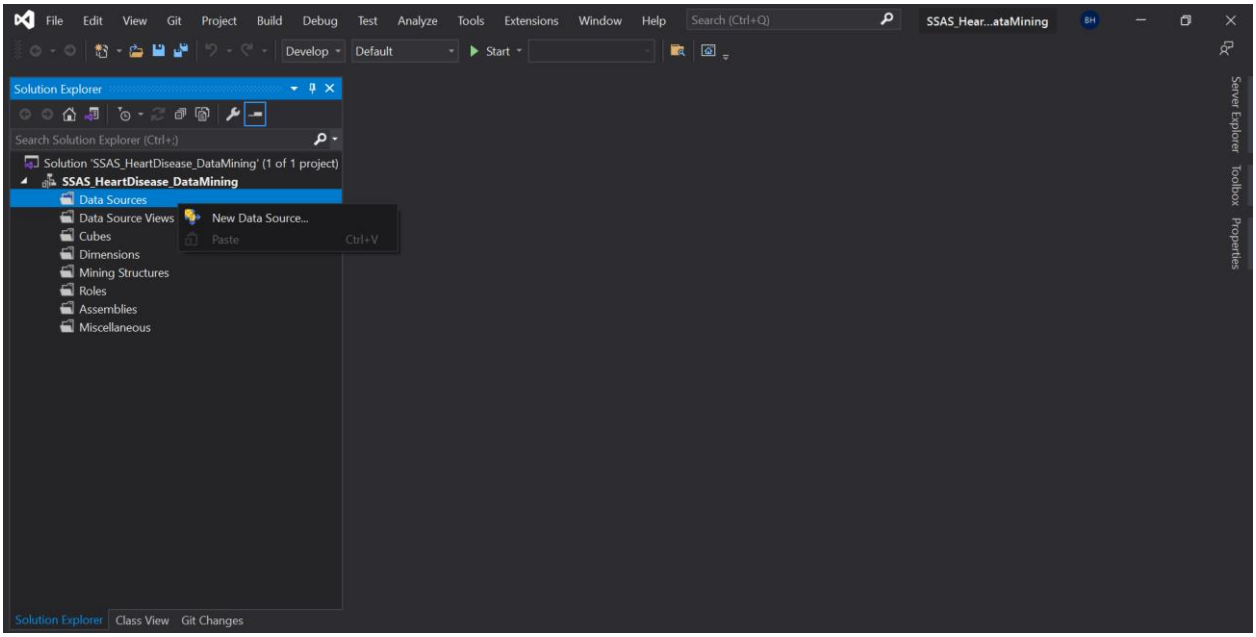
3.2. Quá trình Data Mining bằng công cụ SSAS của Visual Studio 2019 và thuật toán Decision Trees(Cây quyết định):

a. Tạo cấu trúc khai thác dữ liệu và mô hình cây quyết định

Bước 1: Mở công cụ **Visual Studio 2019 → Create new project →** Chọn **Analysis Services Multidimensional and Data Mining Project** để tạo mới một SSAS project → Đặt tên cho dự án → **Create** để hoàn thành tạo dự án.

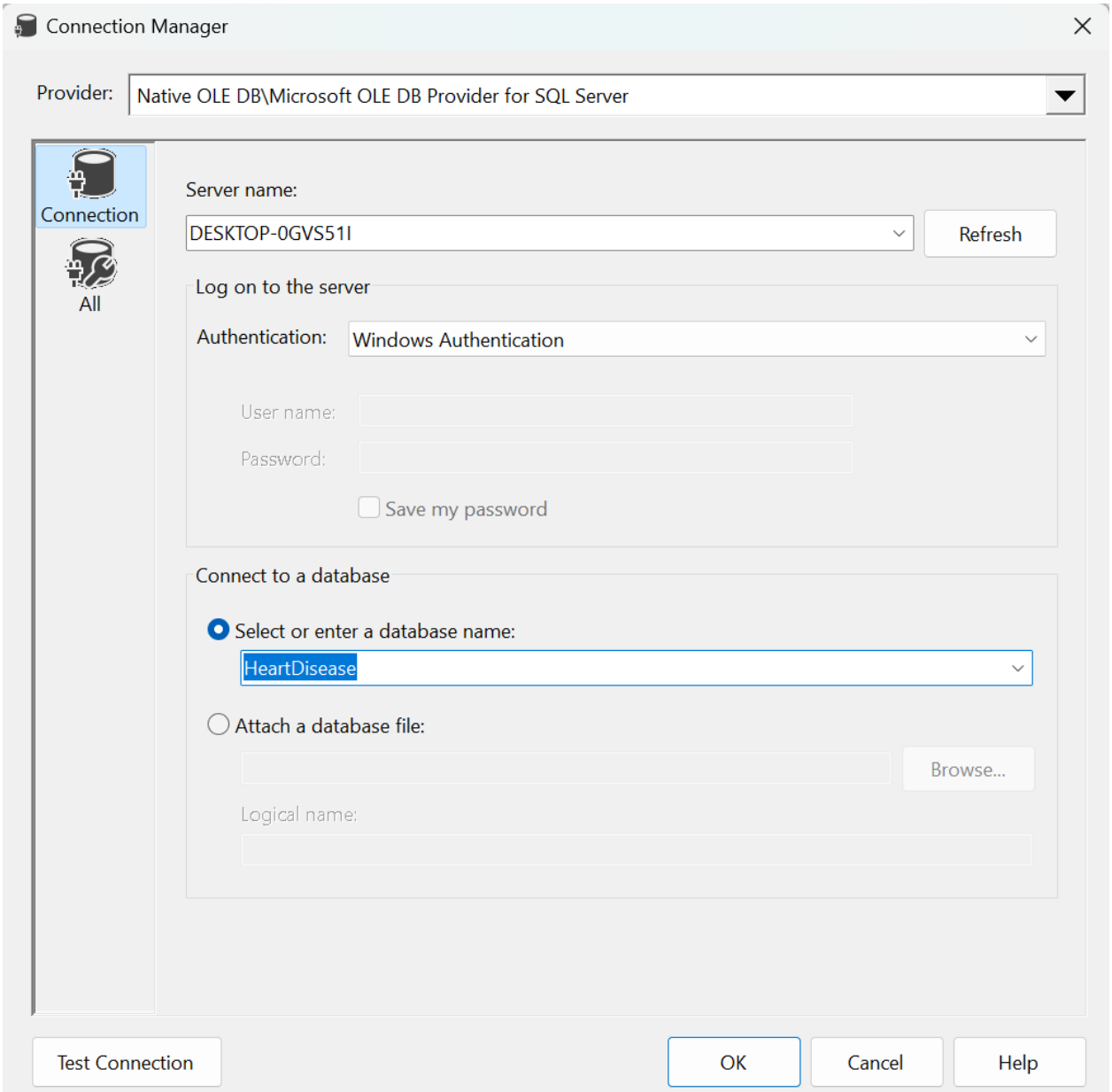


Bước 2: Tạo một Data Source mới kết nối với database HeartDisease vừa được tạo ở bước đổ dữ liệu.



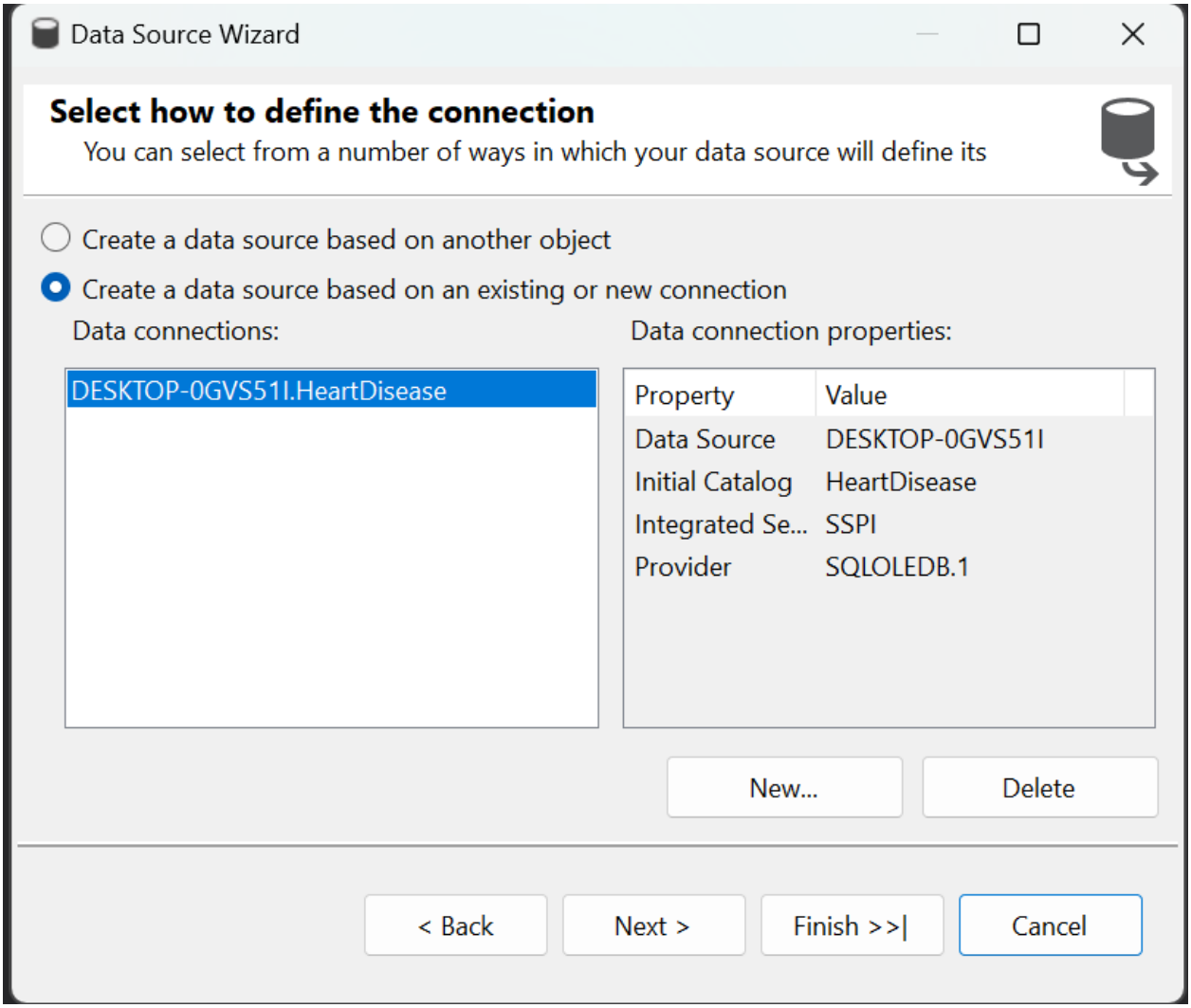
Hình 5.10 Tạo kết nối tới cơ sở dữ liệu HeartDisease

Bước 3: Chọn **Provider** cho kết nối → Nhập tên Server + Loại xác thực → Chọn cơ sở dữ liệu **HeartDisease → OK**.



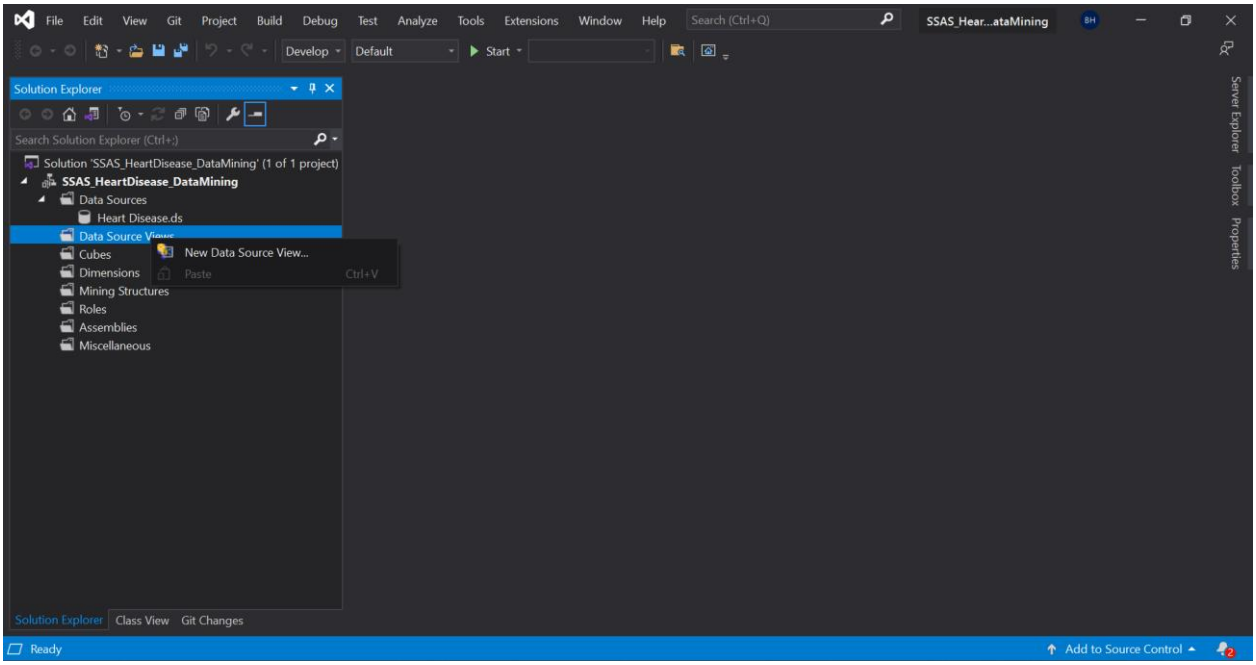
Hình 5.11 Cấu hình kết nối tới cơ sở dữ liệu HeartDisease

Bước 4: Chọn kết nối vừa tạo và nhấn Next để sang bước tiếp theo.



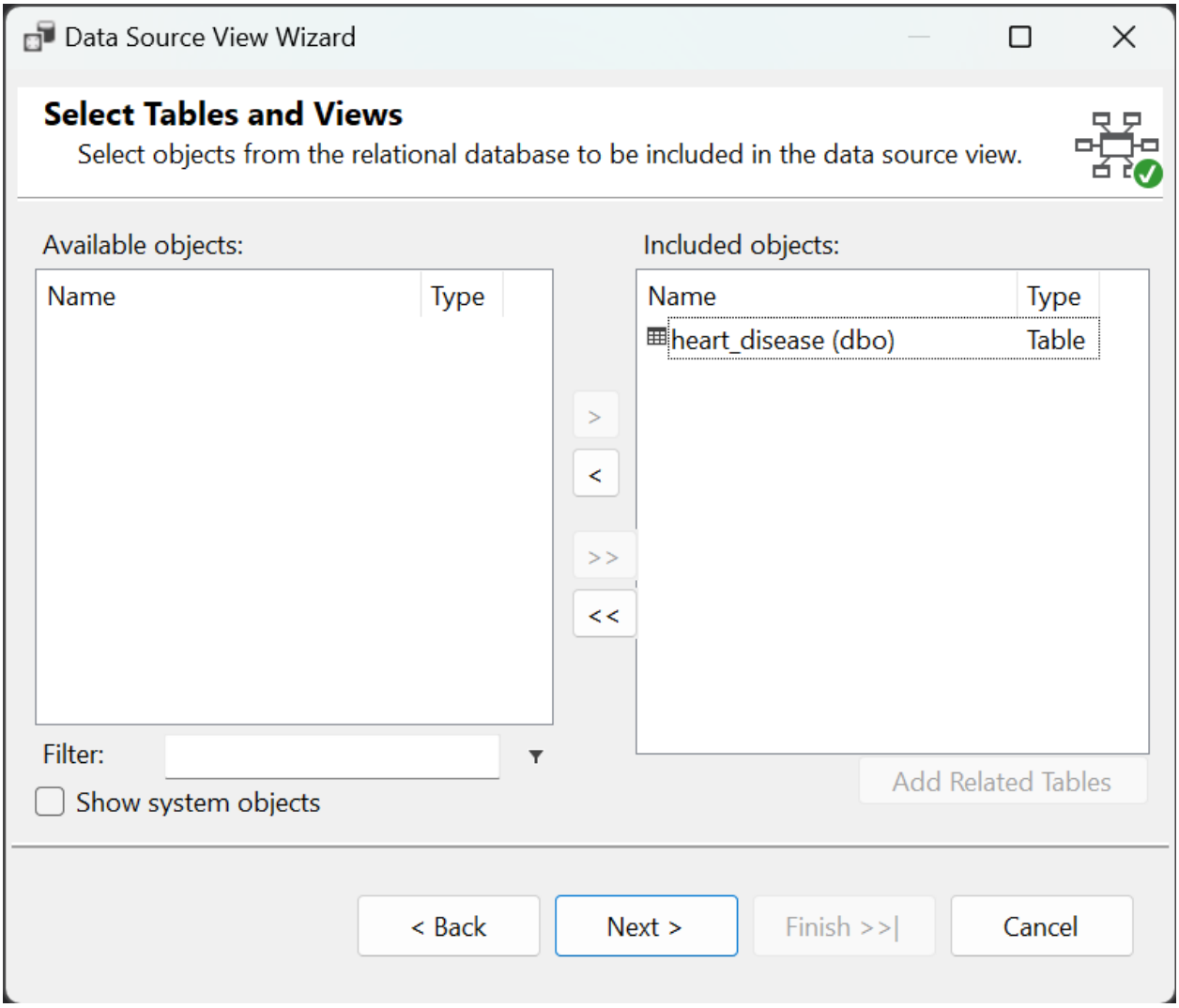
Hình 5.12 Chọn kết nối vừa tạo

Bước 5: Sau khi đã tạo một kết nối tới cơ sở dữ liệu **HeartDisease** thì ta sẽ tiến hành tạo mới một **Data Source View** cho dữ liệu, đây chính là đầu vào của các thuật toán Data Mining chúng ta sẽ dùng để khai thác dữ liệu của **HeartDisease**.



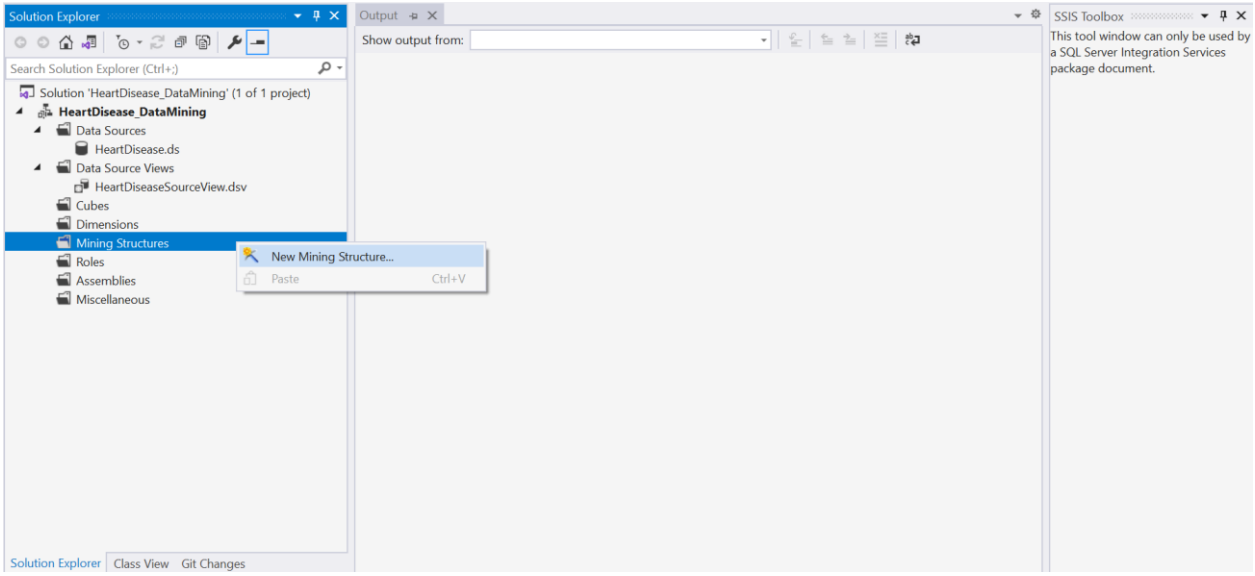
Hình 5.13 Tạo mới một Data Source View

Bước 6: Ở bước chọn đối tượng được thể hiện trong **Data Source View** thì ta chọn **table HeartDisease2022** và nhấn biểu tượng > (hoặc nhấn đúp chuột trái) để đưa bảng vào view.



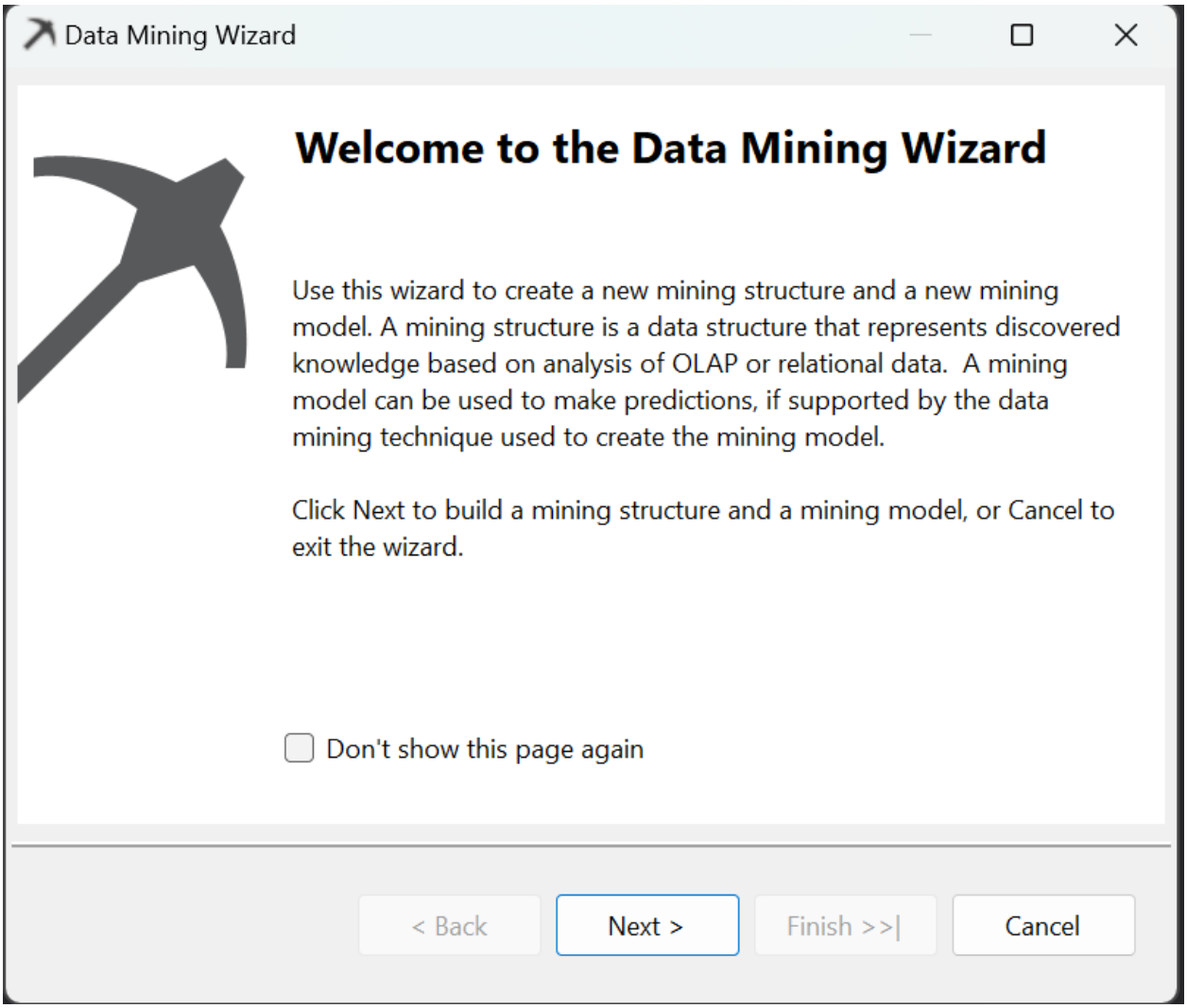
Hình 5.14 Chọn đối tượng bảng thể hiện trong View

Bước 7: Chuột phải vào **Mining Structures → New Mining Structure** để tạo mới cấu trúc dữ liệu để phục vụ cho quá trình khai thác dữ liệu.



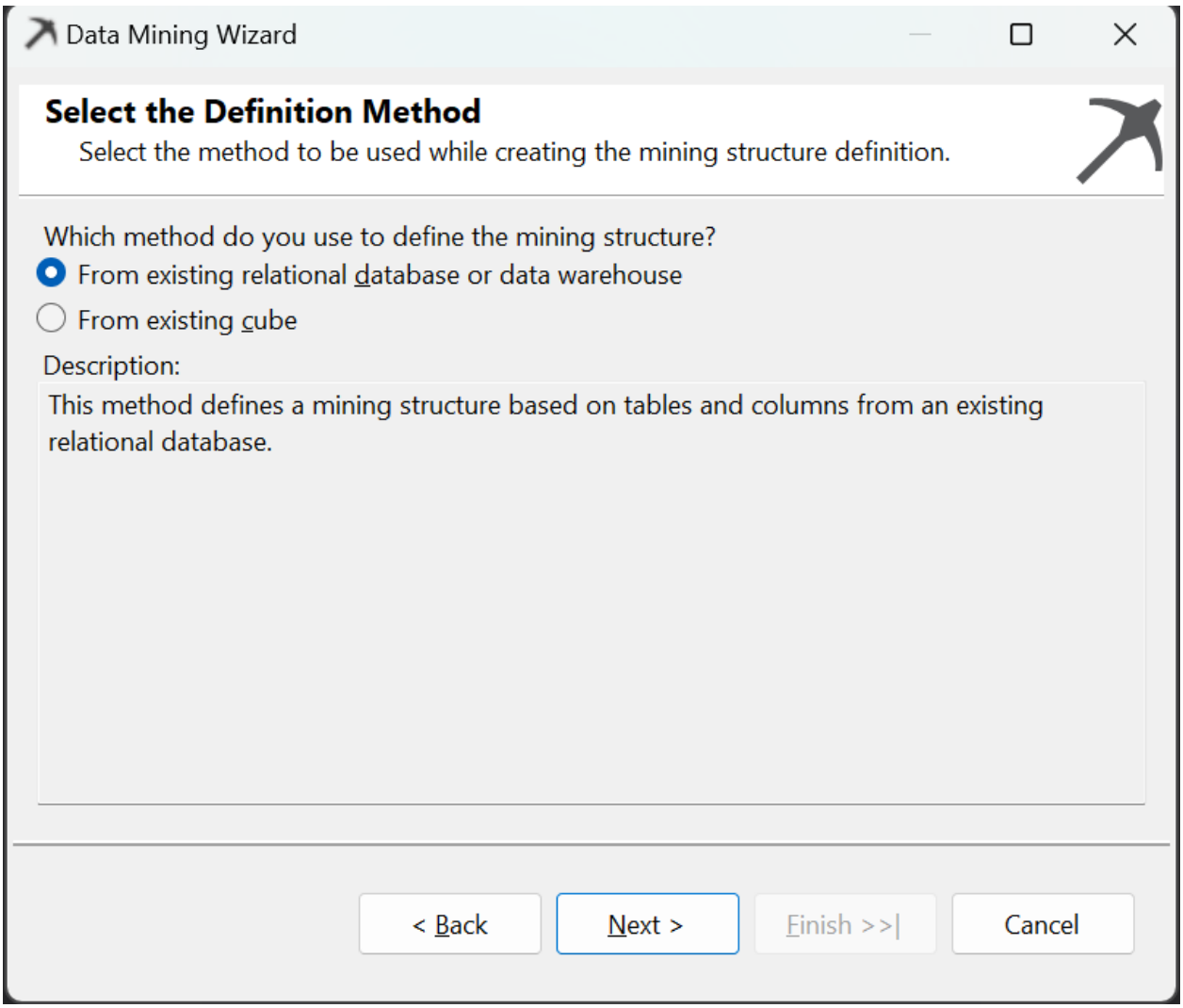
Hình 5.15 Tạo mới Mining Structure

Bước 8: Ở hộp thoại **Data Mining Wizard**, chọn **Next**.



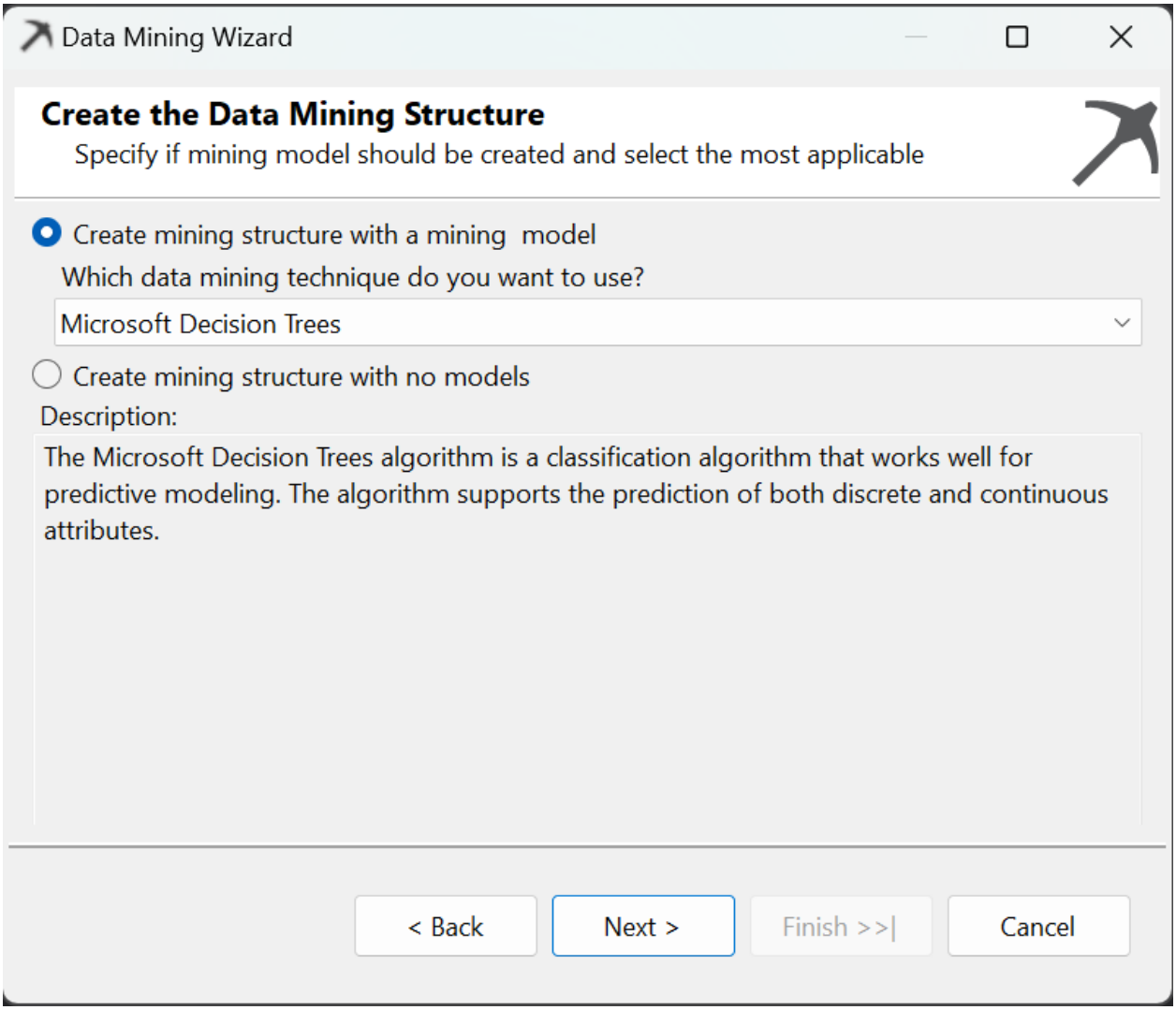
Hình 5.16 Hộp thoại Data Mining Wizard

Bước 9: Chọn dữ liệu nguồn để làm đầu vào của cấu trúc khai thác dữ liệu. Ở đây ta chọn **From existing relational database or data warehouse** để lấy dữ liệu nguồn là **HeartDisease**.



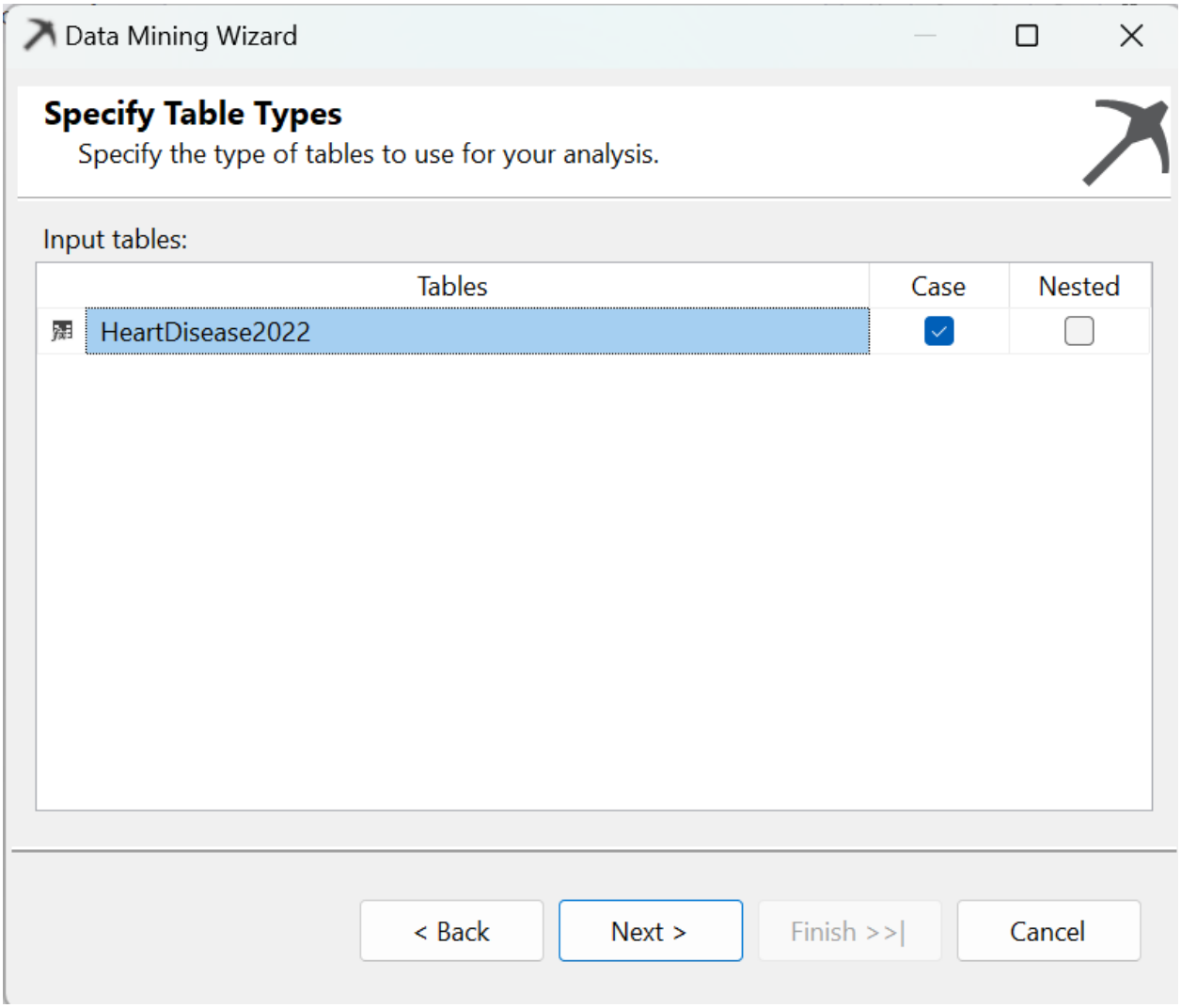
Hình 5.17 Chọn dữ liệu nguồn cho cấu trúc khai thác dữ liệu

Bước 10: Chọn mô hình dùng để khai thác dữ liệu là Microsoft Decision Trees để áp dụng thuật toán cây quyết định cho việc khai thác dữ liệu.



Hình 5.18 Chọn thuật toán cây quyết định cho quá trình khai thác dữ liệu

Bước 11: Chọn bảng **HeartDisease2022** làm case cho quá trình khai thác.

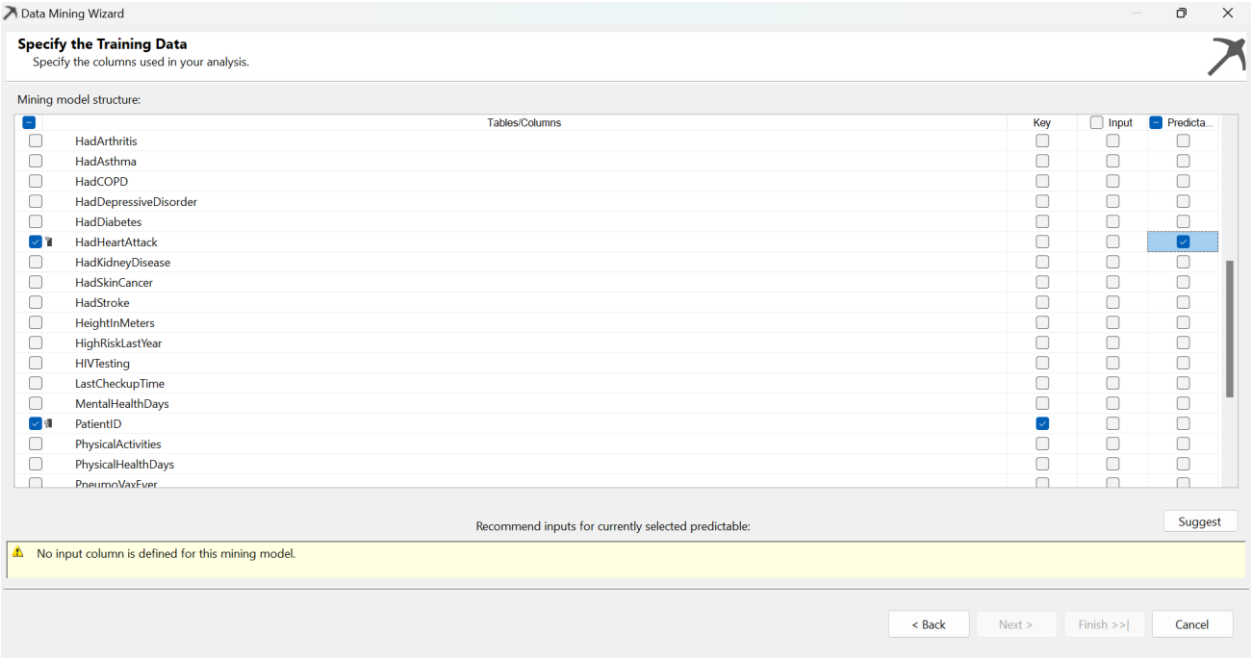


Hình 5.19 Chọn bảng HeatDisease2022 làm đối tượng khai thác

Bước 12: Chọn thuộc tính cho các loại thuộc tính sau:

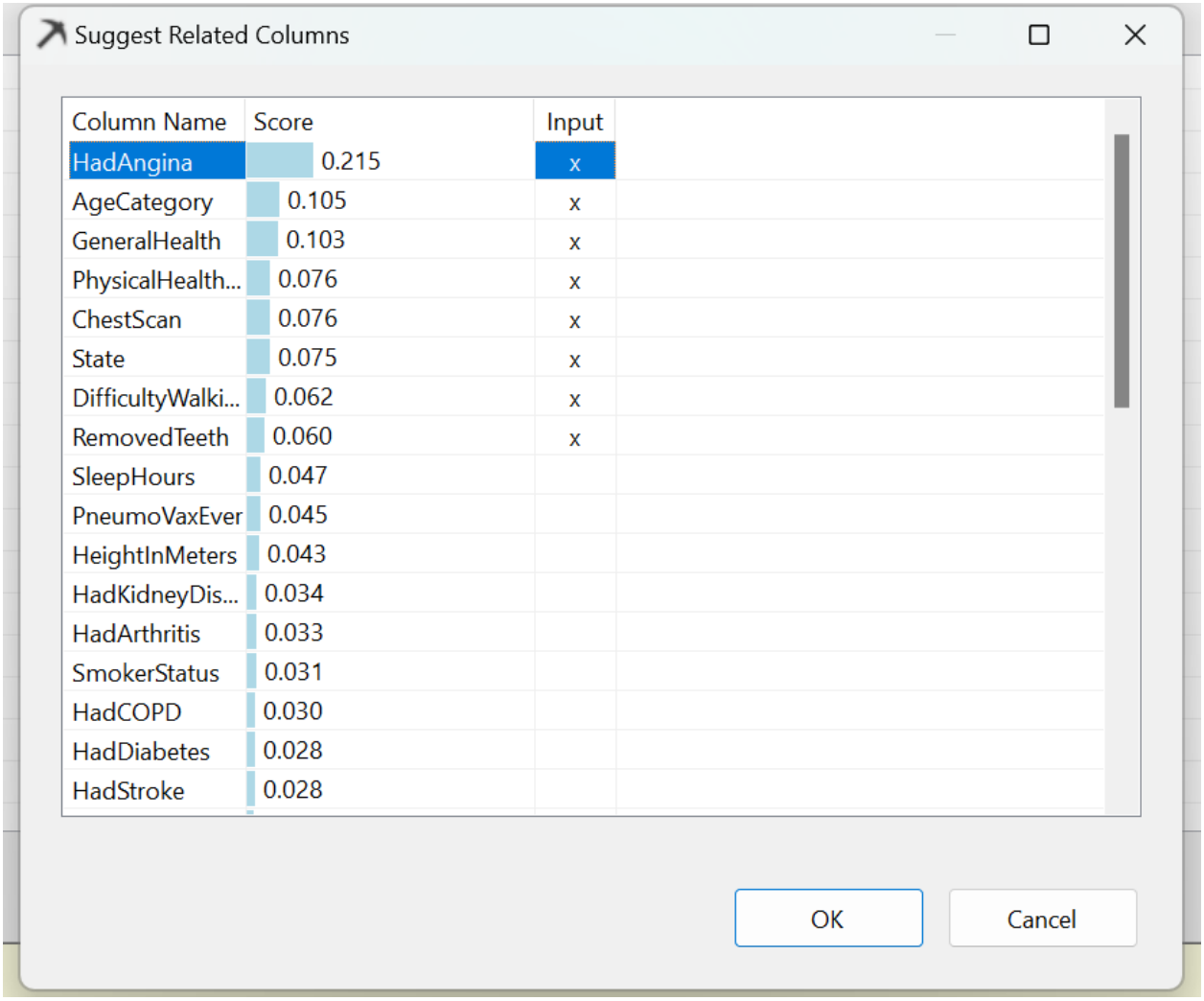
- Key: Thuộc tính khoá chính.
- Input: Thuộc tính đầu vào để từ đó thuật toán có thể sử dụng để phân lớp và đưa ra tỉ lệ tương quan với cột thuộc tính dự đoán.
- Predicts: Thuộc tính cần dự đoán giá trị.

Ở đây thuộc tính dự đoán là **HadHeartAttack** tức là dự đoán xem người đó có



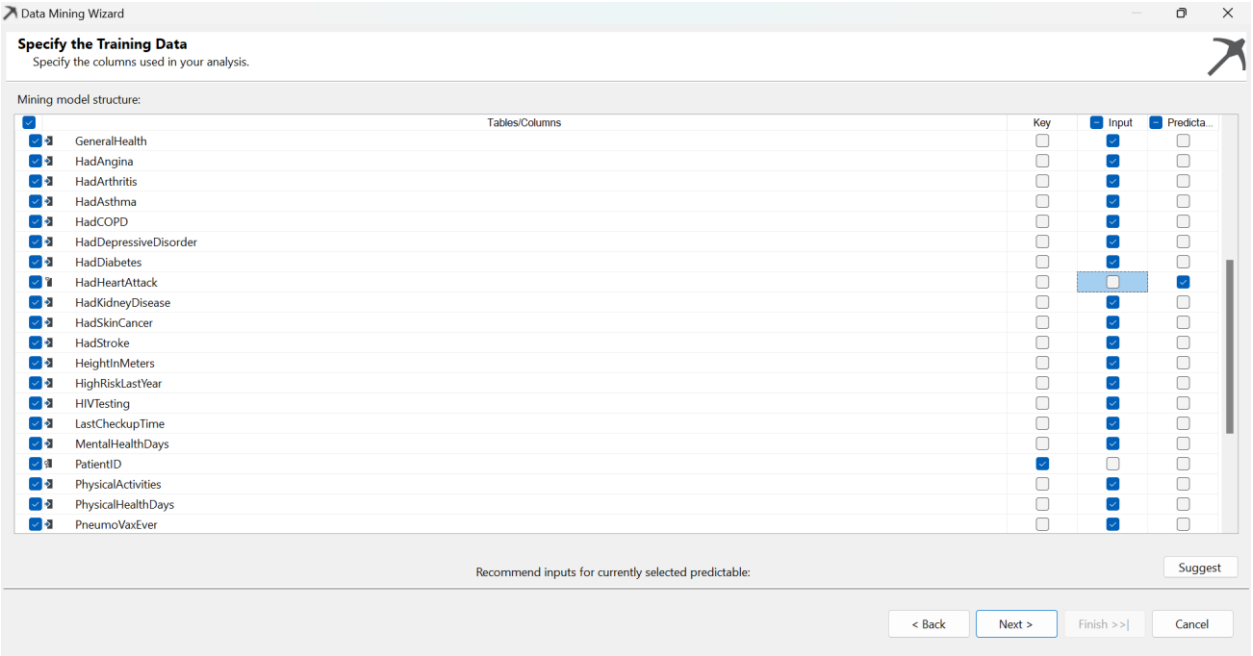
Hình 5.20 Chọn các thuộc tính đầu vào và thuộc tính dự đoán

Khi nhấn vào nút Suggest thì công cụ sẽ gợi ý cho chúng ta các thuộc tính có thể dùng làm đầu vào cho thuật toán, các cột này được hệ thống gợi ý theo thứ tự liên quan từ mạnh nhất đến yếu nhất với cột được dự đoán. Nếu như đồng ý với các cột được gợi ý thì nhấn **OK** để chọn các cột input.



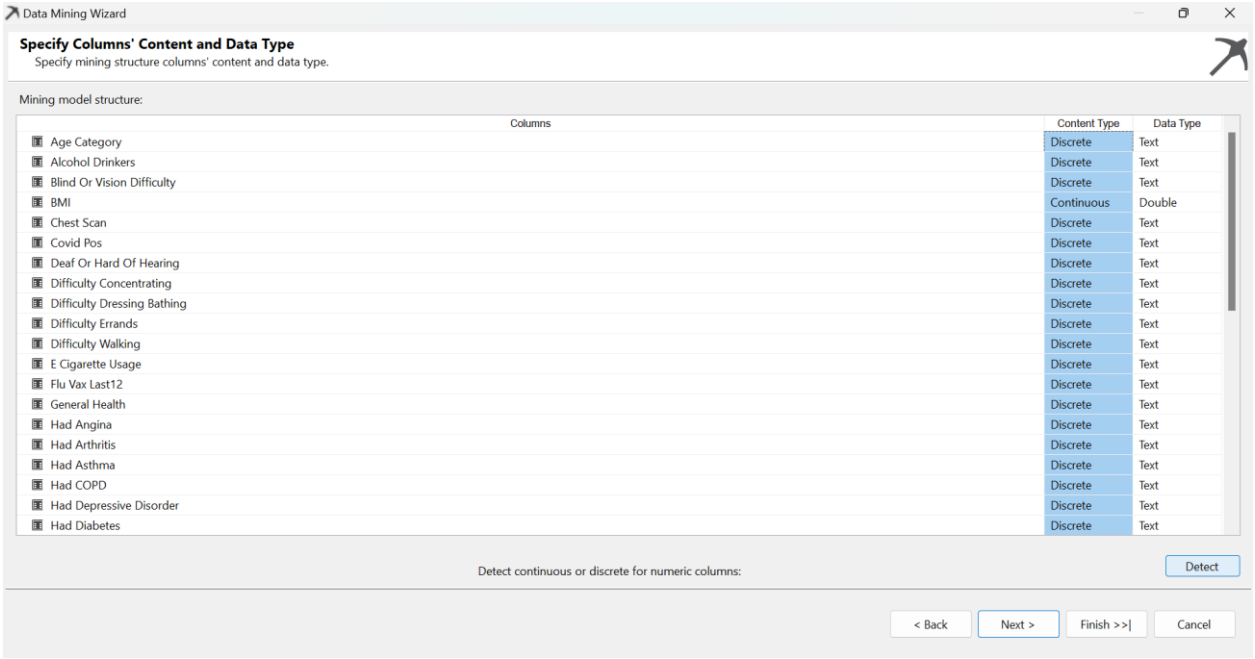
Hình 5.21 Hệ thống gợi ý các cột có thể làm dùng để làm input

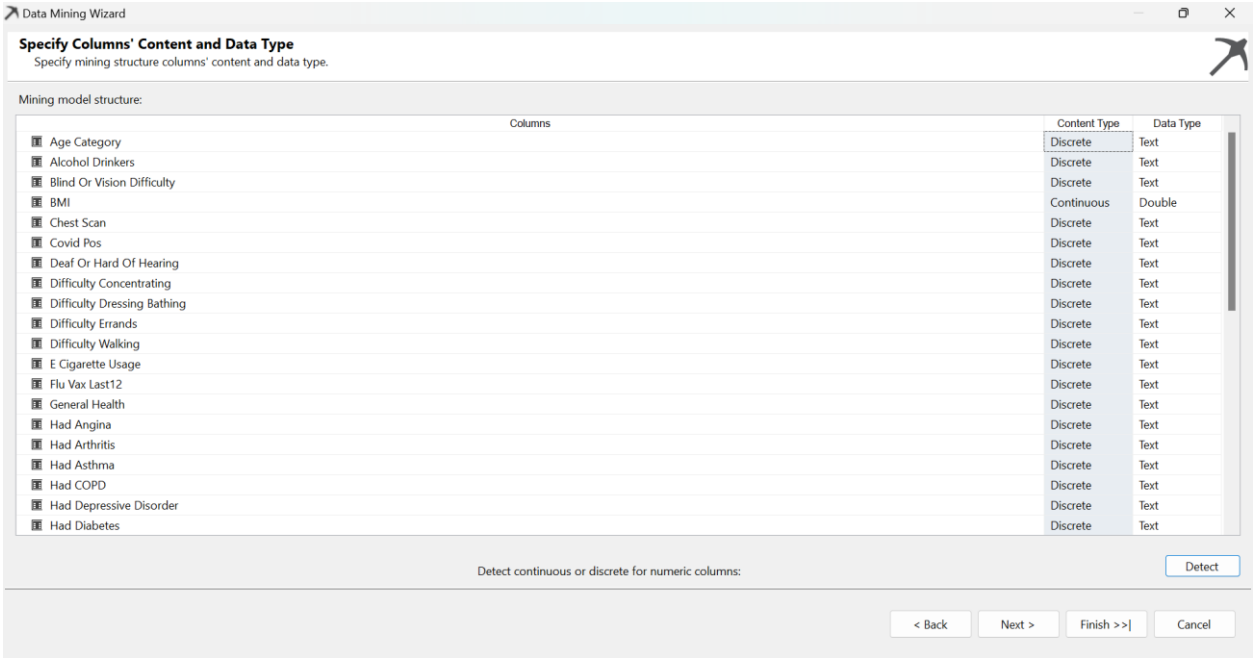
Đồng thời, ở bước này chúng ta có thể chọn thủ công các thuộc tính làm đầu vào cho thuật toán cây. Ngoài khoá chính và thuộc tính dự đoán thì nhóm quyết định chọn tất cả các cột còn lại làm input. Sau đó, nhấn Next để sang bước tiếp theo.



Hình 5.22 Chọn các thuộc tính còn lại ngoại trừ khoá và thuộc tính dự đoán làm input

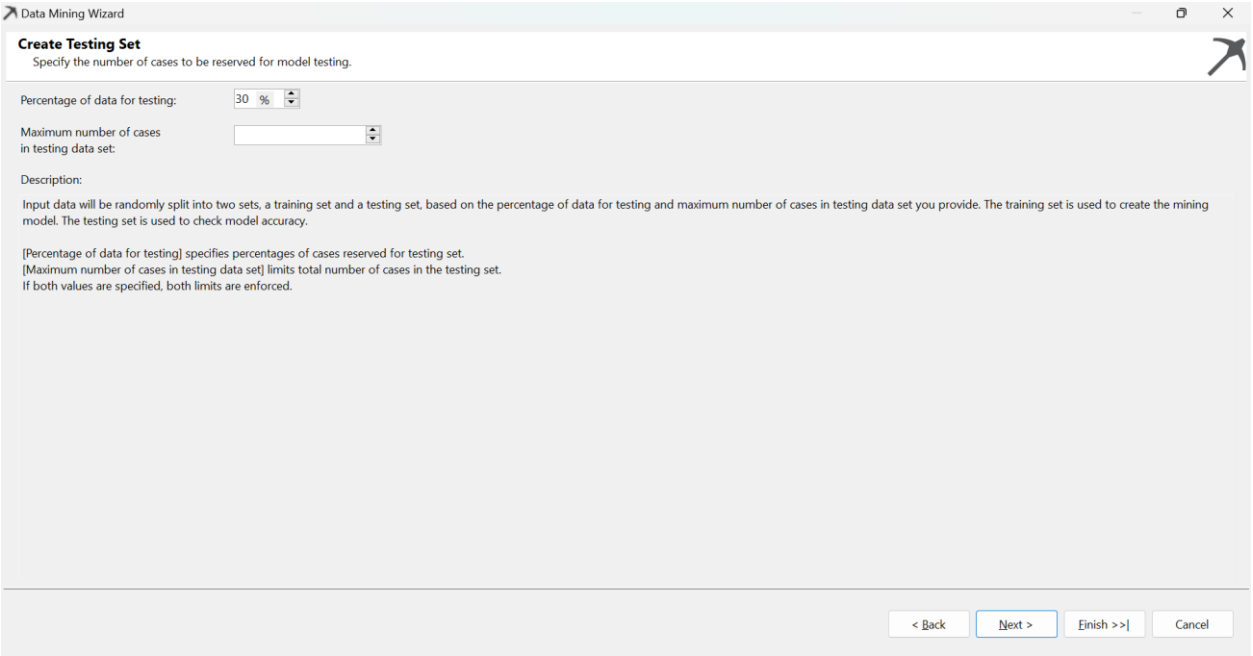
Bước 13: Điều chỉnh lại kiểu nội dung và kiểu dữ liệu các thuộc tính. Ta có thể sử dụng chức năng **Detect** để hệ thống các thuộc tính được cài đặt sao cho phù hợp nhất → **Next** để sang bước tiếp theo.





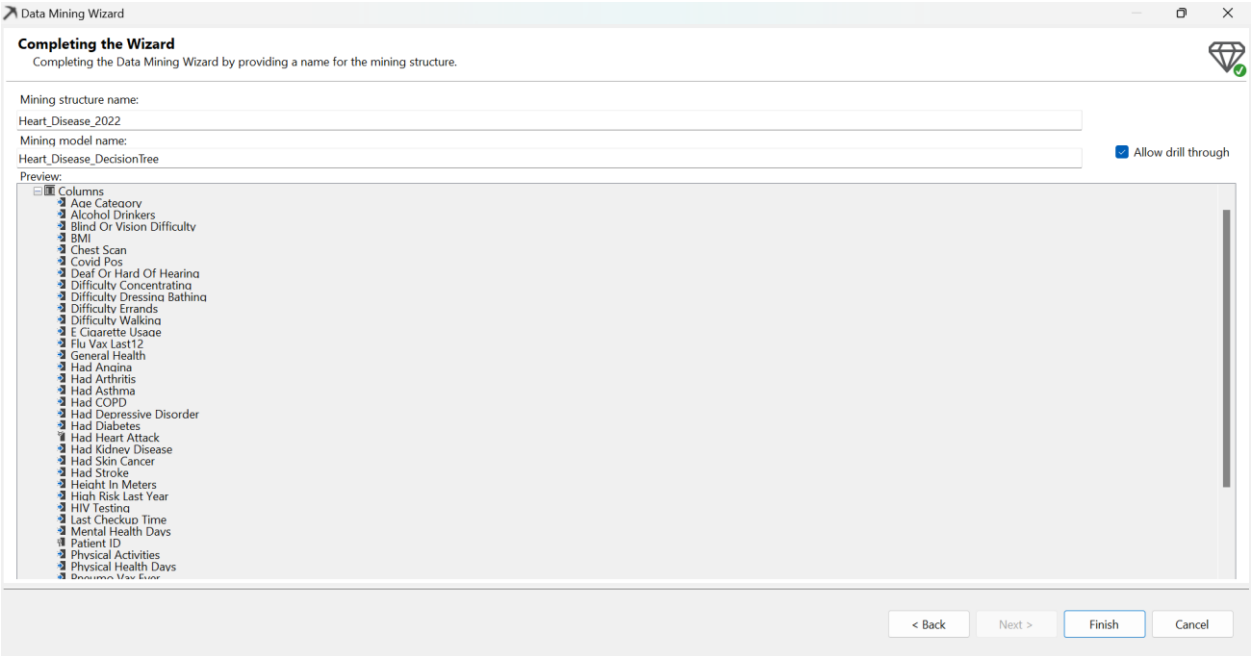
Hình 5.23 Điều chỉnh các thuộc tính của các cột dữ liệu

Bước 14: Cài đặt tỉ lệ chia dữ liệu kiểm thử như thế nào. Ở đây nhóm sẽ chọn tỉ lệ của dữ liệu kiểm thử mặc định là **30% → Next.**

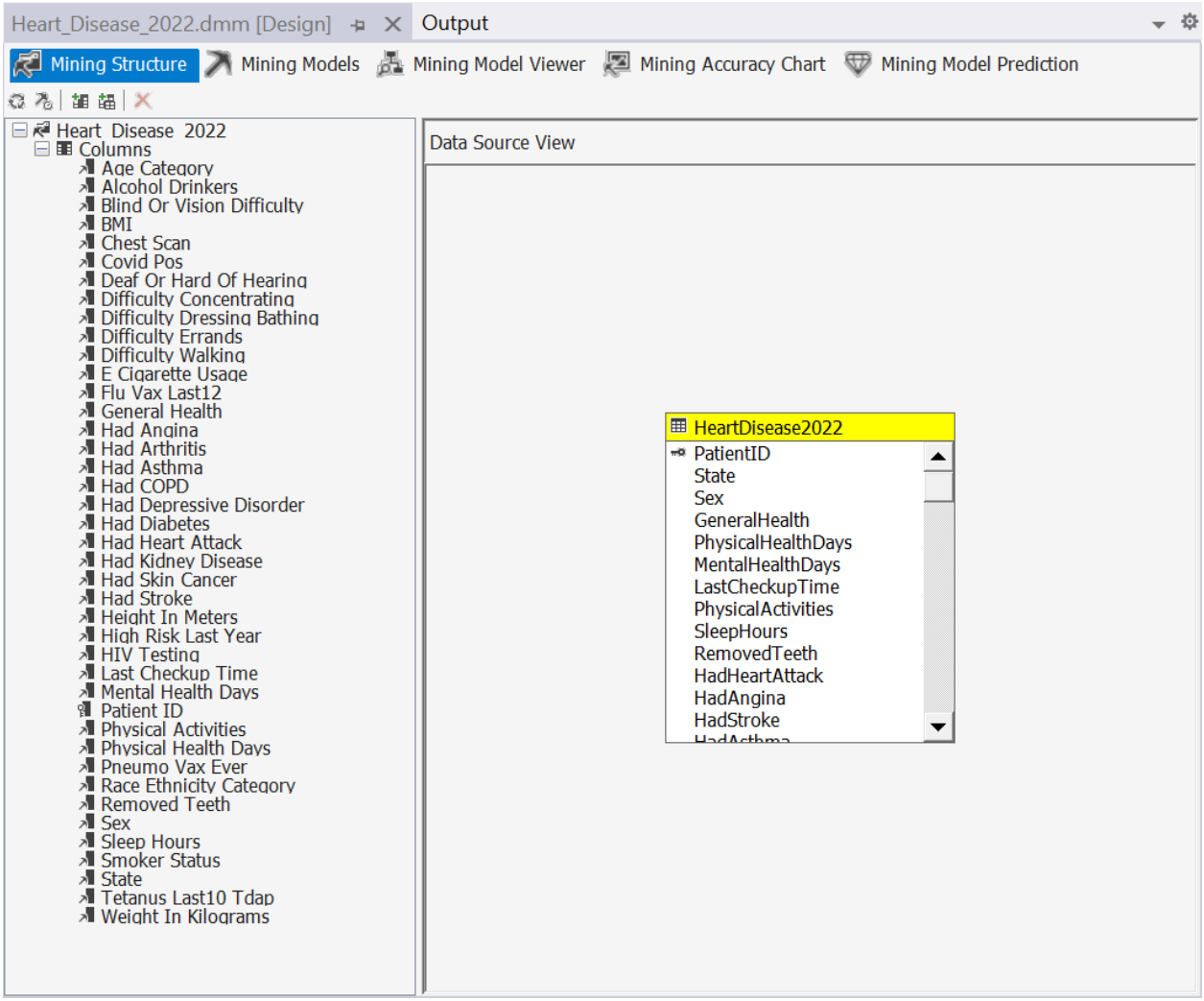


Hình 5.24 Điều chỉnh tỉ lệ dữ liệu kiểm thử

Bước 15: Đặt tên cho mô hình khai thác dữ liệu là **Heart_Disease_2022**, tên của mô hình khai thác là **Heart_Disease_DecisionTree.** Ngoài ra, ta cần tích chọn ô **Allow drill through** để ta có thể tương tác trực tiếp với các mô hình trong quá trình làm việc. Cuối cùng, ta chọn **Finish** để hoàn thành.

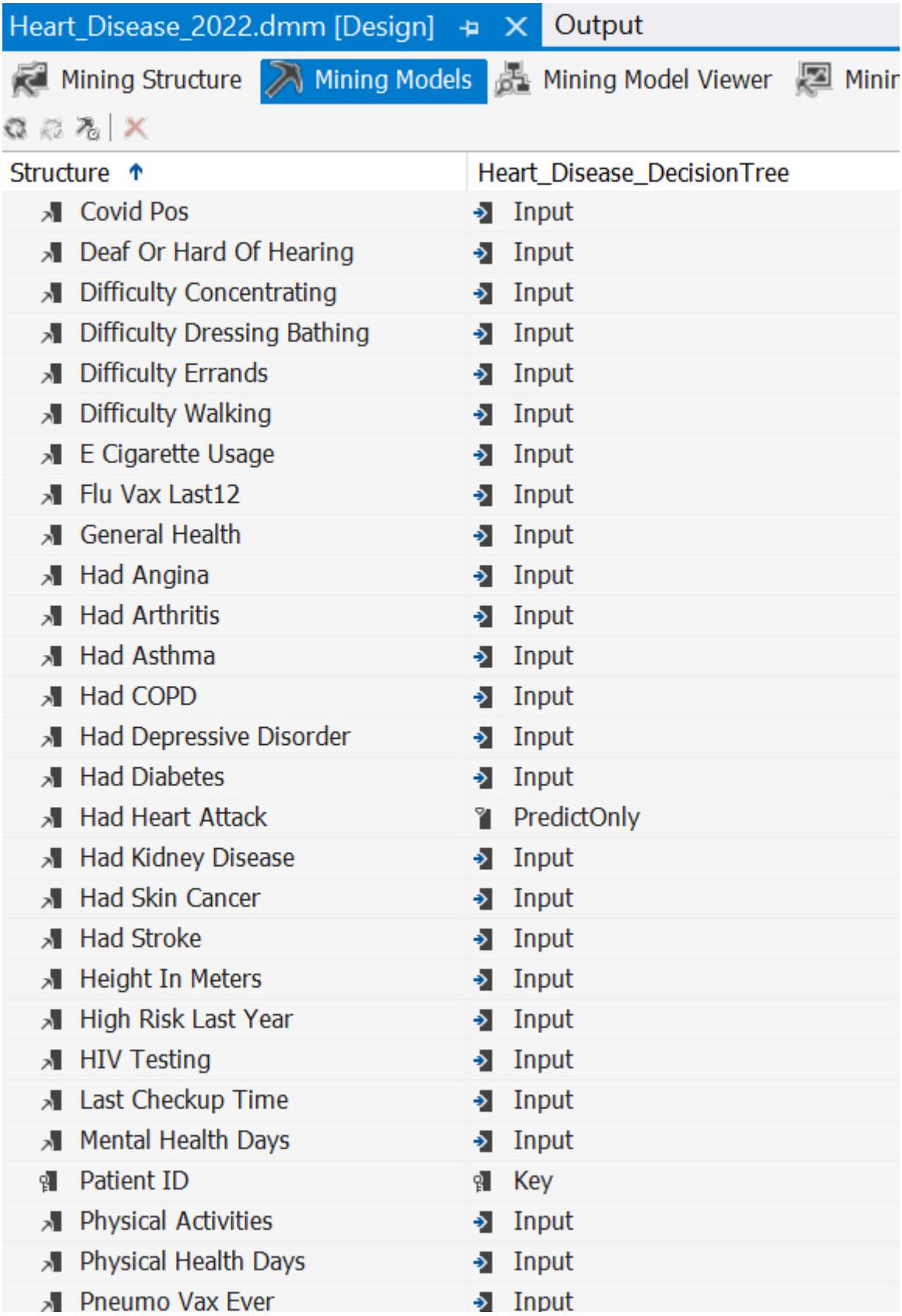


Hình 5.25 Hoàn thiện tạo cấu trúc khai thác dữ liệu



Hình 5.26 Kết quả của quá trình tạo cấu trúc khai thác dữ liệu

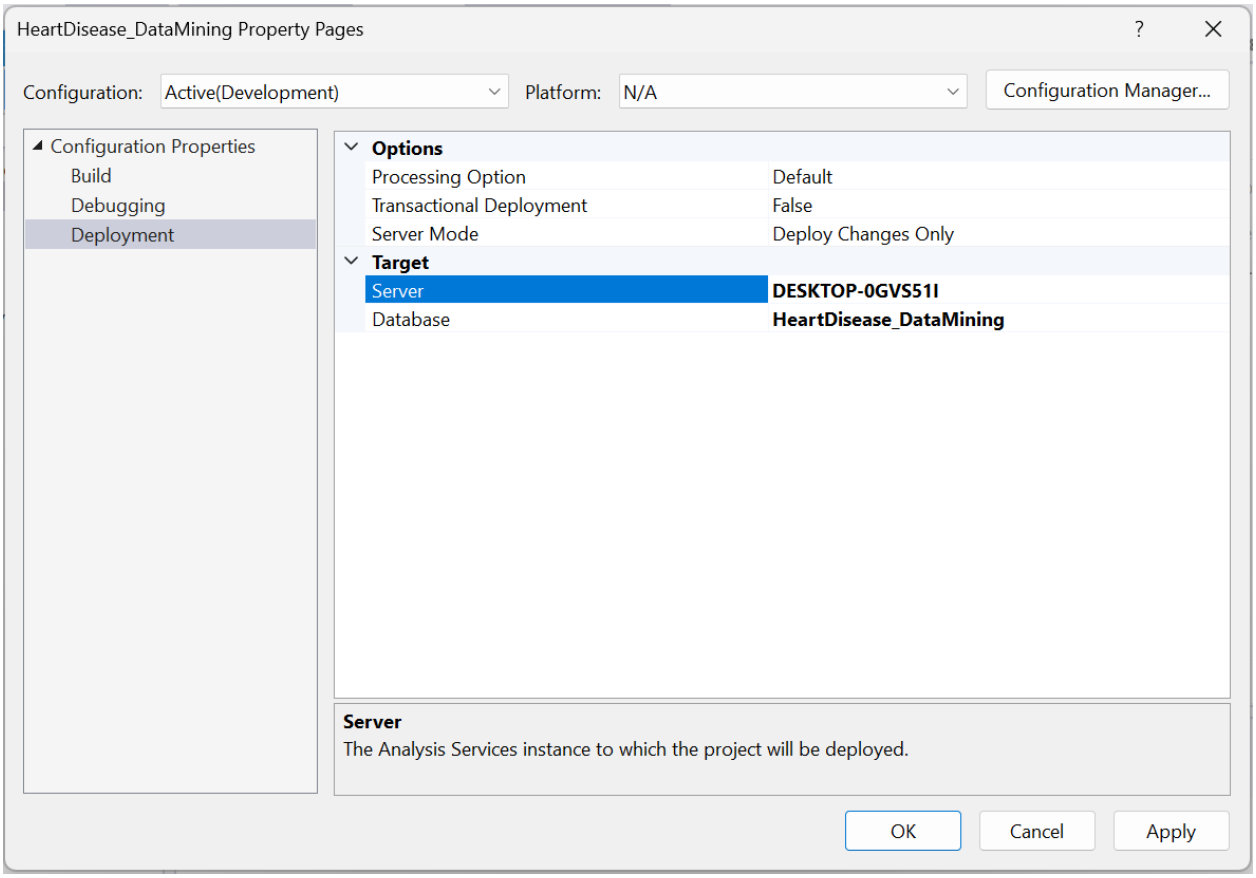
Bước 16: Sau khi hoàn thành bước tạo thì ta chọn sang tab **Mining Models** để xem lại thiết lập các trúc khai thác dữ liệu.



Structure	Heart_Disease_DecisionTree
Covid Pos	Input
Deaf Or Hard Of Hearing	Input
Difficulty Concentrating	Input
Difficulty Dressing Bathing	Input
Difficulty Errands	Input
Difficulty Walking	Input
E Cigarette Usage	Input
Flu Vax Last12	Input
General Health	Input
Had Angina	Input
Had Arthritis	Input
Had Asthma	Input
Had COPD	Input
Had Depressive Disorder	Input
Had Diabetes	Input
Had Heart Attack	PredictOnly
Had Kidney Disease	Input
Had Skin Cancer	Input
Had Stroke	Input
Height In Meters	Input
High Risk Last Year	Input
HIV Testing	Input
Last Checkup Time	Input
Mental Health Days	Input
Patient ID	Key
Physical Activities	Input
Physical Health Days	Input
Pneumo Vax Ever	Input

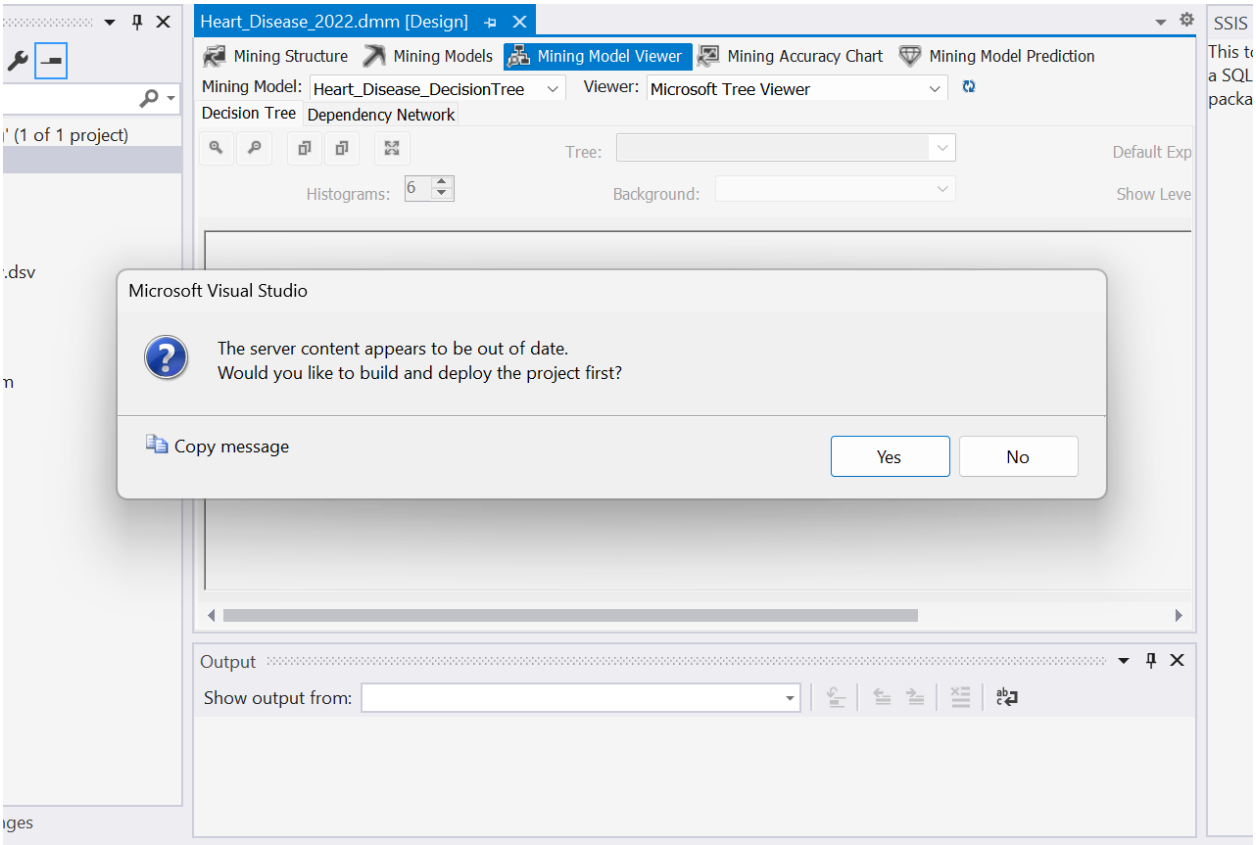
Hình 5.27 Thiết lập cấu trúc khai thác dữ liệu tại tab Mining Models

Bước 17: Để quá trình deploy dự án được thành công thì trong Properties của dự án ta chọn tab **Deployment** và thay tên **Server** từ **localhost** sang thành tên Server mà ta tạo trong **SQL Server Analysis Services**.

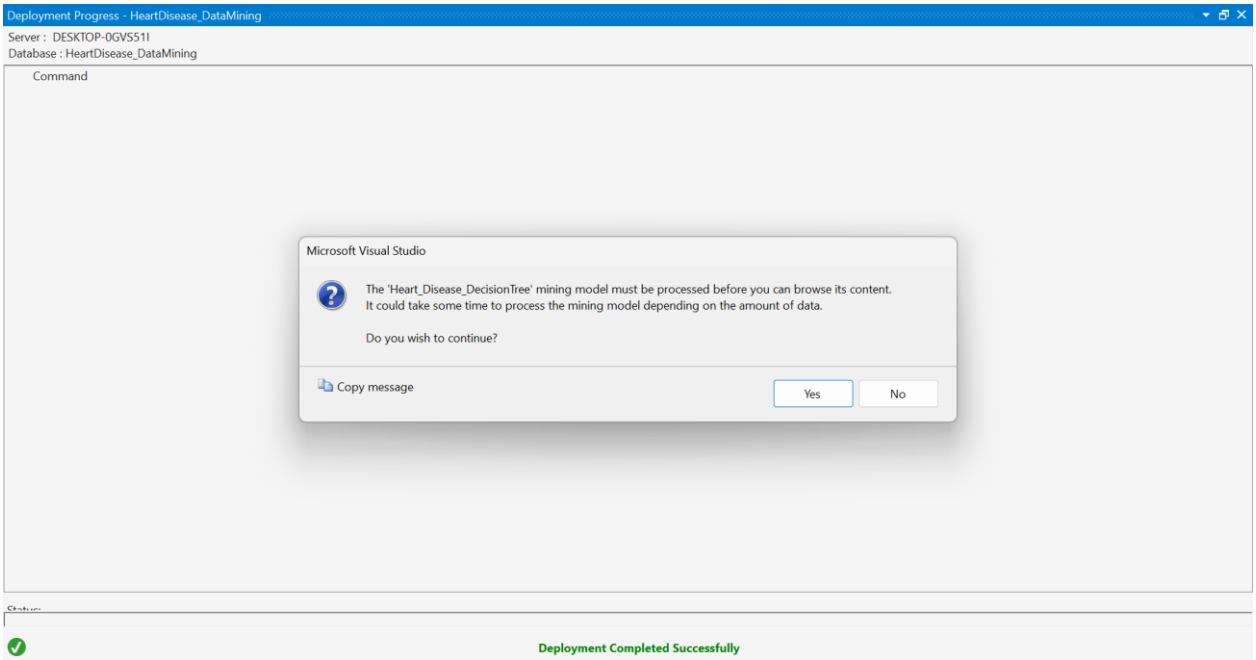


Hình 5.28 Thay tên server

Bước 18: Tiến hành deploy dự án bằng cách chọn sang tab **Mining Model Viewer** → Chọn Yes khi hộp thoại hiện lên.

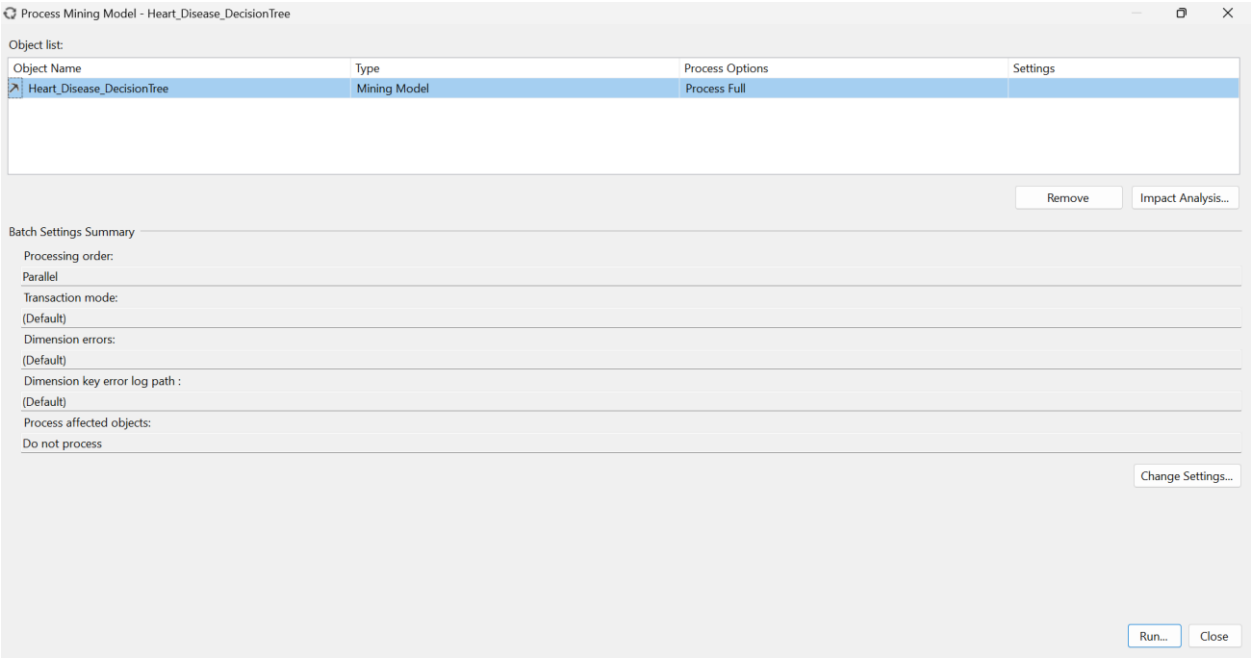


Hình 5.29 Deploy dự án

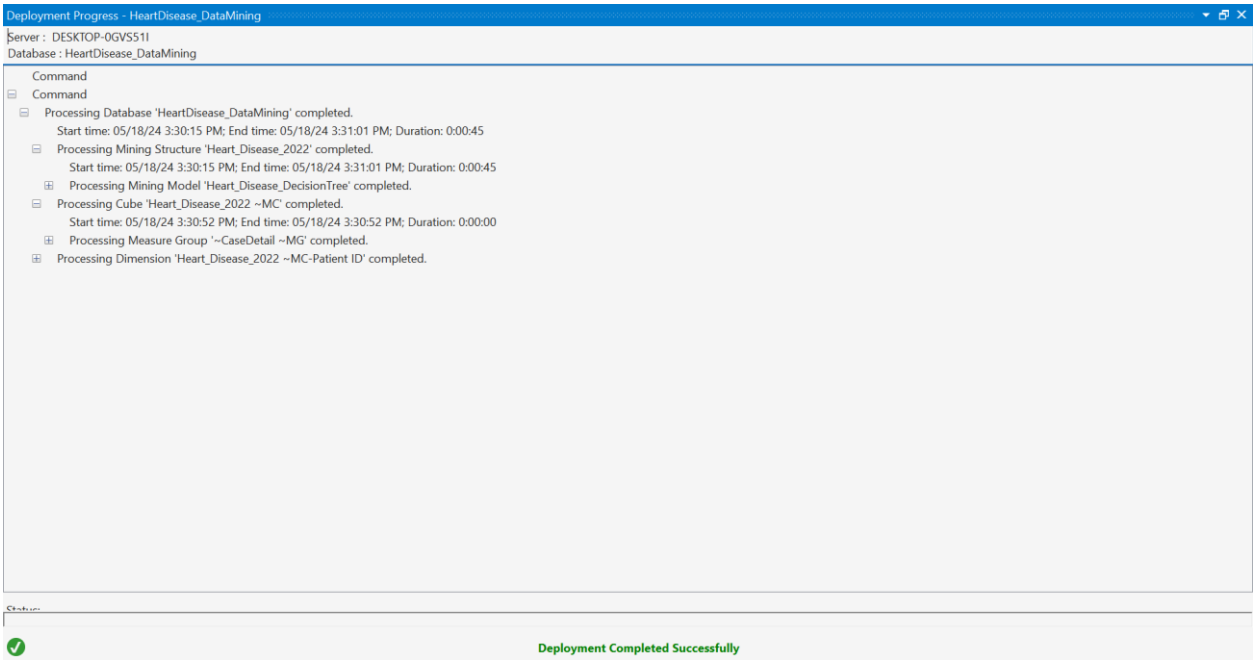


Hình 5.30 Deploy dự án thành công

Bước 19: Tiếp tục chon **Run** để chạy thuật toán cây quyết định.



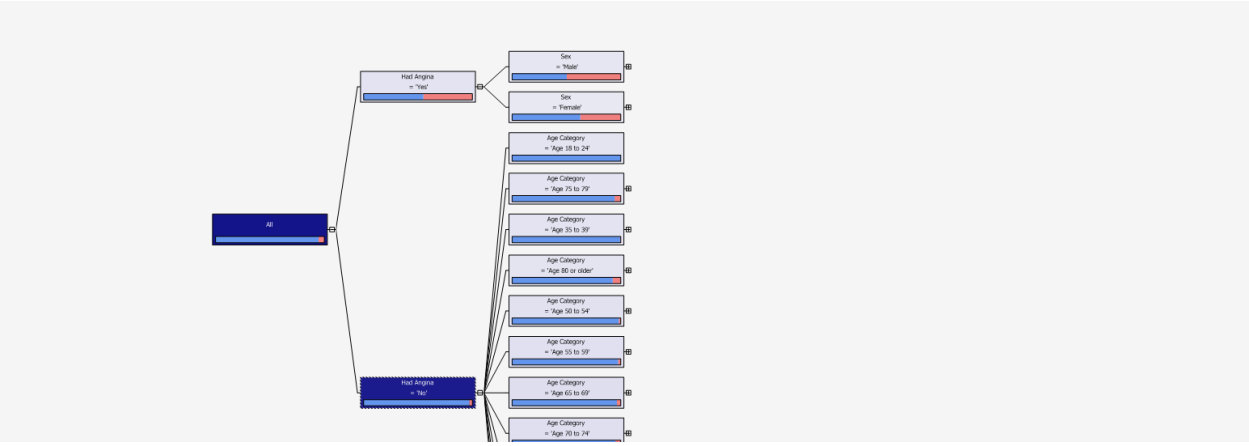
Hình 5.31 Chạy thuật toán cây quyết định



Hình 5.32 Thuật toán chạy thành công

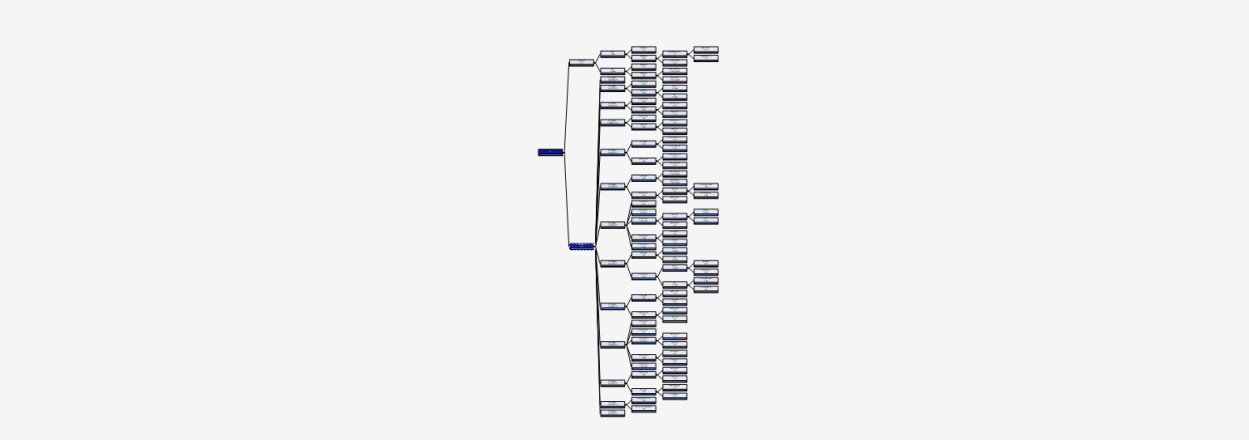
b. Phân tích mô hình Cây quyết định:

Bước 1: Sau khi chạy thuật toán cây quyết định thì mặc định trong tab Mining Model Viewer thì cây sẽ được vẽ với mức hiển thị là 3 và Background là "All".



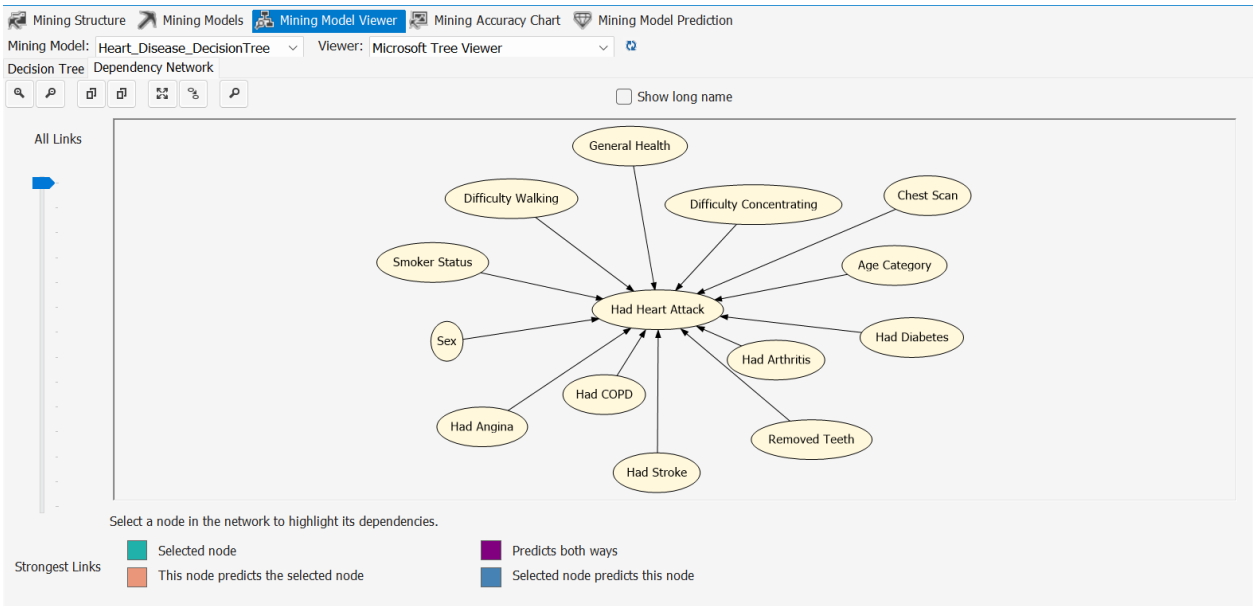
Hình 5.33 Cây quyết định mặc định

Khi ta chọn mức hiển thị cây quyết định lên mức cao nhất là **6** và **Background** ta chọn thuộc tính **HadHeartAttack** làm thuộc tính có giá trị ta muốn dự đoán. Cây quyết định sẽ được thể hiện như sau.



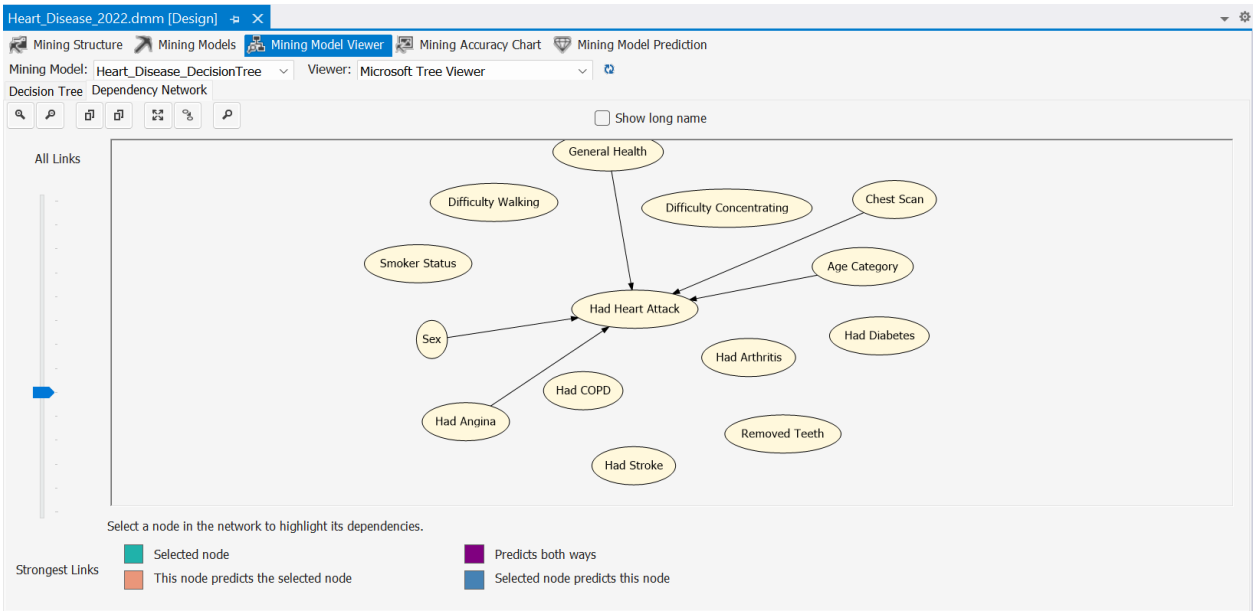
Hình 5.34 Cây quyết định đầy đủ

Bước 2: Khi chọn tab **Dependency Network** để ta có thể quan sát mạng lưới phụ thuộc của các thuộc tính với nhau. Ta có hình dưới mô tả các thuộc tính có quan hệ và có thể được dùng để dự đoán giá trị của thuộc tính dự đoán.



Hình 5.35 Các thuộc tính được dùng để dự đoán giá trị của thuộc tính dự đoán

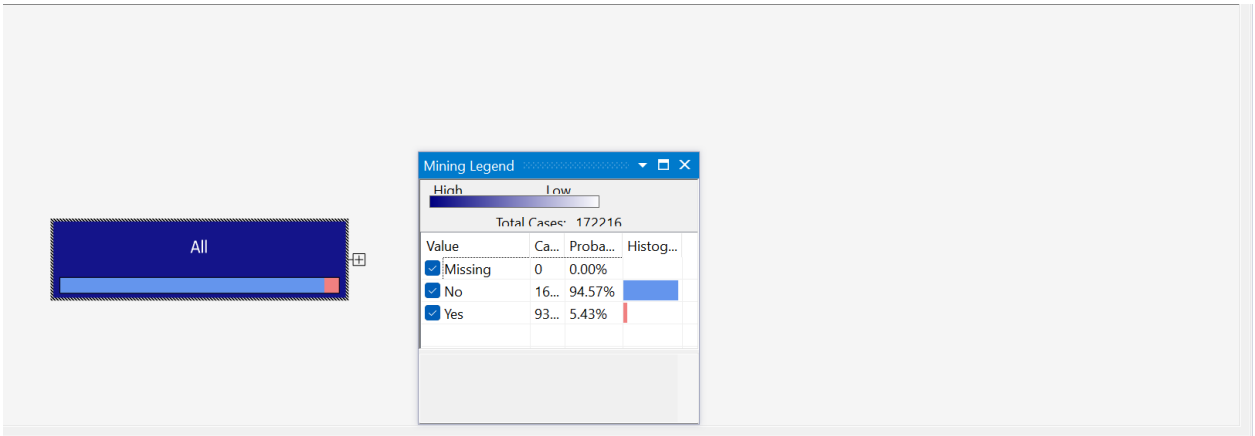
Và, khi điều chỉnh thanh **All Links** nằm dọc bên tay trái thì ta có thể biết được thứ tự các thuộc tính có sự tương quan mạnh trong việc dự đoán từ yếu đến mạnh nhất.



Hình 5.36 Sử dụng All Links để biết được mực độ mạnh yếu của thuộc tính đầu vào

Bước 3: Quay trở lại tab Mining Model View, ta phân tích cây quyết định để từ đó xác định tập luật.

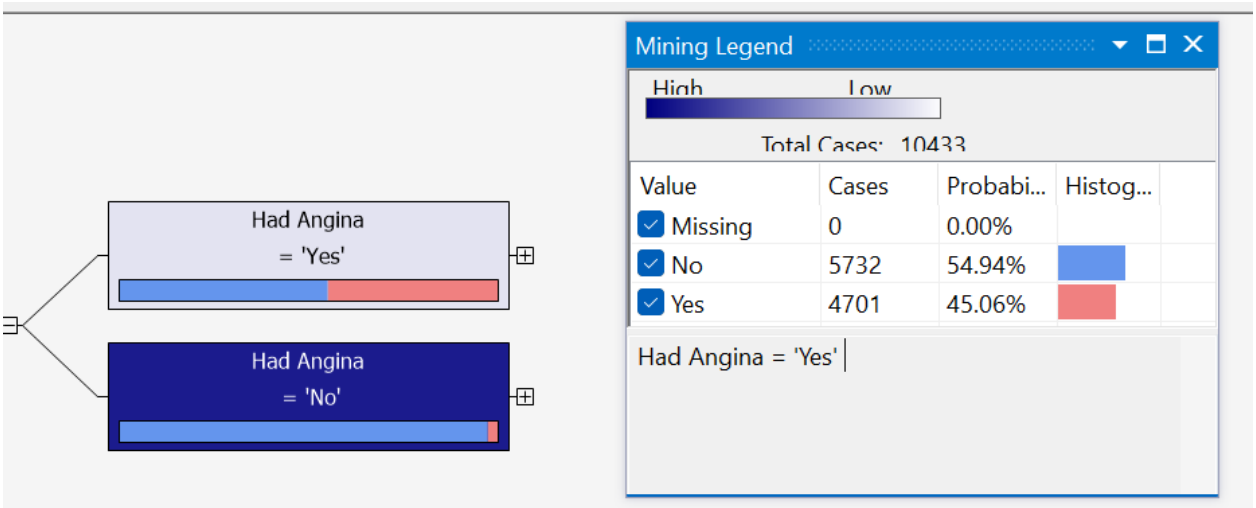
- **Tại cấp cây là 1:**



Hình 5.37 Cây ở cấp 1

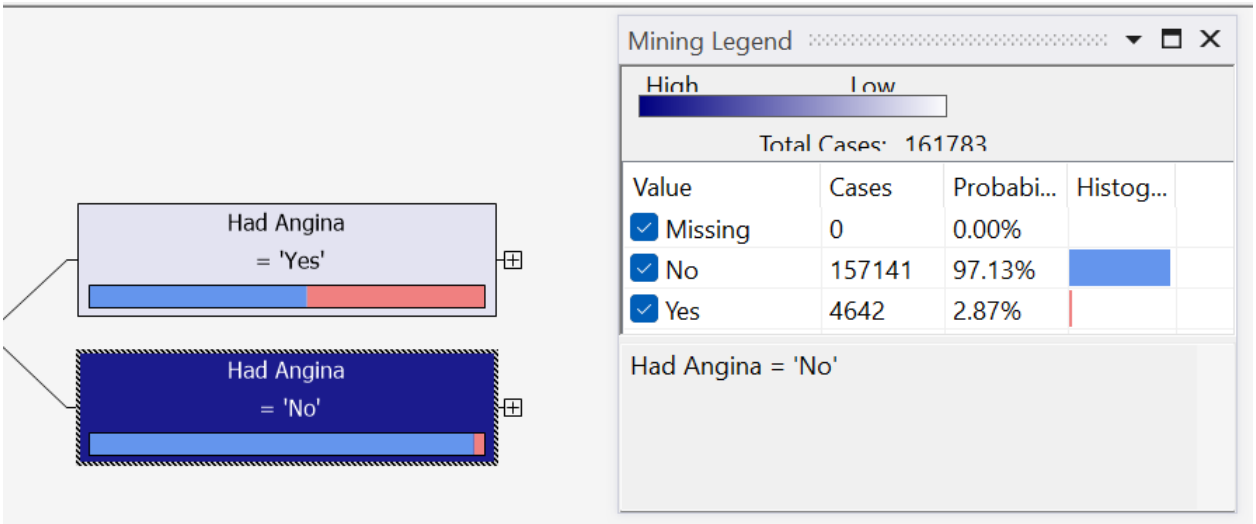
- Ta nhận thấy trong tất cả các trường hợp của tập mẫu quan sát, có **5.43%** tổng số bệnh nhân có lên cơn đau tim trong năm thống kê.

- **Xét cấp cây 2:**



Hình 5.38 Cấp ở cấp 2(1)

- Chọn thể hiện **Legend** của thuộc tính **Had Angina(Từng lên cơn đau thắt vùng ngực)** có giá trị là **Yes** thì số người không lên cơn đau tim là 54.94% và số bệnh nhân lên cơn đau tim là **45.06%.**



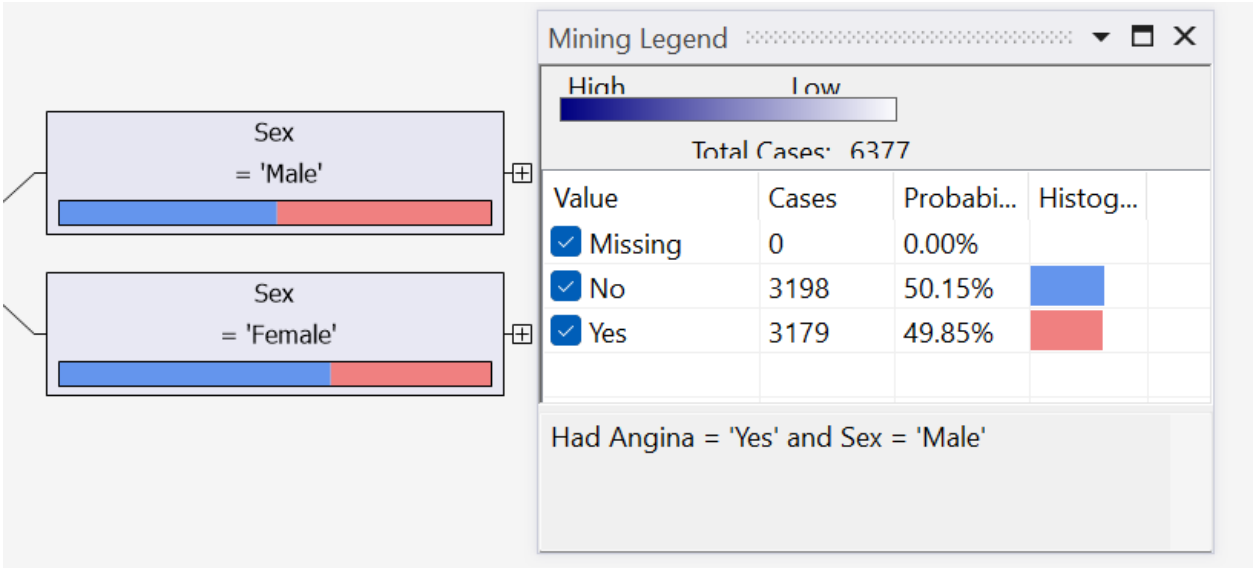
Hình 5.39 Cây ở cấp 2(2)

- Chọn thể hiện **Legend** của thuộc tính **Had Angina** có giá trị là **No** thì số người không lên cơn đau tim là 97.13% và số bệnh nhân lên cơn đau tim là **2.87%.**

⇨ Bệnh nhân có **Had Angina = "Yes"** sẽ có tỉ lệ lên cơn đau tim cao hơn.

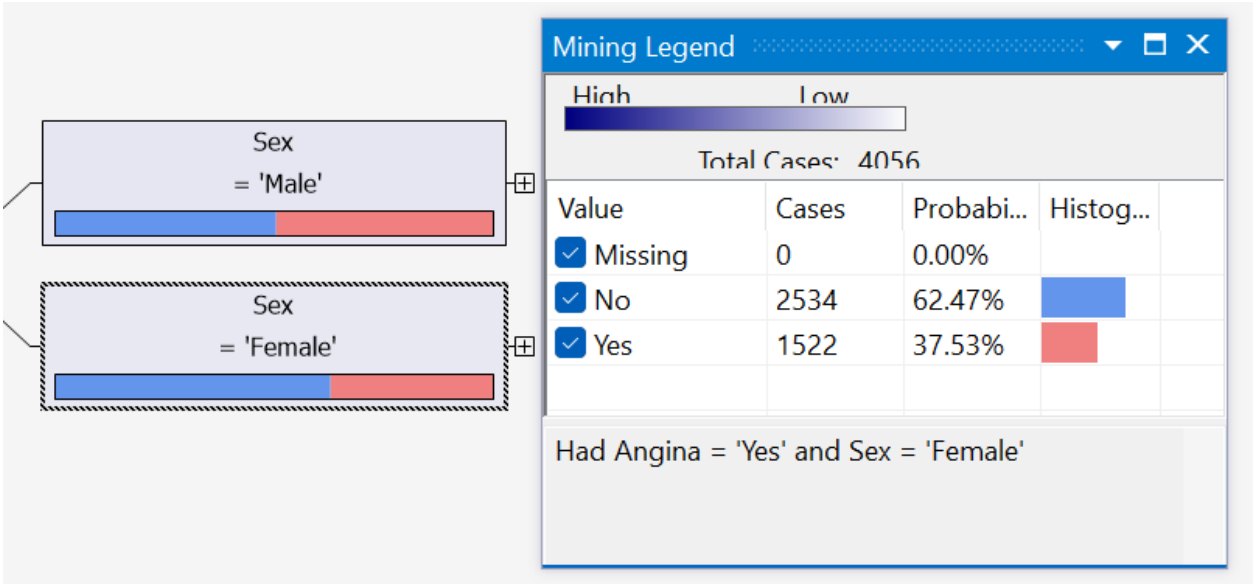
- **Xét cấp cây 3:**

  Ở đây ta sẽ xét tiếp nhánh phía **Had Angina = "Yes".**



Hình 5.40 Cây ở cấp 3(1)

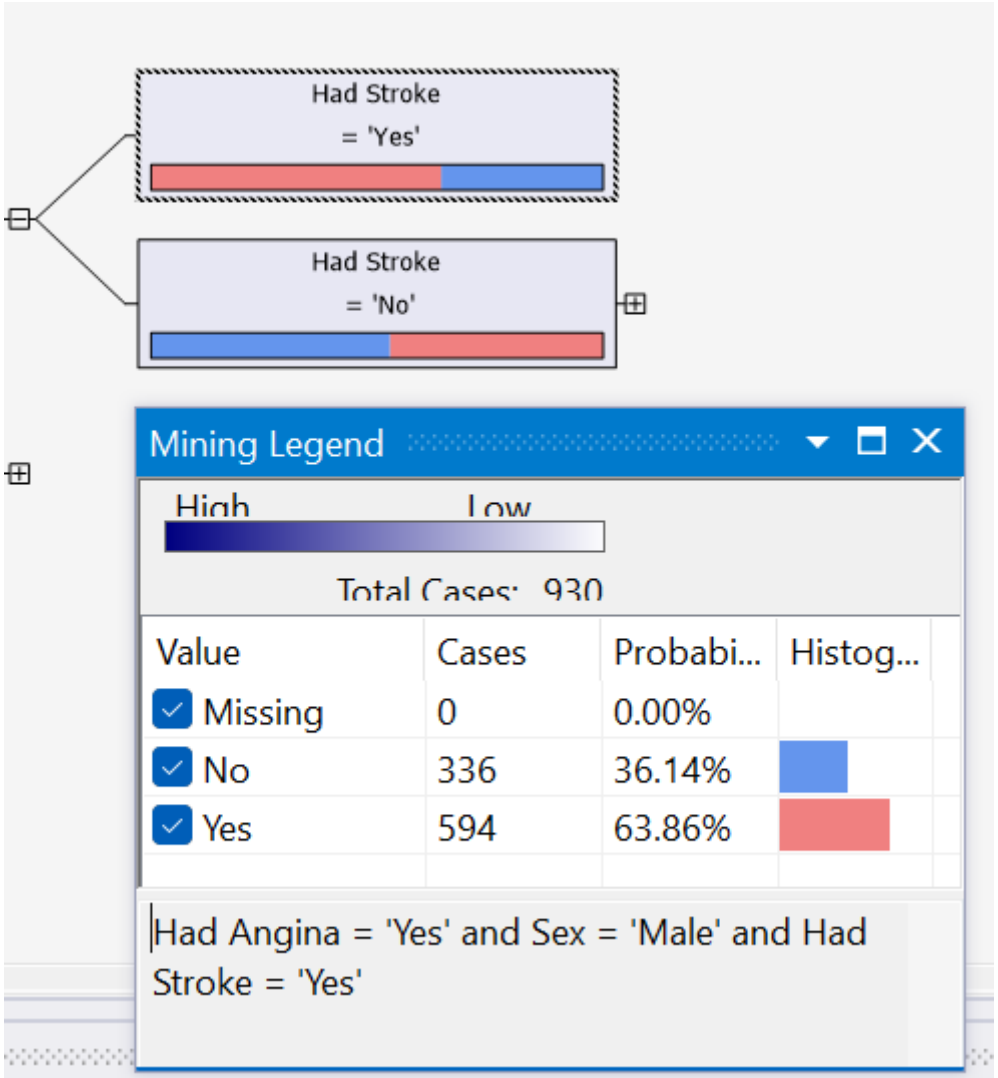
- Khi chọn hiện **Legend** của ô **Sex = "Male"** thì tỉ lệ lên cơn đau tim là **49.85%.**



Hình 5. 41 Cây ở cấp 3(2)

- Trong khi đó **Sex = "Female"** thì tỉ lệ sẽ là **37.53%.**

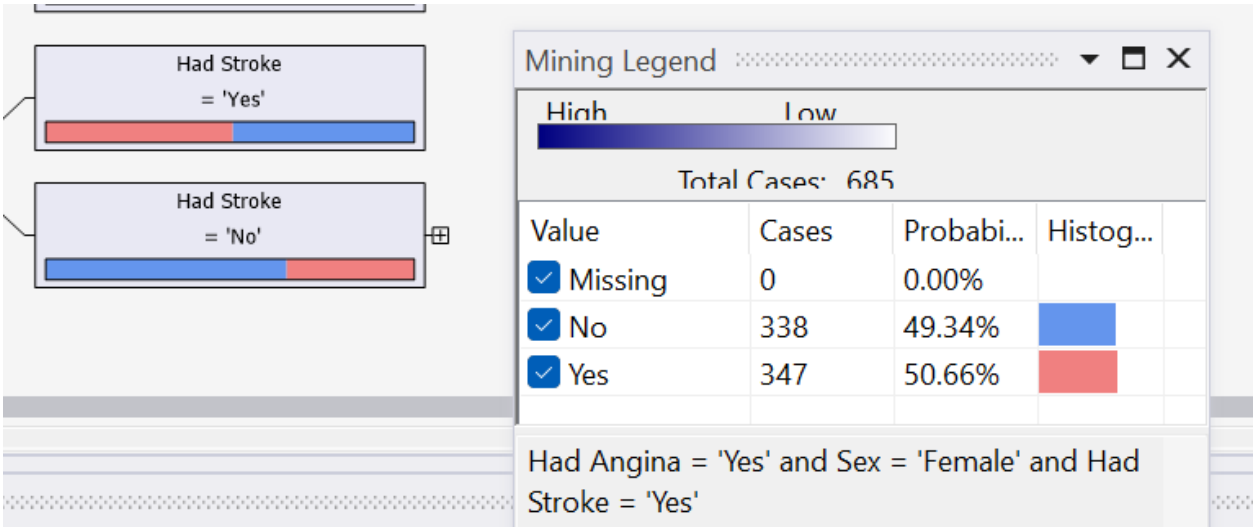
- **Xét cây ở cấp 4:**



Hình 5.42 Cây ở cấp 4(1)

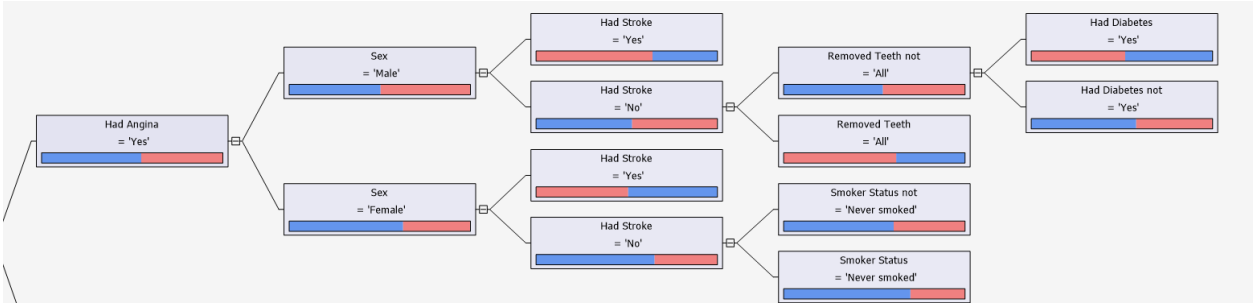
- Khi chọn ô **Had Stroke = "Yes"** của nhánh **Male** thì ta thấy tỉ lệ đau tim là **63.86%.**
- Sau Had Stroke = "Yes" ở thì không còn phân cấp nào, chứng tỏ trong cùng 1 lớp ở các lớp sau thì chỉ có 1 loại giá trị duy nhất. Như vậy ta đã tìm được một trong các tập luật:

  Had Angina = "Yes" và Sex = "Male" và Had Stroke = "Yes" → Had Heart Attack = "Yes" với xác suất thành công là **63.86%.**



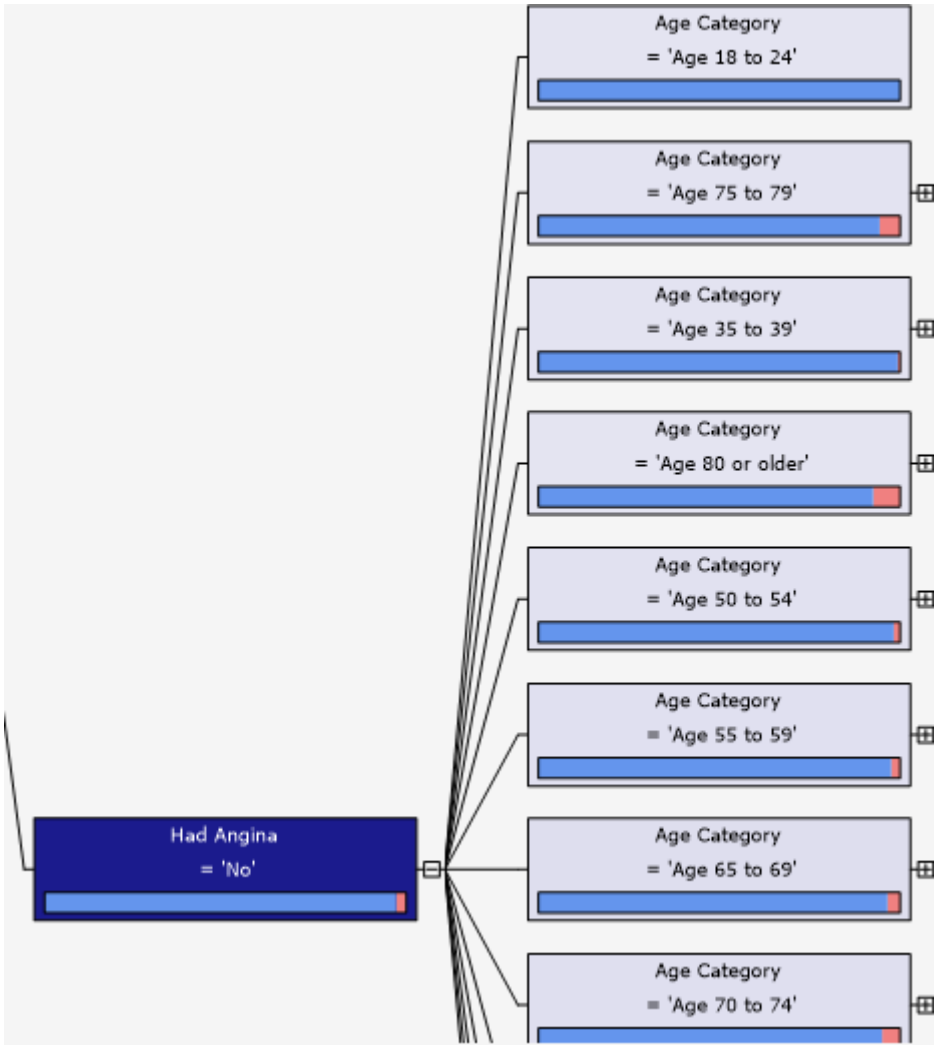
Hình 5.43 Cây ở cấp 4(2)

- Đối với **Had Stroke = "Yes"** ở nhánh **Female thì tỉ lệ bệnh tim là 50.66%.**



Hình 5.44 Cây quyết định ở nhánh Had Agina = "Yes"

- **Ngược lại, nhánh có nguy cơ không bị mắc đau tim** của nhồi máu cơ tim nhất là trường hợp những người chưa từng bị đau thắc vùng ngực và có độ tuổi trong khoảng **18 đến 24** và **35 đến 39**. Như vậy ta tìm được **tập luật**:
  - **Had Angina = "No" và Age Category = "Age 18 to 24" => Had Heart Attack = "No"** với xác suất thành công là **99.7%.**
  - **Had Angina = "No" và Age Category = "Age 35 to 39" => Had Heart Attack = "No"** với xác suất thành công là **99.23%.**



Hình 5.45 Cây ở nhánh Had Angina = "No"

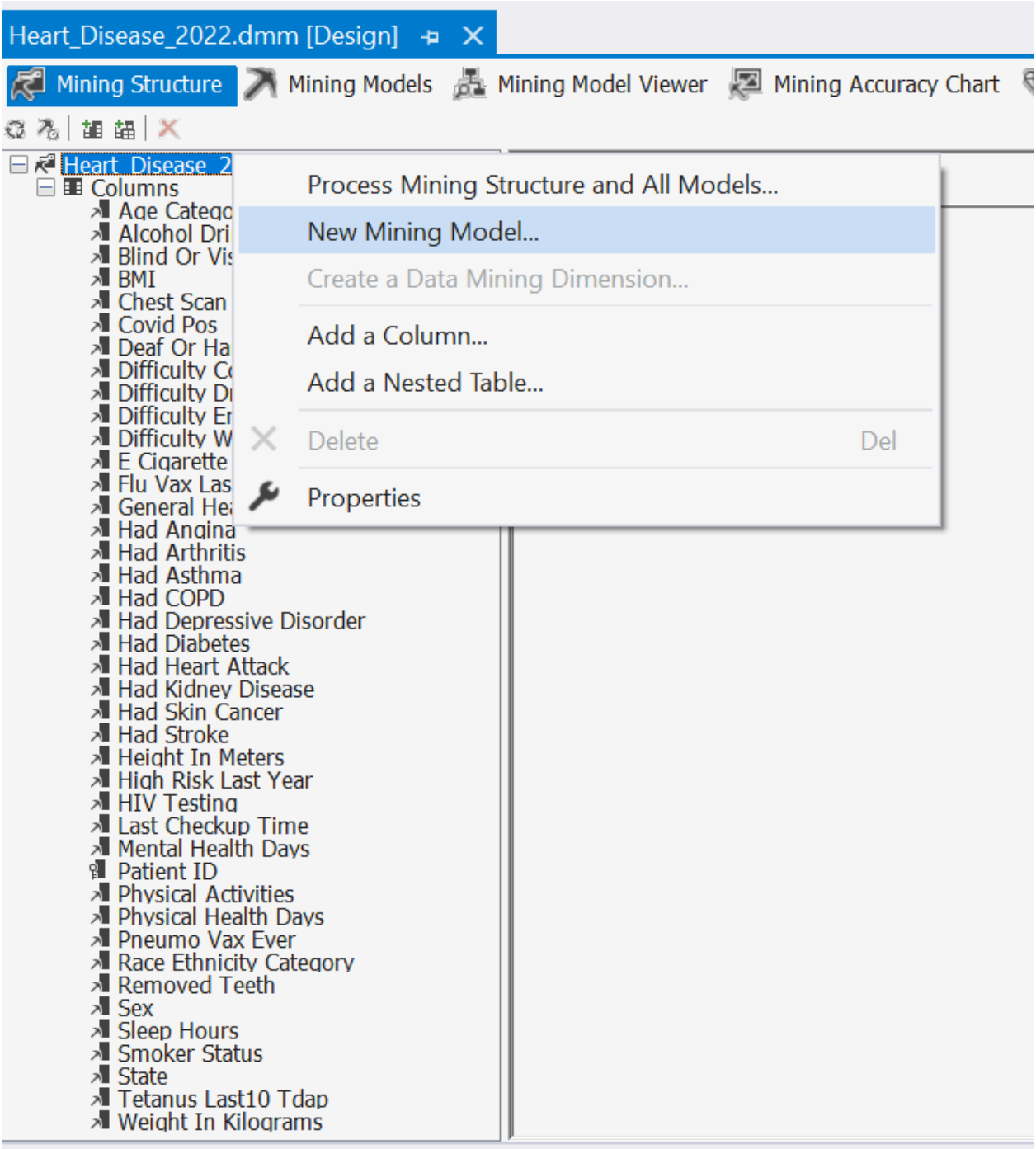
3.2.2. Xác định tập luật:

- Một số tập luật mà nhóm rút ra được đối với thuật toán Decision Trees.
  - **Had Angina = "Yes" và Sex = "Male" và Had Stroke = "Yes" → Had Heart Attack = "Yes"** với xác suất thành công là **63.86%.**
  - **Had Angina = "Yes" và Sex = "Male" và Had Stroke = "No" và Removed Teeth not = "All" và Had Diabetes = "Yes" → Had Heart Attack = "Yes"** với xác suất **51.6%**.
  - **Had Angina = 'Yes' và Sex = 'Female' và Had Stroke = 'No' và Smoker Status = 'Never smoked' → Had Heart Attack = "No"** với xác suất **69.73%**.
  - **Had Angina = 'No' và Age Category = 'Age 25 to 29' → Had Heart Attack = "No"** với xác suất **99.73%**.

3.3. Khai thác dữ liệu bằng thuật toán Clustering(gom cụm)

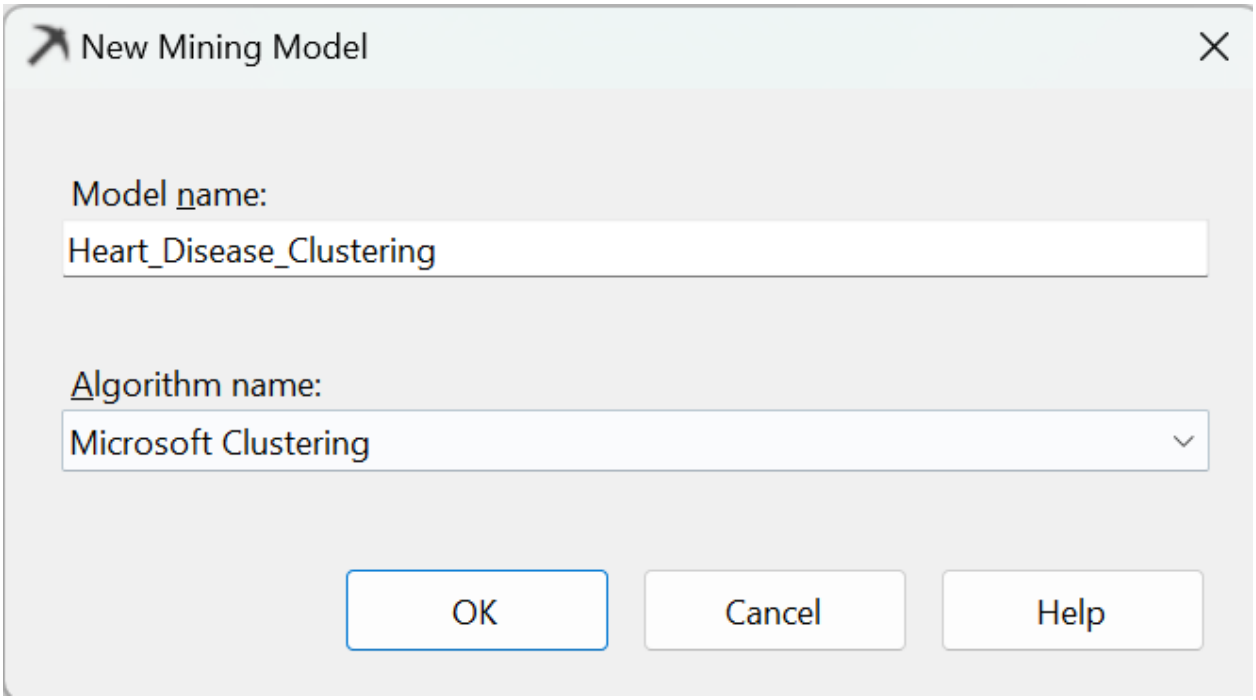
3.3.1. Khai báo mô hình

Bước 1: Trong tab Mining Structure, ta nhấn chuột phải vào tên của **Mining Structure** là **Heart_Disease_2022** và chọn New Mining Model… để thêm mô hình mới.



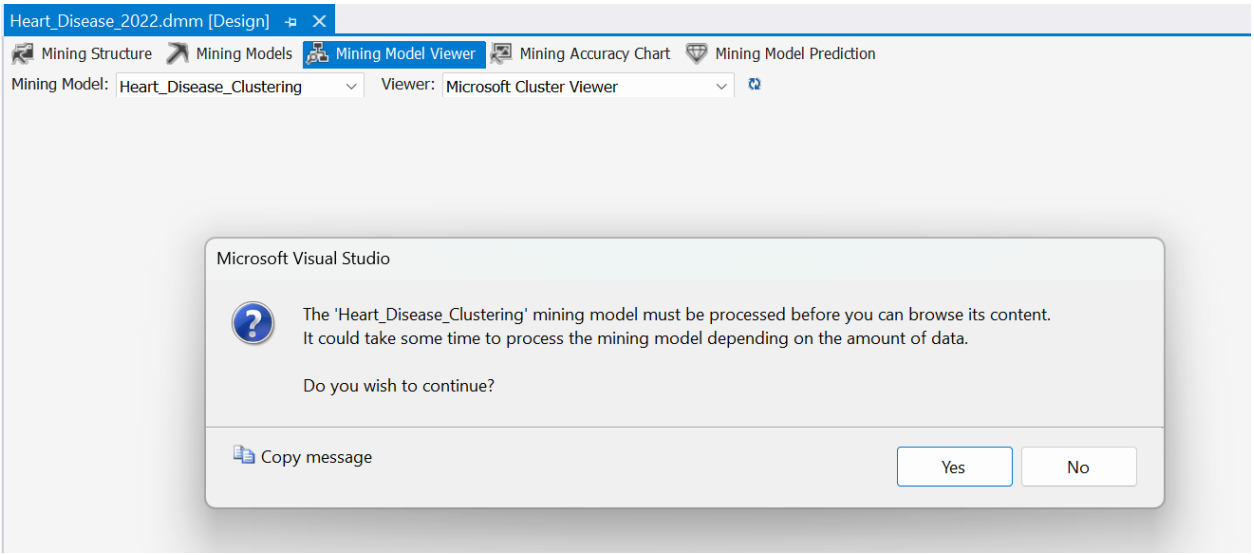
Hình 5.46 Tạo mới một mô hình khai thác dữ liệu

Bước 2: Khi xuất hiện cửa sổ **New Mining Model** thì ta đặt tên cho mô hình là **Heart_Disease_Clustering** và chọn thuật toán sử dụng là Microsoft Clustering → **OK** để hoàn thành tạo mô hình.



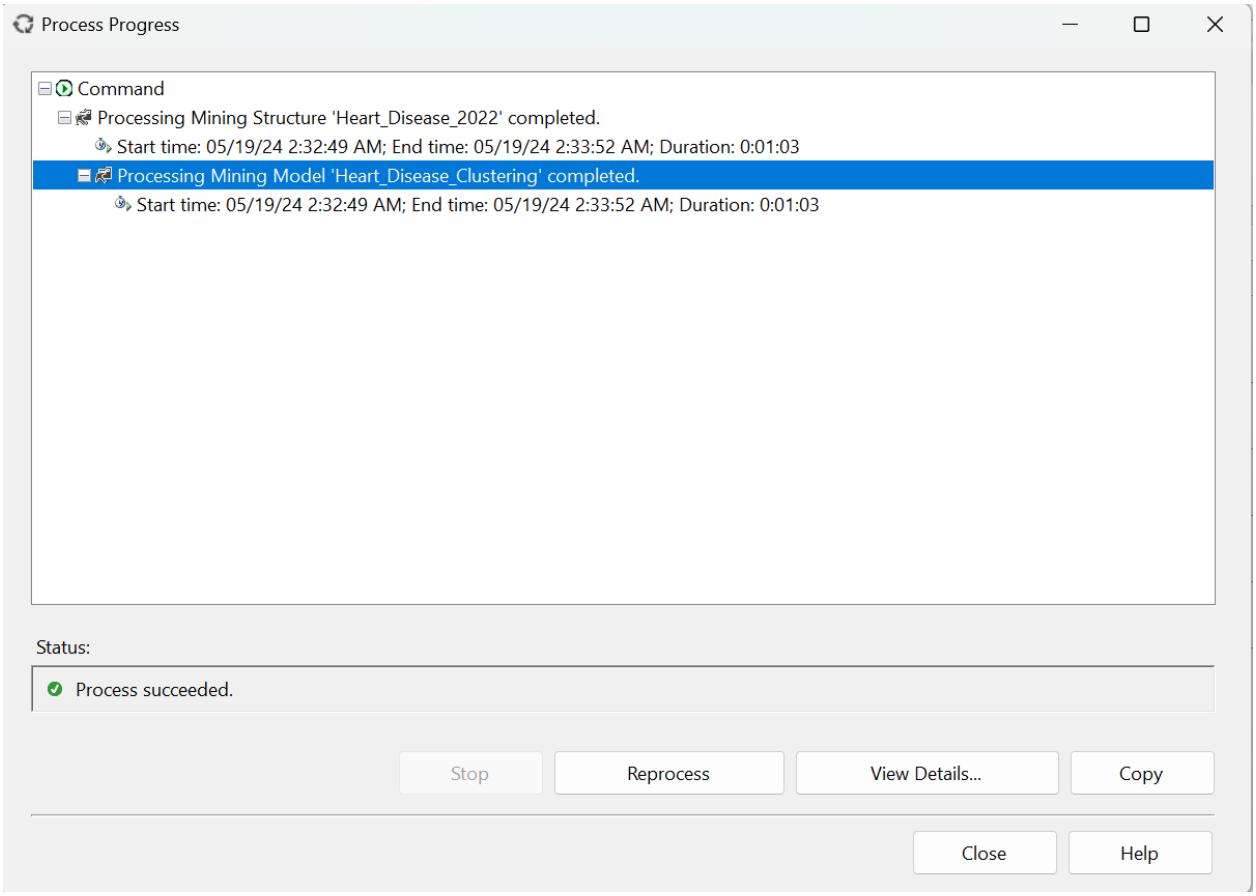
Hình 5.47 Đặt tên và chọn thuật toán

Bước 3: Sau khi tạo thành công mô hình thì khi ta nhấn qua tab Mining Model Viewer và chọn thuật toán mô hình Heart_Disease_Clustering thì hệ thống sẽ yêu cầu chúng ta cần phải chạy mô hình trước. Tại đây ta nhấn **Yes** để tiến hành chạy mô hình.



Hình 5.48 Hệ thống yêu cầu chạy mô hình

Bước 4: Sau đó hộp thoại chạy mô hình sẽ hiện lên và ta nhấn Run để bắt đầu chạy mô hình.

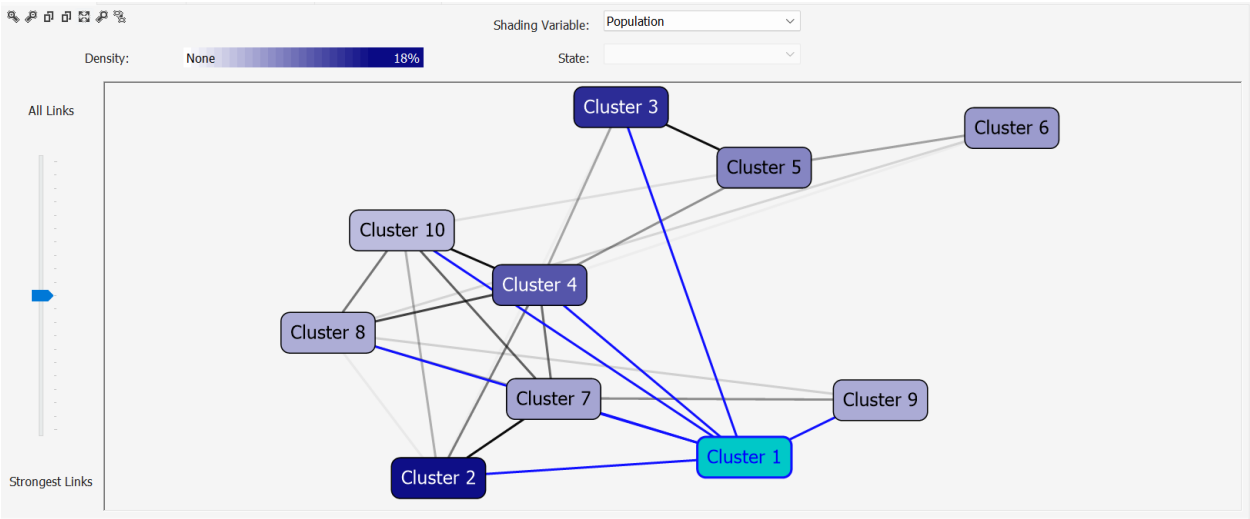


Hình 5.49 Kết quả chạy mô hình thành công

3.3.2. Phân tích mô hình:

Bước 1: Sau thuật toán chạy thành công thì ta có thể xem được kết quả mô hình, tại giao diện **Cluster Diagram.** Tại giao diện này sẽ có một vài đặc điểm ta cần lưu ý:

- **Shading Variable**: Thuộc tính ta dùng để gom cụm. Mặc định sẽ là **Population**
- **State:** Trạng thái của các thuộc tính dùng để gom cụm.
- **Thanh điều chỉnh lượng liên kết:** Thanh kéo nằm ở phía trái, ta có thể tuỳ chỉnh để hiển thị số lượng liên kết theo nhu cầu, vị trí thấp nhất đó chính là các liên kết mạnh nhất.
- **Cluster n:** Là các cụm dữ liệu thu thập được sau khi chạy thuật toán, ngoài ra màu của cụm dữ liệu còn thể hiện số lượng dữ liệu tập trung tại cụm đó → màu cụm càng đậm thì dữ liệu tập trung càng nhiều. **VD: Cluster 2**



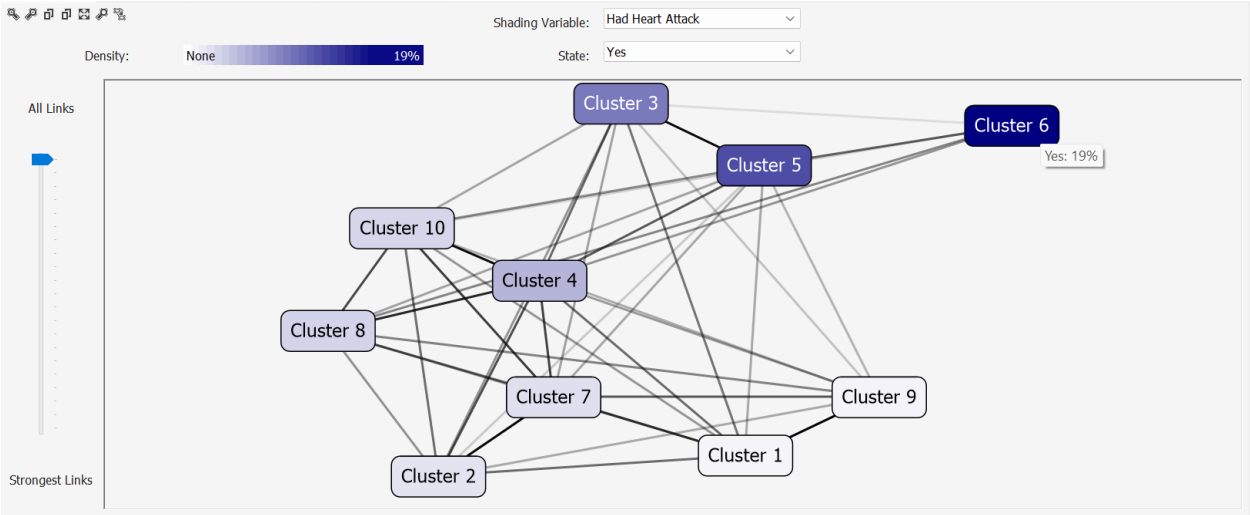
Hình 5.50 Mặc định hiển thị kết quả chạy thuật toán Clustering

Bước 2: Tuy nhiên, bởi vì thuộc tính gom cụm mà ta quan tâm là bệnh nhân có nguy cơ bị đau tim hay không vì vậy ta chọn **Shading Variable** là **Had Heart Attack** và ta quan tâm đến những trường hợp có nguy cơ đau tim vì vậy ta chọn **State** là **Yes.**

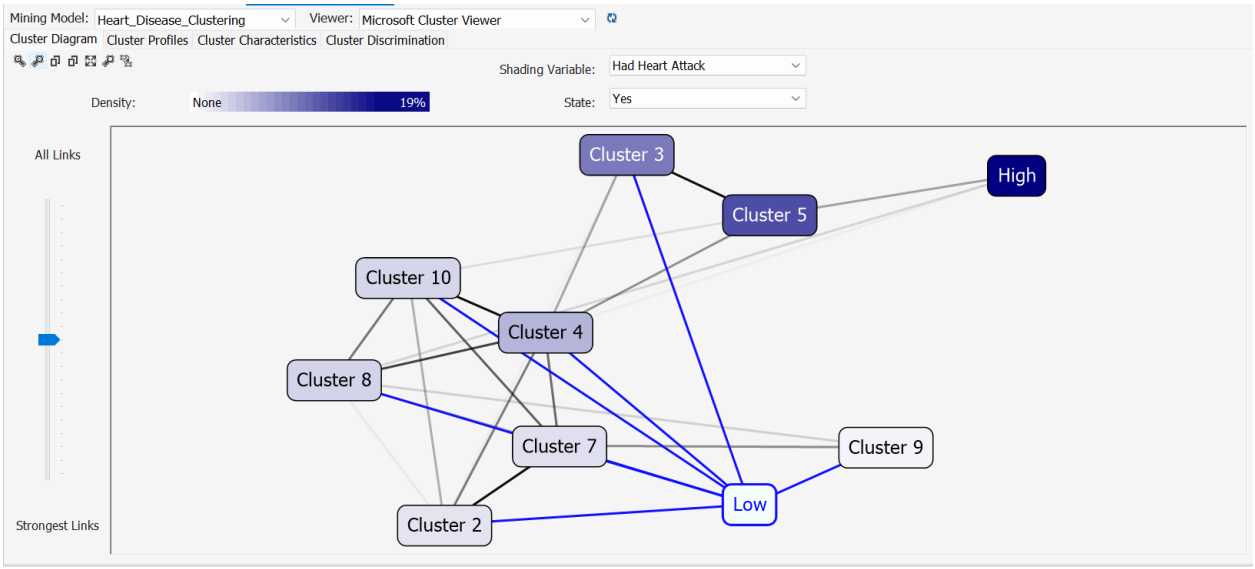
Trong trường hợp này, ta nhận thấy ở **Cluster 6** là có màu đậm nhất với tỉ lệ cao nhất là **19%** vì vậy có thể thấy ở cụm này xác suất trường hợp có đau tim là lớn nhất nên ta sẽ tiến hành nghiên cứu cụm này để rút ra được tập luật.

Ngoài ra, ở **Cluster 1** lại chỉ có tỉ lệ trường hợp có bị đau tim là 1% vì vậy ta cũng có thể nghiên cứu cụm này để rút ra được tập luật của những trường hợp không có nguy cơ đau tim.

⇨ Để thuận tiện thì ta sẽ đặt lần lượt **cụm 6** và **cụm 1** là **High** và **Low**.



Hình 5.51 Các cụm dữ liệu có tỉ lệ dữ liệu nổi bật

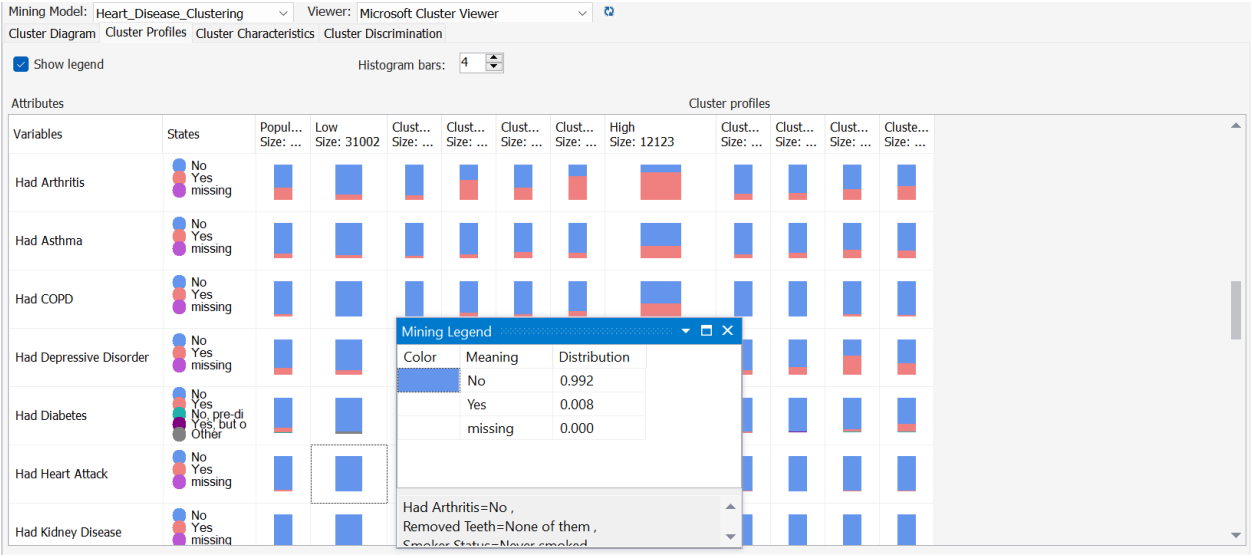


Hình 5.52 Đặt tên lại cho cụm 1 và cụm 6

Bước 3: Khi chọn sang tab **Cluster Profile** thì tại đây ta có thể thấy được số chi tiết của các cụm dữ liệu: Số dòng dữ liệu trong cụm, tỉ lệ chi tiết trong các thuộc tính của mỗi cụm.

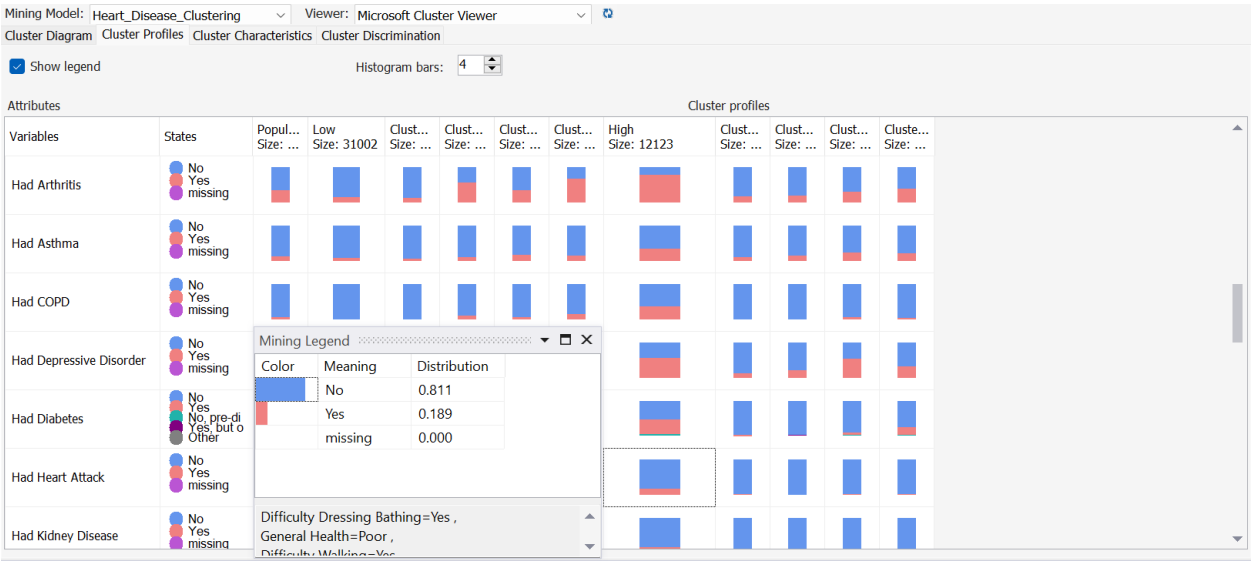
Hình 5.53 Cluster Profile hiển thị thông tin chi tiết các cụm

Bước 4: Khi ta chọn vào thuộc tính **Had Heart Attack** của **cụm Low** thì ta có thể thấy được trong cửa sổ **Mining Legend** là tỉ lệ của nhãn **No** là **99.2%** và phía dưới đó chính là chi tiết giá trị của các thuộc tính khác trong **cụm Low**.



Hình 5.54 Chi tiết của thuộc tính Had Heart Attack và cụm Low

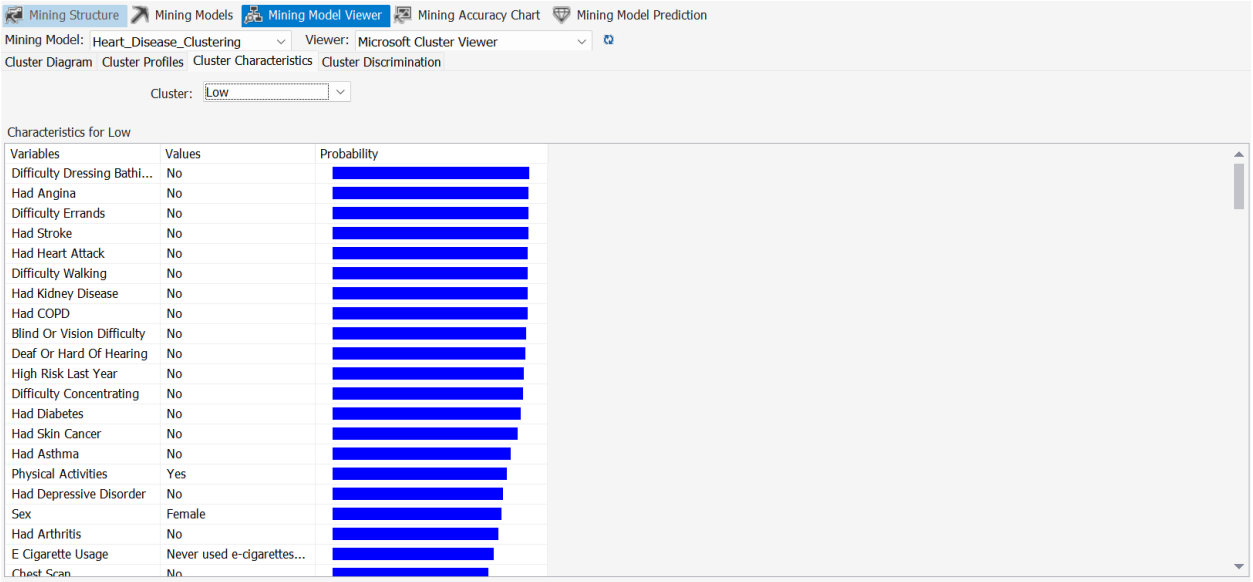
Tương tự với thuộc tính gom cụm **Had Heart Attack** của **cụm High** thì ta thấy tỉ lệ của nhãn **Yes** chiếm lên tới **18.9%** tức là tỉ lệ của người mắc bệnh cao hơn nếu có những giá trị của các thuộc tính nằm trong **cụm High**.



Hình 5.55 Chi tiết của thuộc tính Had Heart Attack và cụm Low

Bước 6: Tiếp theo, trong tab **Cluster Characteristics** ta có thể thấy được tỉ lệ của các nhãn của các thuộc tính được thể hiện chi tết đối với cụm đang được chọn.

Hình 5.56 Cluster Characteristics của cụm High

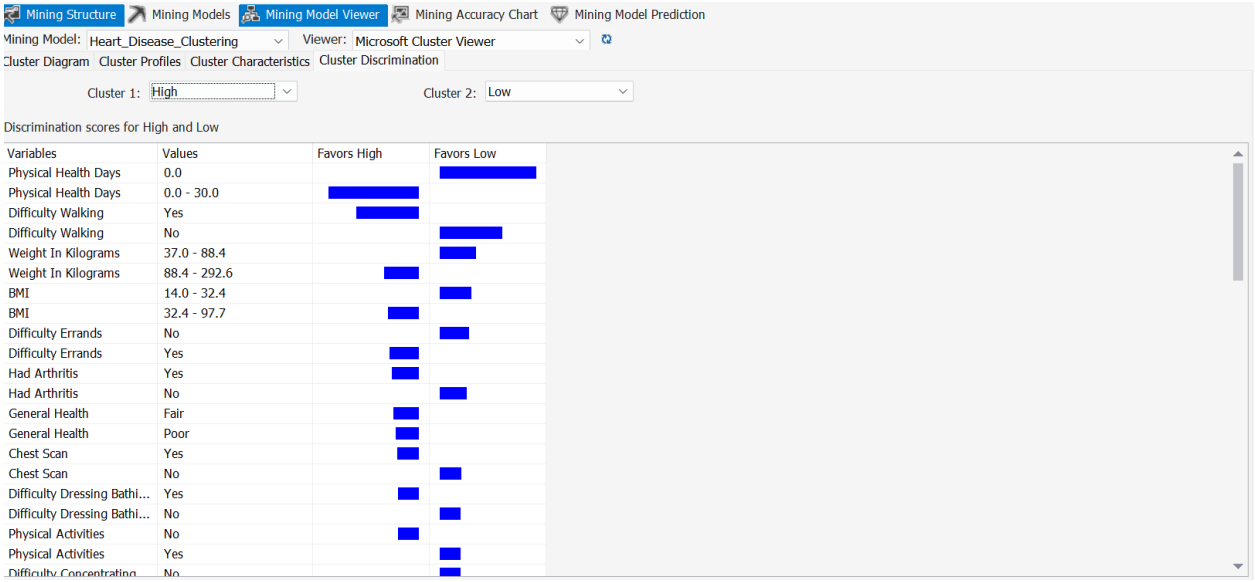


Hình 5.57 Cluster Characteristics của cụm Low

Bước 7: Khi chọn sang tab **Cluster Discrimination** thì ta sẽ thấy được các điểm phân biệt giữa hai cụm dữ liệu mà ta chọn. Ở đây ta chọn hai cụm mà ta cần quan sát sự khác biệt là đó chính là **cụm High** và **cụm Low**.

Một vài điểm khác biệt mà ta có thể nhận thấy đó chính là:

- Đối với thuộc tính **Difficult Walking** trong **cụm High(có tỉ lệ lên cơn đau tim cao hơn)** có giá trị = **"Yes"** sẽ thiên về **cụm High** và ngược lại đối với giá trị **Difficult Walking = "No"** thì cũng thiên về **cụm Low**.



Hình 5.58 Discrimination Scores của 2 của cụm High và cụm Low

3.3.3. Xác định tập luật:

- Đối với cụm High thì ta có tập luật ta có được: một người có 19% nguy cơ bị đau tim nếu có những đặc điểm sau.

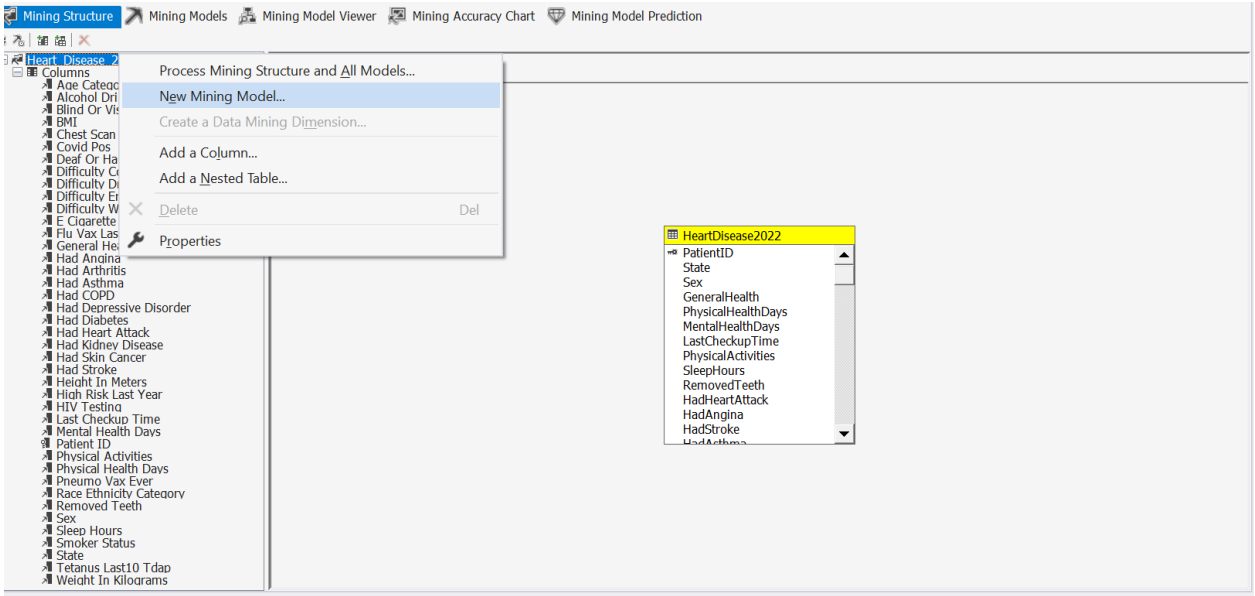
Difficulty Dressing Bathing = Yes	**General Health=Poor | General Health=Fair**	**Difficulty Walking=Yes**	**Difficulty Errands=Yes**
Had COPD=Yes	**Difficulty Concentrating=Yes**	**0 <=Physical Health Days <=30**	**Age Category=Age 60 to 64 | Age Category=Age 55 to 59**
Physical Activities=No	**Mental Health Days=30**	**Had Depressive Disorder=Yes**	**Blind Or Vision Difficulty=Yes**
Blind Or Vision Difficulty=Yes	**Had Arthritis=Yes**	**Had Diabetes=Yes**	**Had Stroke=Yes**
Had Angina=Yes	**Removed Teeth=All**	**Had Kidney Disease=Yes**	**Chest Scan=Yes**

Removed Teeth=6 or more, but not all	**Had Asthma=Yes**	**Alcohol Drinkers=No**	**Deaf Or Hard Of Hearing=Yes**
Pneumo Vax Ever=Yes	**Smoker Status=Current smoker - now smokes every day**	**HIV Testing=Yes**	**Mental Health Days=15**
…	**….**	**….**	**…**

3.4. Thực hiện khai thác dữ liệu với thuật toán Naïve Bayes

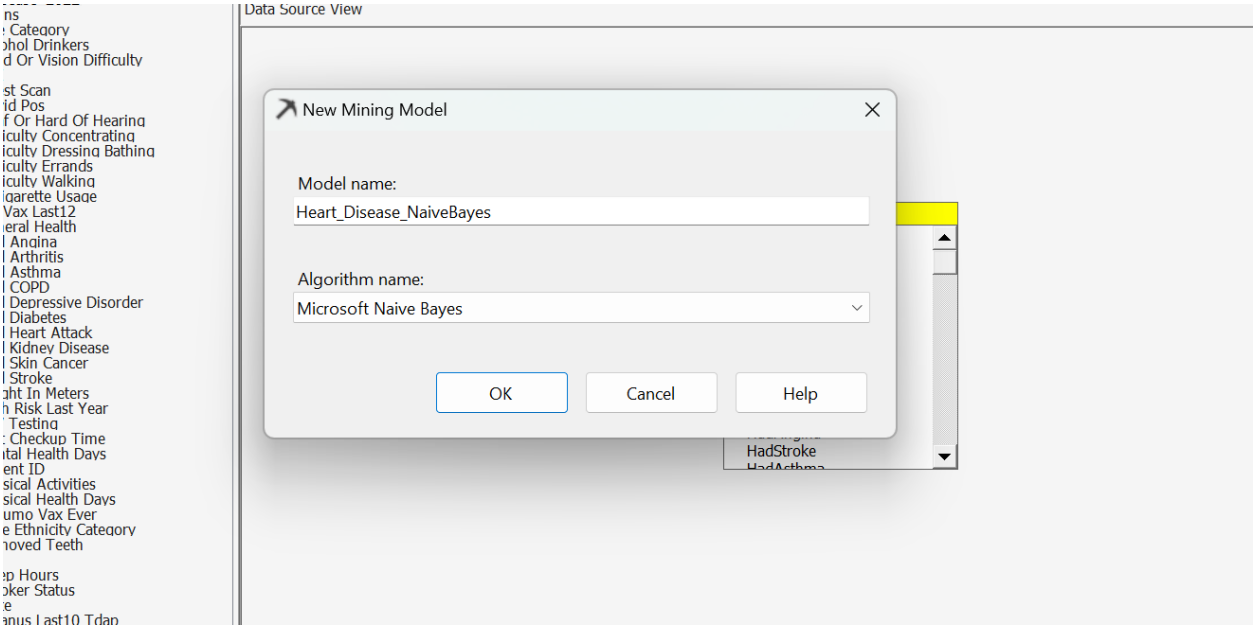
3.4.1. Khai báo mô hình

Bước 1: Trong tab **Mining Structures**, nhấn chuột phải vào tên Mining Structures và chọn New Mining Model để tiến hành thêm mới một mô hình.



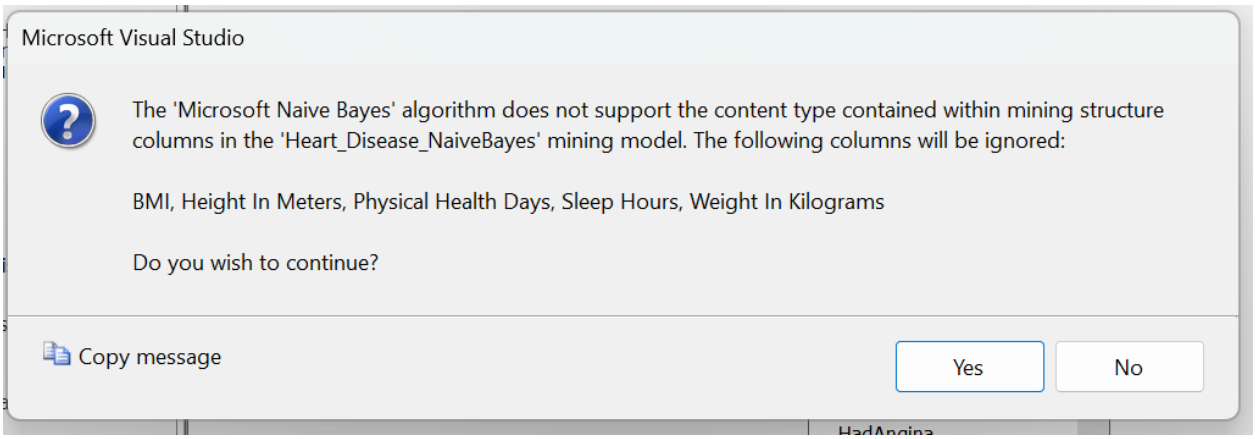
Hình 5.59 Thêm mới một mô hình

Bước 2: Đặt tên cho mô hình là **Heart_Disease_NaiveBayes** và chọn thuật toán cho mô hình là **Naïve Bayes**. Nhấn OK để hoàn thành tạo mô hình.



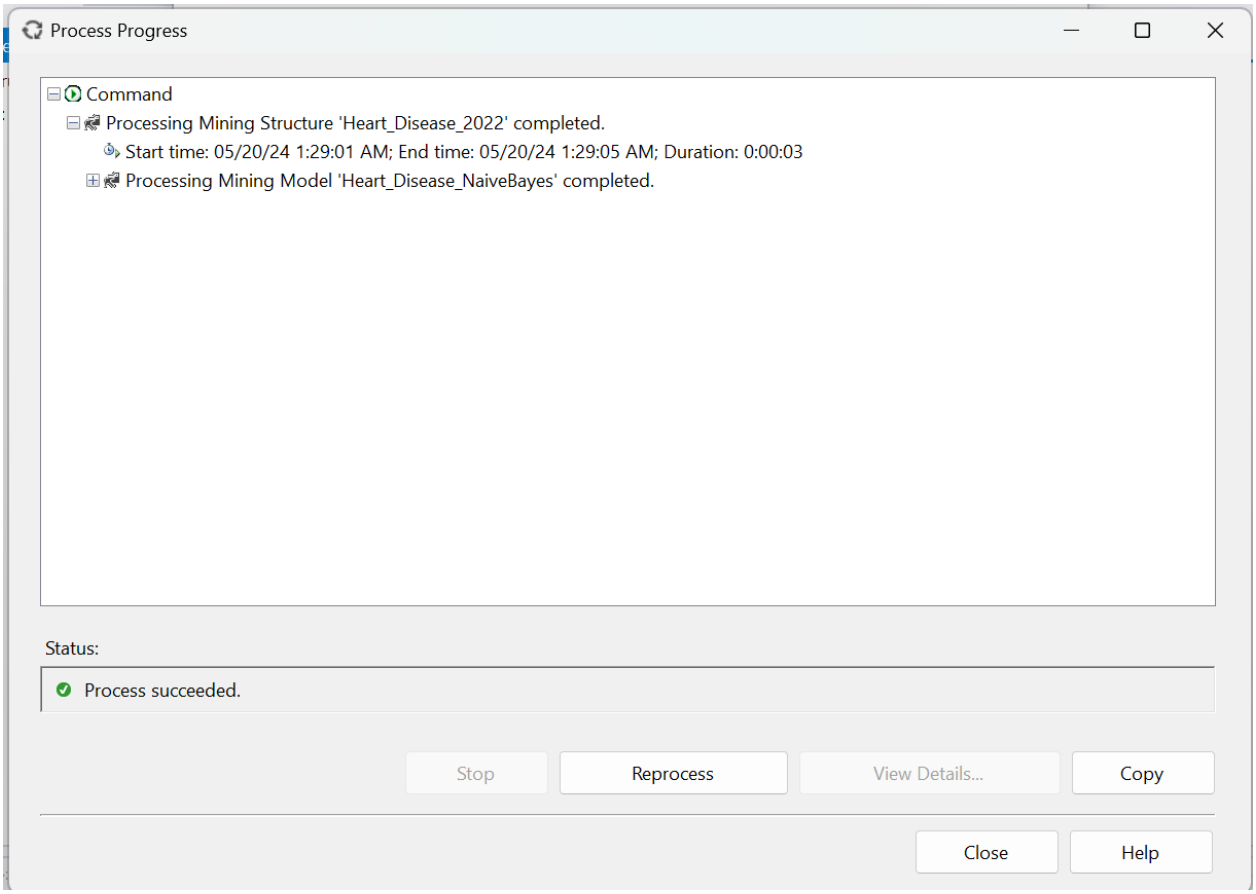
Hình 5.60 Đặt tên và chọn thuật toán cho mô hình

Bước 3: Sau khi tạo mô hình và chuyển sang tab **Mining Model Viewer** thì hệ thống sẽ thông báo chúng ta cần phải tiến hành chạy thuật **Naïve Bayes** cho mô hình trước. Nhấn **Yes** để đồng ý.



Hình 5.61 Hệ thống yêu cầu chạy thuật toán

Bước 4: Sau đó hộp thoại chạy thuật toán sẽ hiển thị và ta sẽ nhấn **Run** để bắt đầu chạy thuật toán trên bộ dữ liệu mà ta đã chọn.



Hình 5.62 Kết quả chạy thuật toán thành công

3.4.2. Phân tích mô hình

Bước 1: Khi nhấn vào tab Mining Model Viewer và chọn Model **Heart_Disease_NaiveBayes**. Khi ở tab **Dependency Network** thì ta sẽ có được kết quả như hình sau.

Hình 5.63 Mạng lưới thể hiện sự tác độngcủa các thuộc tính đầu vào đối với thuộc tính dự đoán

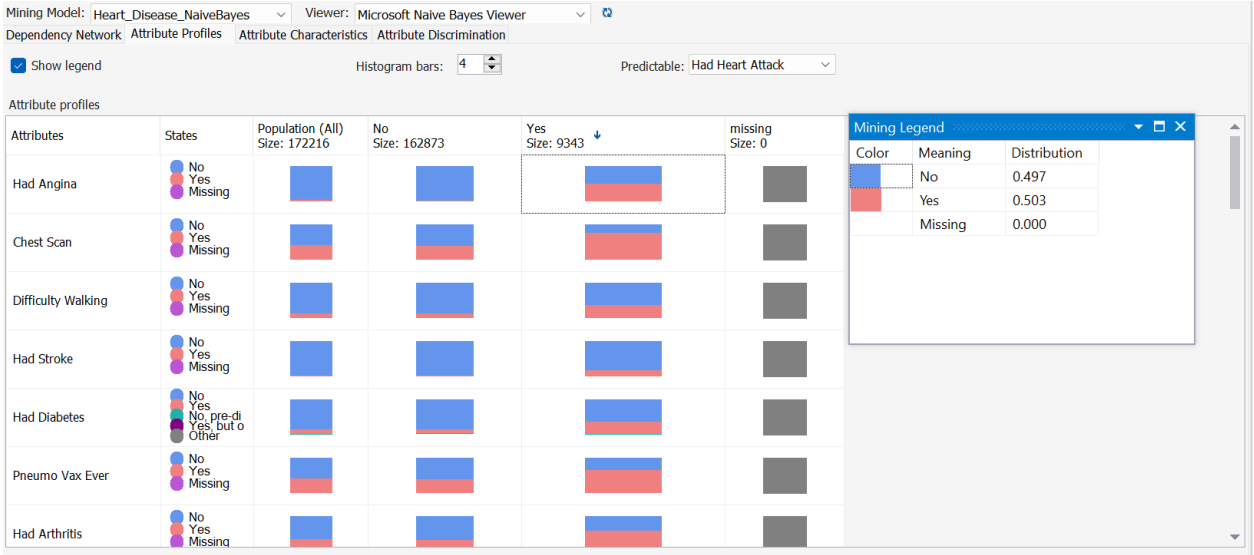
Trong đó, thuộc tính có sự liên kết mạnh nhất(có tác động mạnh nhất) đối với giá trị dự đoán là Had Angina.

Hình 5.64 Thuộc tính có liên kết mạnh nhất

Bước 2: Khi nhấn sang tab Attribute Profiles, ta có thể xem được tỉ lệ các nhãn(các giá trị) của các cột thuộc tính đầu vào khi xét đến thuộc tính dự đoán.

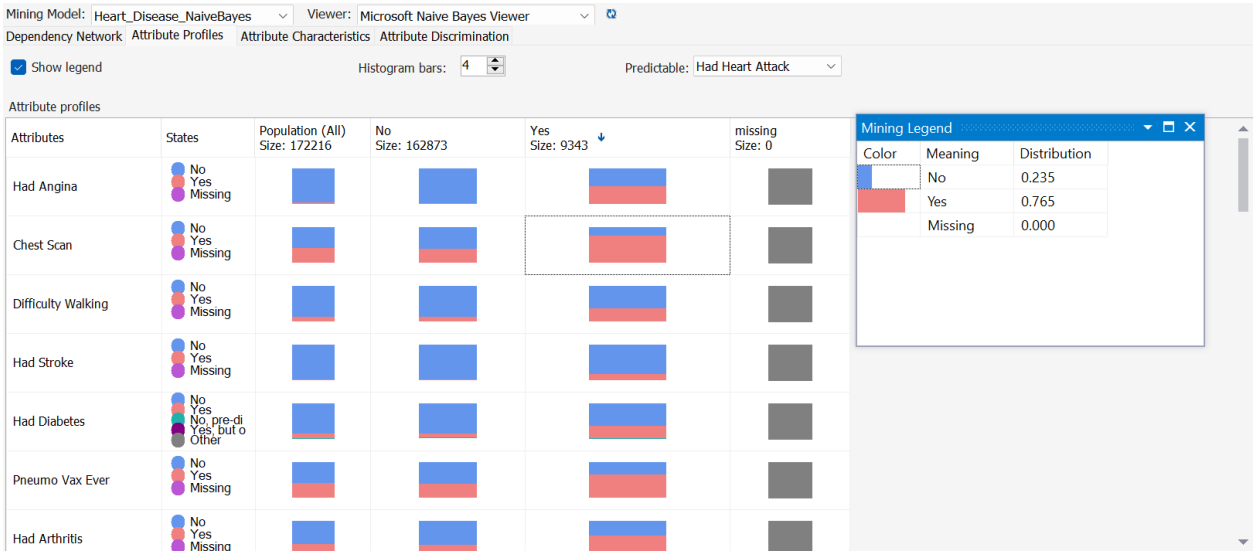
Ví dụ:

- Khi nhấn ô giao giữa **Had Angina** và **Yes** thì ta có thể thấy tỉ lệ những bệnh nhân có lên cơn đau tim thì sẽ thuộc vào nhóm **50,3%** người từng bị đau thắt vùng ngực và **49,7%** là không bị.



Hình 5.65 Tỉ lệ người từng bị đau thắt vùng ngực đối với biến dự đoán

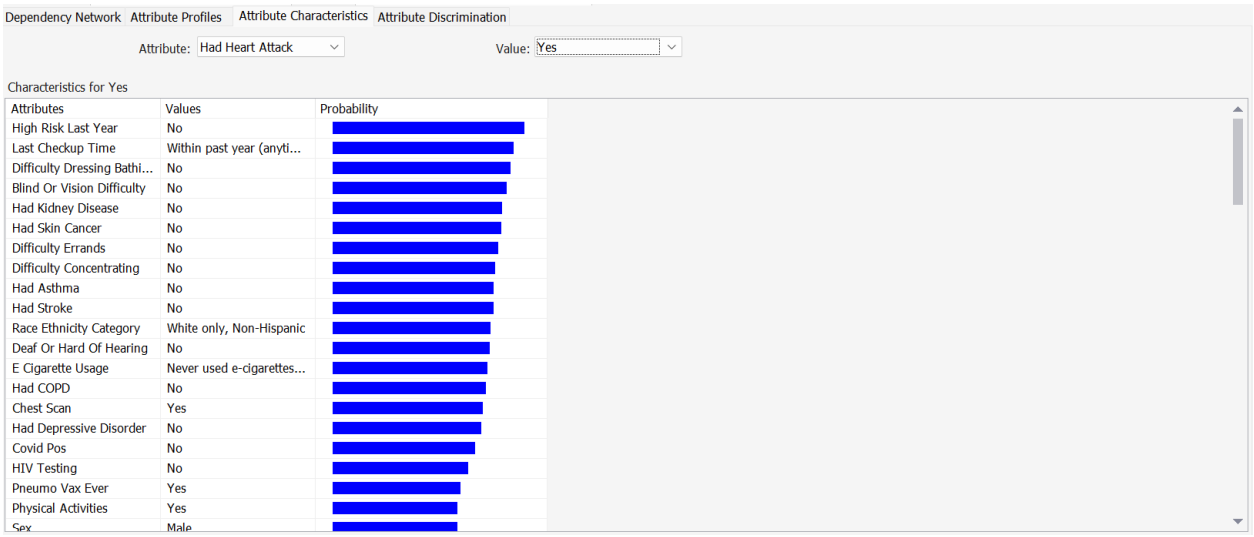
- Tương tự khi ta chọn vào ô Chest Scan thì ta có thể thấy trong số 9343 những người bị lên cơn đau tim thì 76.5% trong số đó đã từng đi chụp cắt lớp vi tính vùng ngực(Chest Scan = "Yes") và 23.5% còn lại là không đi.



Hình 5.66 Tỉ lệ người từng đi chụp cắt lớp vi tính vùng ngực đối với biến dự đoán

Bước 3: Khi chọn sang **tab Attribute Characteristics → Value: Yes** thì ta có thể thấy đối với thuộc tính dự đoán Had Heart Attack có giá trị "Yes" thì các thuộc tính High Risk Last Year = "No", Difficulty Dressing Breathing = "No", Had Kidney Disease = "No",… đều chiếm tỉ lệ rất cao. Một vài đặc điểm có thể rút ra được về những người có lên cơm đau tim đó chính

là: Họ không có nguy cơ bị cao năm trước đó, không bị khó thở vùng ngực, không bị khó mặc quần áo khi tắm, có tiền sử bị bệnh thận.

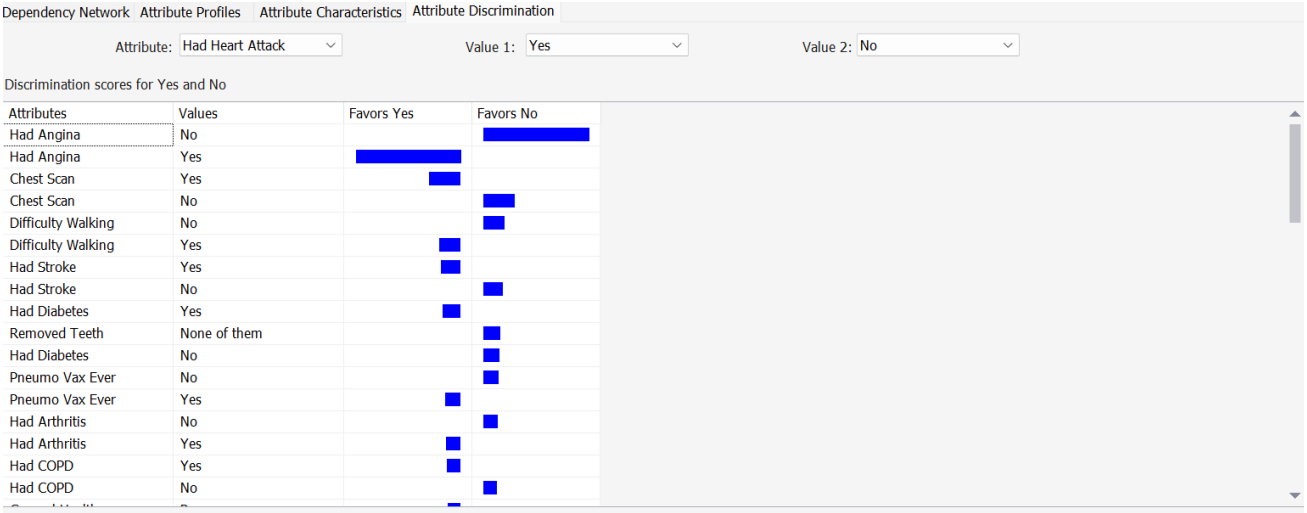


Hình 5.67 Tỉ các giá trị các thuộc tính đầu vào khi xét tới giá trị thuộc tính dự đoán

Bước 4: Tiếp tục chọn qua tab **Attribute Discrimination** để xem xét sự tương quan của các giá trị các thuộc tính đầu vào với giá trị và các giá trị của thuộc tính dự đoán được xét. Tỉ lệ tương quan sẽ được sắp xếp giảm dần.

Quan sát hình bên dưới ta có thể rút được nhận xét:

- Đối với những bệnh nhân có đặc điểm **Had Angina = "Yes", Chest Scan = "Yes", Difficulty Walking = "Yes", Had Stroke = "Yes", Had Diabetes = "Yes"** thì sẽ có nguy cơ đau tim.
- Ngược lại, đối với các bệnh nhân có nguy cơ đau tim thấp thường sẽ có những đặc điểm ngược lại đối với các giá trị các thuộc tính của trường hợp **Had Heart Attack = "Yes".**



Hình 5.68 Tương quan giữa các giá trị các thuộc tính với giá trị thuộc tính dự đoán

3.4.3. Xác định tập luật:

- Từ chức năng **Dependcy Network** và **Attribute Discrimination,** ta rút ra được tập luật đối với những bệnh nhân có nguy cơ lên cơn đau tim(Had Heart Attack = "Yes") khi có những đặc điểm sau:

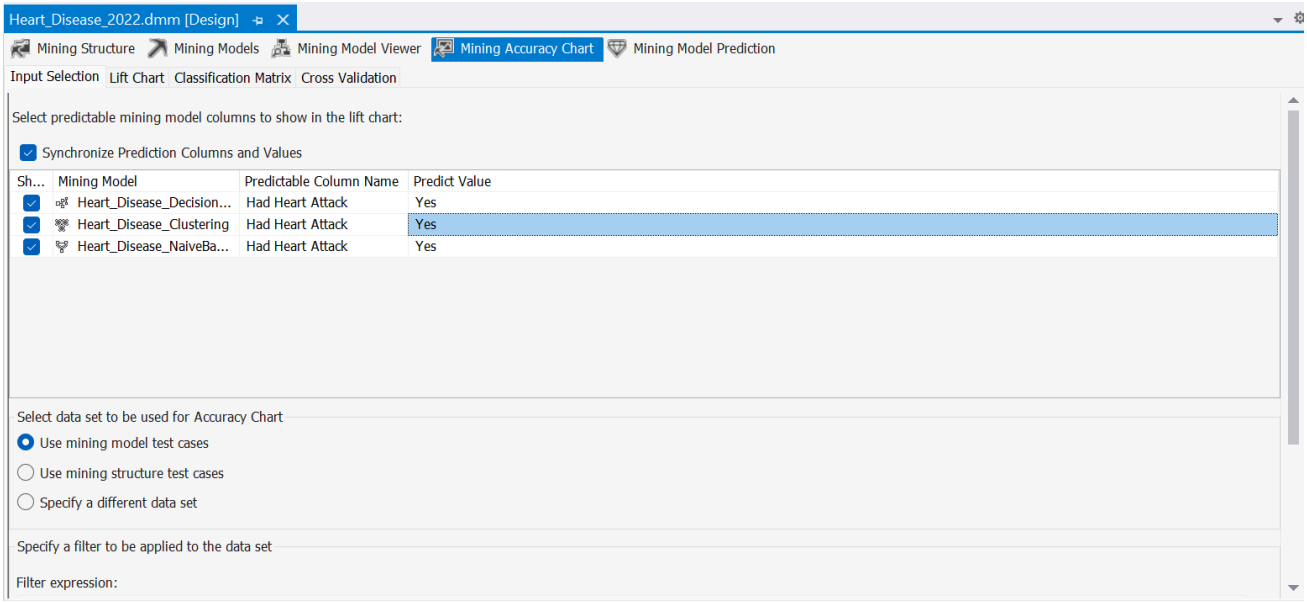
Had Angina = Yes	**Chest Scan = Yes**	**Difficulty Walking = Yes**	**Had Stroke = Yes**
Had Diabetes = Yes	**Pneumo Vax Ever = Yes**	**Had Arthritis = Yes**	**Had COPD = Yes**
General Health = Poor	**Removed Teeth = All**	**General Health = Fair**	**Had Kidney Disease = Yes**
Deaf Or Hard Of Hearing = Yes	**Age Category = Age 80 or older**	**Removed Teeth = 6 or more, but not all**	**Physical Activities = No**
…	…		

3.5. So sánh mức độ độ hiệu quả của các thuật toán:

3.5.1. Sử dụng Lift Chart

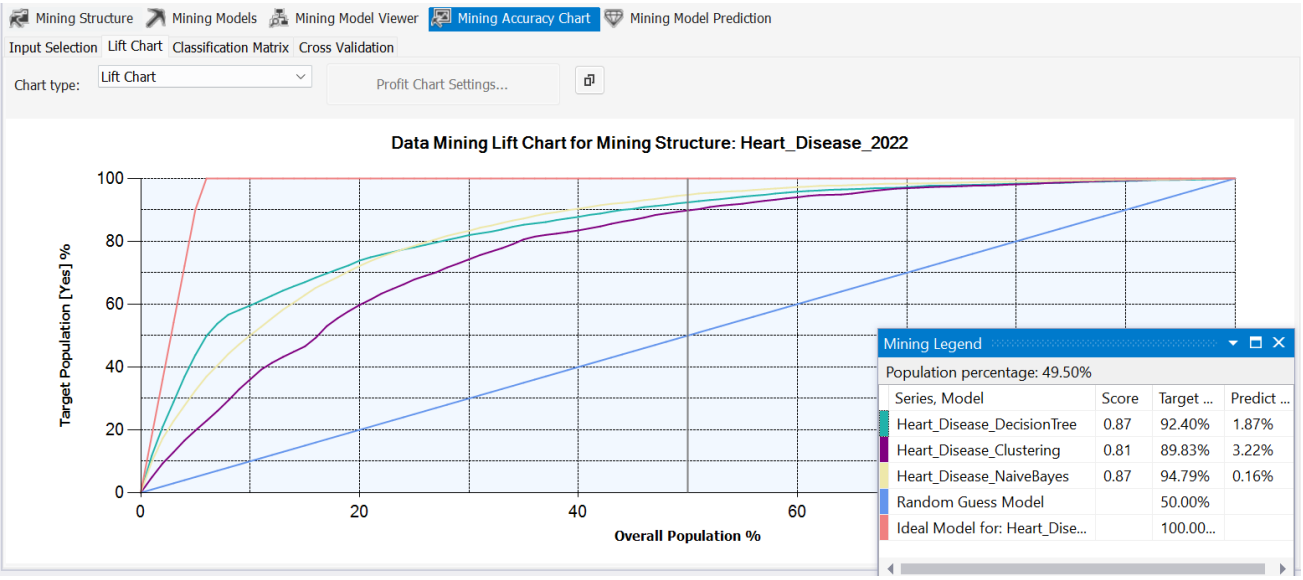
Bước 1: Sau khi đã có thực hiện chạy đầy đủ 3 mô hình thì ta có thể tiến hành đánh giá và hiệu suất giữa 3 thuật toán với nhau bằng chức năng **Lift Chart**. Đầu tiên, ta chọn vào **tab**

Mining Accuracy Chart của **Mining Structures** hiện tại. Tại **tab Input Selection**, ta có thể tuỳ chọn thông tin đầu vào của biểu đồ. Vì đối với mục đích cần biết được về tỉ lệ dự đoán khả năng bị đau tim nên ta giữ nguyên, chỉ có cột **Predict Value** ta chọn **"Yes"**.



Hình 5.69 Chọn thông số đầu vào cho biểu đồ

Bước 2: ta chọn vào tab Lift Chart để tiến hành vẽ biểu đồ so sánh 3 thuật toán.



Hình 5.70 Biểu đồ so sánh 3 thuật toán

Bước 3: Dựa trên biểu đồ ta có thể đưa ra một vài nhận xét sau:

- Mô hình sử dụng thuật toán **Decision Tree** có số điểm là **0.87**(một trong hai thuật toán có điểm cao nhất), phần tram chính xác là **92.4%**
- Mô hình sử dụng thuật toán **Clustering** có số điểm là **0.81**(một trong hai thuật toán có điểm cao nhất), phần tram chính xác là **89.83%**

- Mô hình sử dụng thuật toán **Naïve Bayes** có số điểm là **0.87**(một trong hai thuật toán có điểm cao nhất), phần tram chính xác là **94.79%**

⇨ Như vậy ta có thuật toán **Decision Trees** và **Naïve Bayes** có số điểm bằng nhau và là cao nhất trong 3 thuật toán nên việc chọn ra thuật toán tốt nhất thì cần phải xét thêm một vài yếu tố khác:

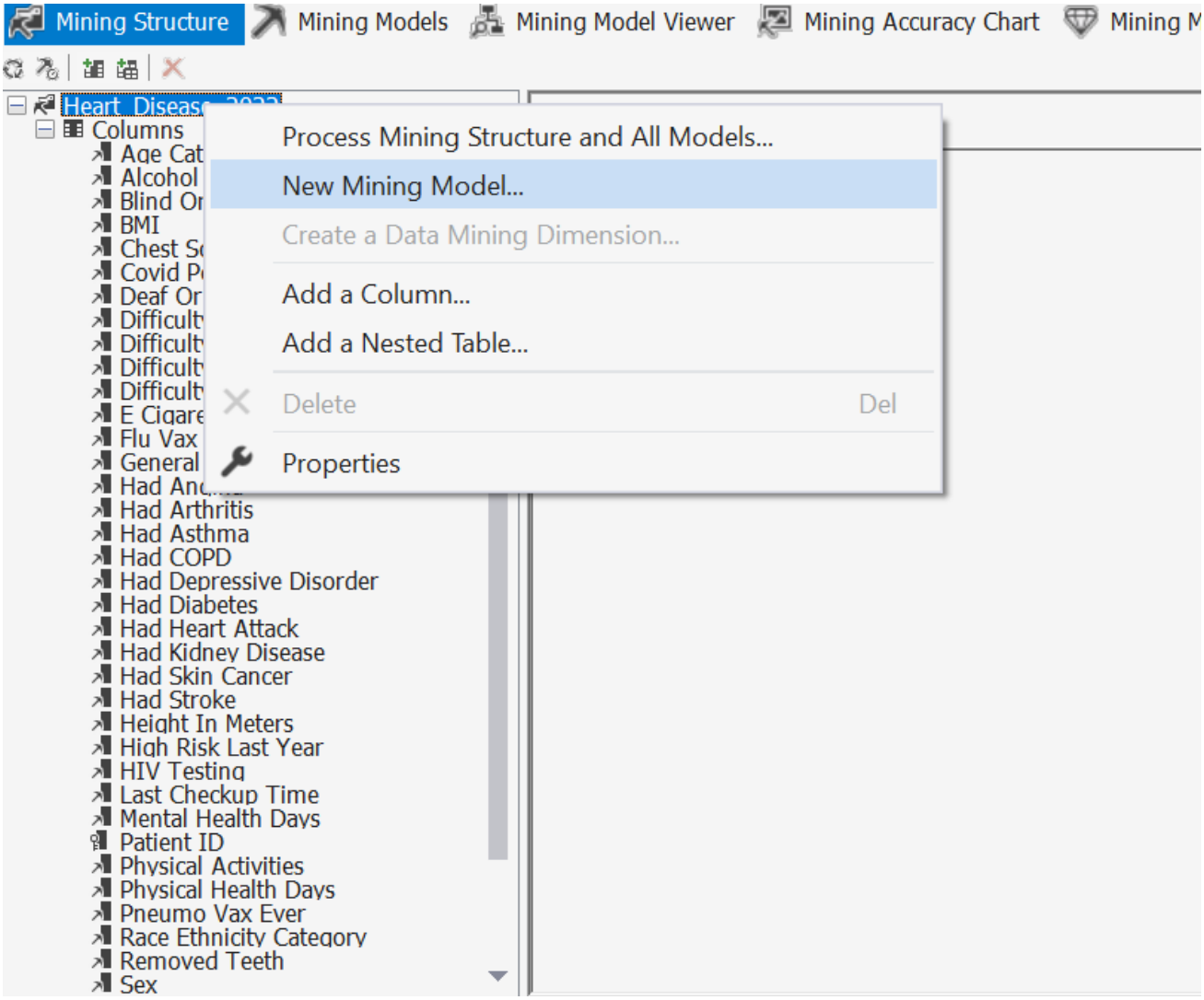
  - Thuật toán **Decision Trees** có một vấn đề đó chính là mô hình tạo ra có thể bị overfitting(quá phù hợp) khiến cho độ chính xác của dự đoán trên dữ liệu test hoặc dữ liệu mới không được cao → Khó phù hợp với yêu cầu dự đoán nguy cơ bị đau tim.
  - Trong khi đó mặc dù thuật toán Naïve Bayes cũng có nhược điểm riêng đó chính là giả thiết về tính độc lập điều kiện của thuộc tính làm giảm độ chính xác mô hình nhưng bù lại thuật toán này lại phù hợp với tập dữ liệu lớn và phức tạp

⇨ **Kết luận:** Ta sẽ chọn thuật toán **Naïve Bayes** là thuật toán tối ưu nhất đối với bộ dữ liệu này.

3.6. Kiểm tra chức năng Filtered Model

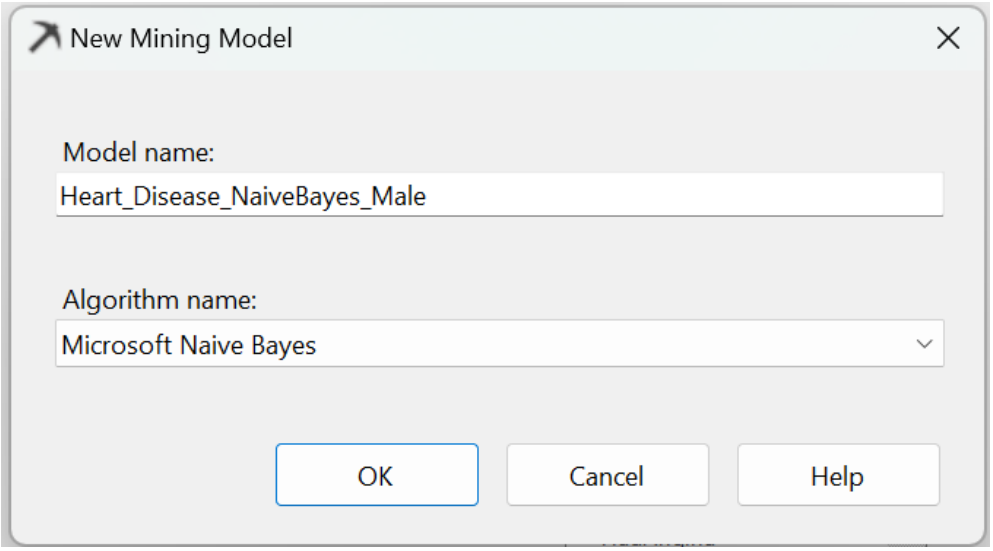
- Ta đã xác định được thuật toán cho độ chính xác cao nhất đối với bộ dữ liệu này là **Naïve Bayes**.
- Ở phần này, nhóm sẽ sử dụng chức năng **Filtered Model** để tìm hiểu xem liệu có sự khác nhau giữa thuộc tính các mô hình khi áp dụng dự đoán khả năng bị đau tim dựa trên giới tính(**Nam** hoặc **Nữ**).

Bước 1: Để tạo mô hình với điều kiện lọc thì trước tiên ta tạo thêm một mô hình mới bằng chức New Mining Model…



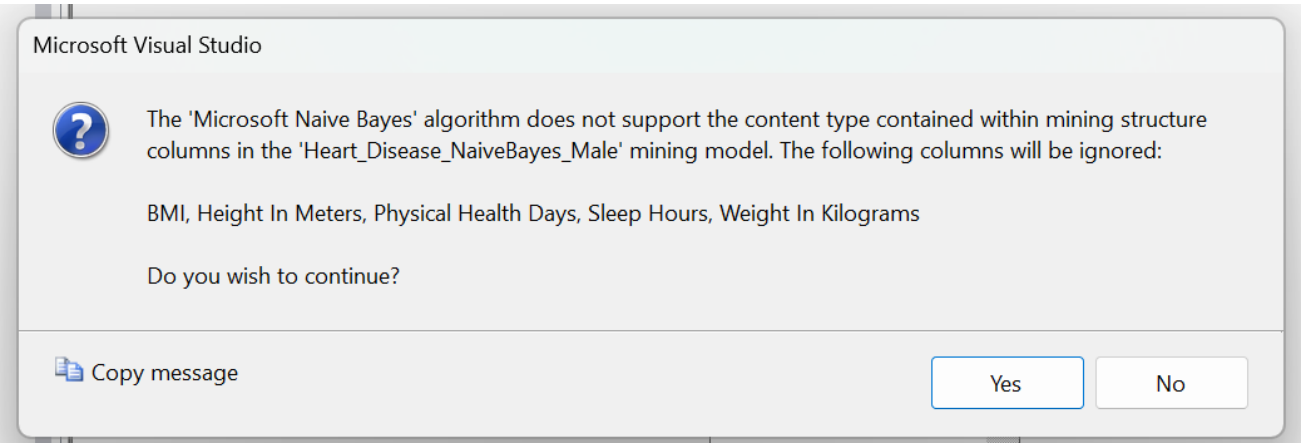
Hình 5.71 Tạo mới mô hình

Bước 2: Ở đây trước tiên chúng ta sẽ tạo mô hình sử dụng cho điều kiện lọc là Sex = "Male" vì vậy ta đặt tên và chọn thuật toán **Naïve Bayes** theo như hình.



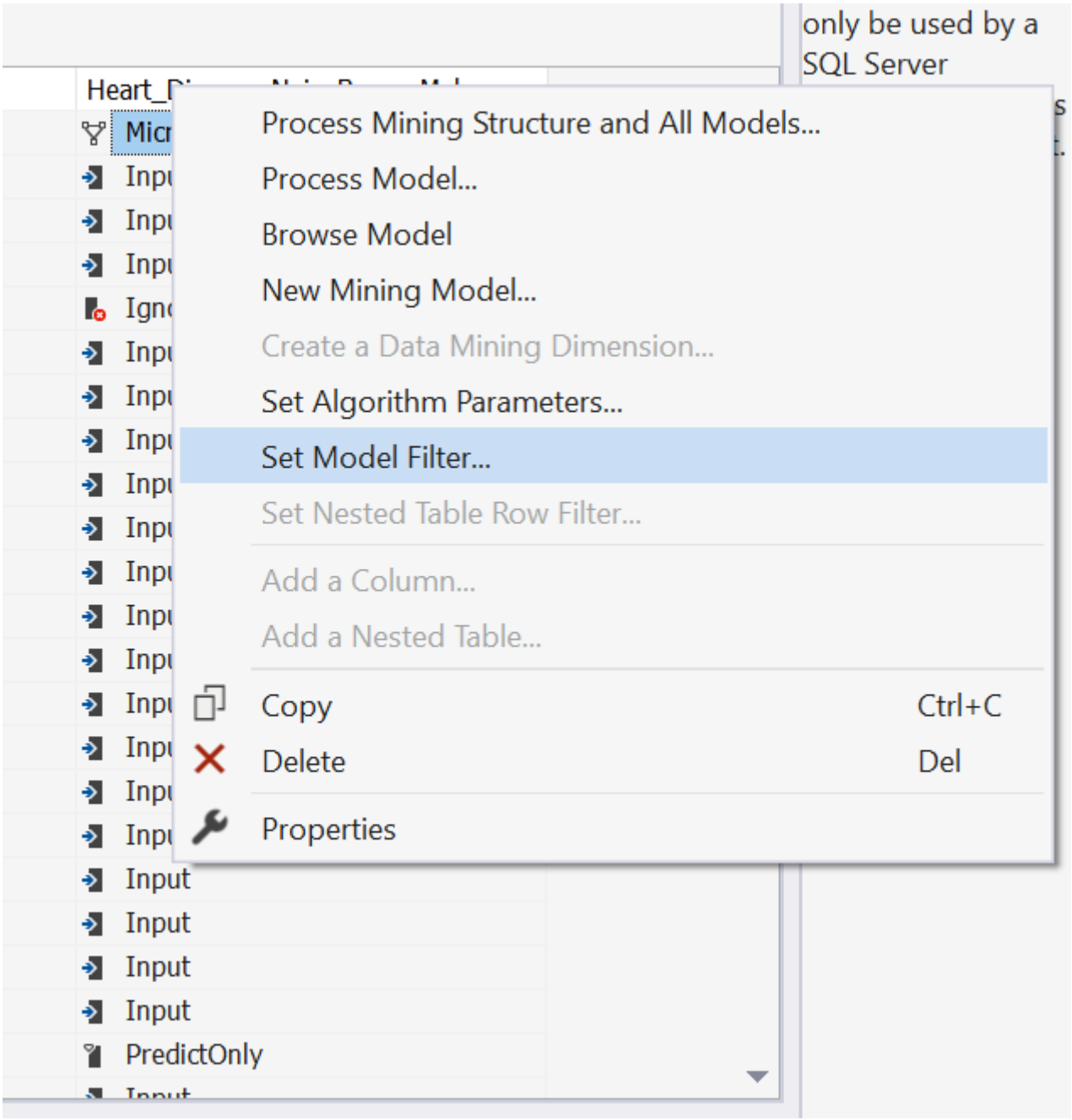
Hình 5.72 Đặt tên và chọn thuật toán mô hình

Vì thuật toán **Naïve Bayes** không sử dụng các thuộc tính có dạng số làm đầu vào vì vậy hệ thống sẽ thông báo bỏ các qua các thuộc tính với đặc điểm ấy, ta nhấn **Yes** để chấp nhận.



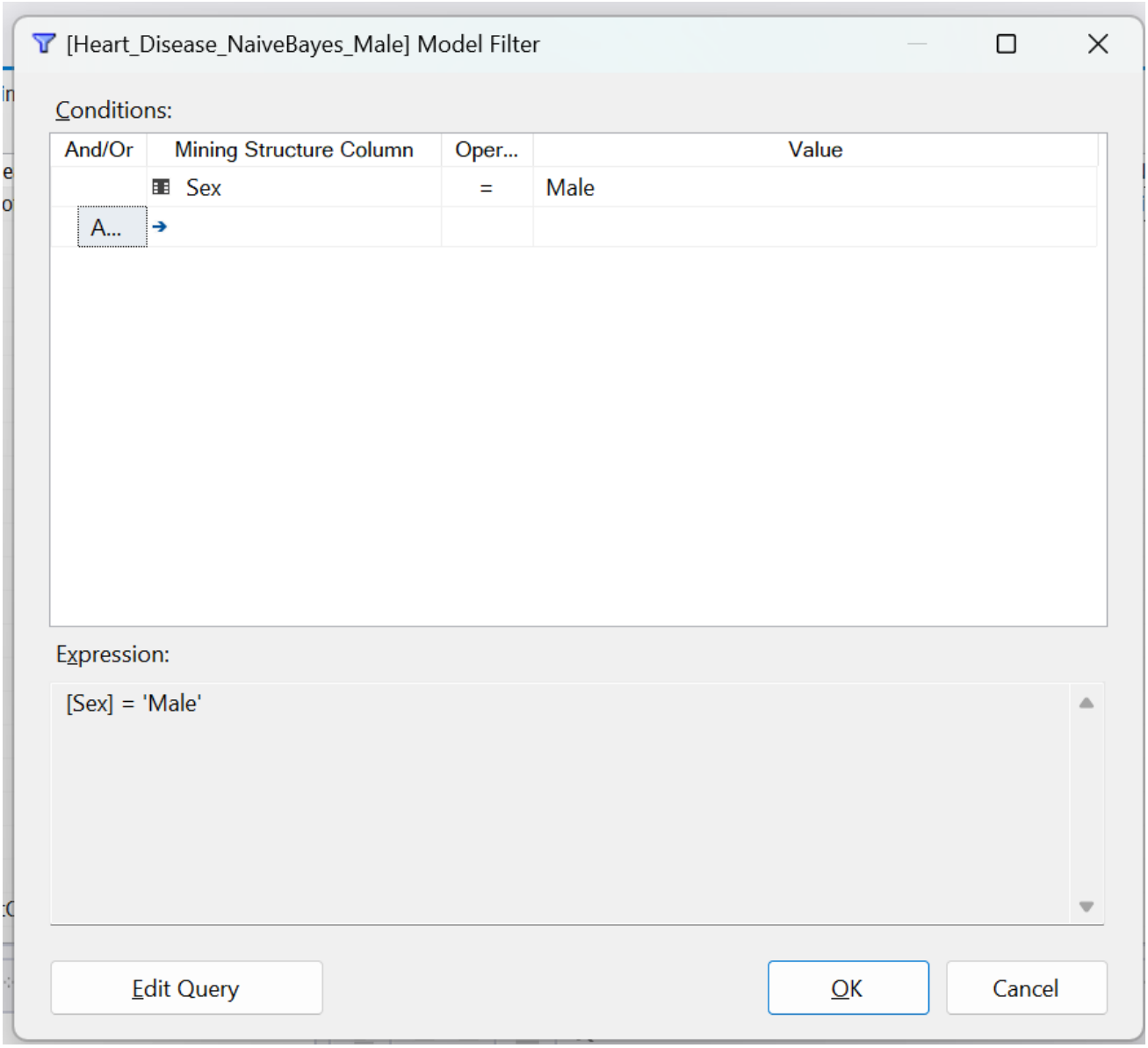
Hình 5.73 Bỏ qua các thuộc tính có giá trị số

Bước 3: Và khi chuyển qua cửa sổ **Mining Models** thì ta sẽ thấy mô hình mà ta vừa tạo. Ta tiếp tục nhấn chuột phải vào cột mô hình vừa tạo → chọn **Set Model Filter** để tạo bộ lọc cho mô hình.



Hình 5.74 Thêm bộ lọc cho mô hình

Bước 4: Khi hộp thoại **Model Filter** hiện lên thì ta tiến hành chọn thuộc tính tướng ứng với các cột là: **Mining Structure Column** là **Sex**(giới tính), **Operator**(phép toán) là **"="**, **Value** là **Male**(giới tính Nam). Trong quá trình tinh chỉnh, ở ô **Expression** phía dưới cũng sẽ hiển thị cho chúng ta phép toán mà chúng ta vừa định nghĩa. Nhấn **OK** để áp dụng tạo bộ lọc.



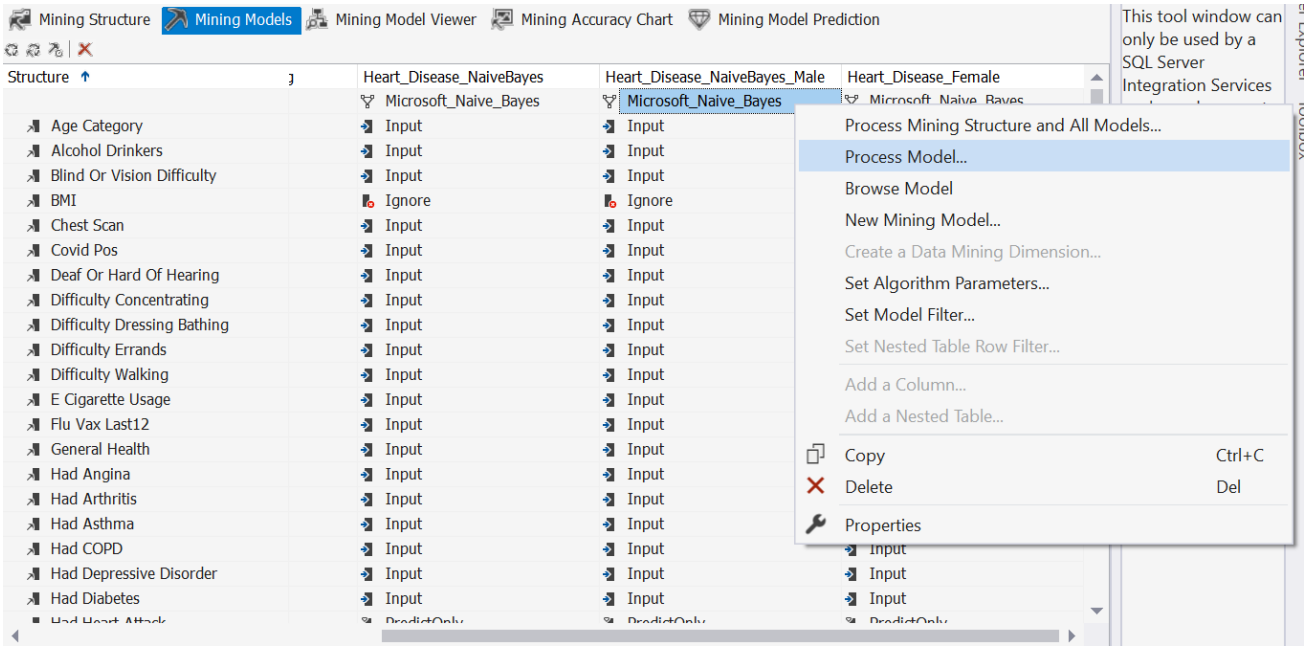
Hình 5.75 Cấu hình bộ lọc

Bước 5: Tương tự với việc thiết lập mô hình với điều kiện Sex = "Female", ta thực hiện lại các **bước 2, 3, 4**.

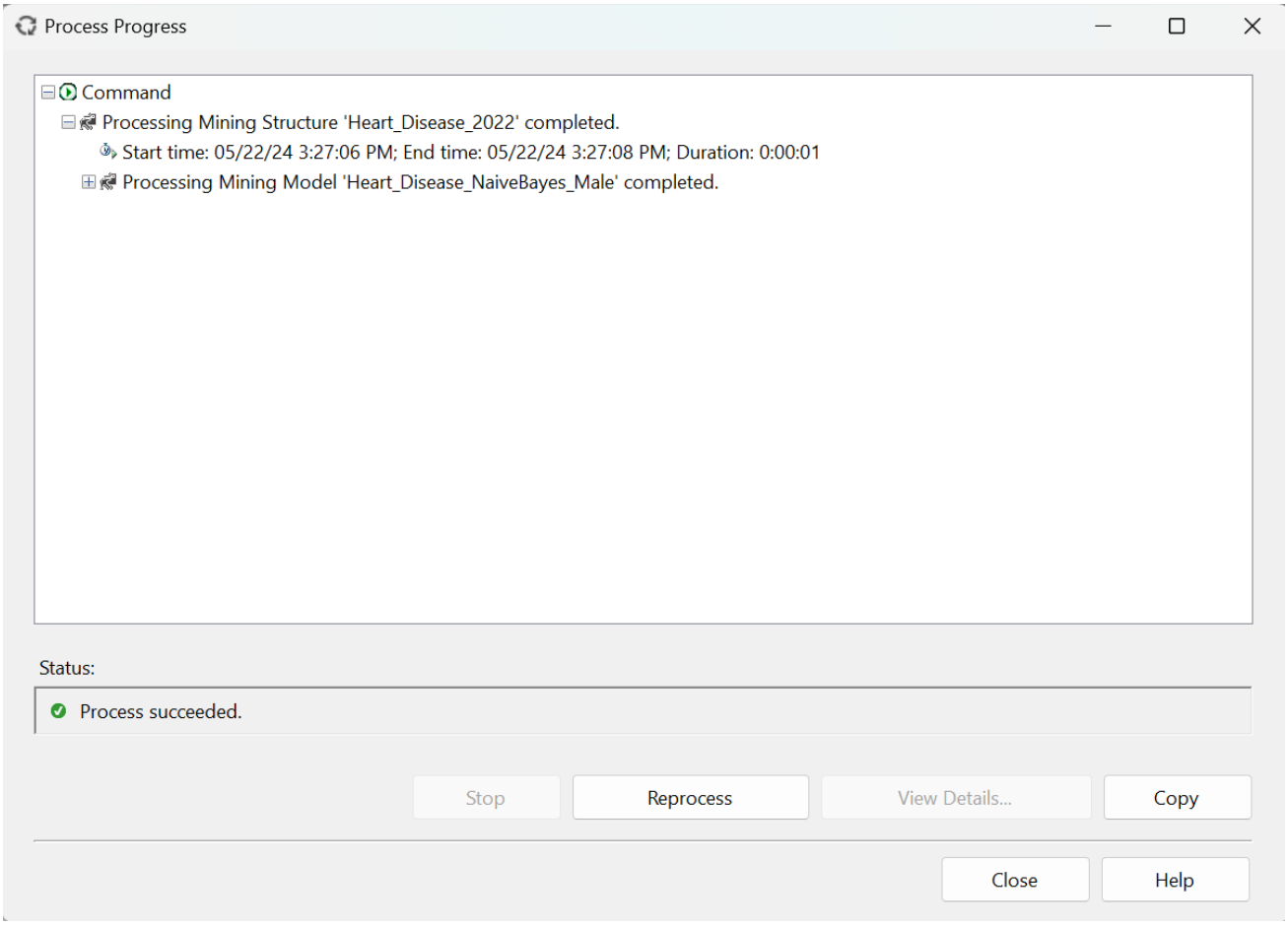
Heart_Disease_NaiveBayes_Male	Heart_Disease_Female
Microsoft_Naive_Bayes	Microsoft_Naive_...
Input	Input
Input	Input
Input	Input
Ignore	Ignore
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
Input	Input
PredictOnly	PredictOnly

Hình 5.76 Tạo thêm mô hình với thuộc tính lọc là Sex = "Feale"

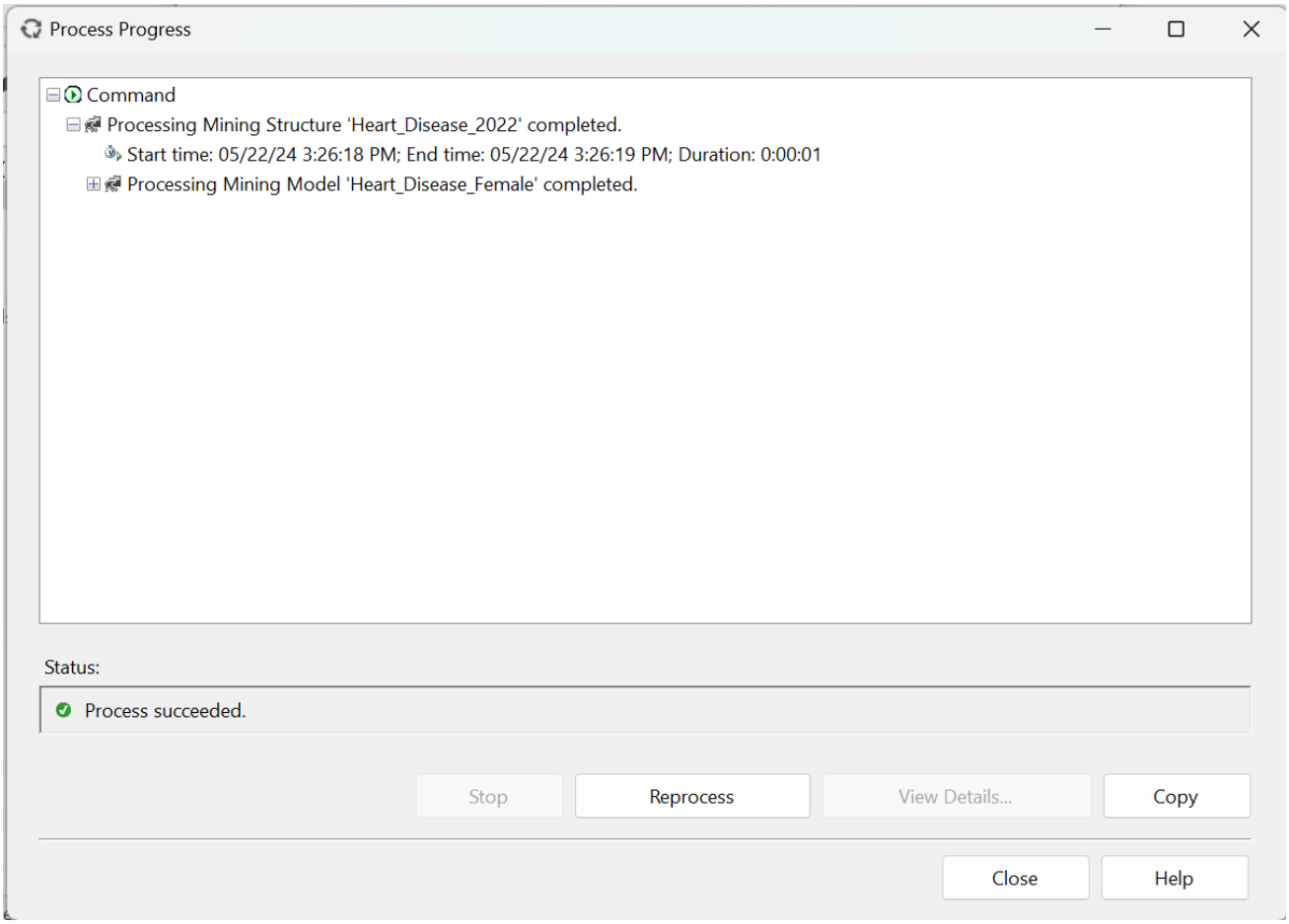
Bước 6: Ta tiến hành chạy thuật toán của hai mô hình vừa tạo.



Hình 5.77 Chạy thuật toán của mô hình



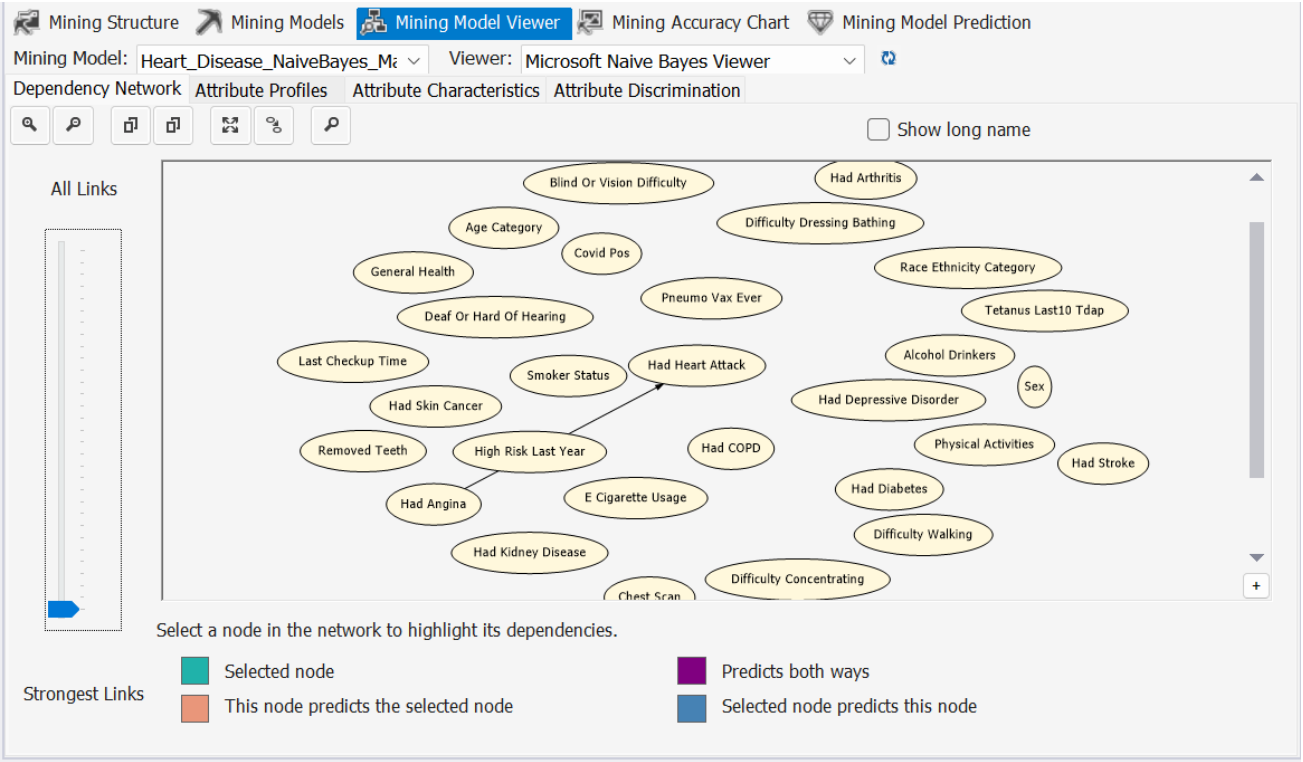
Hình 5.78 Kết quả chạy mô hình với bộ lọc Sex = "Male"



Hình 5.79 Kết quả chạy thuật toán của mô hình với bộ lọc Sex = "Female"

- Đối với mô hình Heart_Disease_NaiveBayes_Male:

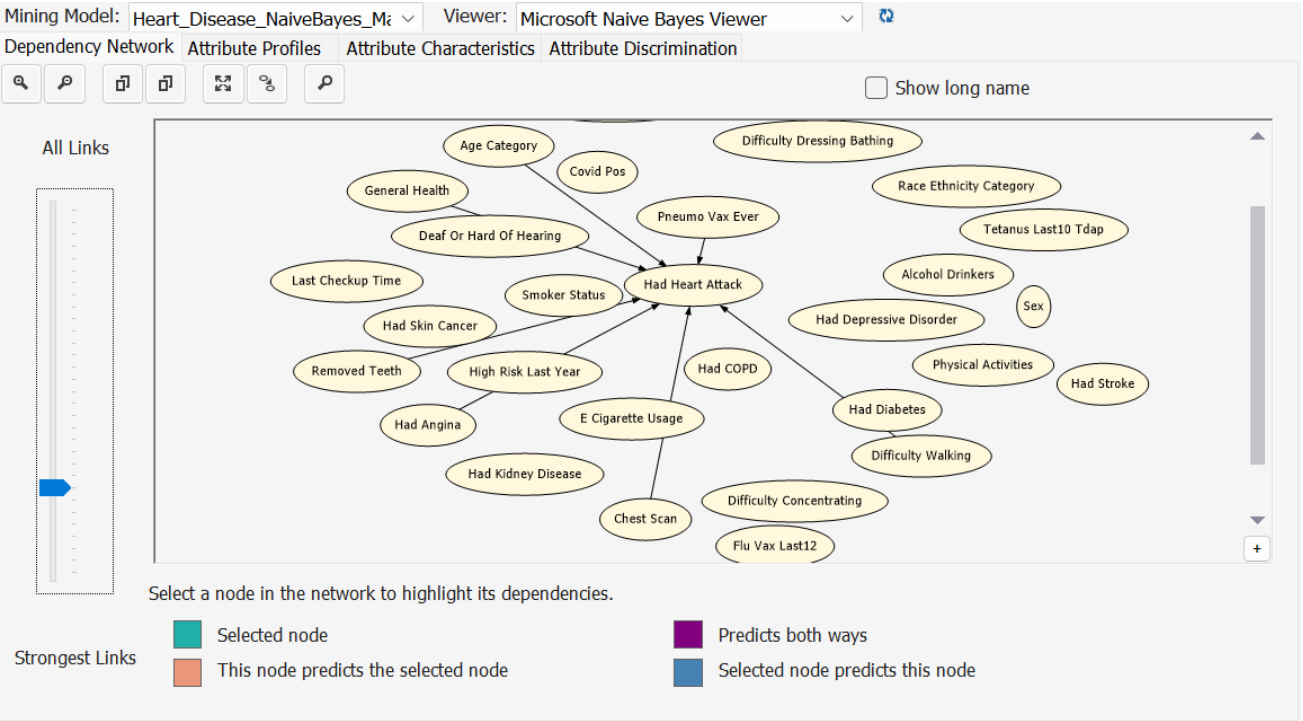
  Thuộc tính có liên kết mạnh nhất tới biến dự đoán là **Had Angina**.



Hình 5.80 Thuộc tính có liên kết mạnh nhất

Thứ tự 7 liên kết từ có độ mạnh giảm dần là: **Age Category → General Health → Chest Scan → Removed Teeth → Difficulty Walking → Pneumo Vax Ever**.

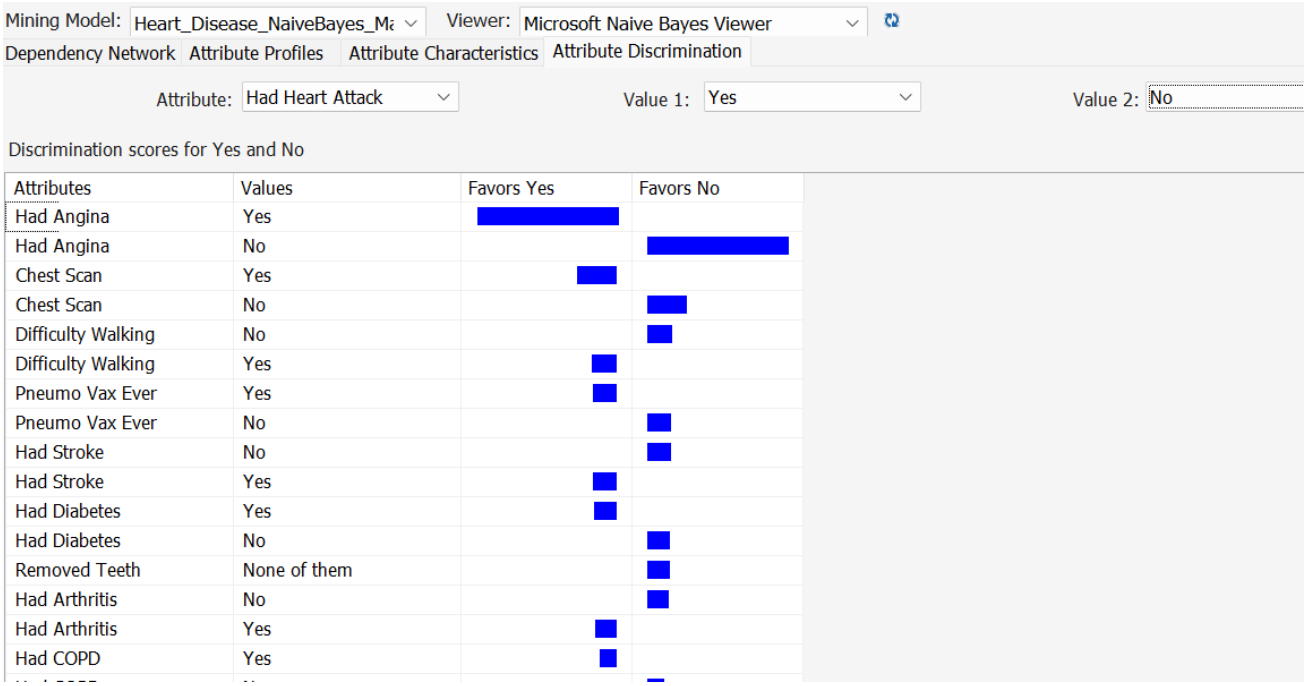
⇨ Đã có sự thay đổi khi chạy thuật toán khi không có bộ lọc.



Hình 5.81 7 thuộc tính có liên kết từ mạnh đến yếu tới thuộc tính dự đoán

Khi chuyển sang tab Attribute Discrimination thì dựa trên điểm phân biệt của các giá trị các thuộc tính đầu vào thì ta có thể rút ra được tập luật sau:

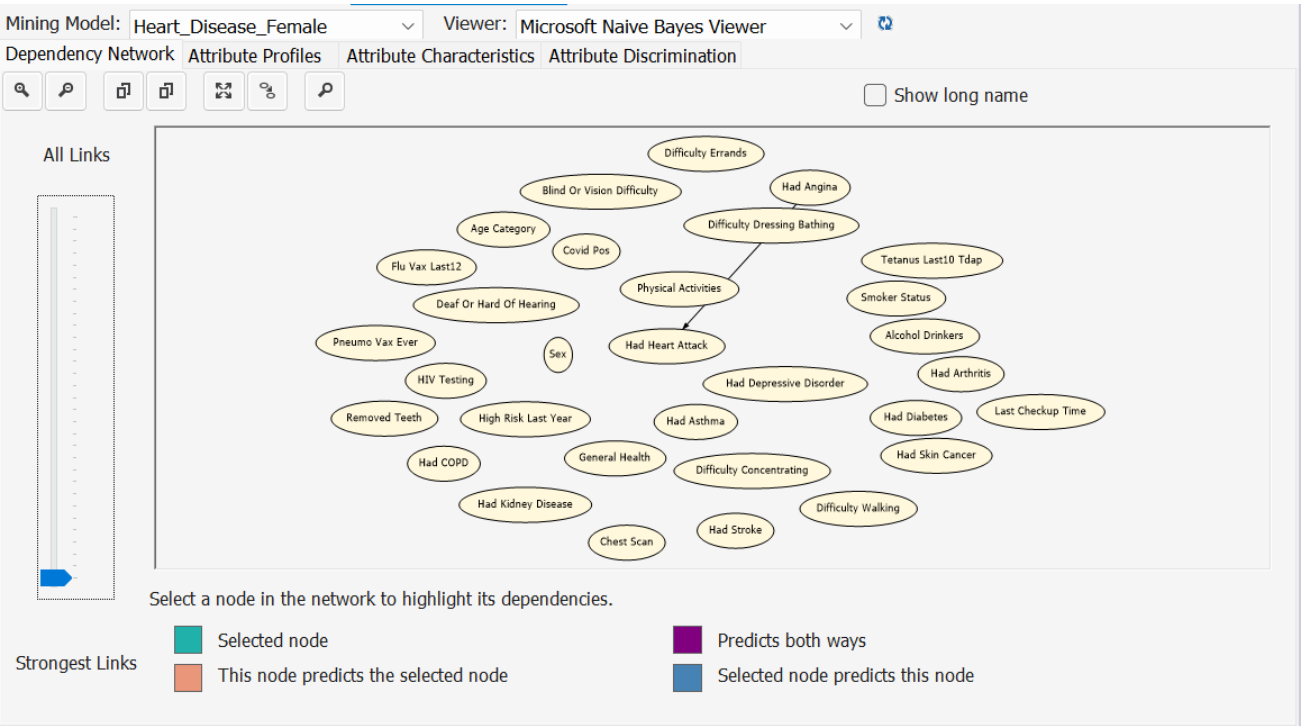
- Những người có giới tính nam có **bị đau thắt vùng ngực(Had Angina = "Yes"), có chụp cắt lớp vi tính vùng ngực(Chest Scan = "Yes"), gặp khó khăn khi đi bộ(Difficulty Walking = "Yes"), có tiêm vắc-xin phế cầu(Pneumo Vax Ever = "Yes"), từng bị đột quỵ(Had Stroke = "Yes"), bị tiểu đường(Had Diabetes)** sẽ có nguy cơ bị **lên cơn đau tim**.



Hình 5.82 Các điểm phân biệt cho hai giá trị Had Heart Attack

- Tương tự đối với mô hình **Heart_Disease_Male**:

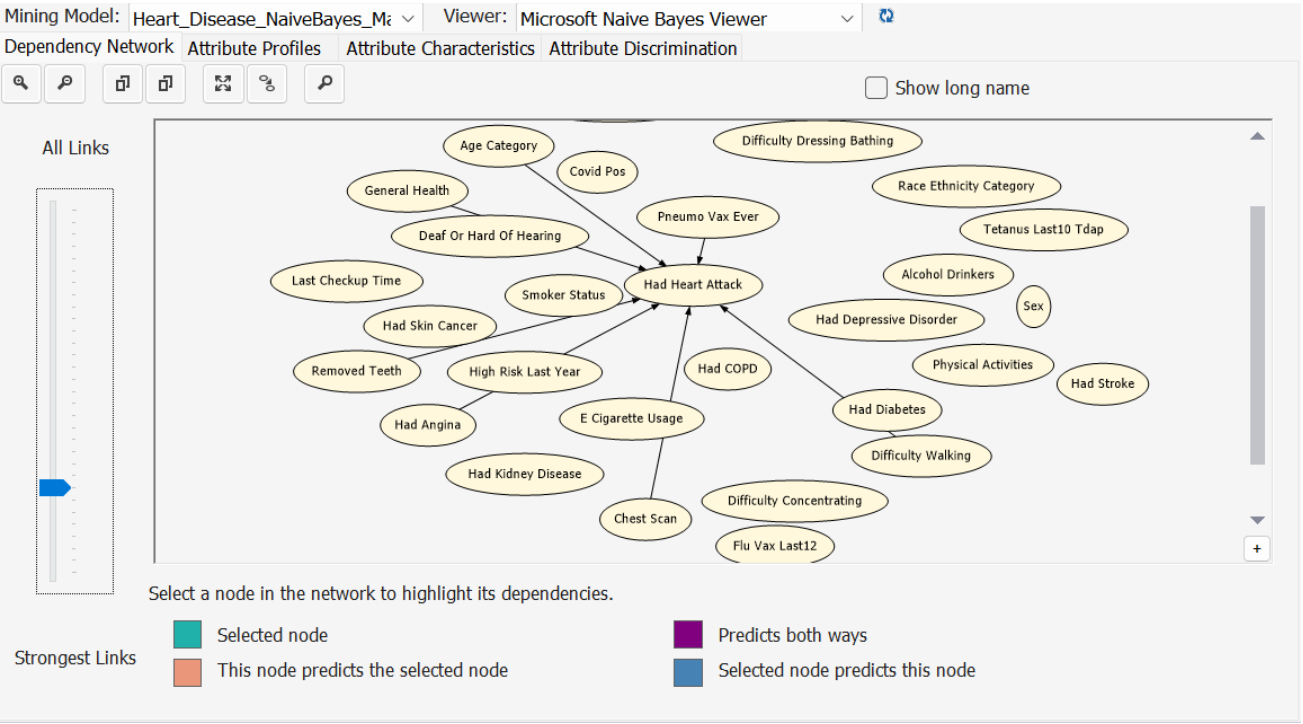
Thuộc tính có liên kết mạnh nhất tới biến dự đoán là **Had Angina**.



Hình 5.83 Thuộc tính có liên kết mạnh nhất

Thứ tự 7 liên kết từ có độ mạnh giảm dần là: **Had Angina → General Health → Difficulty Walking → Chest Scan → Removed Teeth → Age Category → Had Stroke**.

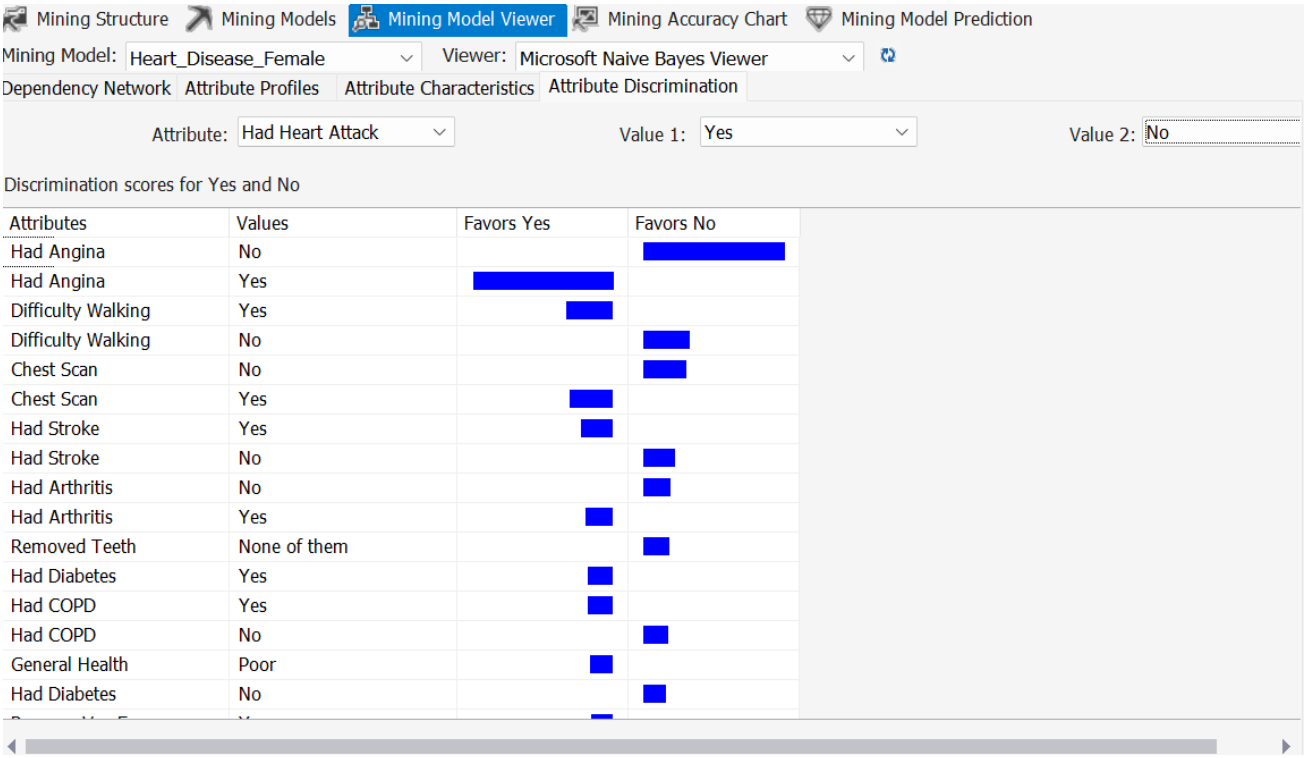
⇨ Đã có sự thay đổi khi chạy thuật toán khi không có bộ lọc và khác với mô hình của giới tính **Nam**.



Hình 5.83 7 thuộc tính có liên kết từ mạnh đến yếu tới thuộc tính dự đoán

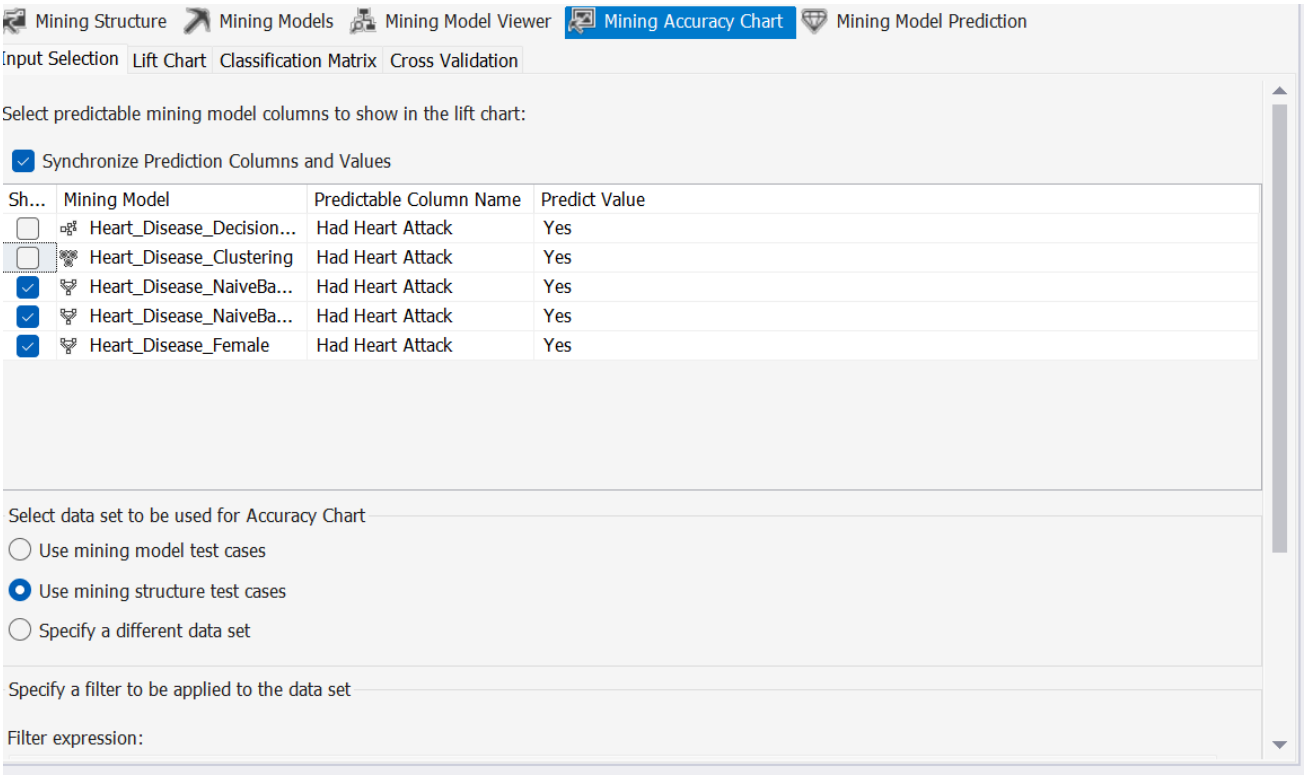
Khi chuyển sang tab **Attribute Discrimination** thì dựa trên điểm phân biệt của các giá trị các thuộc tính đầu vào thì ta có thể rút ra được tập luật sau:

- Những người có giới tính nữ có **bị đau thắt vùng ngực(Had Angina = "Yes"), có chụp cắt lớp vi tính vùng ngực(Chest Scan = "Yes"), gặp khó khăn khi đi bộ(Difficulty Walking = "Yes"), từng bị đột quỵ(Had Stroke = "Yes"), Bị đau xương khớp(Had Arthritis = "Yes"), bị tiểu đường(Had Diabetes)** sẽ có nguy cơ bị **lên cơn đau tim**

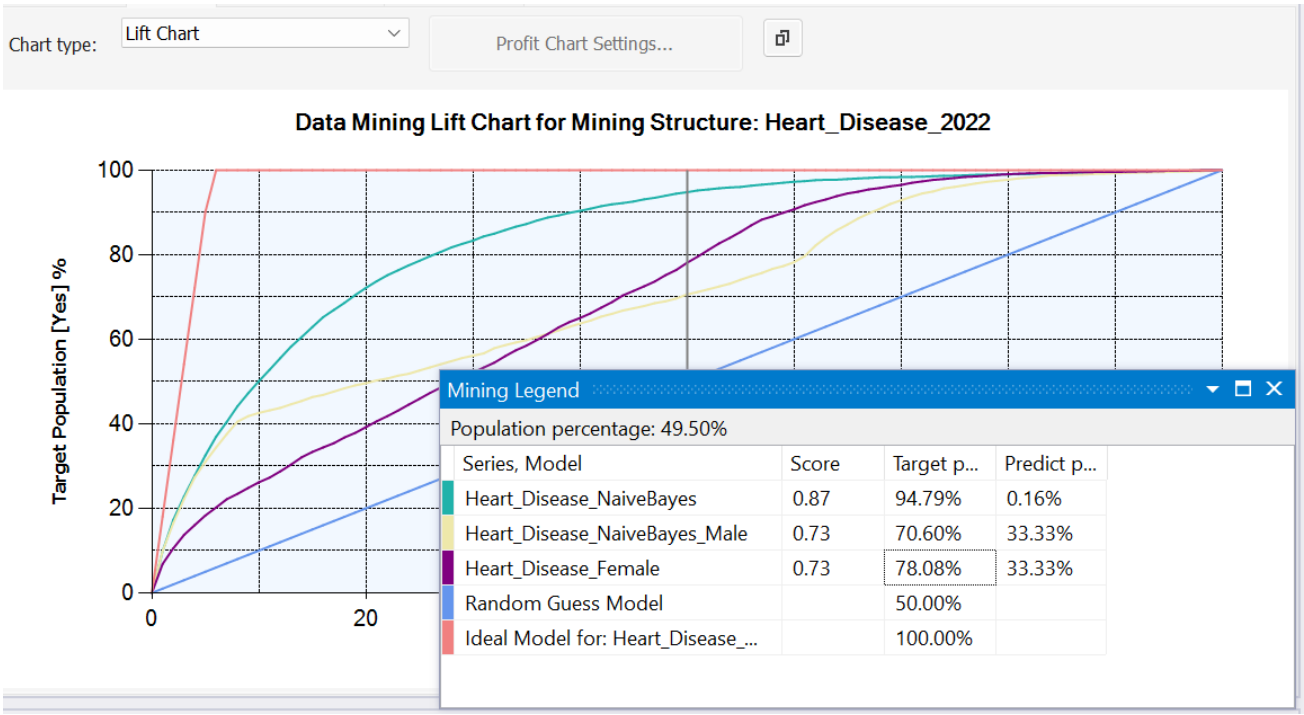


Hình 5.84 Các điểm phân biệt cho hai giá trị Had Heart Attack đối với mô hình giới tính Nữ

- Tiến hành sử dụng **LiftChart** để so sánh 3 thuật toán: Naïve Bayes, Naïve Bayes cho **Nam** và Naïve Bayes cho **Nữ**.



Hình 5.85 Chọn thuật toán làm đầu vào biểu đồ



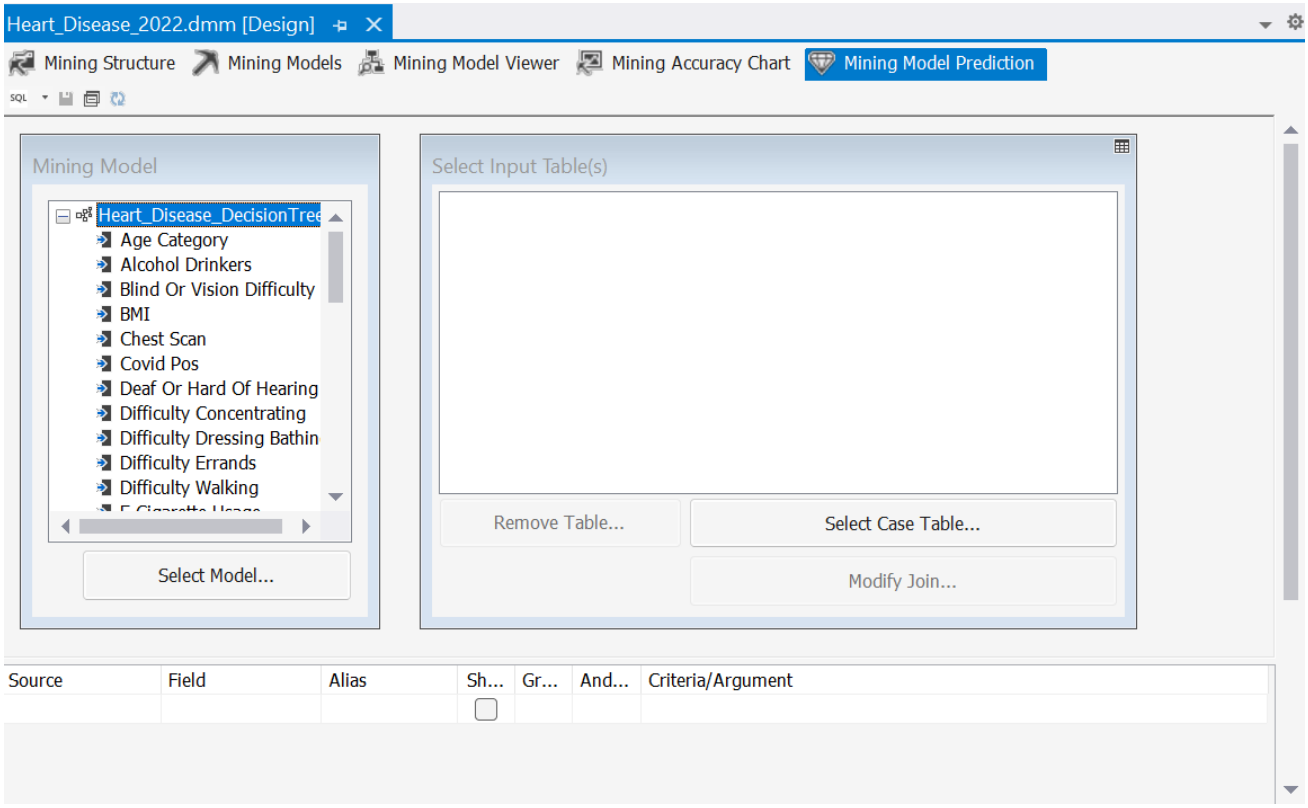
Hình 5.86 Biểu đồ phân tích độ hiệu quả 3 mô hình

Từ biểu đồ trên có thể nhận thấy rằng đối với thuật toán Naïve Bayes áp dụng lên cho tổng thể bộ dữ liệu kiểm thử vẫn sẽ cho ra số điểm cao hơn và đồng thời là hiệu suất dự đoán cũng cao hơn so với 2 mô hình còn lại(**94,79%**).

3.7. Dự đoán với các giá trị tương lai

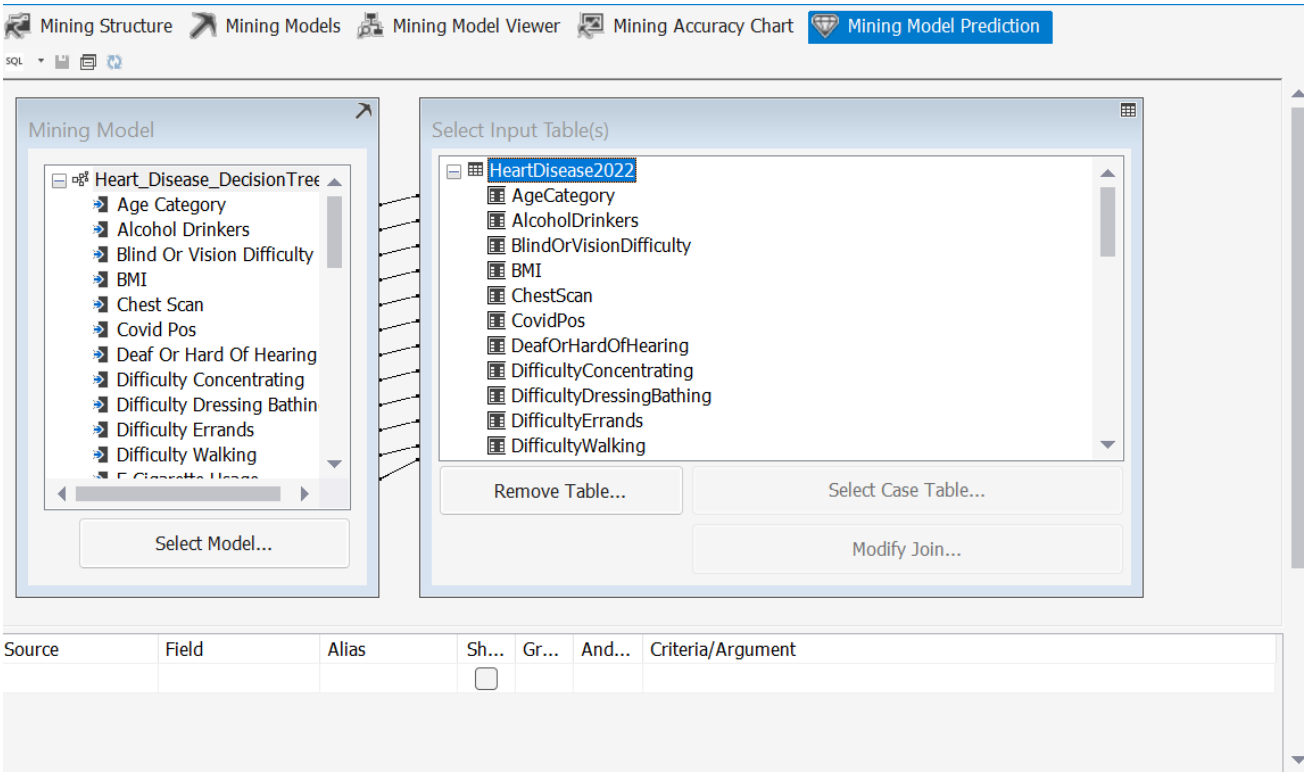
- Sau khi đã trải qua bước huấn luyện, kiểm tra và so sánh các mô hình khác nhau thì bây giờ ta có thể sử dụng mô hình vừa tạo được để dự đoán các giá trị trong tương lai. Ở đây bài toán chúng ta đặt ra là dự đoán bệnh nhân có nguy cơ lên cơn đau tim(**Had Heart Attack = "Yes")** và thuật toán sử dụng sẽ là **Decision Trees**.

Bước 1: Chọn tab Mining Model Prediction để mở giao diện.



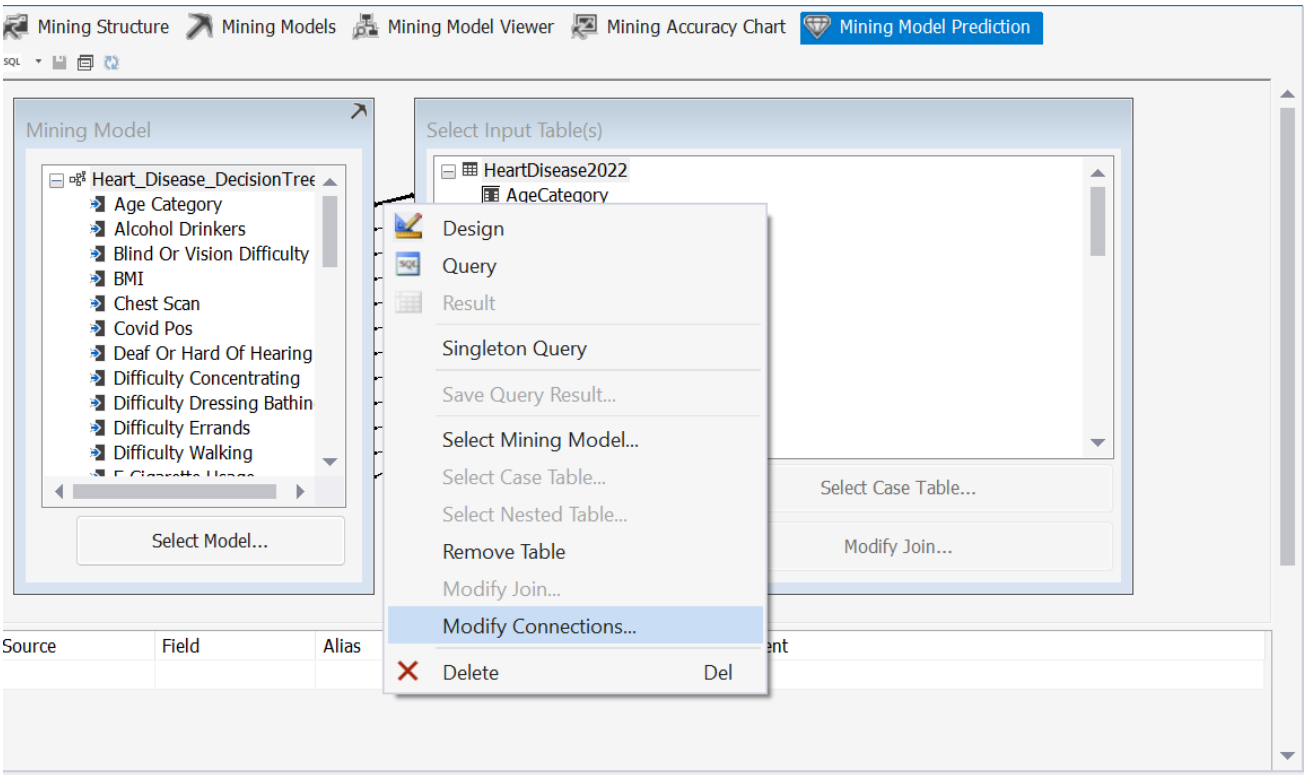
Hình 5.87 Giao diện dự đoán

Bước 2: chọn mô hình là **Heart_Disease_DecisionTree** và Case Table là HeartDisease2022.

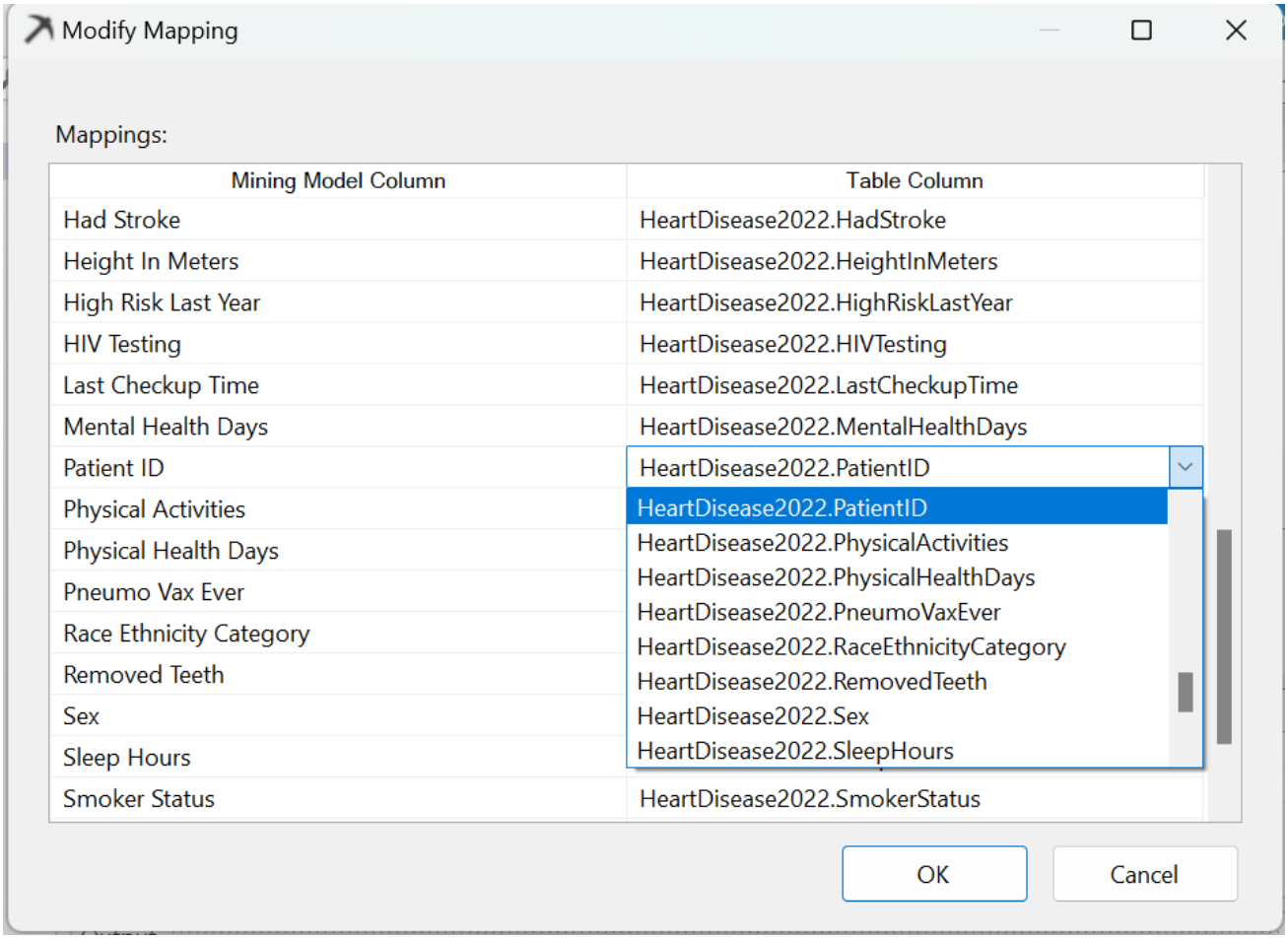


Hình 5.88 Chọn bảng để làm mục tiêu dự đoán

Bước 2: Ta tiến hành tinh chỉnh ánh xạ các cột mô hình với cột của bảng dữ liệu dự đoán bằng cách nhấn chuột phải một liên kết → **Modify Connections**.

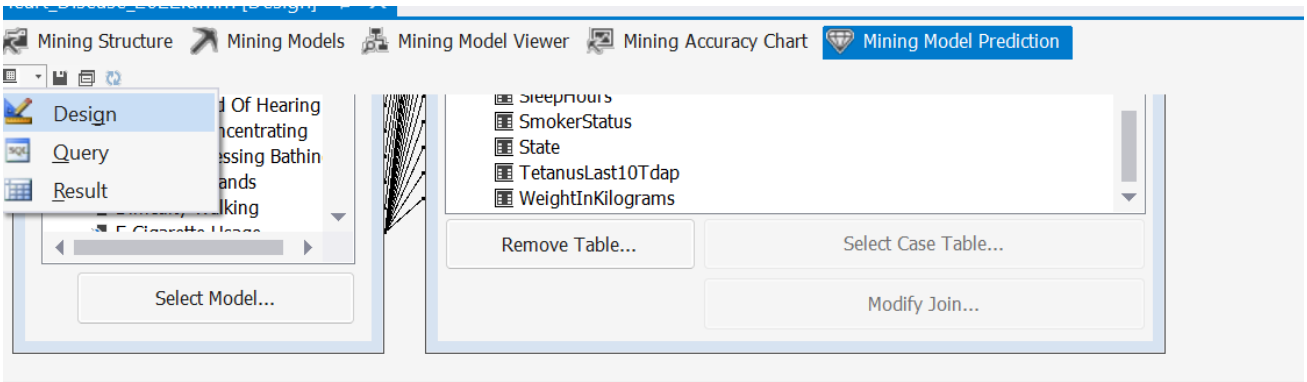


Hình 5.89 Tinh chỉnh ánh xạ



Hình 5.90 Tinh chỉnh ánh xạ(1)

Bước 3: Để có xác định dữ liệu thuộc tính cần dự đoán thì ta nhấn chuột phải vào góc trên bên trái của giao diện **Mining Model Prediction**(có chữ **SQL**) → Chọn **Design**.



Hình 5.91 Thiết lập thuộc của dự đoán

Bước 4: Các tuỳ chỉnh sẽ nằm dưới Mining Models. Ở đây, ta sẽ thực hiện 2 hành động:

- Chọn **Source** là **Prediction Function**, **Field** là **PredictProbability**, Đặt Alias là **Predict Probability**, kéo thả thuộc tính **Had Heart Attack** từ trên xuống

Criteria/Argument → Cấu hình thuộc tính cần dự đoán và giá trị dự đoán là xác suất xảy ra của giá trị thuộc tính dự đoán.

- Chọn **Soucre** là mô hình cây quyết định, **Field** là **Had Heart Attack** và **Criteria/Argument** là = "Yes" → Chỉ lấy những giá trị có nguy cơ bị bệnh tim.

Source	Field	Alias	Sh...	Group	And...	Criteria/Argument
Heart_Disease_DecisionTree	Had Heart Attack		☑			='Yes'
Prediction Function	PredictProbability		☑			[Heart_Disease_Tree].[Had Heart Attack]
			☐			

Source	Field	Alias	Sh...	Group	And...	Criteria/Argument
Heart_Disease_DecisionTree	Had Heart Attack		☑			='Yes'
Prediction Function	PredictProbability	Predict Probability	☑			[Heart_Disease_DecisionTree].[Had He...

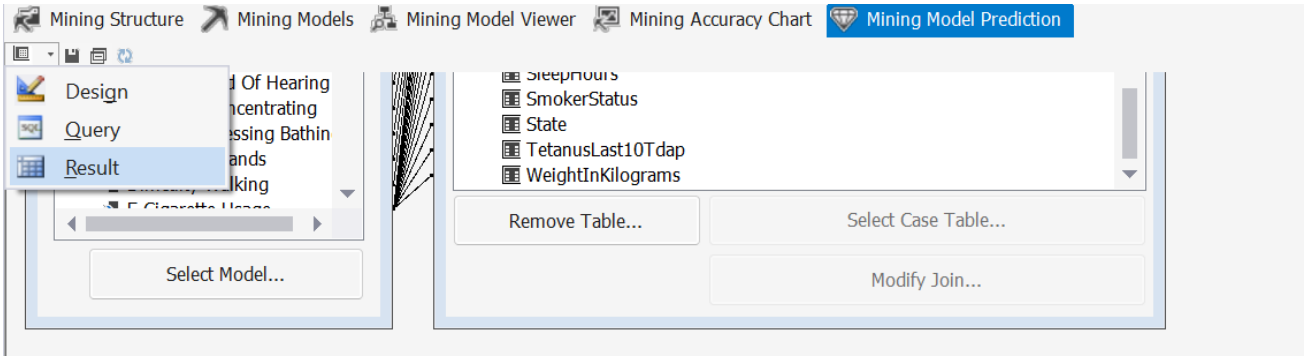
Hình 5.92 Cấu hình mô hình dự đoán

Bước 5: Thêm các cột thông tin truy vấn, phải có cột thuộc tính khoá chính. Tất cả sẽ được lấy từ trong **Source.**

Source	Field	Alias	Sh...	Group	And...	Criteria/Argument
Heart_Disease_DecisionTree	Had Heart Attack		☑			='Yes'
Prediction Function	PredictProbability		☑			[Heart_Disease_Tree].[Had Heart Attack]
HeartDisease2022	PatientID		☑			
HeartDisease2022	HadAngina		☑			
HeartDisease2022	Sex		☑			
HeartDisease2022	DifficultyWalking		☑			
HeartDisease2022	AgeCategory		☑			
HeartDisease2022	ChestScan		☑			
			☐			

Hình 5.93 Thêm thuộc tính truy vấn

Bước 6: Xem kết quả truy vấn bằng cách chọn chức năng **Result** trong cùng vùng với chọn chức năng **Design.**



Hình 5.94 Hiển thị kết quả

Bước 7: Với kết quả hiển thị thì xác suất dự đoán khả năng bị bệnh tim của các bệnh nhân nằm trong khoảng **50 – 64%** độ chính xác đối với các bệnh nhân thực sự có bị đau tim.

Mining Structure | Mining Models | Mining Model Viewer | Mining Accuracy Chart | Mining Model Prediction

Had Heart Attack	Predict Probability	PatientID	HadAngina	Sex	Difficulty...	AgeCateg...	ChestScan
Yes	0.516045236632773	9	Yes	Male	No	Age 80 o...	Yes
Yes	0.516045236632773	24	Yes	Male	No	Age 55 t...	Yes
Yes	0.506563664847928	33	Yes	Female	Yes	Age 80 o...	Yes
Yes	0.516045236632773	119	Yes	Male	No	Age 80 o...	Yes
Yes	0.516045236632773	155	Yes	Male	No	Age 65 t...	Yes
Yes	0.506563664847928	189	Yes	Female	No	Age 45 t...	Yes
Yes	0.506563664847928	212	Yes	Female	No	Age 70 t...	Yes
Yes	0.516045236632773	328	Yes	Male	Yes	Age 75 t...	No
Yes	0.506563664847928	334	Yes	Female	No	Age 35 t...	Yes
Yes	0.618954671187651	352	Yes	Male	No	Age 70 t...	Yes
Yes	0.506563664847928	362	Yes	Female	Yes	Age 60 t...	Yes
Yes	0.506563664847928	440	Yes	Female	Yes	Age 80 o...	No
Yes	0.618954671187651	443	Yes	Male	Yes	Age 60 t...	Yes
Yes	0.516045236632773	466	Yes	Male	Yes	Age 80 o...	Yes
Yes	0.638621348404044	555	Yes	Male	Yes	Age 80 o...	Yes
Yes	0.516045236632773	613	Yes	Male	No	Age 60 t...	No
Yes	0.516045236632773	636	Yes	Male	No	Age 80 o...	Yes
Yes	0.506563664847928	645	Yes	Female	Yes	Age 60 t...	Yes
Yes	0.516045236632773	669	Yes	Male	No	Age 65 t...	No
Yes	0.516045236632773	673	Yes	Male	No	Age 65 t...	Yes
Yes	0.516045236632773	681	Yes	Male	No	Age 65 t...	Yes

Query run completed with 5512 rows fetched

Hình 5.95 Kết quả dự đoán

Tài liệu tham khảo

1 Tài liệu hướng dẫn Khai thác dữ liệu bằng Analysis Service – SQL Server, Microsoft.

2 Báo cáo đồ án tham khảo Kho dữ liệu và OLAP: Nguyễn Minh Nhựt – Trần Quang Phúc.

3 Báo cáo đồ án tham khảo Kho dữ liệu và OLAP: Nguyễn Thuý Mơ.

4 Báo cáo đồ án tham khảo Kho dữ liệu và OLAP: Dương Quốc Cường – Mạc Huy Tú.