

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM



HCMUTE

**BÁO CÁO CUỐI KỲ MÔN KHO DỮ LIỆU
ĐỀ TÀI
XÂY DỰNG KHO DỮ LIỆU QUẢN LÝ BÁN HÀNG**

GVHD: ThS Nguyễn Văn Thành

Môn học: KHO DỮ LIỆU

Mã lớp học phần: DAWH430784_22_2_02

Lớp: Thứ 7 (tiết 1-4)

Sinh viên thực hiện: Nhóm 11

Nguyễn Hiếu Gia Cường

MSSV: 20133027

Huỳnh Công Hậu

MSSV: 20133039

TP Hồ Chí Minh, tháng 05 năm 2023

Lời Nhận Xét Của Giảng Viên

*Ngày ... tháng 05 năm 2023
Giảng viên chấm điểm*

Ths. Nguyễn Văn Thành

Mục lục

CHƯƠNG 1: GIỚI THIỆU VỀ ĐỀ TÀI.....	1
1. Lời mở đầu.....	2
2. Tổng quan về DataSet.....	2
2.1. Nguồn dữ liệu được sử dụng.....	2
2.2. Mô tả chi tiết dữ liệu.....	3
3. Công cụ hỗ trợ.....	4
3.1. Visual Studio 2022.....	4
3.2. SSIS.....	5
3.3. SSAS.....	6
3.4. SQL.....	6
3.5. Power BI.....	7
CHƯƠNG 2: TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS).....	9
1. Quá trình tạo mới project SSIS.....	10
2. Quá trình đổ dữ liệu từ excel vào database.....	11
2.1. Quá trình tạo Excel Connection Manager.....	11
2.2. Quá trình tạo OLE DB Connection.....	12
2.3. Quá trình tạo Control Flow.....	15
2.4. Quá trình tạo Data Flow.....	17
2.4.1. Bảng Dim_Time.....	17
2.4.2. Bảng Dim_Location.....	31
2.4.3. Bảng Dim_Customer.....	38
2.4.4. Bảng Dim_Product.....	42
2.4.5. Bảng Fact_Orders.....	47
2.5. Quá trình tạo Execute SQL Task.....	56
CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU (SSAS).....	59
1. Danh sách câu truy vấn.....	60
2. Quá trình xây dựng mô hình.....	60
2.1. Data Sources.....	60
2.2. Data Sources View.....	64
2.3. Cubes.....	66
2.4. Dimension.....	71
3. Các câu truy vấn.....	75
3.1. Tính tỷ lệ đơn hàng được giao ở nước Mỹ so với tổng thể.....	75

3.2.	Thống kê tỉ lệ giao hàng được người dùng chọn.....	75
3.3.	Thống kê doanh số mang lại từ 3 loại hình sản phẩm.....	77
3.4.	Thống kê chi tiết doanh thu mang lại từ từng loại sản phẩm theo Sub-Category	78
3.6.	Liệt kê các sản phẩm bán chạy nhất.....	82
3.7.	Thống kê số sản phẩm bán chạy nhất theo bảng.....	83
3.8.	Liệt kê những khu vực có số lượng mua hàng nhiều nhất.....	84
3.9.	So sánh tổng doanh thu với chi phí ship.....	86
3.10.	So sánh tổng số lượng bán hàng giữa các danh mục.....	88
3.11.	Thống kê những khách hàng mua nhiều sản phẩm nhất.....	89
3.12.	Tìm xu hướng số lượng đơn hàng trong năm.....	90
4.	Dashboard đã thực hiện.....	92

LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin được gửi lời cảm ơn đặc biệt đến Thầy – Nguyễn Văn Thành - Giảng viên phụ trách môn Kho dữ liệu – trường đại học Sư Phạm Kỹ Thuật Tp Hồ Chí Minh .

Trong thời gian nhóm em làm đồ án , tụi em đã nhận được nhiều sự giúp đỡ từ thầy. Thầy đã cung cấp đầy đủ kiến thức, chỉ bảo và đóng góp những ý kiến quý báu giúp tụi em có thể hoàn thành được đồ án môn học của mình một cách tốt nhất.

Xuất phát từ mục đích học tập, tìm hiểu sâu hơn các kiến thức về dữ liệu và các thuật toán, cũng như tìm hiểu kỹ về quy trình lên ý tưởng, xây dựng dashboard. Nhóm chúng em đã thực hiện đồ án “Xây dựng kho dữ liệu để phân tích quá trình bán hàng.”. Trong quá trình thực hiện đồ án, dựa trên kiến thức được Thầy cung cấp qua các buổi học lý thuyết cũng như thực hành trên lớp, kết hợp với việc tự tìm hiểu những công cụ và kiến thức mới , nhóm đã cố gắng thực hiện đồ án một cách tốt nhất .Tuy nhiên, đồ án còn chưa được hoàn hiện và có nhiều sai sót.

Nhóm rất mong nhận được sự góp ý từ thầy nhằm rút ra những kinh nghiệm quý báu và hoàn thiện vốn kiến thức để nhóm có thể hoàn thành những đồ án, dự án khác trong tương lai .

Nhóm chúng em xin chân thành cảm ơn thầy!

CHƯƠNG 1: GIỚI THIỆU VỀ ĐỀ TÀI

Chương này sẽ giới thiệu về tập dữ liệu, lý do chọn đề tài, và các công cụ hỗ trợ

1. Lời mở đầu

Tập dữ liệu "Global Superstore" là một tài nguyên quan trọng cho những ai quan tâm đến phân tích dữ liệu và kinh doanh toàn cầu. Tập dữ liệu này bao gồm thông tin về hơn 51.000 đơn đặt hàng từ 2014 đến 2017, ghi lại các thông tin về khách hàng, sản phẩm, doanh thu và lợi nhuận. Nhóm đã chọn đề tài này vì nó cung cấp một cái nhìn tổng quan về hoạt động kinh doanh toàn cầu của một doanh nghiệp và cho phép phân tích xu hướng, định hướng và chiến lược kinh doanh. Bên cạnh đó, tập dữ liệu này cũng cung cấp một số thách thức thú vị đối với các nhà phân tích dữ liệu, như phát hiện ra nguyên nhân gây ra các khoản lỗ, tìm kiếm các sản phẩm có doanh số cao nhất và đánh giá các chiến lược kinh doanh có hiệu quả hay không. Tập dữ liệu này cung cấp rất nhiều thông tin hữu ích và sẽ cho phép nhóm phân tích và tìm hiểu nhiều hơn về các hoạt động kinh doanh toàn cầu.

2. Tổng quan về DataSet

2.1. Nguồn dữ liệu được sử dụng

Tập dữ liệu "Global Superstore" được cung cấp trên trang Kaggle - một nơi tuyệt vời để tìm kiếm và chia sẻ tập dữ liệu, mô hình và các dự án liên quan đến khoa học dữ liệu và trí tuệ nhân tạo. Kaggle được thành lập vào năm 2010 và trở thành một cộng đồng trực tuyến lớn với hơn 4 triệu người dùng trên toàn thế giới. Tại đây, các nhà nghiên cứu, chuyên gia dữ liệu và các nhà phát triển có thể tìm kiếm, chia sẻ và tham gia các cuộc thi liên quan đến khoa học dữ liệu và trí tuệ nhân tạo. Bên cạnh đó, Kaggle cũng cung cấp một loạt các tài nguyên và công cụ hỗ trợ cho các nhà phân tích dữ liệu như học máy, xử lý dữ liệu và trực quan hóa dữ liệu. Với tập dữ liệu "Global Superstore" được cung cấp trên Kaggle, người dùng có thể phân tích và tìm hiểu nhiều hơn về hoạt động kinh doanh toàn cầu và ứng dụng trong nhiều lĩnh vực khác nhau như marketing, bán lẻ và chuỗi cung ứng.

Link download dataset: <https://www.kaggle.com/datasets/shekpaul/global-superstore>

2.2. Mô tả chi tiết dữ liệu

Tập dữ liệu "Global Superstore" là một bộ dữ liệu về doanh số bán hàng và lợi nhuận của một chuỗi cửa hàng bán lẻ toàn cầu trong khoảng thời gian từ năm 2014 đến 2017. Bộ dữ liệu này được cung cấp trên trang Kaggle và bao gồm hơn 51.000 bản ghi về các đơn đặt hàng, bao gồm thông tin về khách hàng, sản phẩm, doanh thu và lợi nhuận.

Mô tả chi tiết các thuộc tính trong dataset:

Tên thuộc tính	Mô tả
RowID	Số thứ tự của bản ghi trong tập dữ liệu
OrderID	Mã số đơn hàng
Order Date	Ngày đặt hàng
Ship Date	Ngày giao hàng
Ship Mode	Phương thức vận chuyển
Customer ID	Mã số khách hàng
Customer Name	Tên khách hàng
Segment	Phân khúc thị trường khách hàng (cá nhân, doanh nghiệp, chính phủ)
Country	Quốc gia
City	Thành phố
State	Bang hoặc tỉnh
Postal Code	Mã bưu chính
Region	Khu vực địa lý
Product ID	Mã số sản phẩm
Category	Loại sản phẩm
Sub-Category	Phân loại sản phẩm
Product Name	Tên sản phẩm

Sales	Doanh số bán hàng
Quantity	Số lượng sản phẩm bán ra
Discount	Tỷ lệ chiết khấu
Profit	Lợi nhuận
Shipping Cost	Chi phí vận chuyển
Order Priority	Mức độ ưu tiên của đơn hàng

3. Công cụ hỗ trợ

- Visual Studio 2022 tích hợp SSIS (SQL Server Integration Services) và SSAS (SQL Server Analysis Services).
- SQL Server Management Studio, SQL Server 2019
- Power BI.

3.1. Visual Studio 2022

Visual Studio 2022 là một phiên bản mới nhất của môi trường phát triển tích hợp (IDE) phổ biến của Microsoft dành cho các nhà phát triển phần mềm. Nó được phát hành vào tháng 4 năm 2021 và được cung cấp với nhiều tính năng mới và cải tiến so với các phiên bản trước đó.

Một trong những tính năng mới đáng chú ý của Visual Studio 2022 là việc hỗ trợ đa nền tảng, cho phép người dùng phát triển ứng dụng cho các nền tảng như Windows, Linux và macOS. Ngoài ra, Visual Studio 2022 cũng cải tiến hiệu suất và khả năng tương thích, giúp tăng tốc độ xử lý và tải ứng dụng.

Visual Studio 2022 cũng có nhiều tính năng mới hỗ trợ cho lập trình viên, bao gồm khả năng tự động hoàn tất mã nguồn, gợi ý mã, định dạng mã nguồn tự động và kiểm tra lỗi trong quá trình phát triển. Ngoài ra, Visual

Studio 2022 cũng cung cấp một loạt các công cụ hỗ trợ cho việc kiểm tra, gỡ lỗi và triển khai ứng dụng của bạn.

Với nhiều tính năng mới và cải tiến, Visual Studio 2022 là một công cụ hữu ích cho các nhà phát triển phần mềm trong việc phát triển và triển khai ứng dụng trên nhiều nền tảng khác nhau.

3.2. SSIS

SSIS (SQL Server Integration Services) là một công cụ ETL (Extract, Transform, Load) được phát triển bởi Microsoft để quản lý quá trình tích hợp dữ liệu giữa các nguồn khác nhau. Nó được tích hợp sẵn trong SQL Server và là một phần quan trọng của nền tảng dữ liệu của Microsoft.

SSIS cho phép người dùng trích xuất dữ liệu từ các nguồn khác nhau, chuyển đổi và tùy chỉnh dữ liệu, sau đó tải dữ liệu đã được xử lý vào các kho dữ liệu hay hệ thống khác. Với SSIS, người dùng có thể tạo các gói tích hợp dữ liệu để quản lý dữ liệu từ nhiều nguồn khác nhau, bao gồm các cơ sở dữ liệu, file dữ liệu, API và các dịch vụ web khác.

SSIS cung cấp nhiều tính năng hỗ trợ cho việc quản lý quá trình tích hợp dữ liệu, bao gồm khả năng kiểm soát đồng bộ hóa, quản lý lỗi và thời gian chạy. Ngoài ra, SSIS cũng hỗ trợ tích hợp với các công cụ và dịch vụ khác của Microsoft, bao gồm SQL Server Management Studio và Visual Studio.

Với khả năng tích hợp sâu sắc với SQL Server và các tính năng mạnh mẽ, SSIS là một công cụ phổ biến trong việc quản lý quá trình tích hợp dữ liệu và được sử dụng rộng rãi trong các doanh nghiệp và tổ chức lớn.

3.3. SSAS

SSAS (SQL Server Analysis Services) là một công cụ phân tích dữ liệu và khai thác dữ liệu được phát triển bởi Microsoft. Nó được tích hợp sẵn trong SQL Server và là một phần quan trọng của nền tảng dữ liệu của Microsoft.

SSAS cho phép người dùng tạo các khối dữ liệu OLAP (Online Analytical Processing) để phân tích dữ liệu và cung cấp cho người dùng thông tin dựa trên các tập dữ liệu khác nhau. Với SSAS, người dùng có thể tạo các báo cáo phân tích dữ liệu và tùy chỉnh chúng để phù hợp với nhu cầu của mình.

SSAS cung cấp nhiều tính năng hỗ trợ cho việc quản lý và phân tích dữ liệu, bao gồm khả năng tạo các chiều, kỳ hạn và phân loại dữ liệu. Nó cũng hỗ trợ nhiều kiểu dữ liệu khác nhau, bao gồm dữ liệu số, dữ liệu ngày và dữ liệu văn bản. Ngoài ra, SSAS còn hỗ trợ tính năng phân tích dữ liệu thời gian thực và khả năng phân tích dữ liệu động.

Với khả năng phân tích dữ liệu mạnh mẽ và tính năng linh hoạt, SSAS là một công cụ quan trọng trong việc phân tích và khai thác dữ liệu và được sử dụng rộng rãi trong các doanh nghiệp và tổ chức lớn.

3.4. SQL

SQL (Structured Query Language) là một ngôn ngữ lập trình được sử dụng để truy xuất và quản lý cơ sở dữ liệu quan hệ. Nó được sử dụng rộng rãi trong các hệ thống quản lý cơ sở dữ liệu phổ biến như MySQL, Microsoft SQL Server, Oracle và PostgreSQL.

SQL cung cấp các câu lệnh để truy xuất và tương tác với cơ sở dữ liệu quan hệ. Các câu lệnh SQL cho phép người dùng tạo, sửa đổi và xóa dữ liệu

trong cơ sở dữ liệu, cũng như tạo các bảng, chỉnh sửa cấu trúc và các ràng buộc của cơ sở dữ liệu.

SQL được coi là một trong những ngôn ngữ lập trình quan trọng nhất trong việc quản lý và truy xuất dữ liệu. Với khả năng truy xuất và quản lý dữ liệu một cách nhanh chóng và hiệu quả, SQL được sử dụng rộng rãi trong các doanh nghiệp và tổ chức lớn để quản lý dữ liệu và phân tích dữ liệu. Ngoài ra, SQL cũng được sử dụng rộng rãi trong các ứng dụng web và mobile để lưu trữ và truy xuất dữ liệu của người dùng.

3.5. Power BI

Power BI là một trong những công cụ phân tích dữ liệu hàng đầu của Microsoft, cho phép người dùng kết nối, phân tích và hiển thị dữ liệu một cách dễ dàng và nhanh chóng. Với Power BI, người dùng có thể tạo các báo cáo và trang tổng quan để phân tích dữ liệu từ nhiều nguồn khác nhau, bao gồm cả dữ liệu trong bảng tính Excel, cơ sở dữ liệu SQL Server, các ứng dụng đám mây như Salesforce và nhiều nguồn dữ liệu khác.

Power BI cung cấp một giao diện đồ họa thân thiện và trực quan, giúp người dùng dễ dàng tạo ra các báo cáo và trang tổng quan với các biểu đồ, đồ thị và bản đồ tương tác. Nó cũng có tính năng trích xuất dữ liệu và phân tích dữ liệu tự động, giúp người dùng tiết kiệm thời gian và tối ưu hóa quá trình phân tích dữ liệu.

Power BI được sử dụng rộng rãi trong các doanh nghiệp và tổ chức lớn để theo dõi và phân tích dữ liệu kinh doanh, quản lý dữ liệu và tổng hợp thông tin từ nhiều nguồn khác nhau. Nó cũng được sử dụng trong các ứng dụng web và mobile để hiển thị thông tin dữ liệu cho người dùng cuối. Với

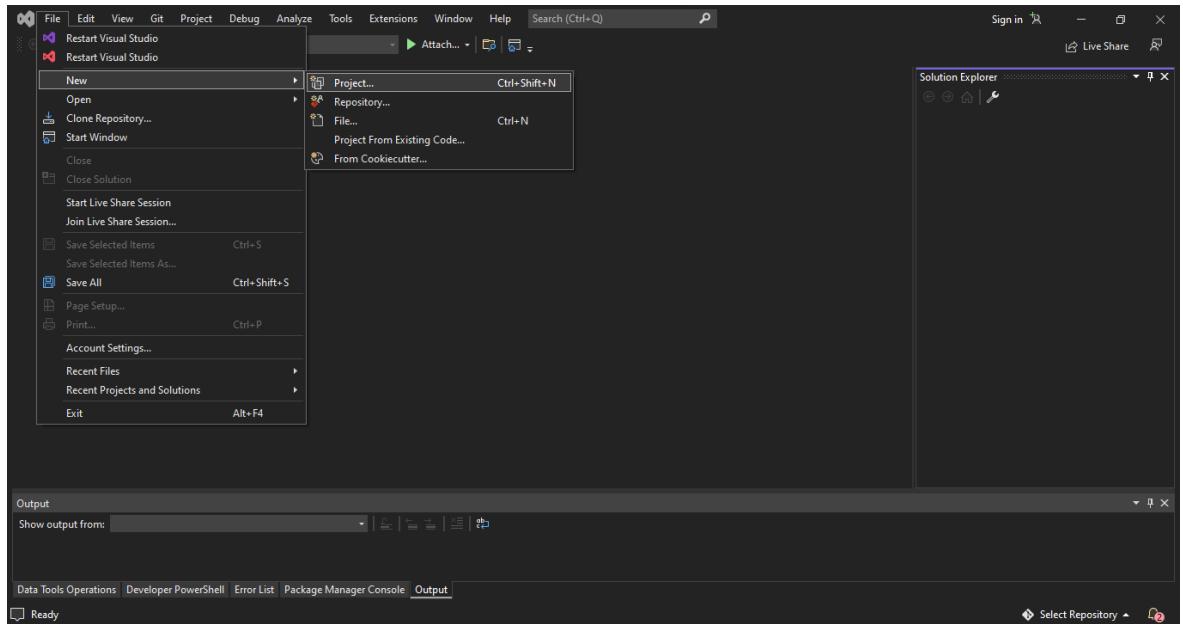
tính năng mạnh mẽ và tính năng linh hoạt, Power BI là một công cụ phân tích dữ liệu quan trọng và được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau.

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

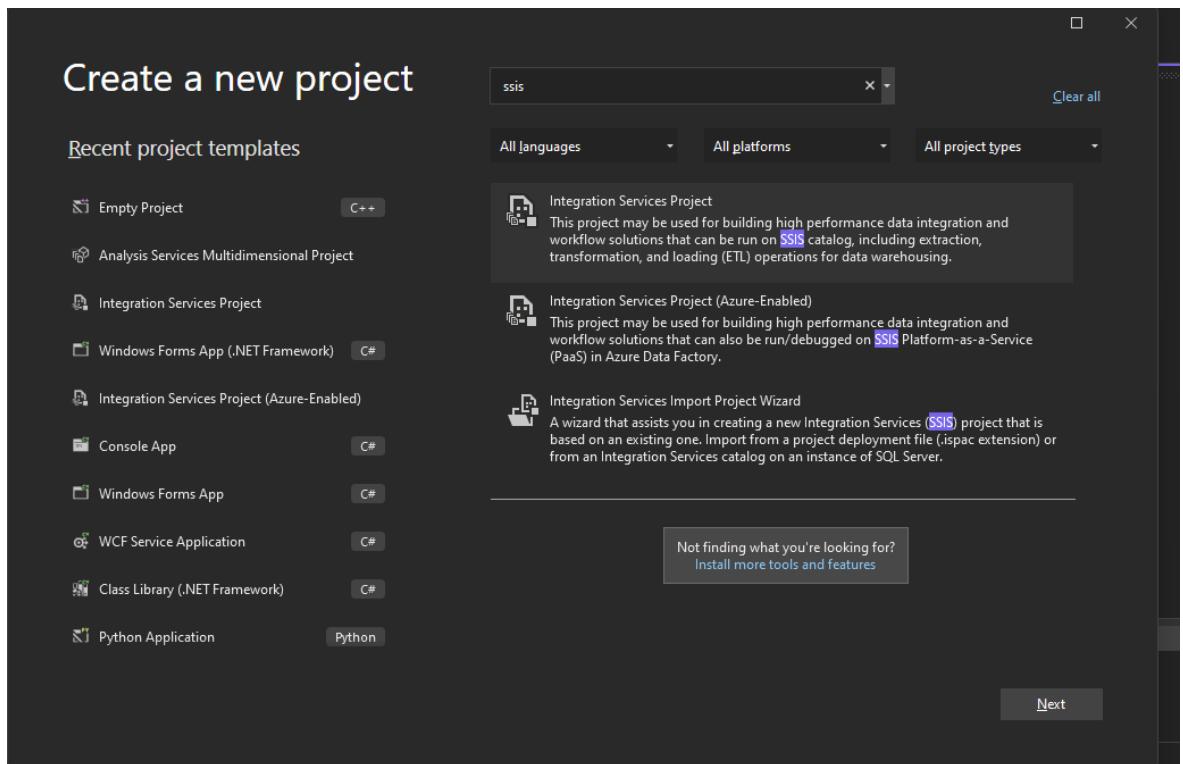
Ở chương này, Sử dụng công cụ SSIS xây dựng các gói tích hợp dữ liệu, cài đặt các gói chạy tự động. Viết truy vấn đưa dữ liệu đến kho dữ liệu.

1. Quá trình tạo mới project SSIS

Chọn tag File – New - Project



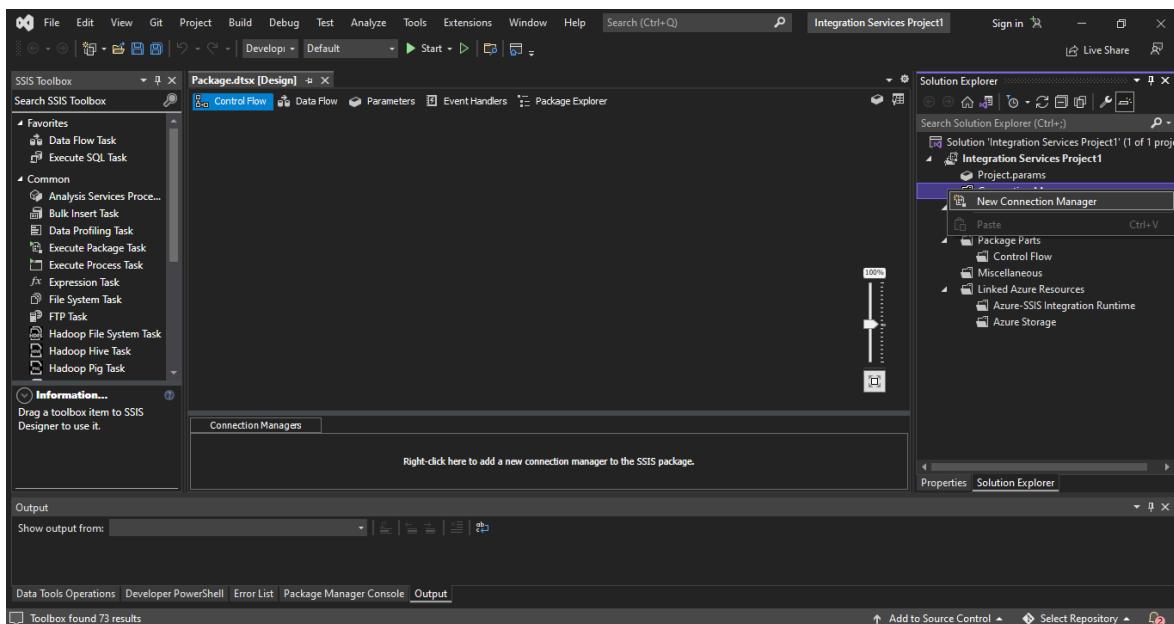
Một dialog xuất hiện. Nhập **SSIS - Intergration Services Project - Nhập Name và Solution name. - OK**



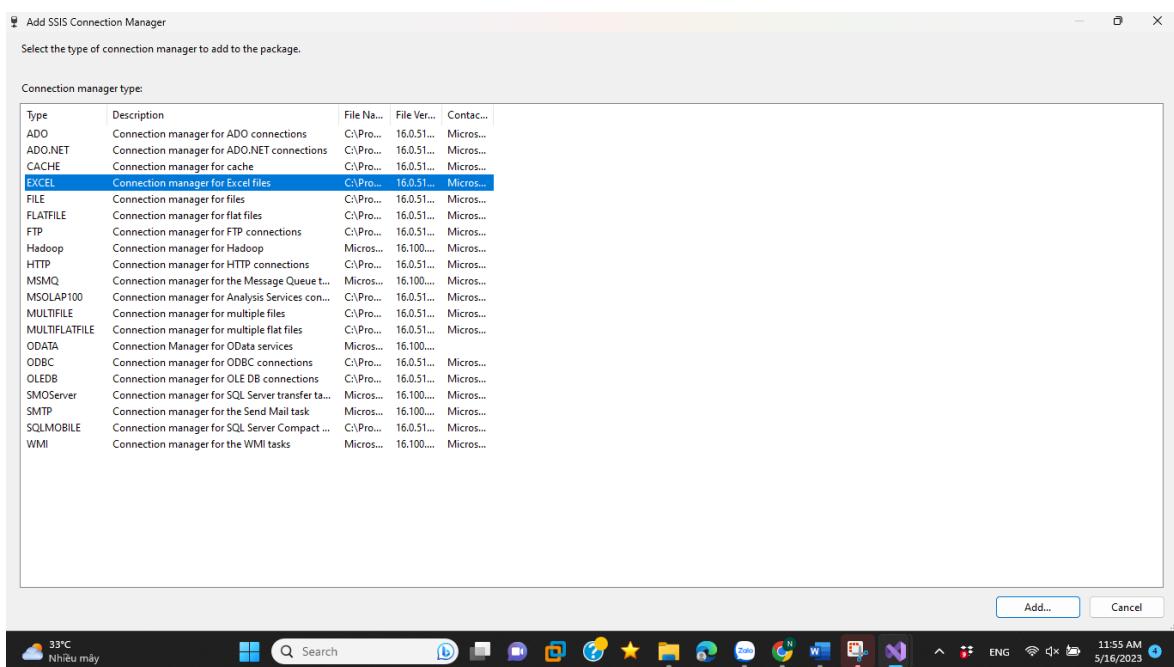
2. Quá trình đổ dữ liệu từ excel vào database

2.1. Quá trình tạo Excel Connection Manager

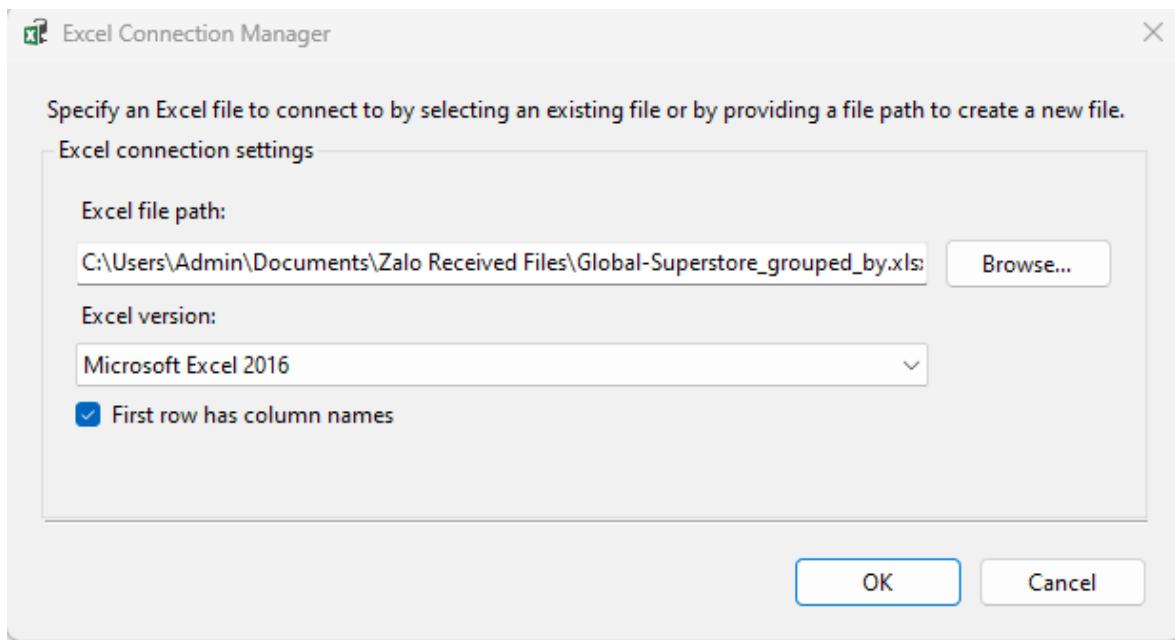
Nhấn chuột phải chọn New Connection:



Dialog config hiện ra chọn Excel – Add ...

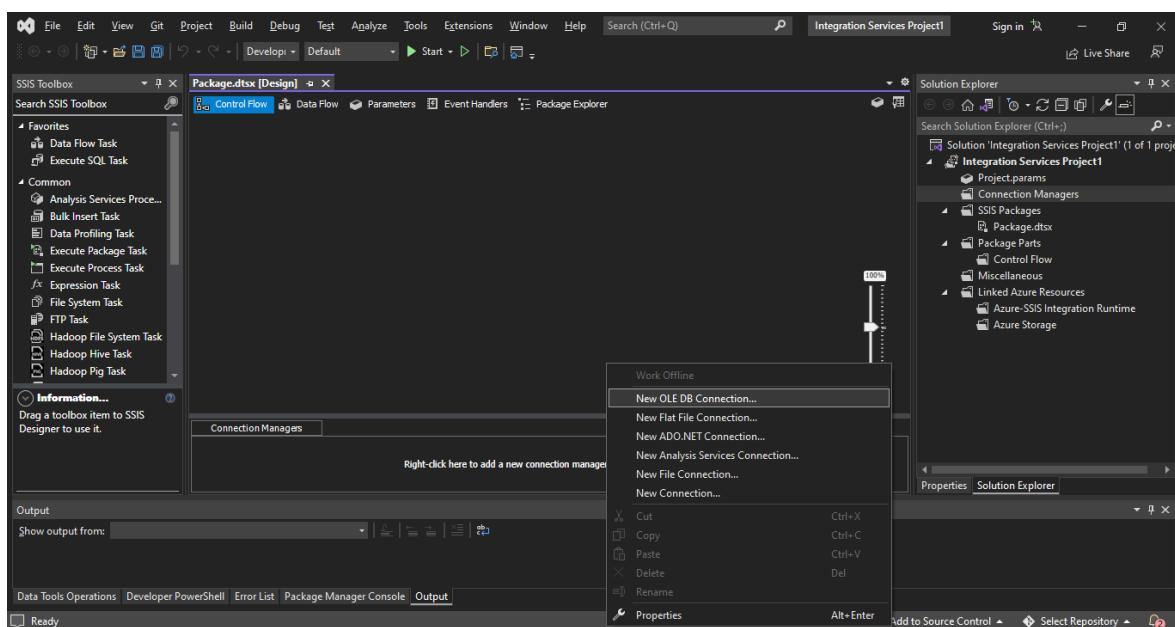


Chọn đường dẫn tới file excel và phiên bản excel đó - Ok

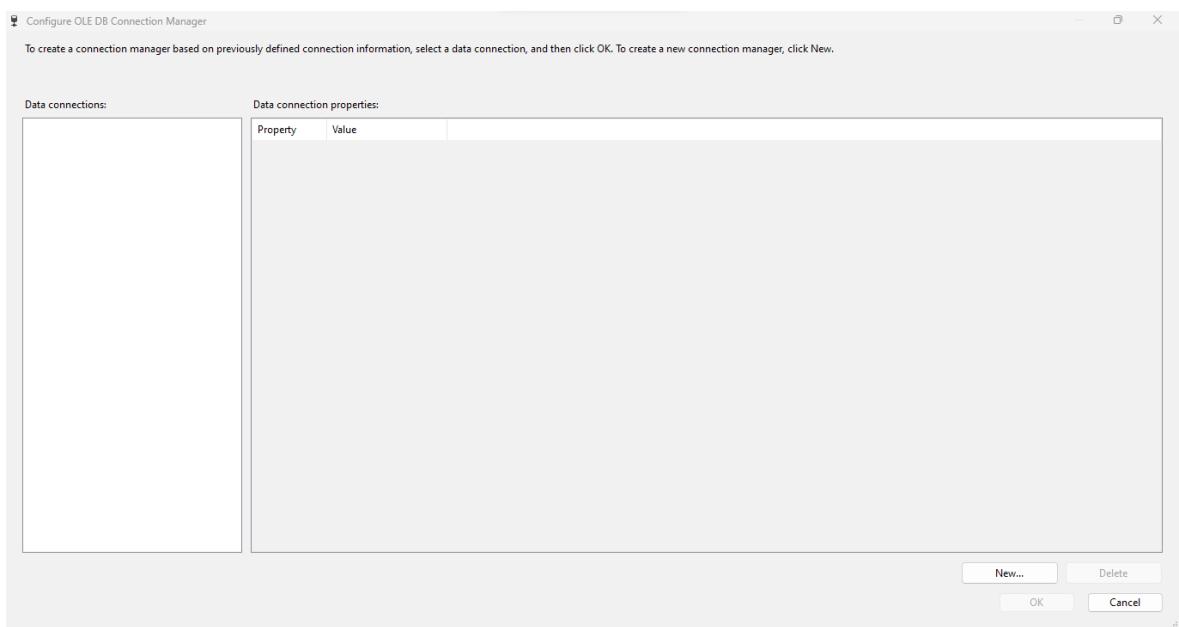


2.2. Quá trình tạo OLE DB Connection

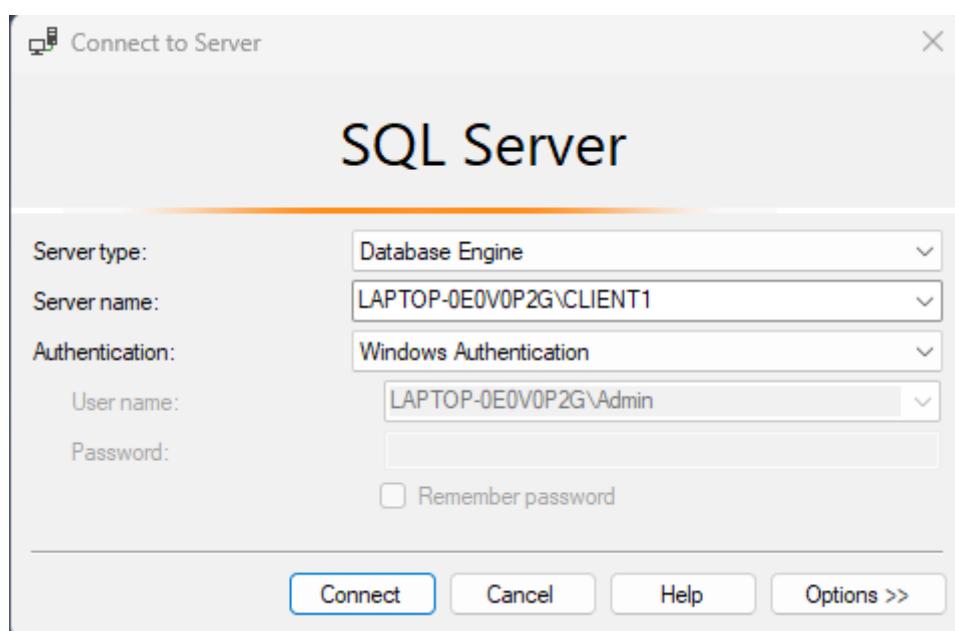
Nhấn chuột phải – chọn New OLE DB Connection

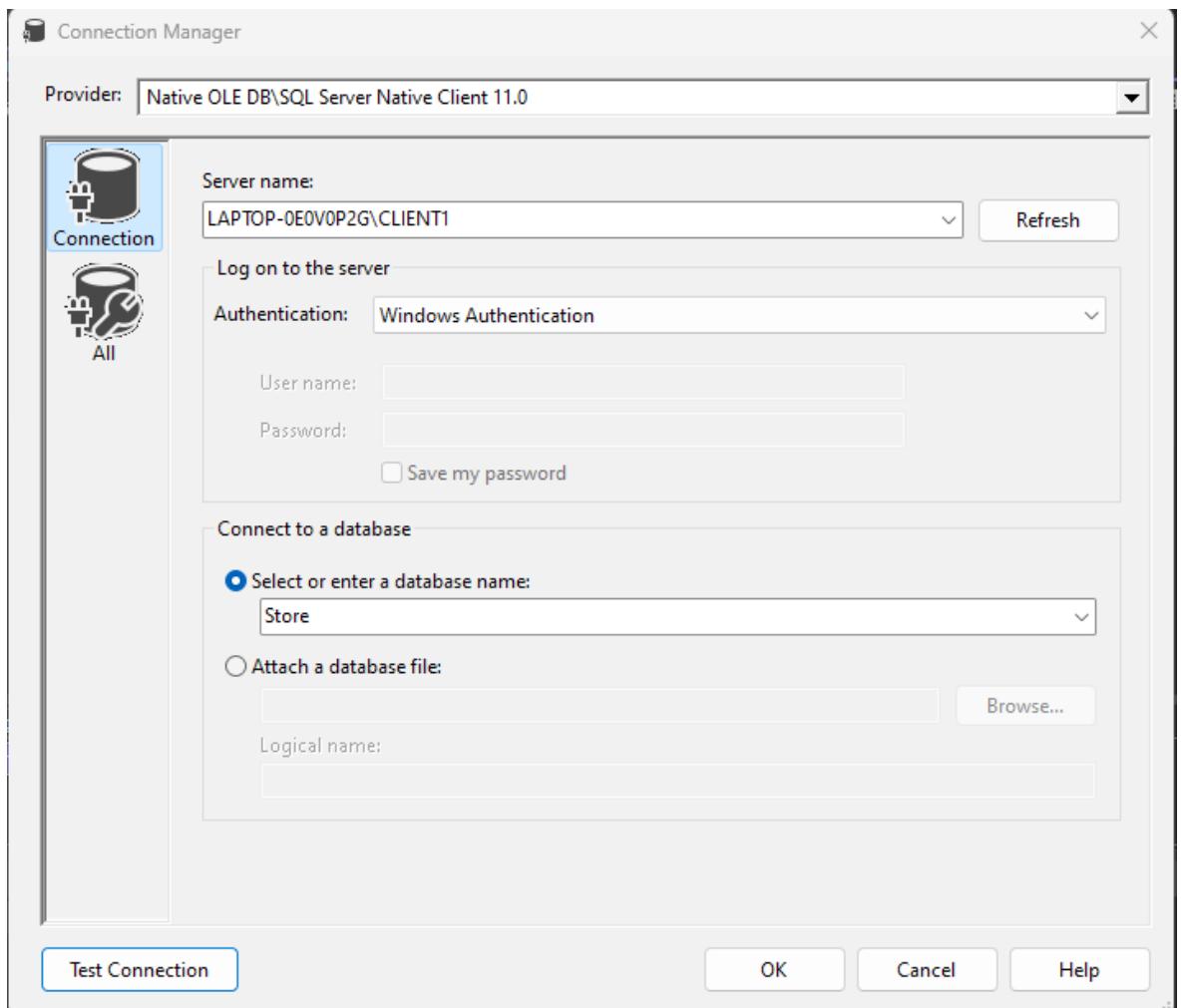


Một cửa sổ mới hiện ra như sau:

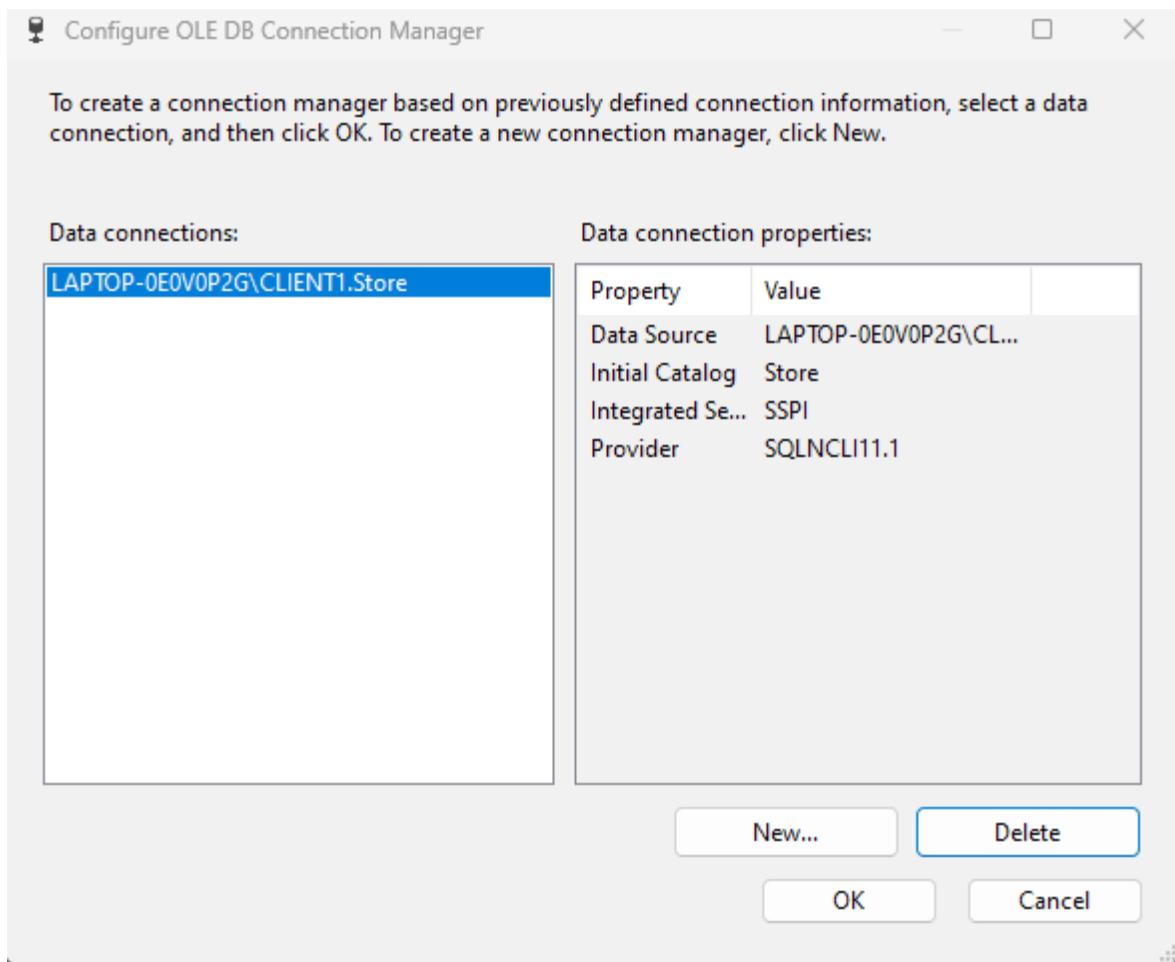


Chọn New ... một cửa sổ mới hiện ra, chọn kết nối tới SQL Server, nhập database đã tạo, chọn OK.





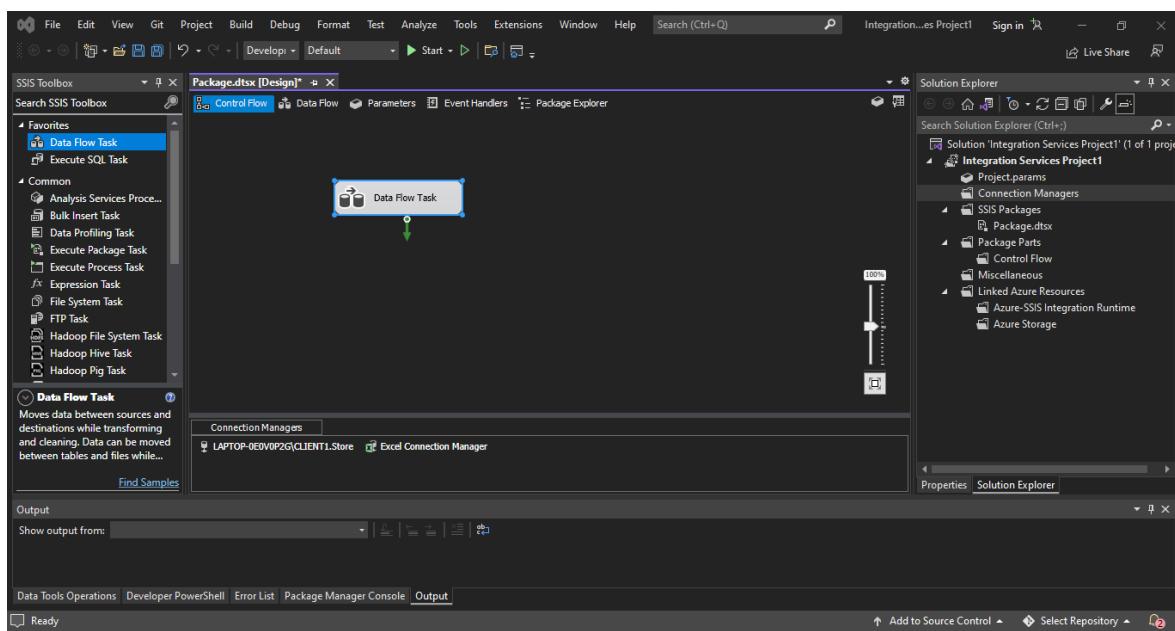
Chọn OK.

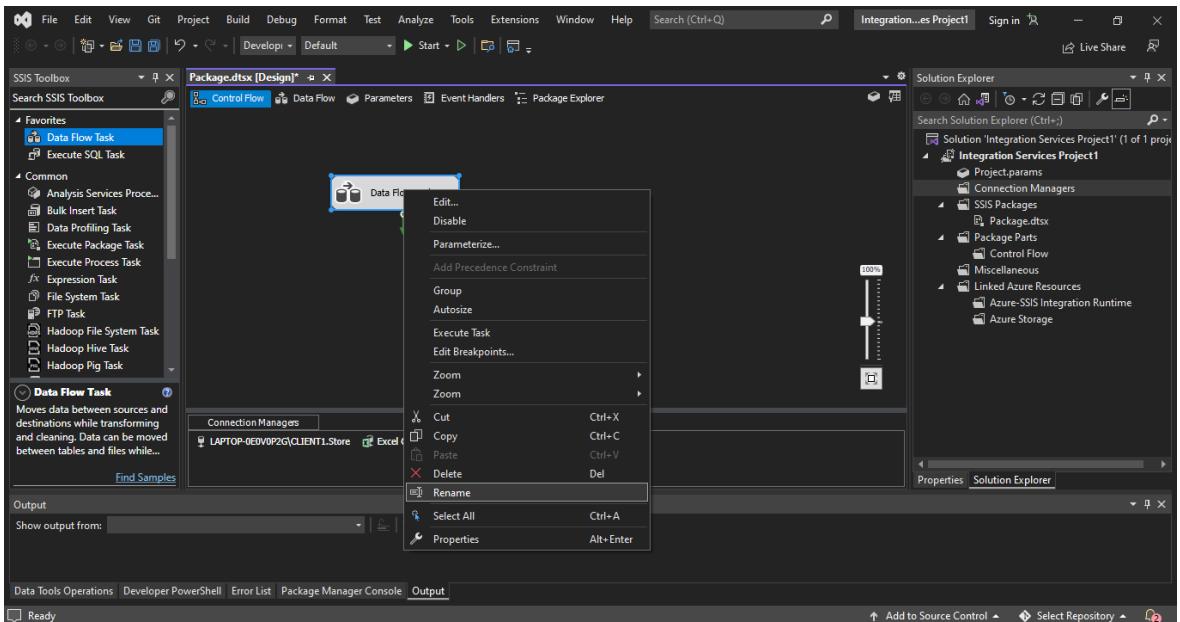


Chọn OK

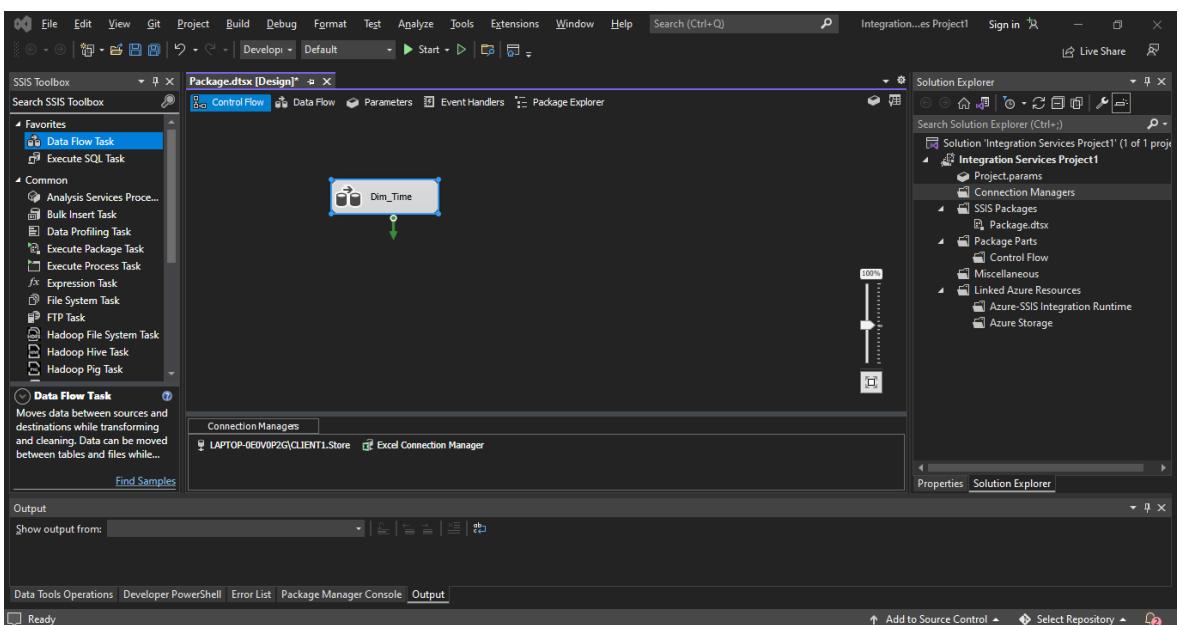
2.3. Quá trình tạo Control Flow

Kéo 1 Data Flow Task mới

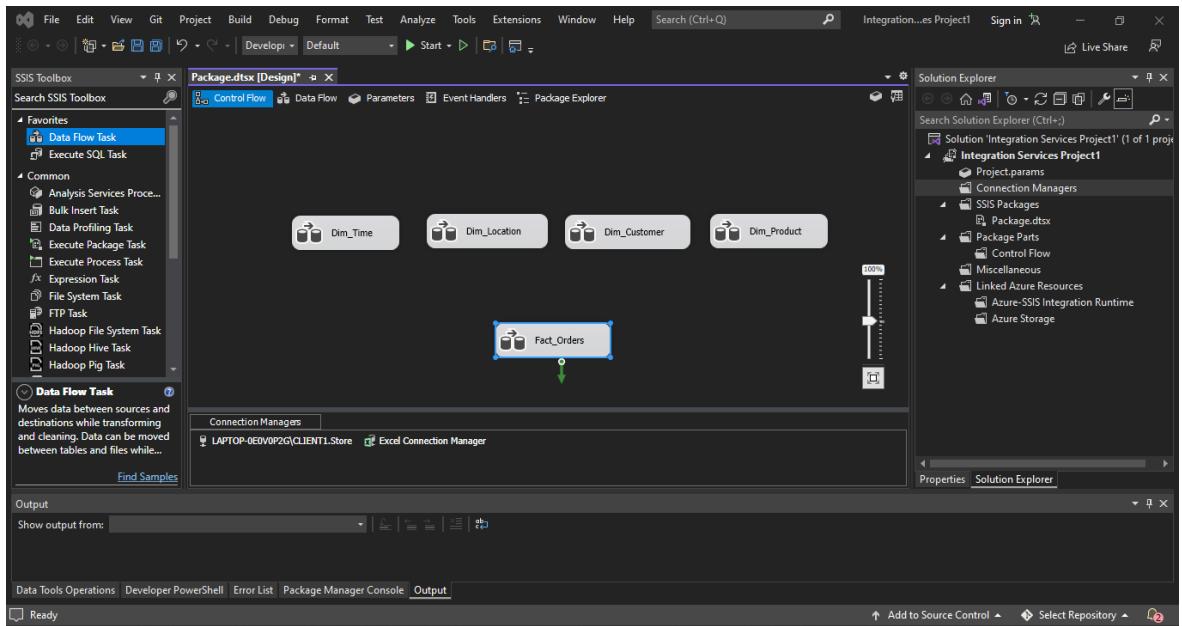




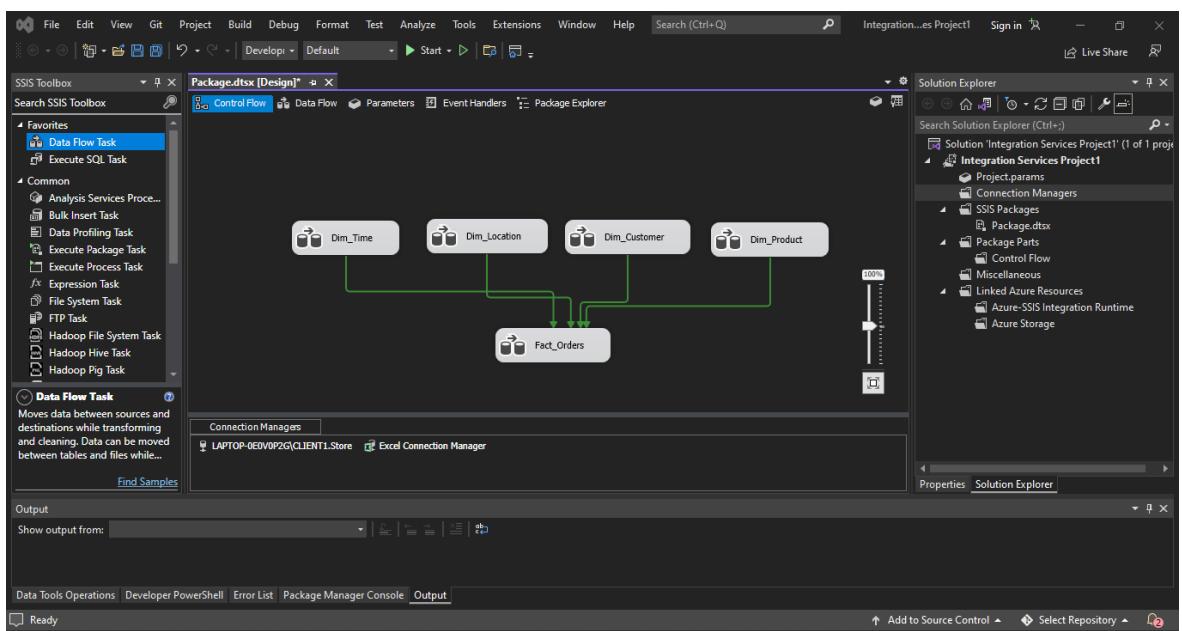
Chọn Rename cho Data Task Flow trên và đổi tên thành Dim_Time



Tạo thêm 4 cái nữa, lần lượt đặt tên là Dim_Location, Dim_Customer và Dim_Product, Fact_Orders



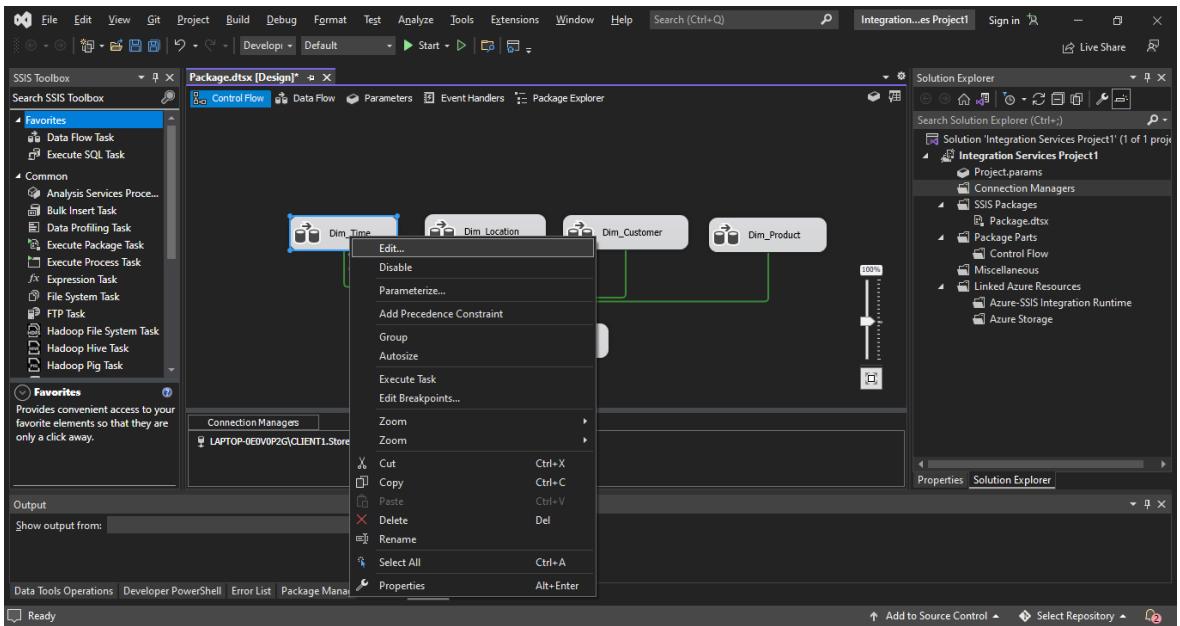
Nối 4 bảng Dim với bảng Fact_Orders



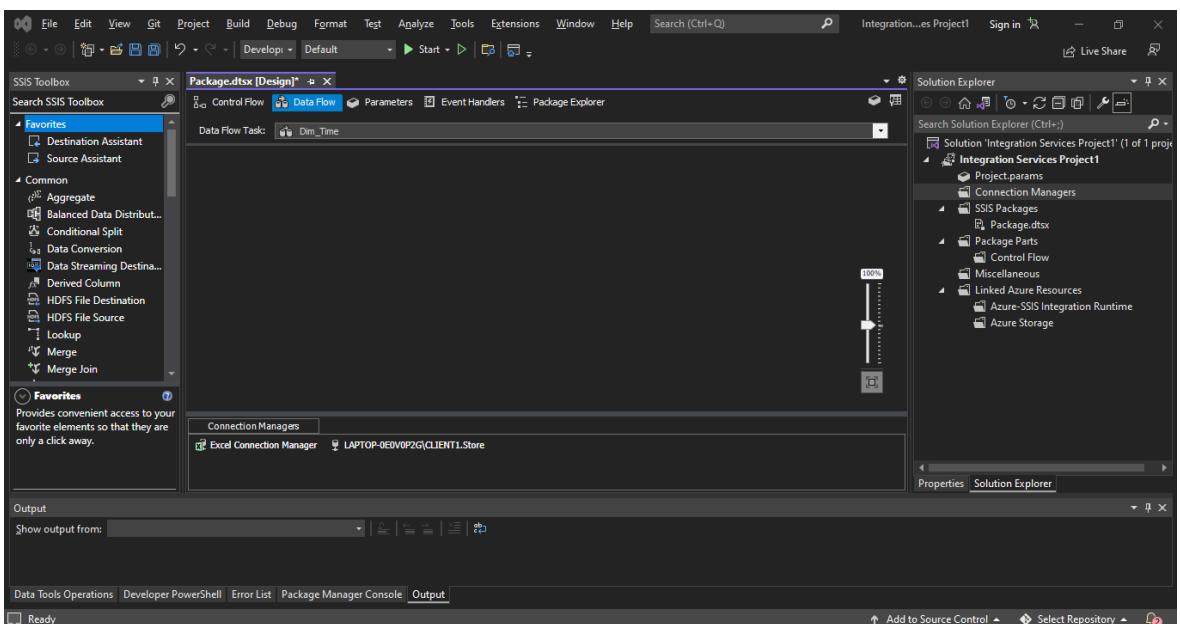
2.4. Quá trình tạo Data Flow

2.4.1. Bảng Dim_Time

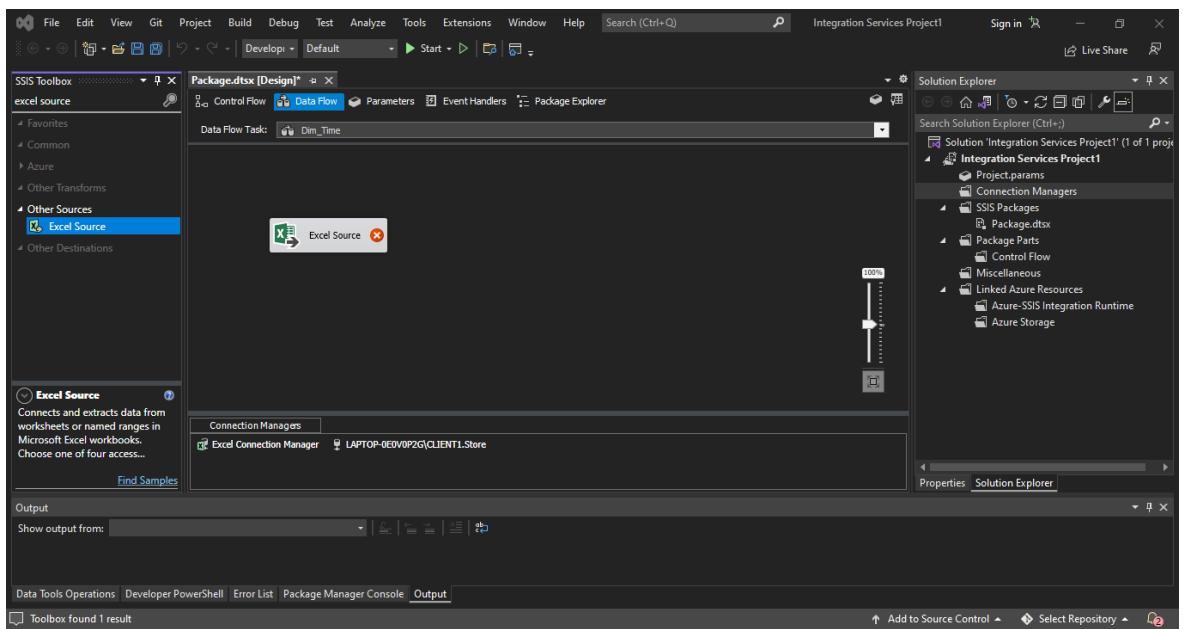
Chuột phải vào bảng Dim_Time chọn Edit



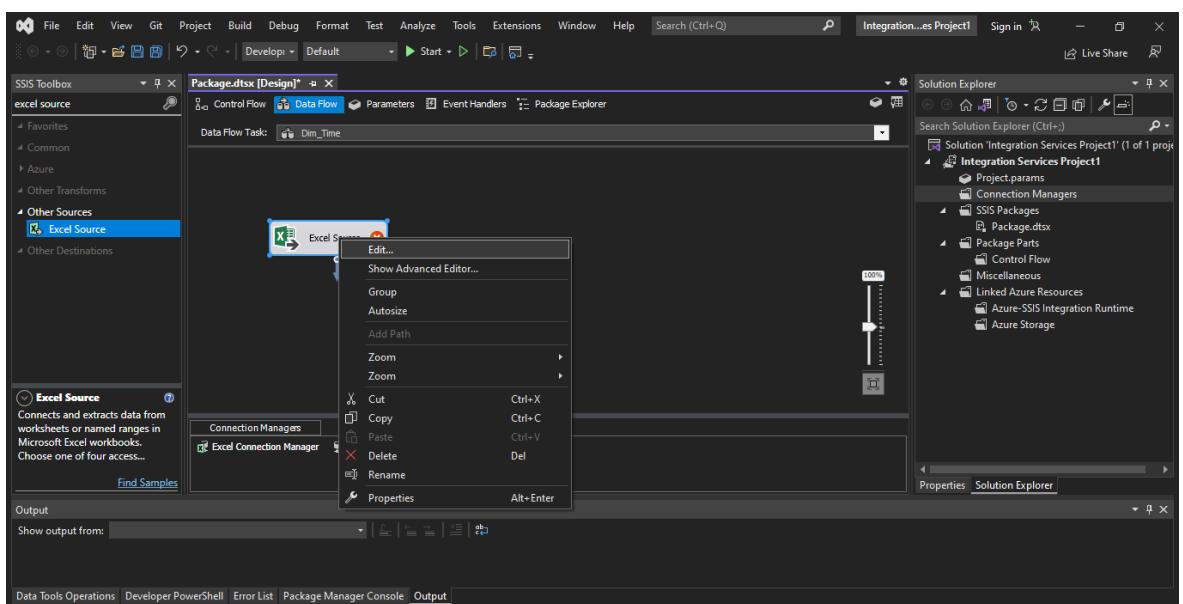
1 cửa sổ Data Flow sẽ hiện ra



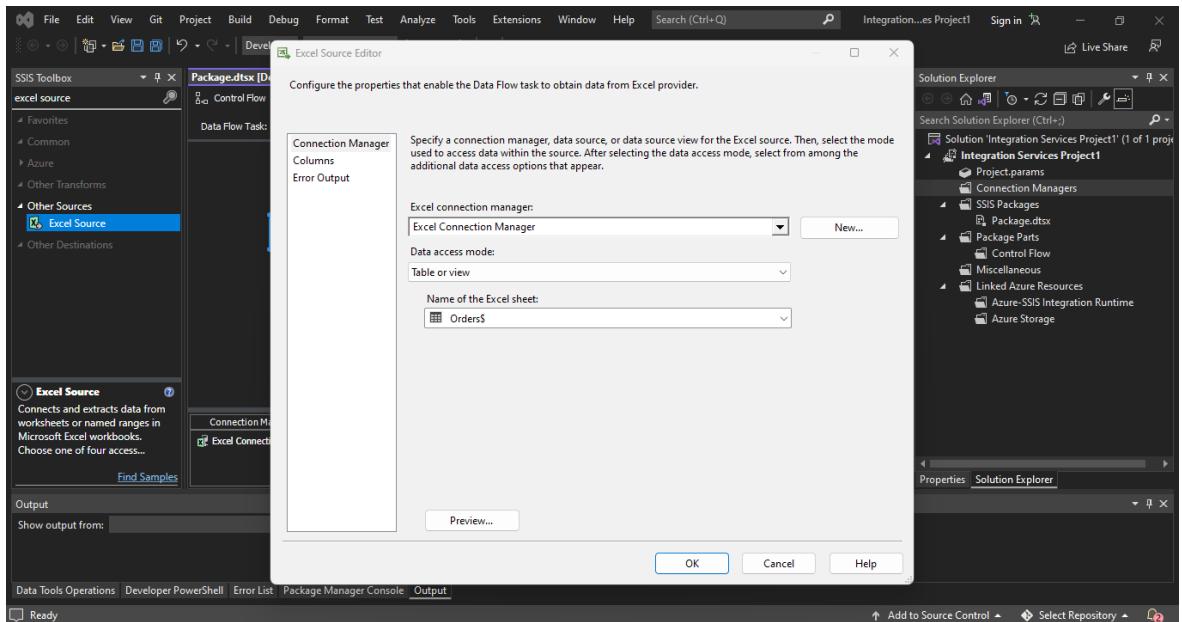
Search trong SSIS Toolbox tên Excel Source rồi kéo thả vào Data Flow



Click chuột phải chọn Edit



Cửa sổ Excel Source Editor sẽ hiện ra



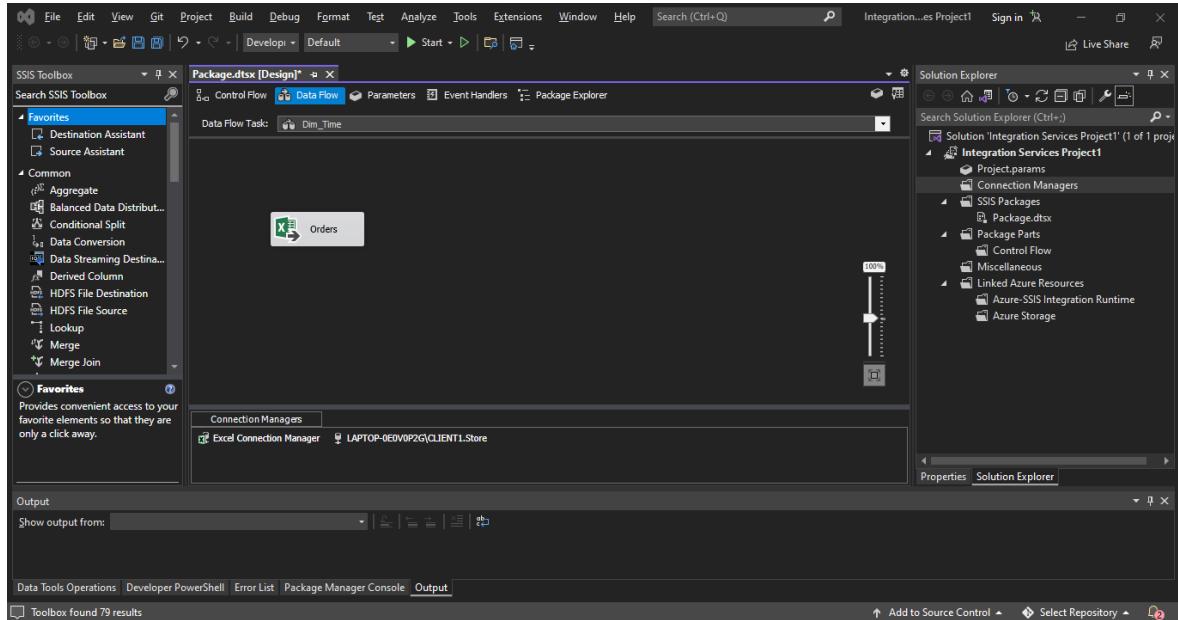
Chọn bảng Order\$

Chọn tùy chọn **Columns** bên trái để xem kiểm tra data

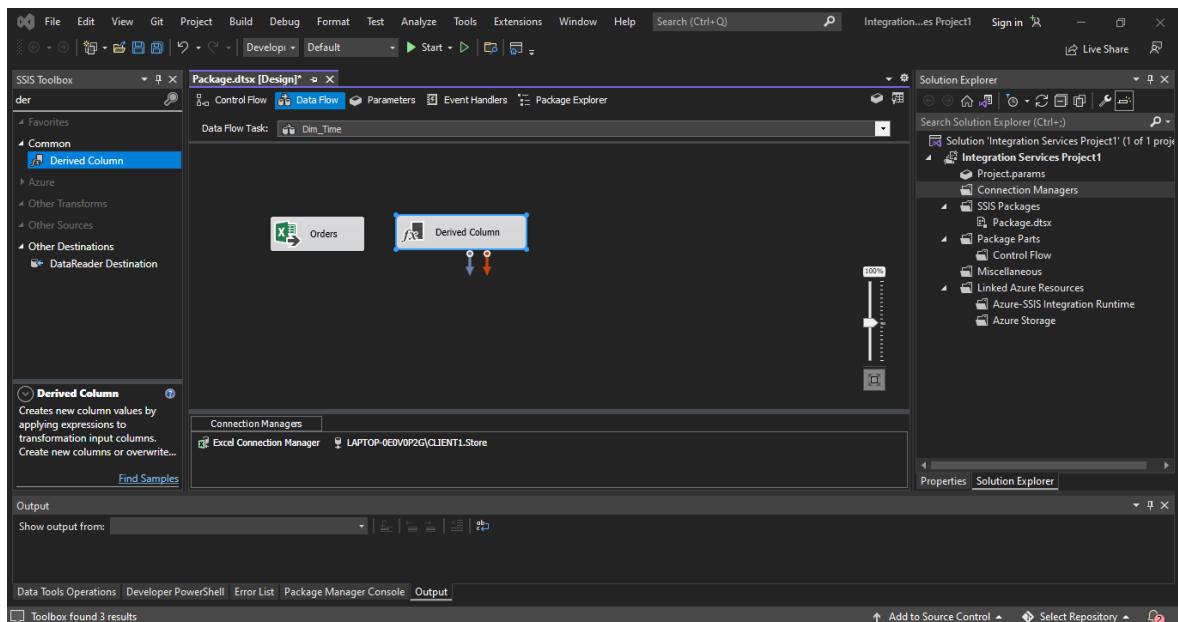
External Column	Output Column
Row ID	Row ID
Order ID	Order ID
Order Date	Order Date
Ship Date	Ship Date
Ship Mode	Ship Mode
Customer ID	Customer ID
Customer Name	Customer Name
Segment	Segment
City	City

Click OK

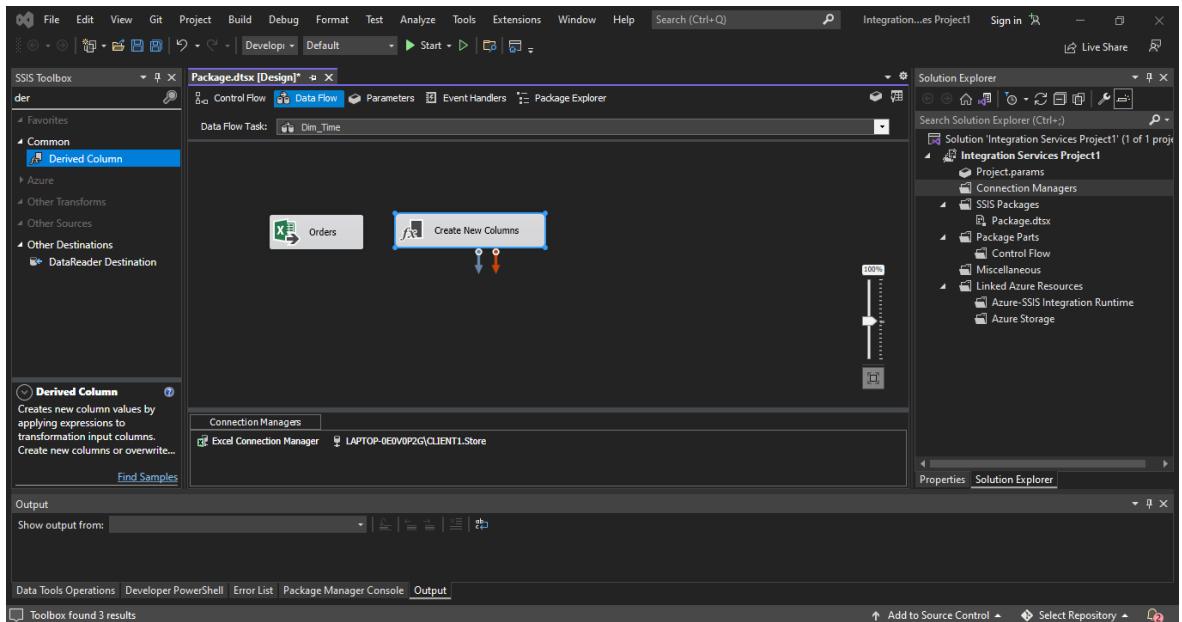
Đổi tên Excel Source lại thành Orders



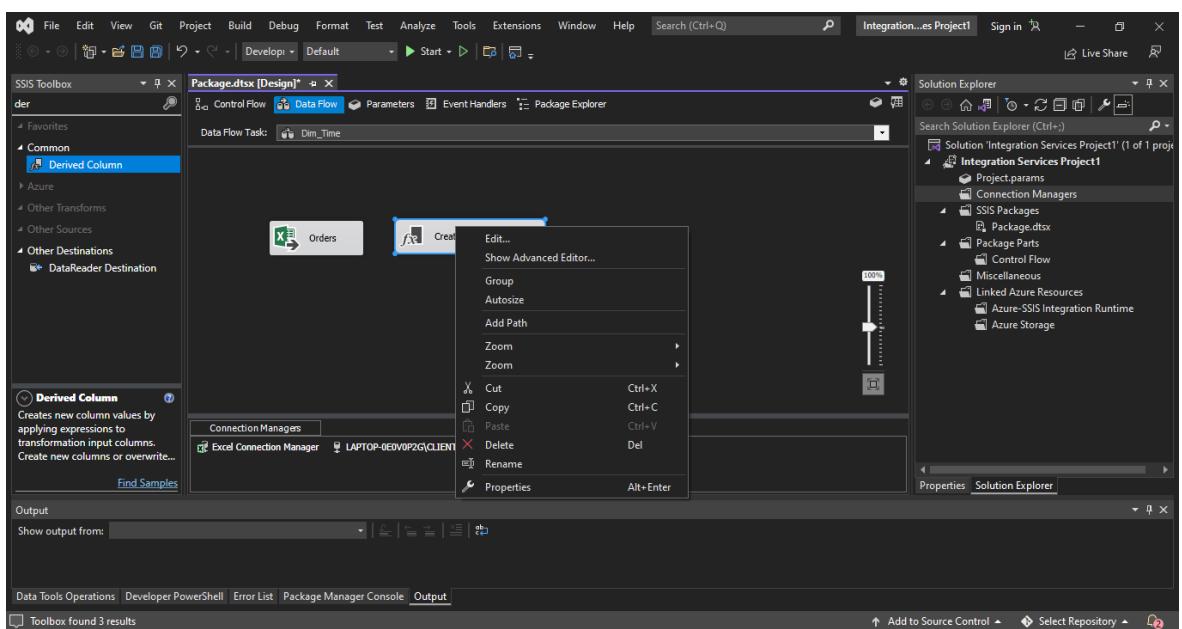
Kéo thả Derived Column vào Data Flow



Đổi tên lại thành Create New Columns

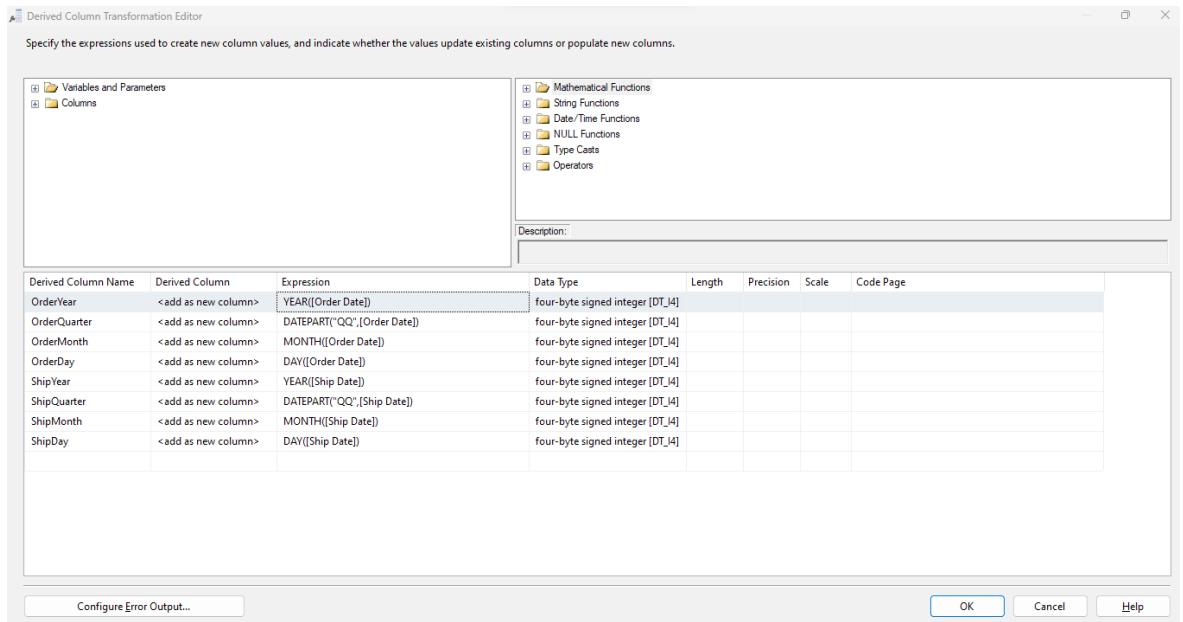


Click chuột phải chọn **Edit**, 1 cửa sổ mới sẽ hiện ra

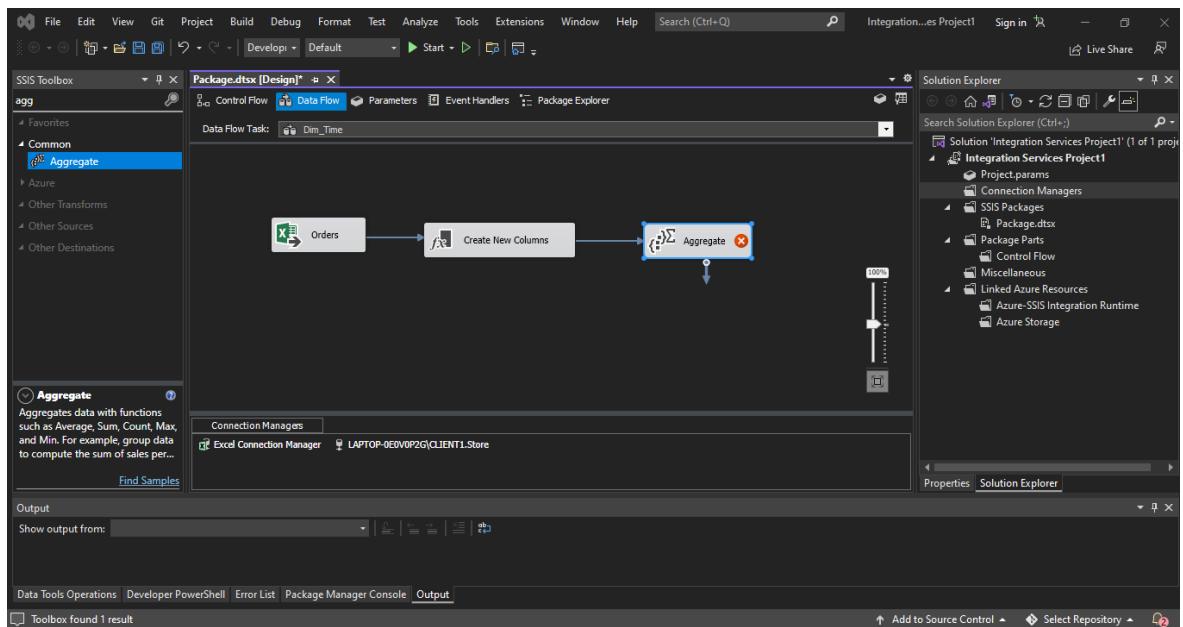


Đặt tên cho các cột mới lần lượt là **OrderYear, OrderQuater, OrderMonth, OrderDay, ShipYear, ShipQuater, ShipMonth, ShipDay**

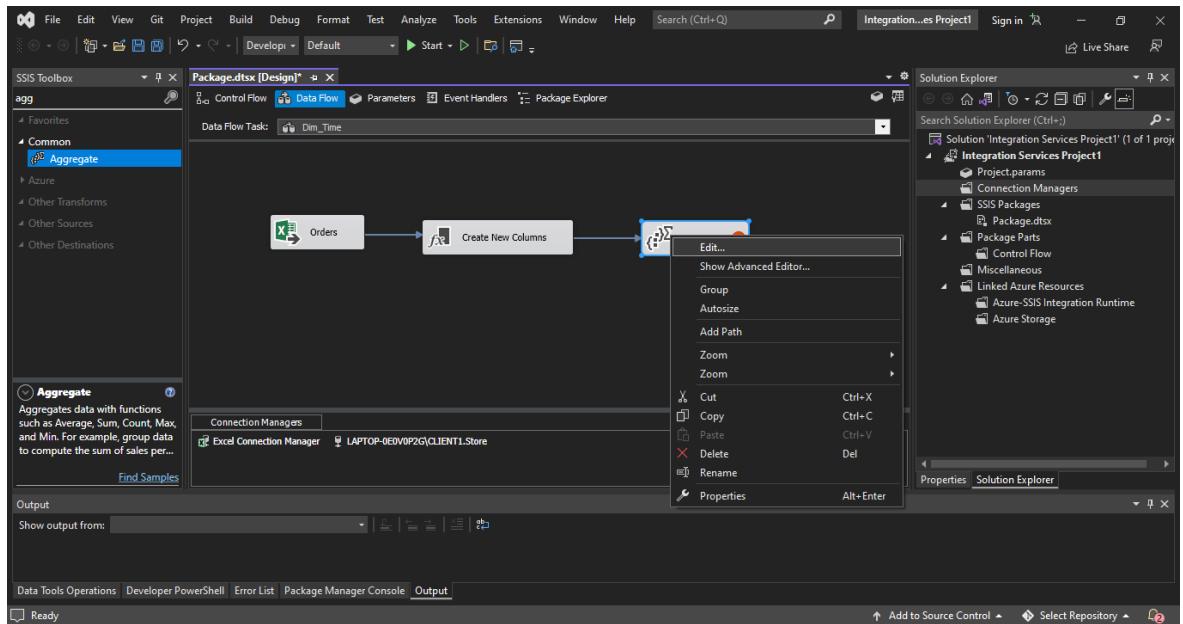
Các Expression sử dụng các câu lệnh SQL như sau:



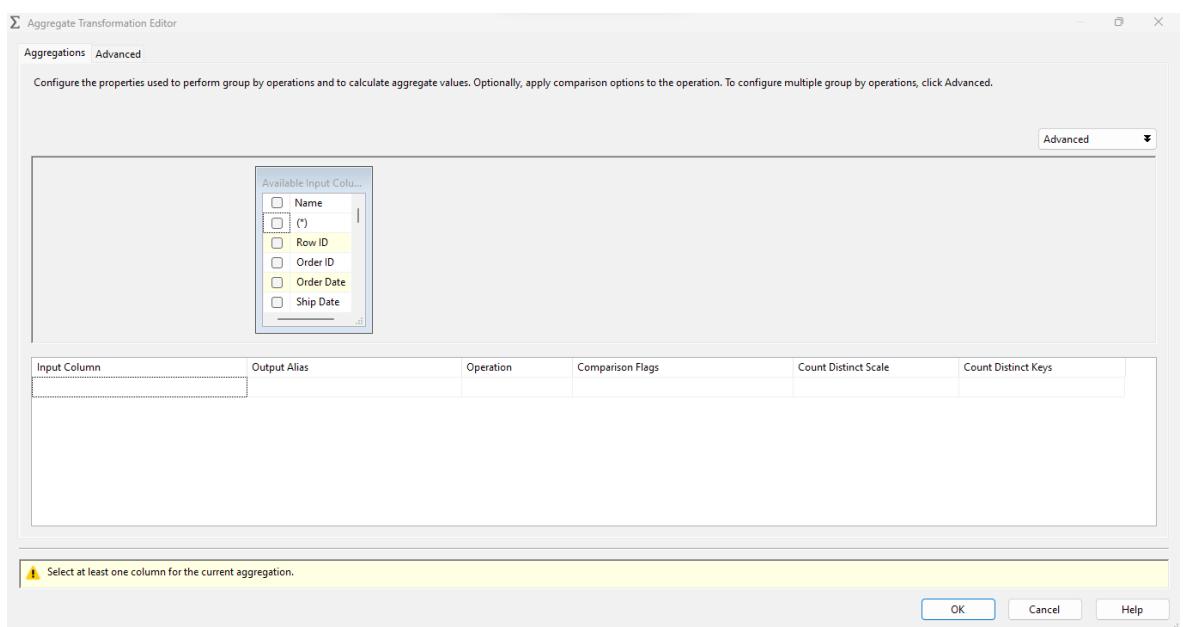
Kéo thả Aggregate và nối với các Excel Source và Derived Column Transformation Editor



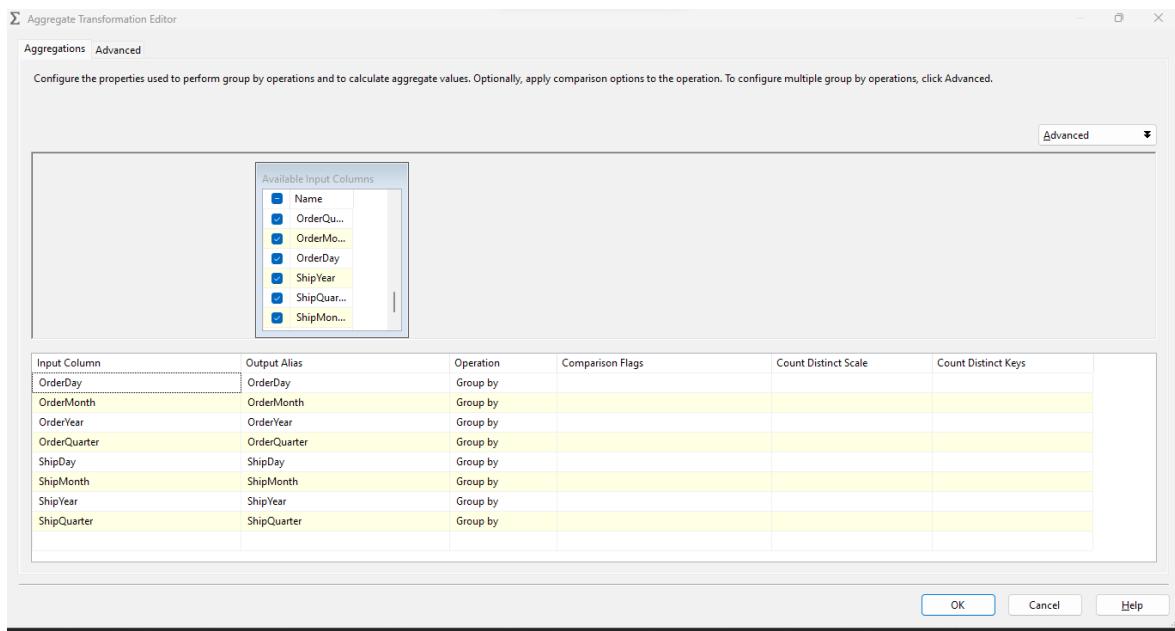
Click chuột phải chọn Edit



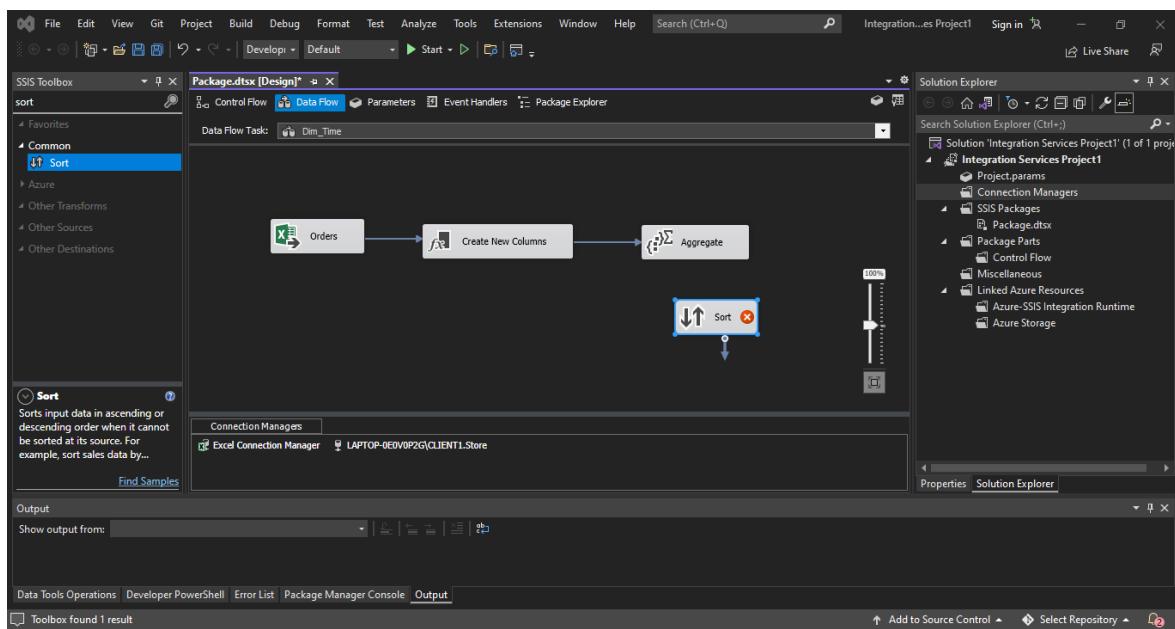
Cửa sổ Aggregate Transformation Editor hiện ra



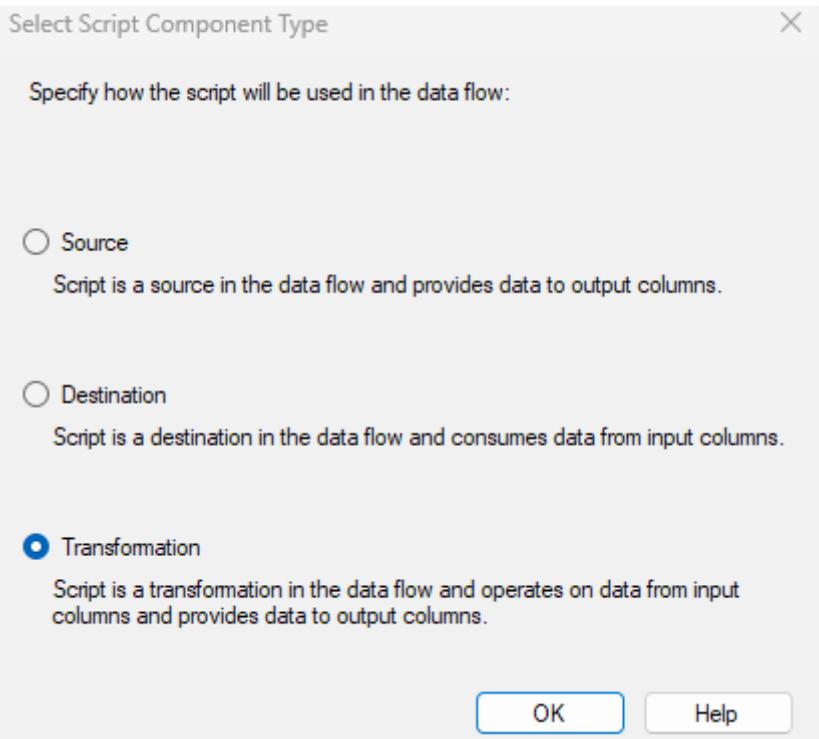
Nhóm các cột trùng nhau



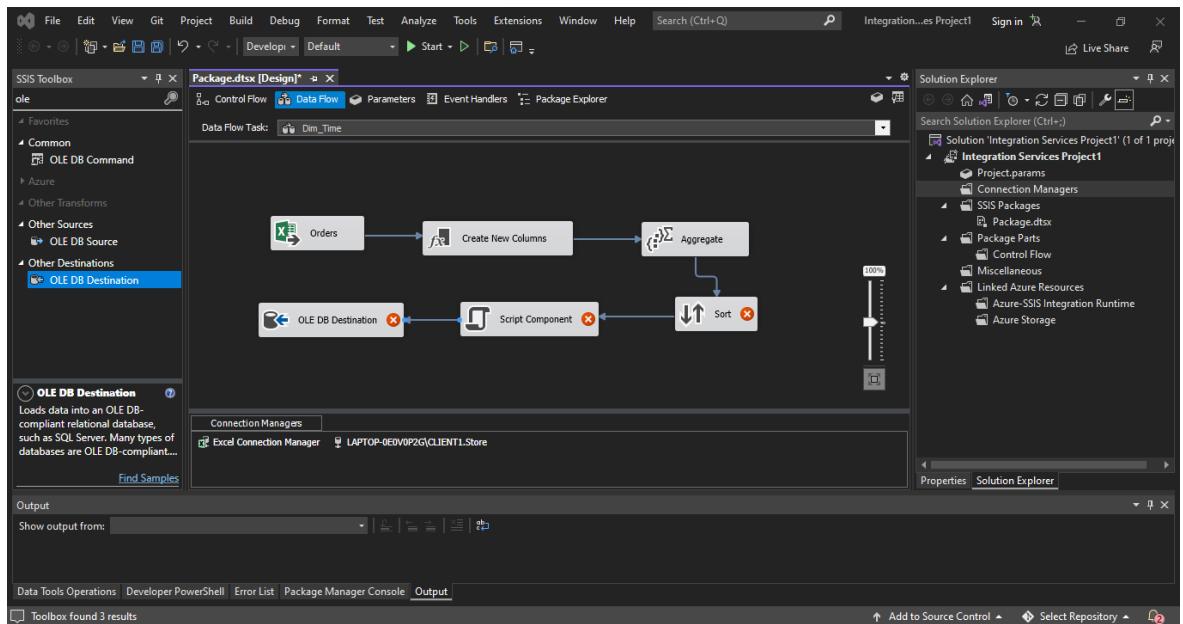
Kéo thả Sort vào Data Flow



Kéo thả Script Component vào Data Flow - chọn Transformation - click OK

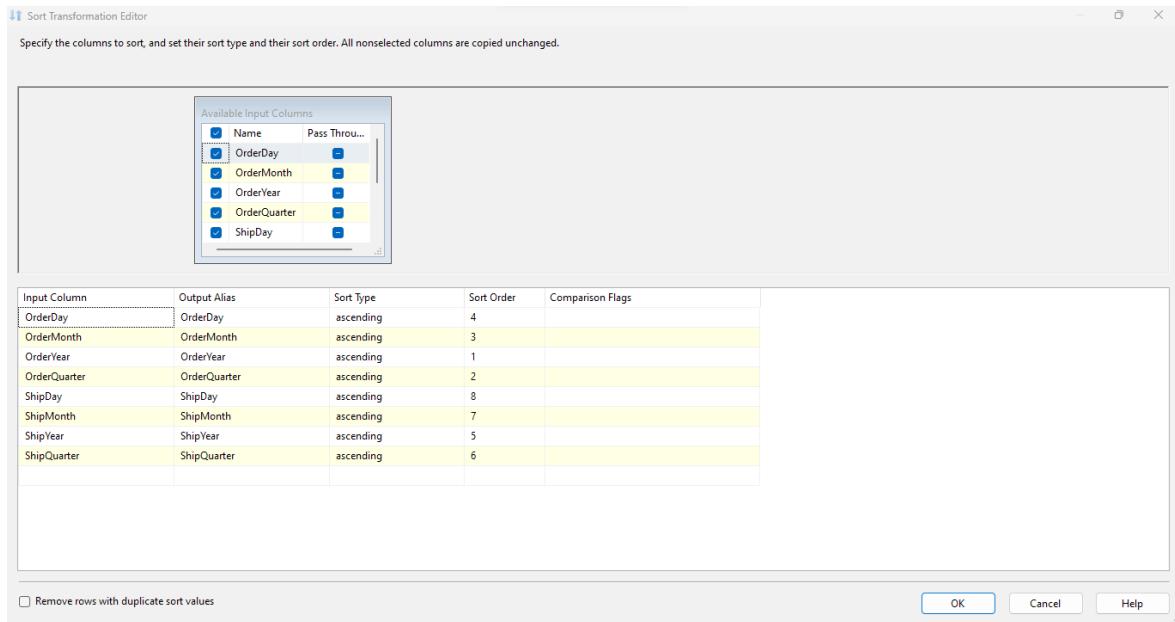


Kéo thả **OLE DB Destination** vào **Data Flow** và nối các công cụ lại với nhau

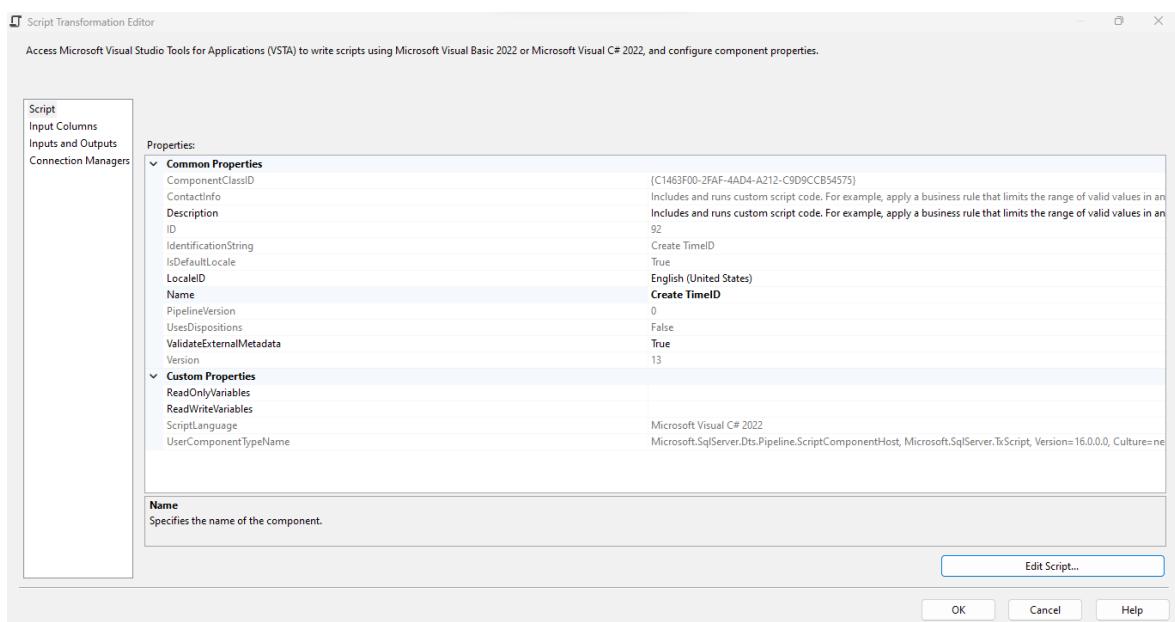


Click chuột phải vào **Sort** chọn **Edit**, cửa sổ **Sort Transformation Editor** sẽ hiện lên

Chọn các cột trên **Available Input Columns** như hình – click **OK**



Click chuột phải vào **Script Component** chọn **Edit**, cửa sổ Script Transformation Editor sẽ hiện ra. Rename giá trị **Name** thành “**Create TimeID**”



Chọn mục “**Edit Script ...**” bên dưới góc phải, 1 file mới sẽ được mở ra

```

102     */
103 }
104 /**
105 * <summary>
106 * This method is called once for every row that passes through the component from Input0.
107 *
108 * Example of reading a value from a column in the the row:
109 * string zipCode = Row.ZipCode
110 *
111 * Example of writing a value to a column in the row:
112 * Row.ZipCode = zipCode
113 * </summary>
114 * <param name="Row">The row that is currently passing through the component</param>
115 public override void Input0_ProcessInputRow(Input0Buffer Row)
116 {
117     /*
118     * Add your code here
119     */
120 }
121
122 }
123

```

Output

Show output from:

Data Tools Operations Developer PowerShell Error List Package Manager Console Output

Tạo biến count kiểu dữ liệu int và gán giá trị = 0

Thêm vào hàm **Input0_ProcessInputRow** 2 dòng code sau:

```

102     */
103 }
104 /**
105 * <summary>
106 * This method is called once for every row that passes through the component from Input0.
107 *
108 * Example of reading a value from a column in the the row:
109 * string zipCode = Row.ZipCode
110 *
111 * Example of writing a value to a column in the row:
112 * Row.ZipCode = zipCode
113 * </summary>
114 * <param name="Row">The row that is currently passing through the component</param>
115 public override void Input0_ProcessInputRow(Input0Buffer Row)
116 {
117     /*
118     * Add your code here
119     */
120     count++;
121     Row.TimeID = count;
122 }
123

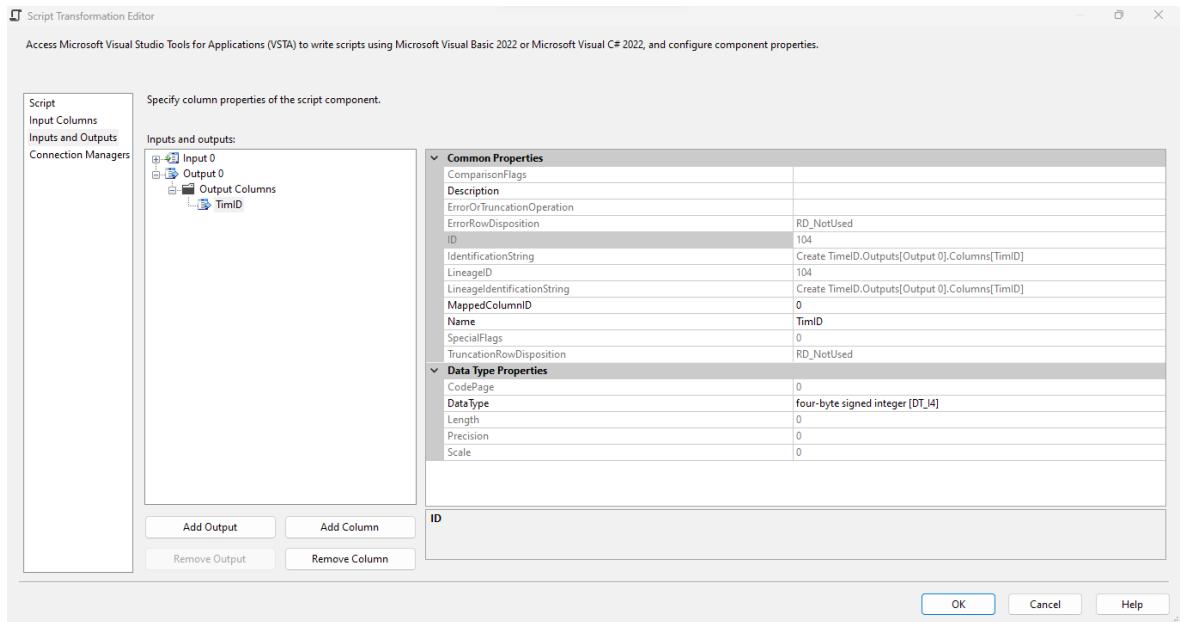
```

Output

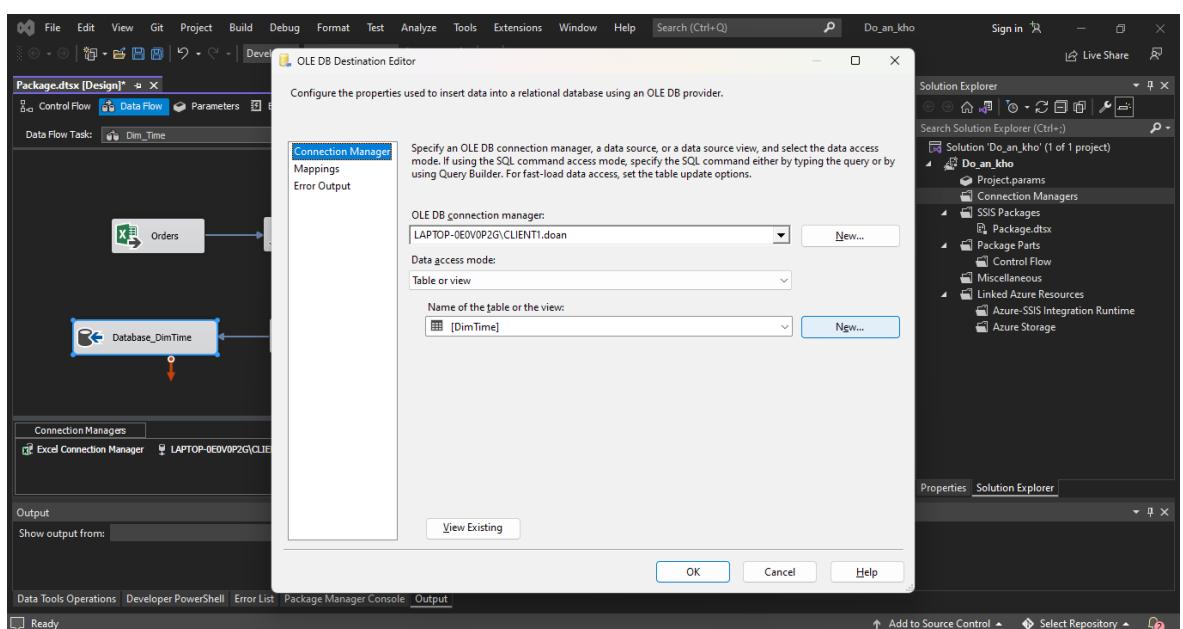
Show output from:

Data Tools Operations Developer PowerShell Error List Package Manager Console Output

Quay lại cửa sổ **Script Transformation Editor**, chọn **Inputs and Outputs – Add Column**. Đặt tên là **TimeID**

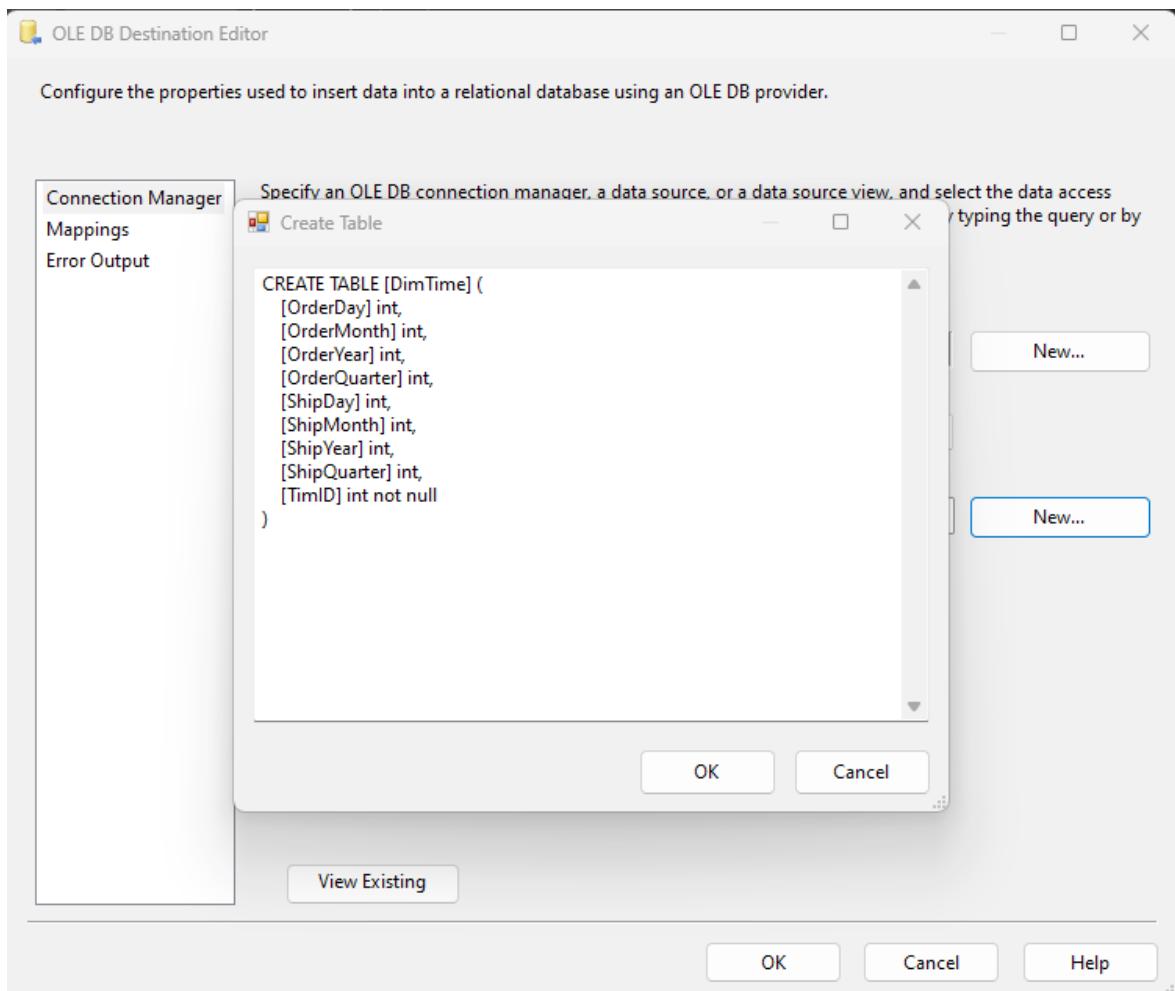


Click chuột phải vào **OLE DB Destination** chọn **Edit**, một cửa sổ mới sẽ hiện ra

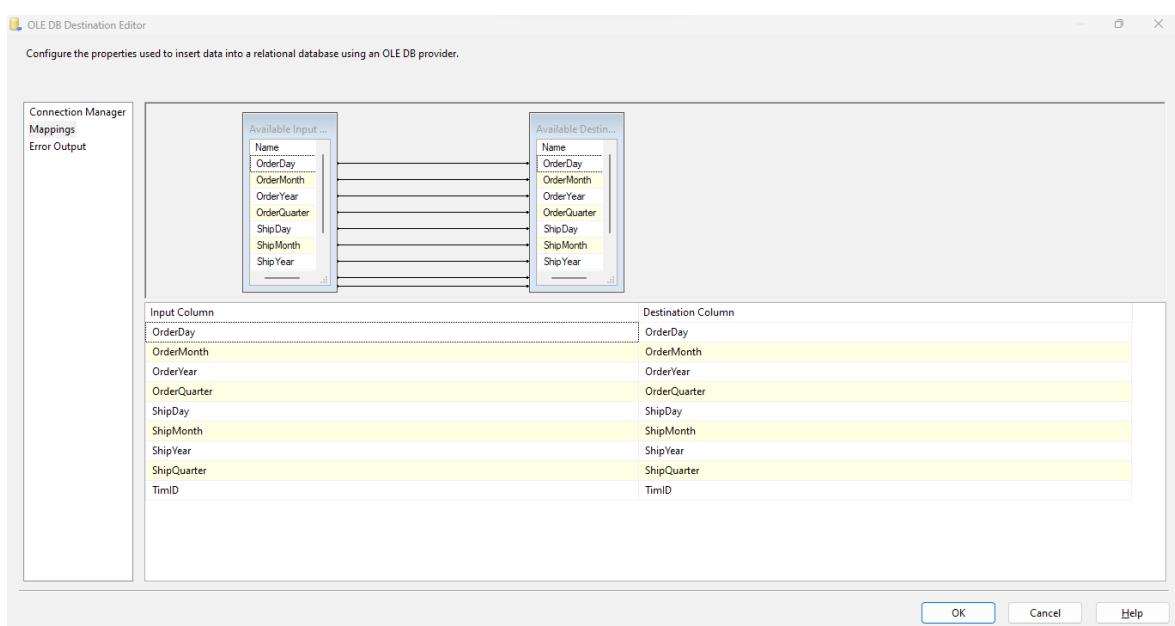


Ở **Connection Manager**, chọn **New ...**

Tạo bảng **DimTime** giống hình bên dưới, click **OK**

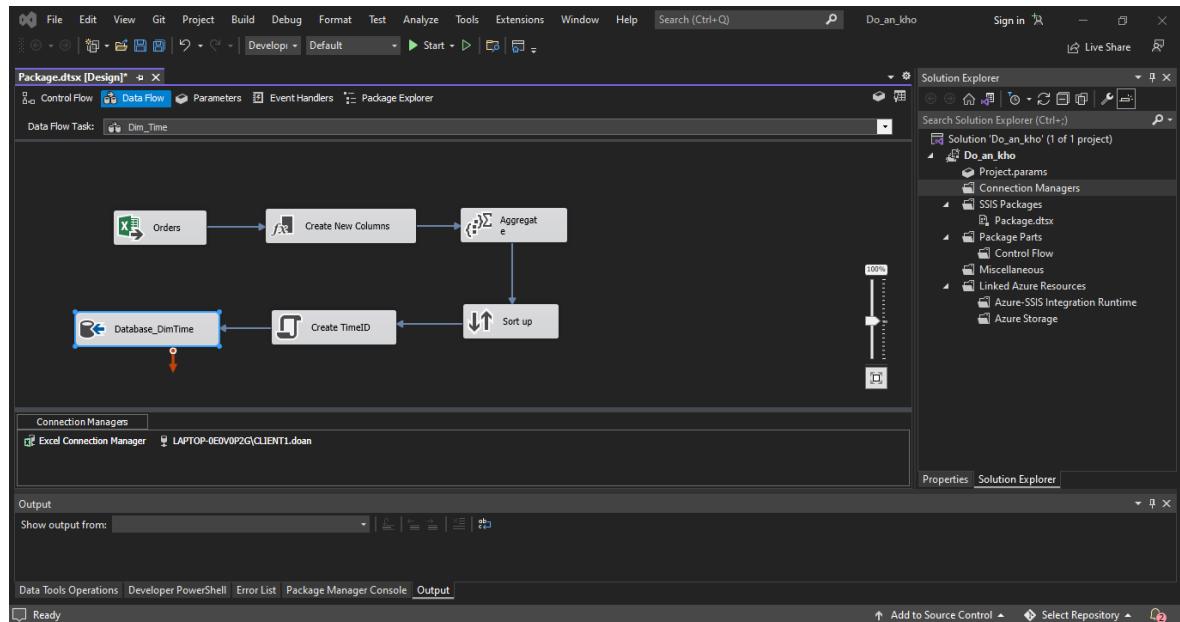


Chọn Mappings để kiểm tra các cột

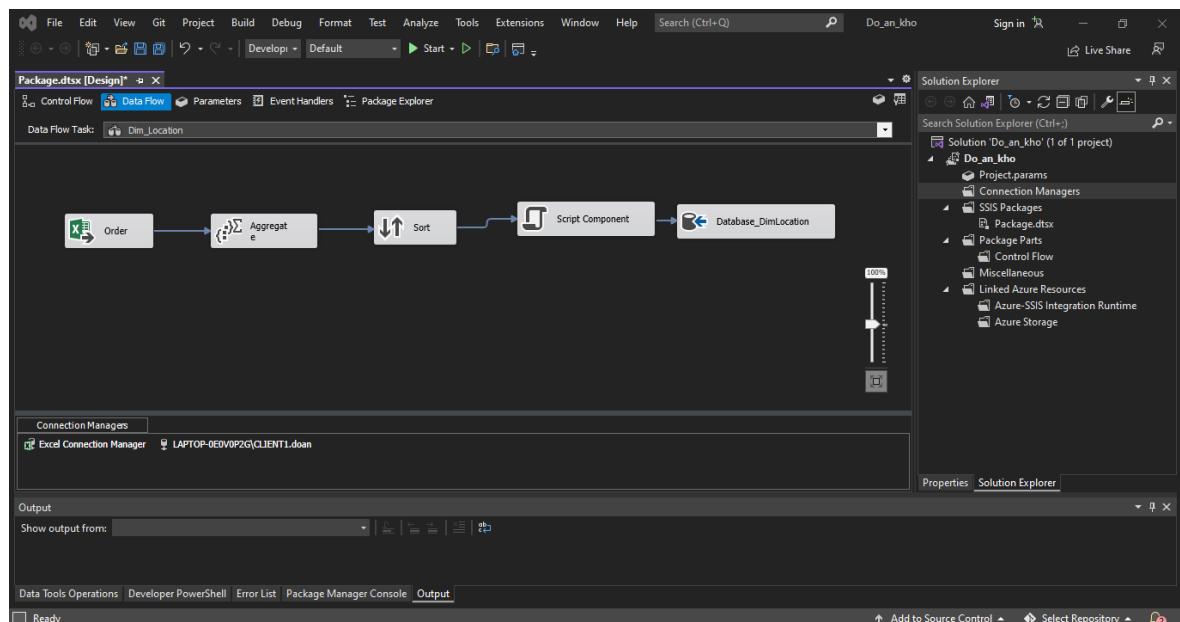


Click OK

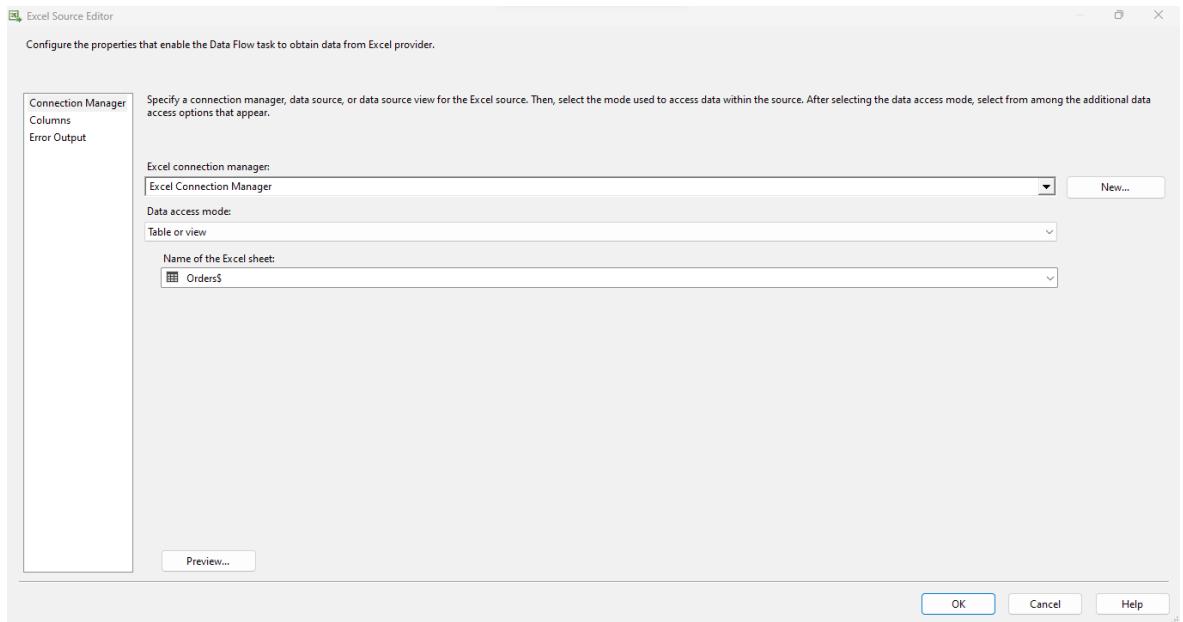
Data Flow sẽ nhìn như sau:



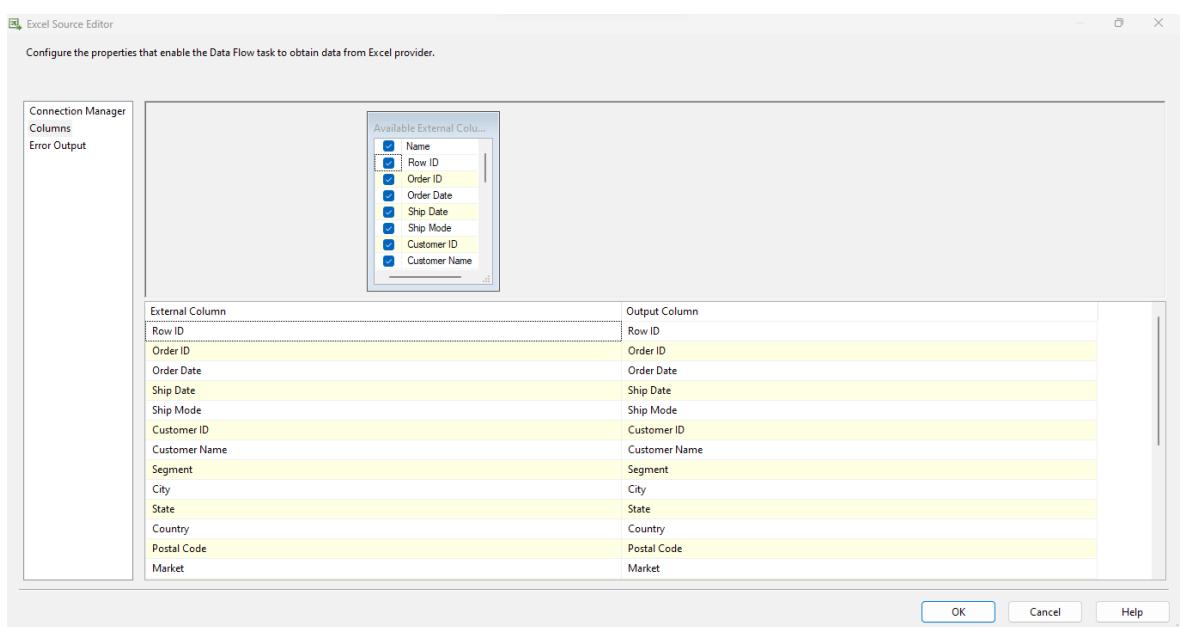
2.4.2. Bảng Dim_Location
Thiết kế Data Flow của Dim_Location giống như hình sau:



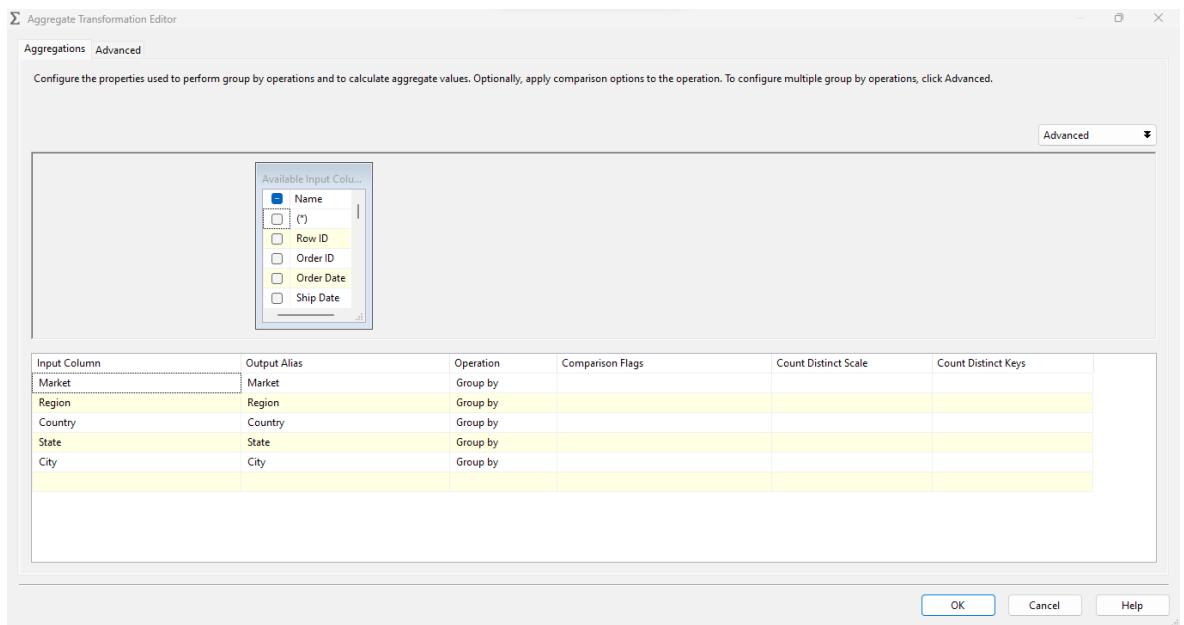
Chuột phải vào **Excel Source**



Chọn Columns

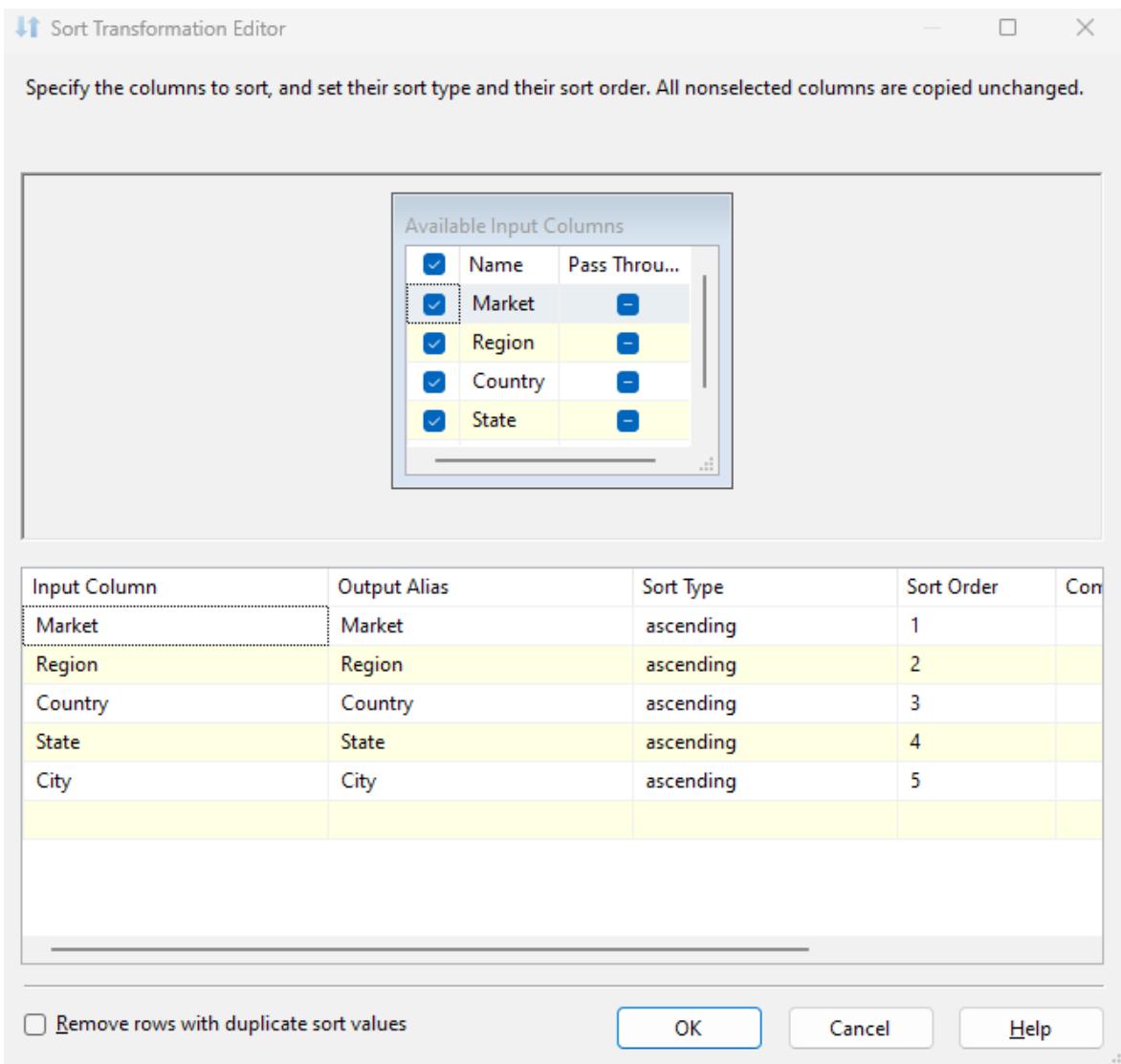


Click chuột phải vào Aggregate chọn Edit



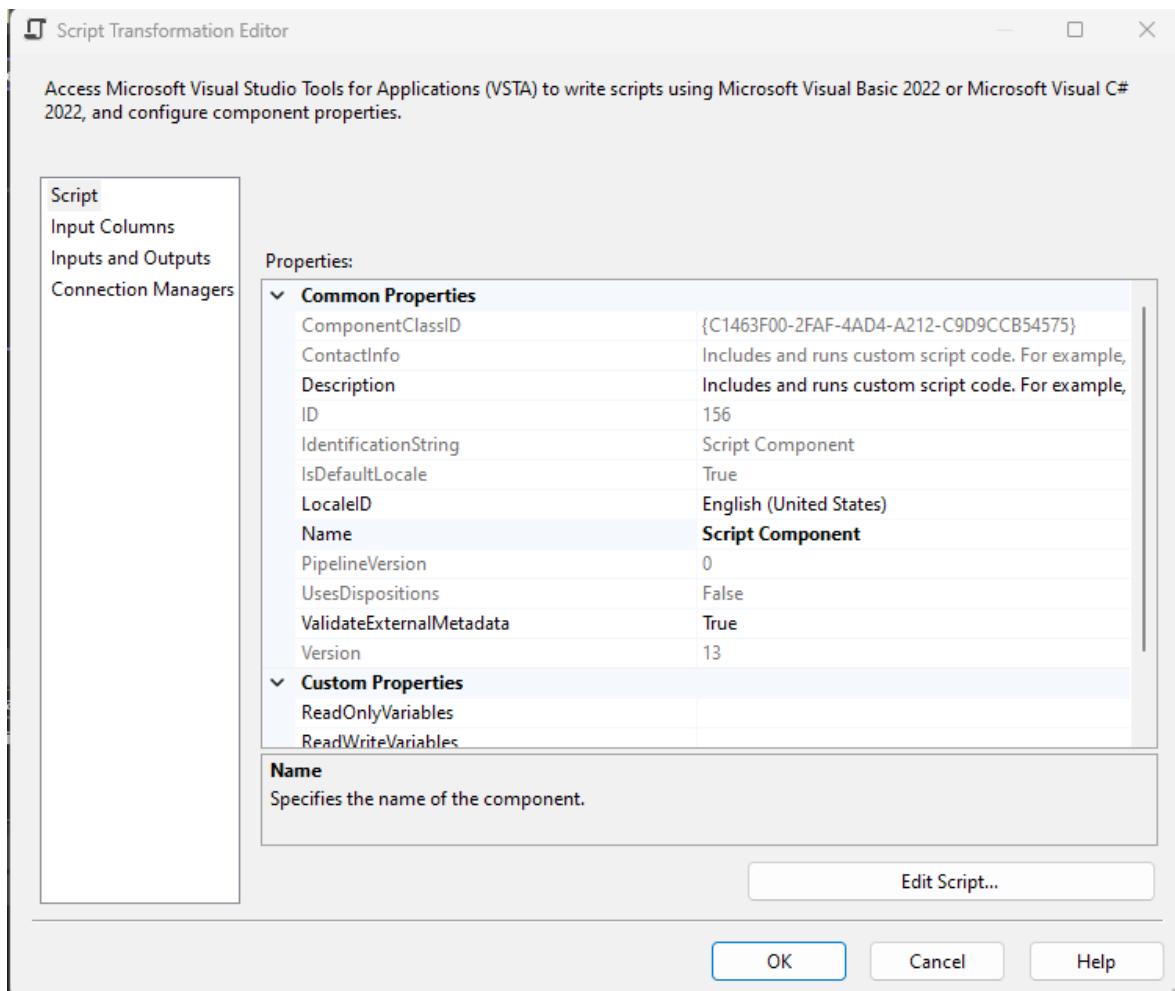
Chọn **OK**

Chuột phải vào **Sort** chọn **Edit**



Chọn các giá trị như hình, chọn OK

Chuột phải vào **Script Component** – chọn **Edit**



Quay lại tùy chọn **Script** chọn **Edit Script**

Một file mới sẽ hiện ra

Thêm vào code sau:

The screenshot shows the Microsoft Visual Studio interface with the VSTA Project Explorer open. The project 'SC_00b6c23392104d62b772fb3d66da320a' is selected. In the code editor, the 'main.cs' file is displayed, containing C# script code for a component named 'Input0'. The code includes comments explaining how to read and write values to a column named 'LocationID'.

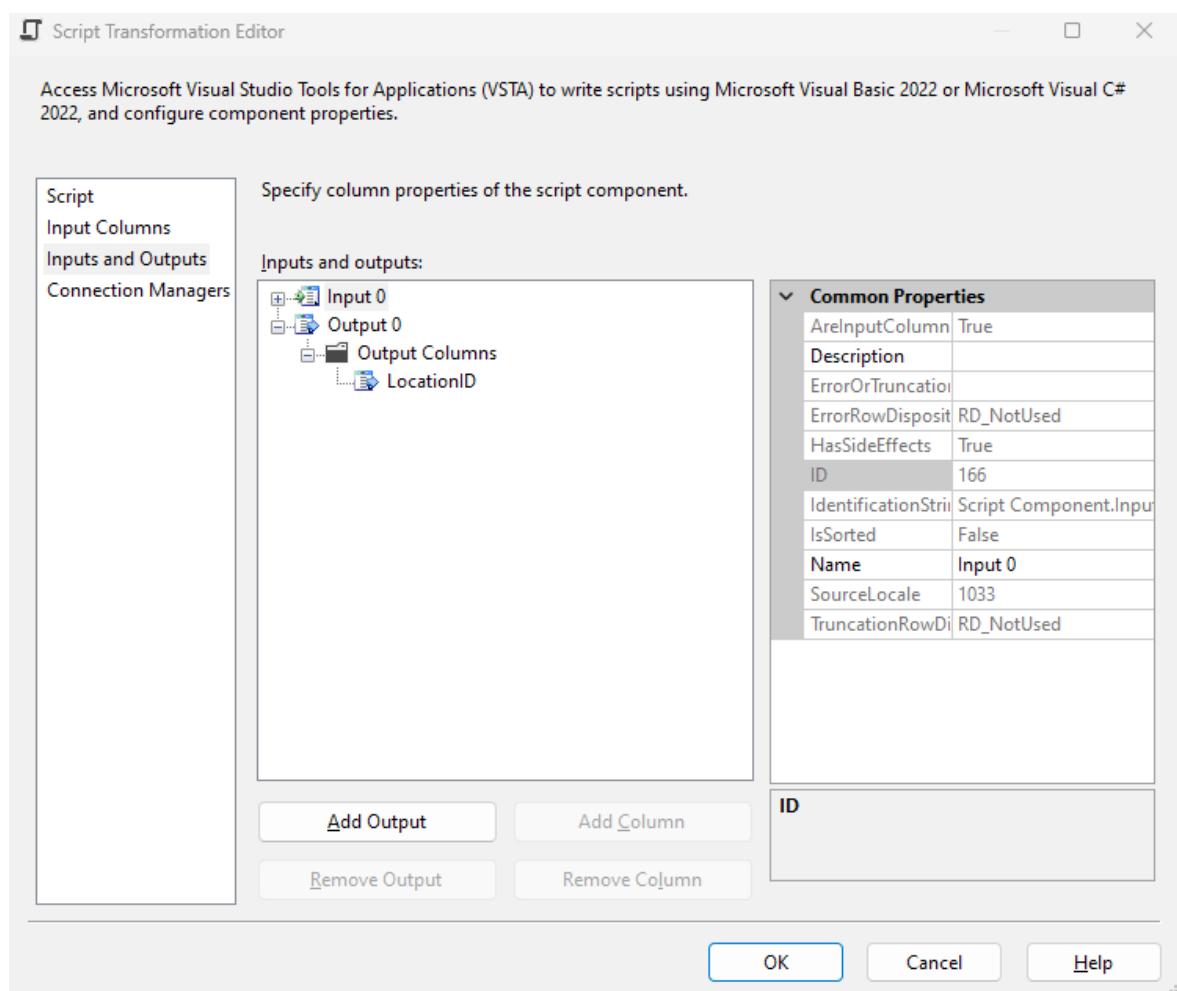
```

main.cs 9 X
SC_00b6c23392104d62b772fb3d66da320a -> ScriptMain -> PreExecute()
Data Sources
183     }
184     int count;
185     /// <summary>
186     /// This method is called once for every row that passes through the component from Input0.
187     ///
188     /// Example of reading a value from a column in the the row:
189     /// string zipCode = Row.ZipCode
190     ///
191     /// Example of writing a value to a column in the row:
192     /// Row.ZipCode = zipCode
193     /// </summary>
194     /// <param name="Row">The row that is currently passing through the component</param>
195     2 references
196     public override void Input0_ProcessInputRow(Input0Buffer Row)
197     {
198         /*
199          * Add your code here
200          */
201         count++;
202         Row.LocationID = count;
203     }
105 % No issues found | Output | Ln: 1 Ch: 1 SPC LF Properties Solution Explorer
Output
Show output from: | □ | □ | □ | □ | □ | □ |
Data Tools Operations Developer PowerShell Error List Package Manager Console Output
Ready

```

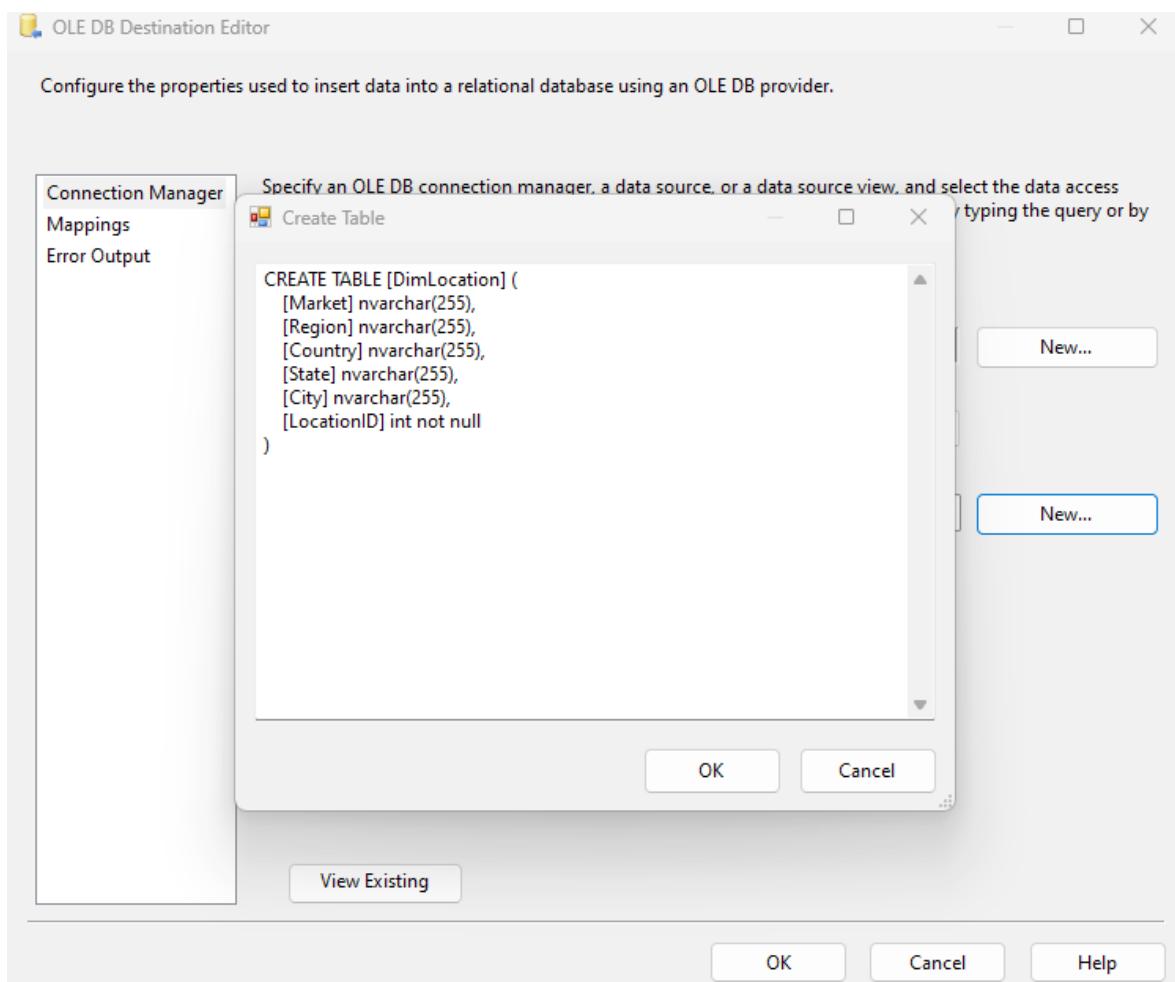
Vào tùy chọn Inputs and Outputs

Add Column mới đặt tên là LocationID



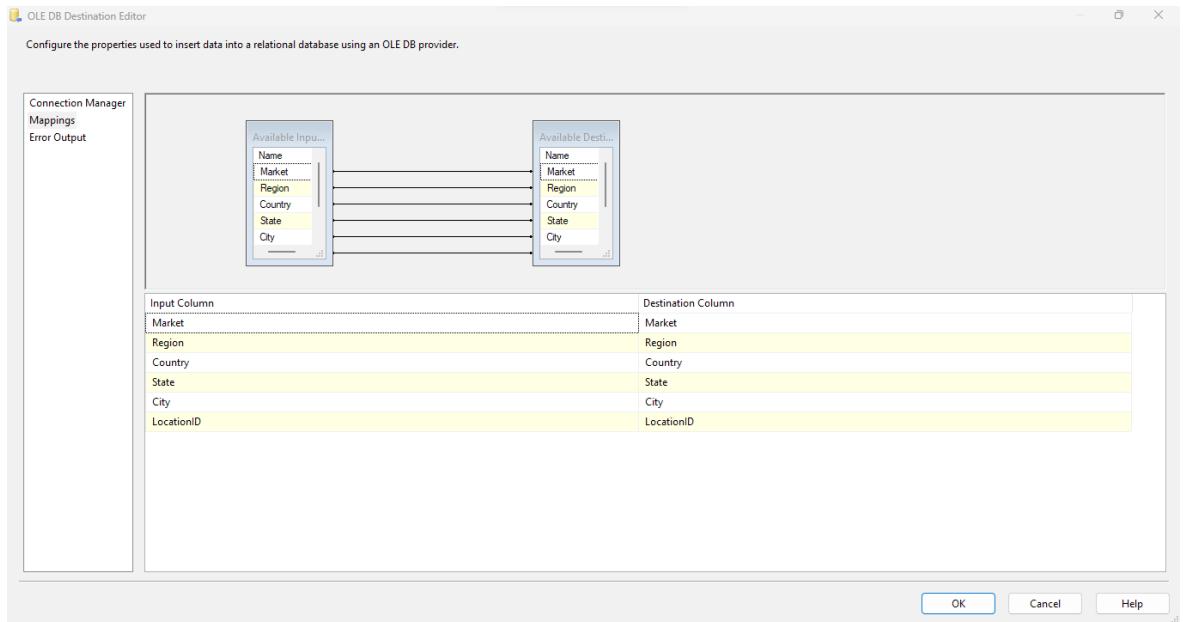
Chuột phải vào **OLE DB Destination**

Ở tùy chọn **Connection Manager** tạo bảng **DimLocation** như hình sau:



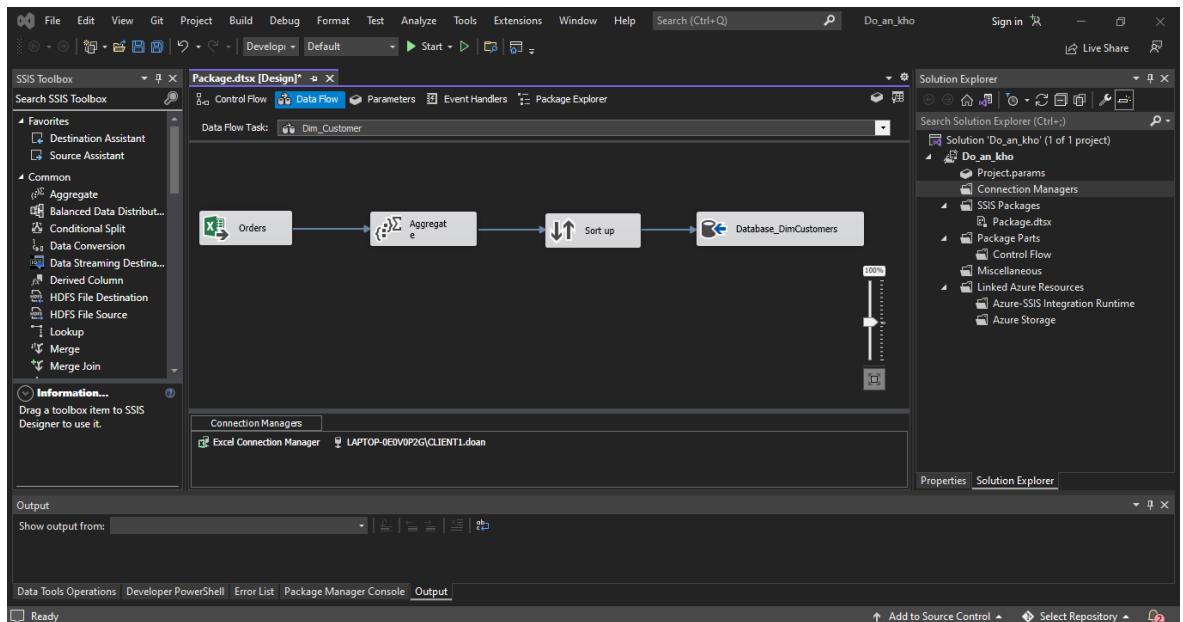
Click **OK**

Ở tùy chọn **Mappings**, kiểm tra các cột

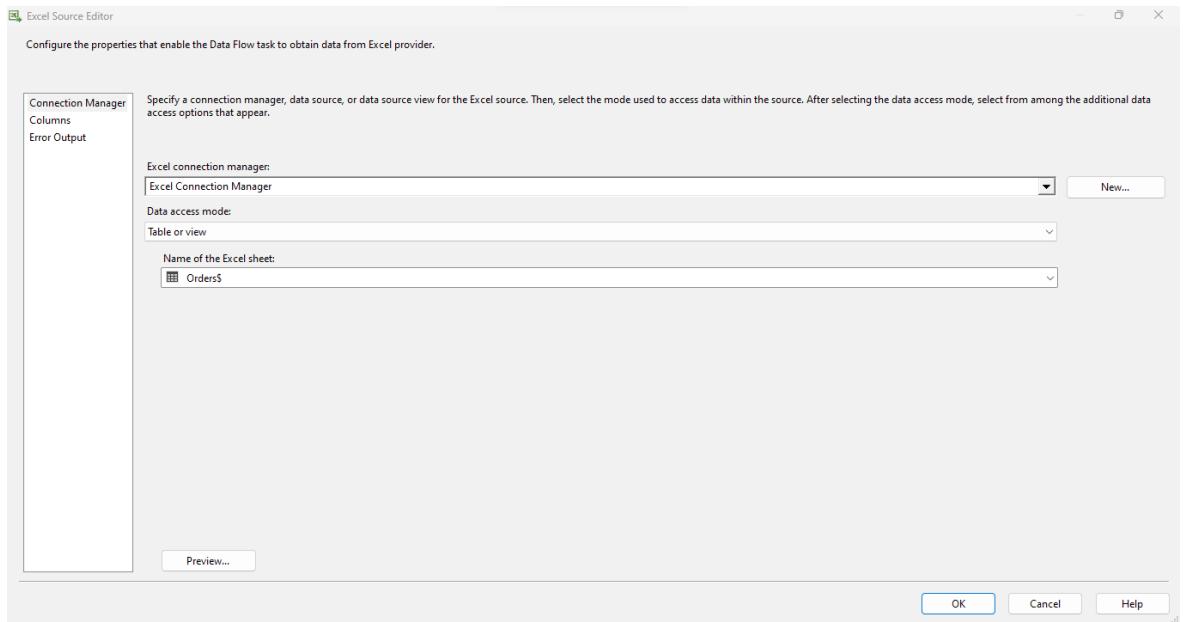


2.4.3. Bảng Dim_Customer

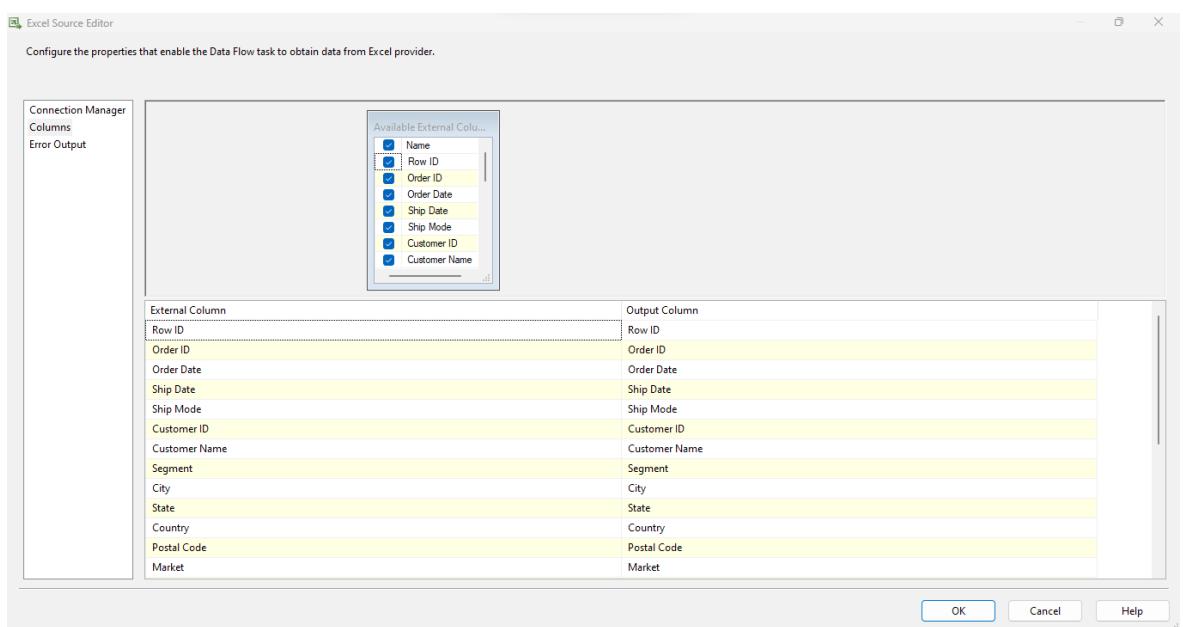
Tạo và thiết kế bảng Dim_Customer như hình



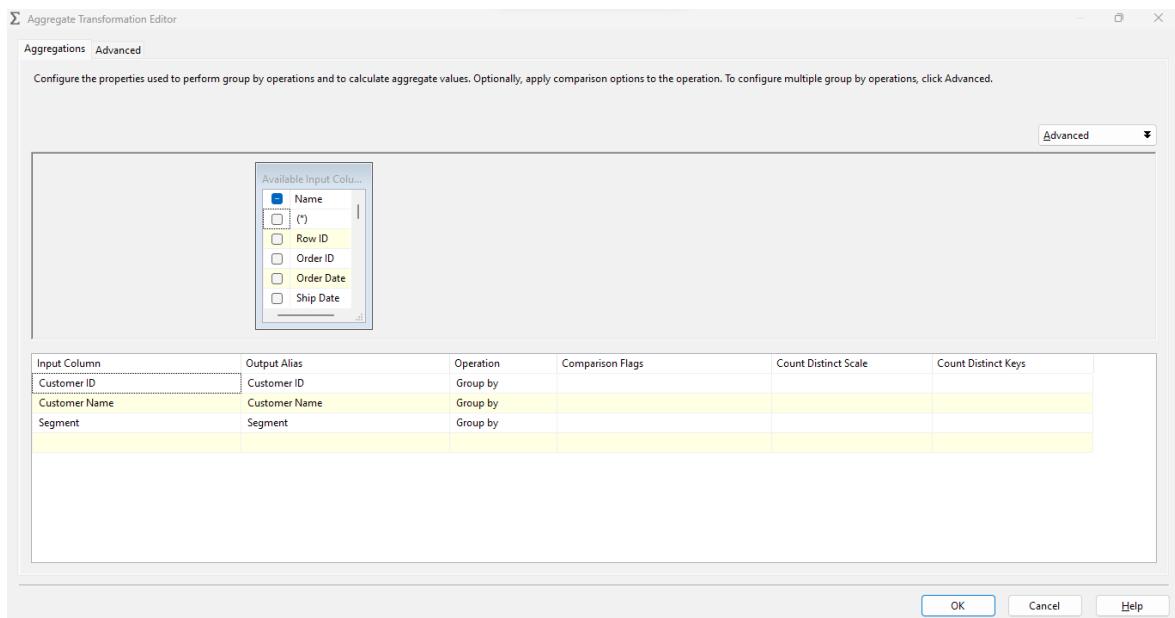
Click chuột phải vào Excel Source – Edit



Chọn tùy chọn Columns

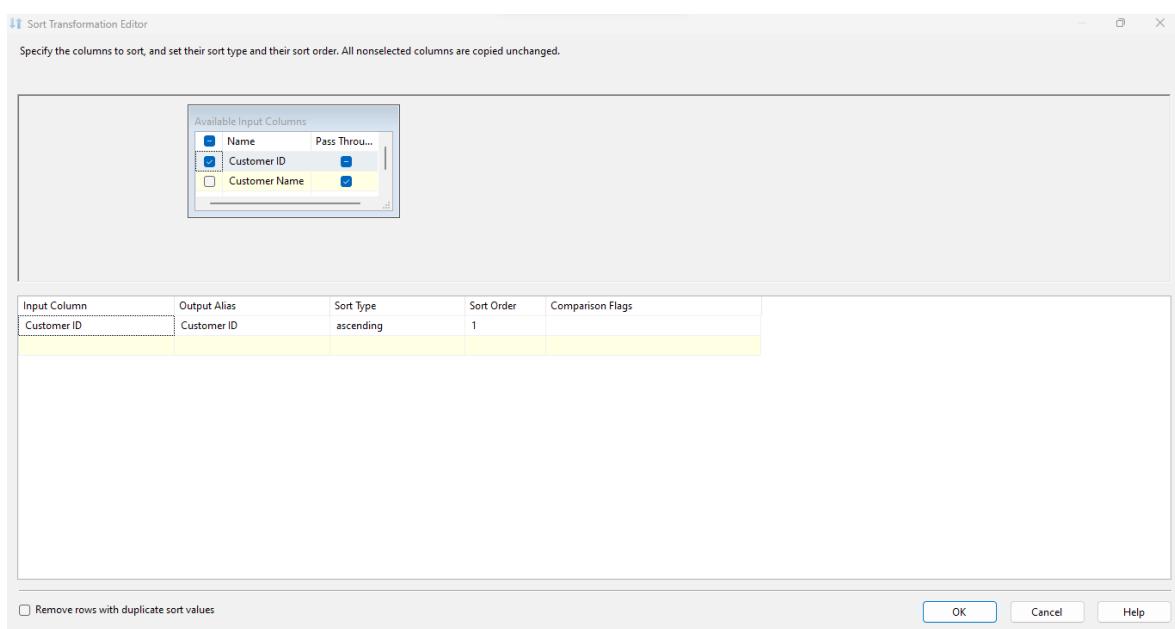


Nhấn chuột phải vào Aggregate – Edit

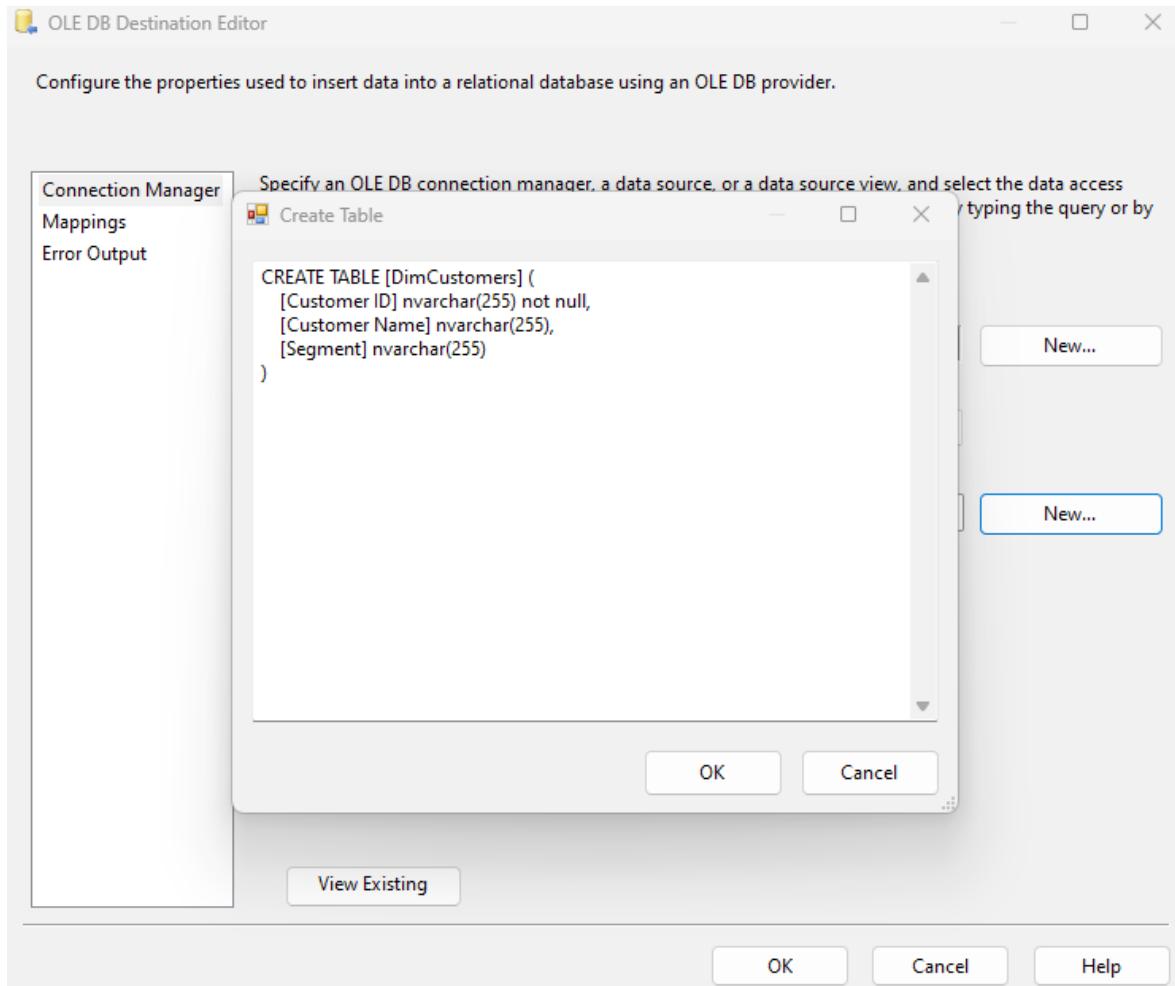


Click chuột phải vào **Sort – Edit**

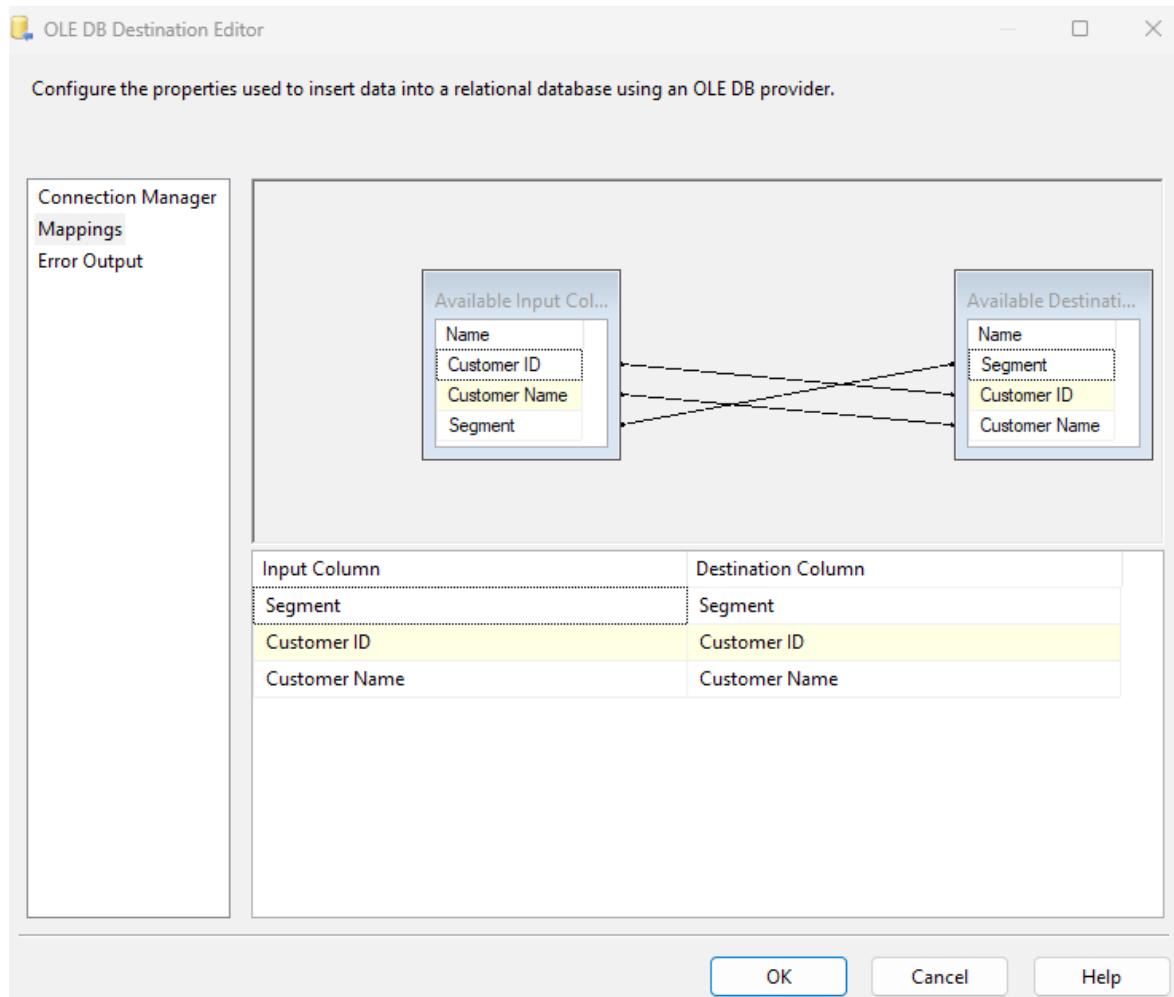
Sort theo CustomerID



Click chuột phải vào **OLE DB Destination - Edit**

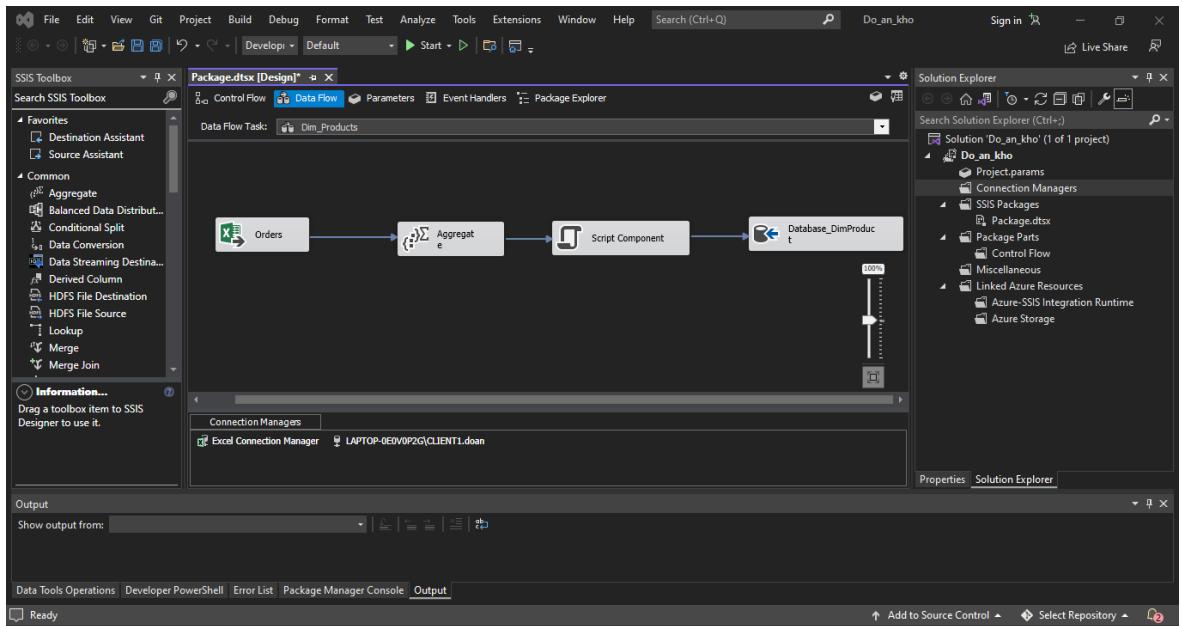


Chọn **Mappings** kiểm tra các cột trong excel với SQL

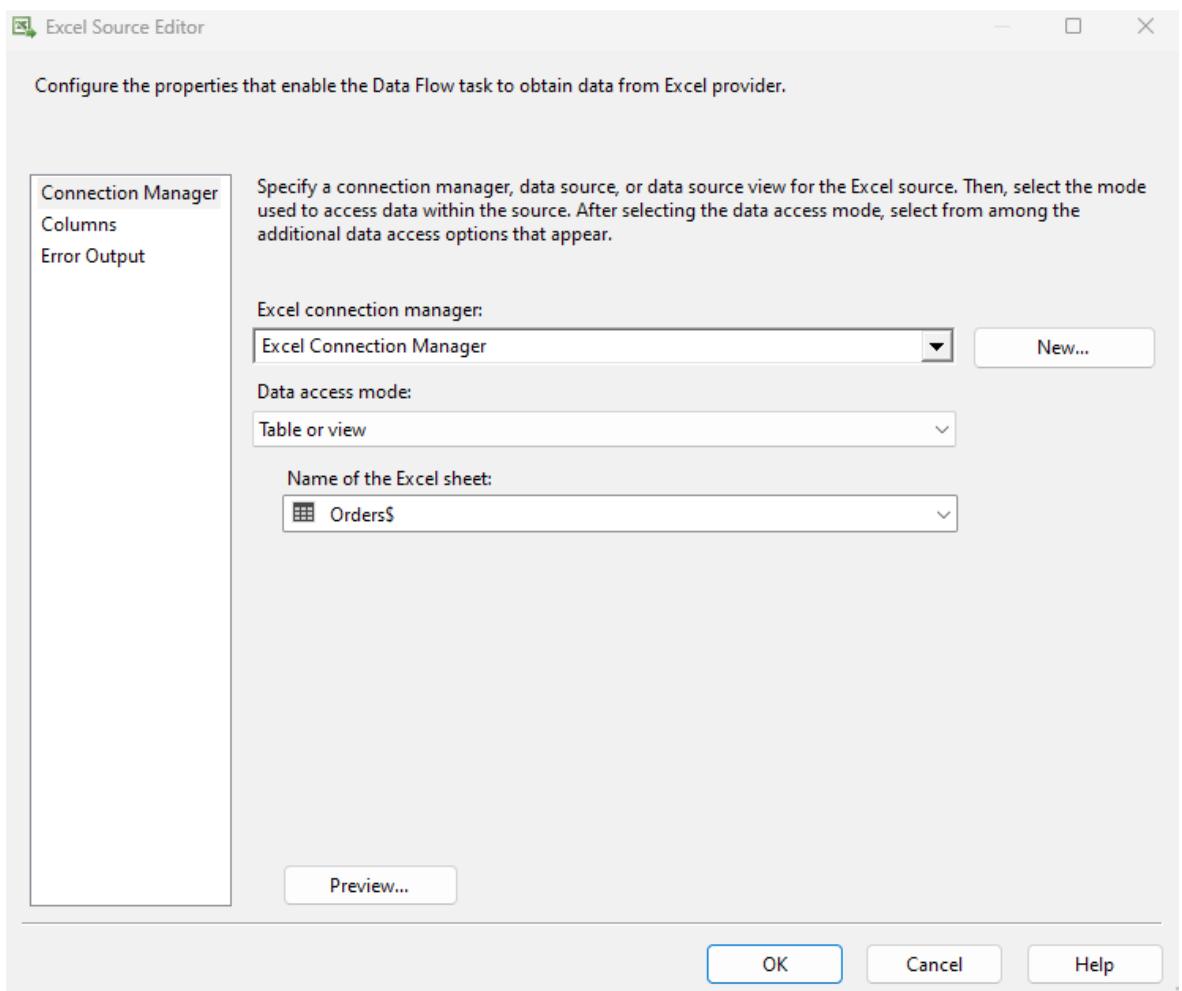


2.4.4. Bảng Dim_Product

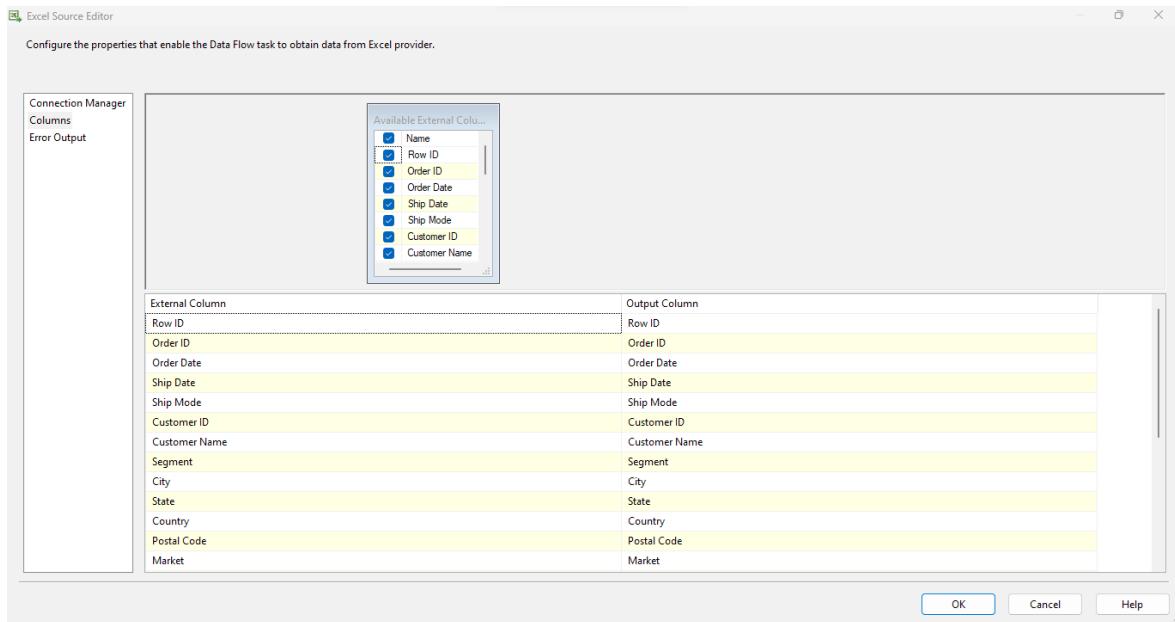
Thiết kế Data Flow như hình



Click chuột phải vào Excel Source - Edit

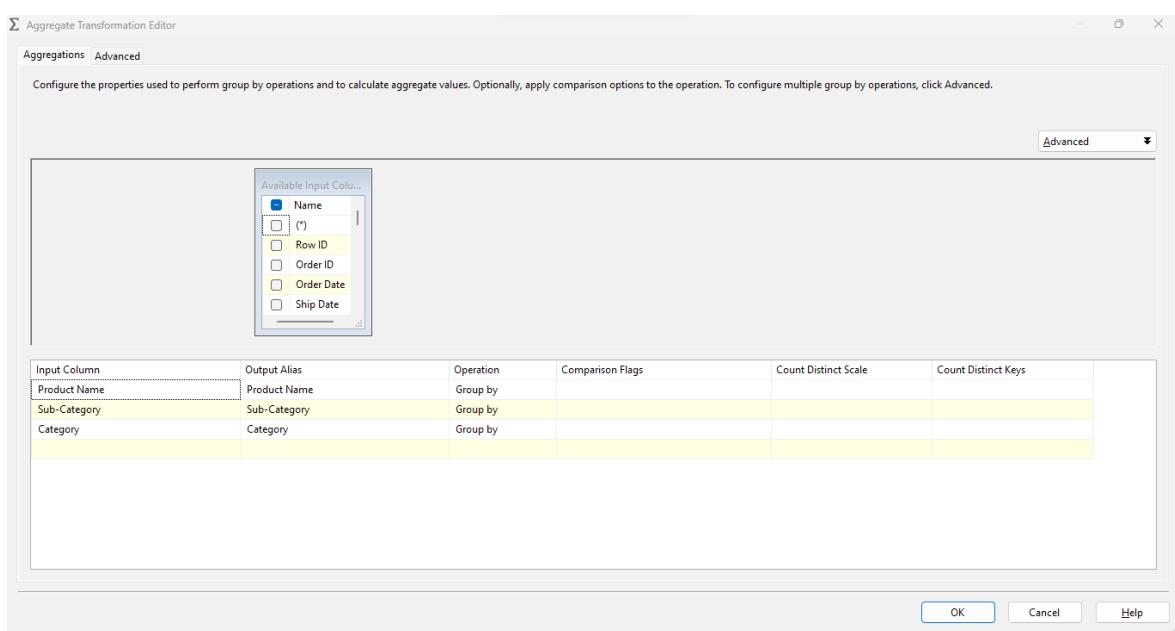


Chọn tùy chọn **Columns**



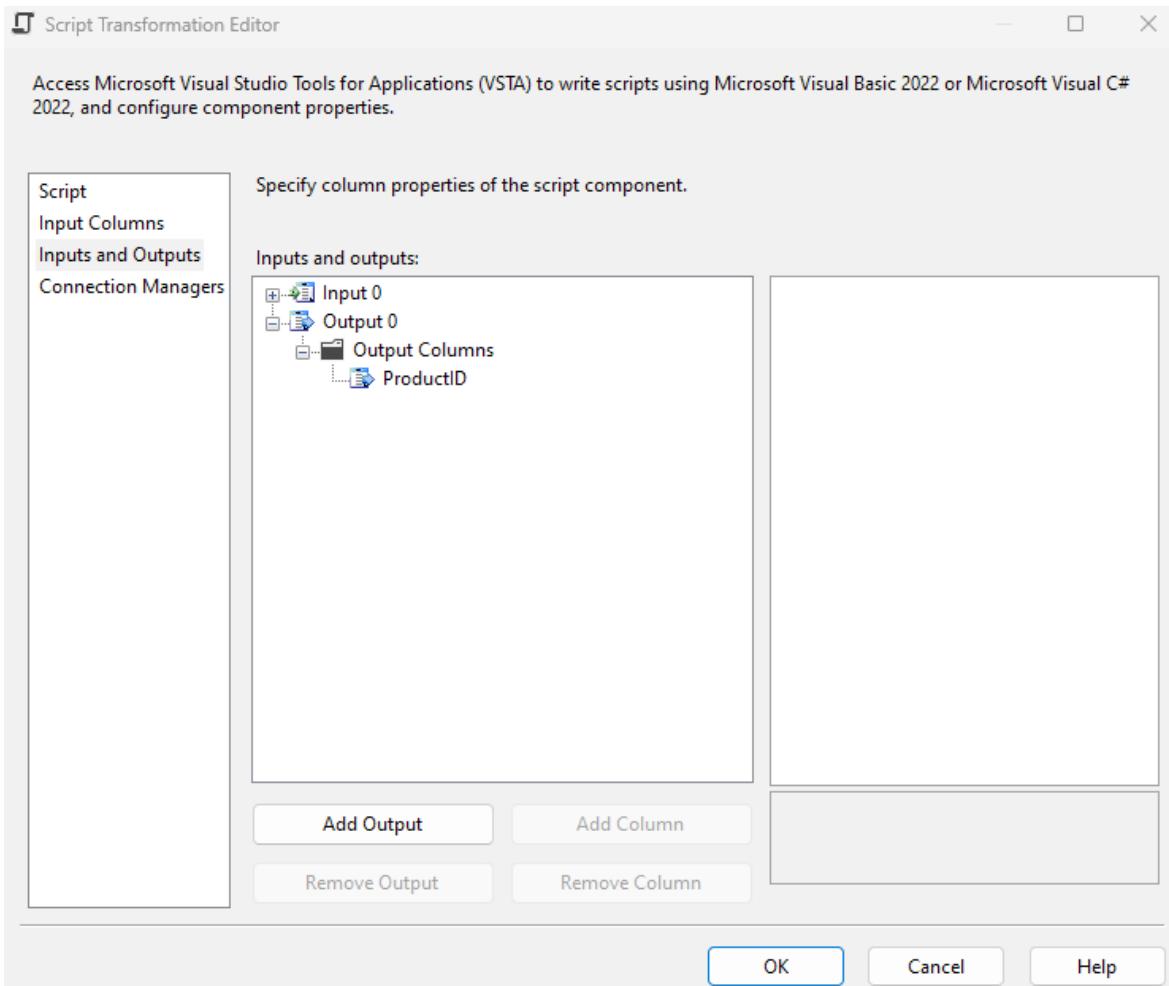
Click OK

Click chuột phải vào Aggregate – Edit



Click chuột phải vào Script Component – Edit

Ở tùy chọn Inputs and Outputs thêm vào cột ProductID để làm khóa chính



Quay lại tùy chọn Script chọn Edit Script

Thêm vào code sau:

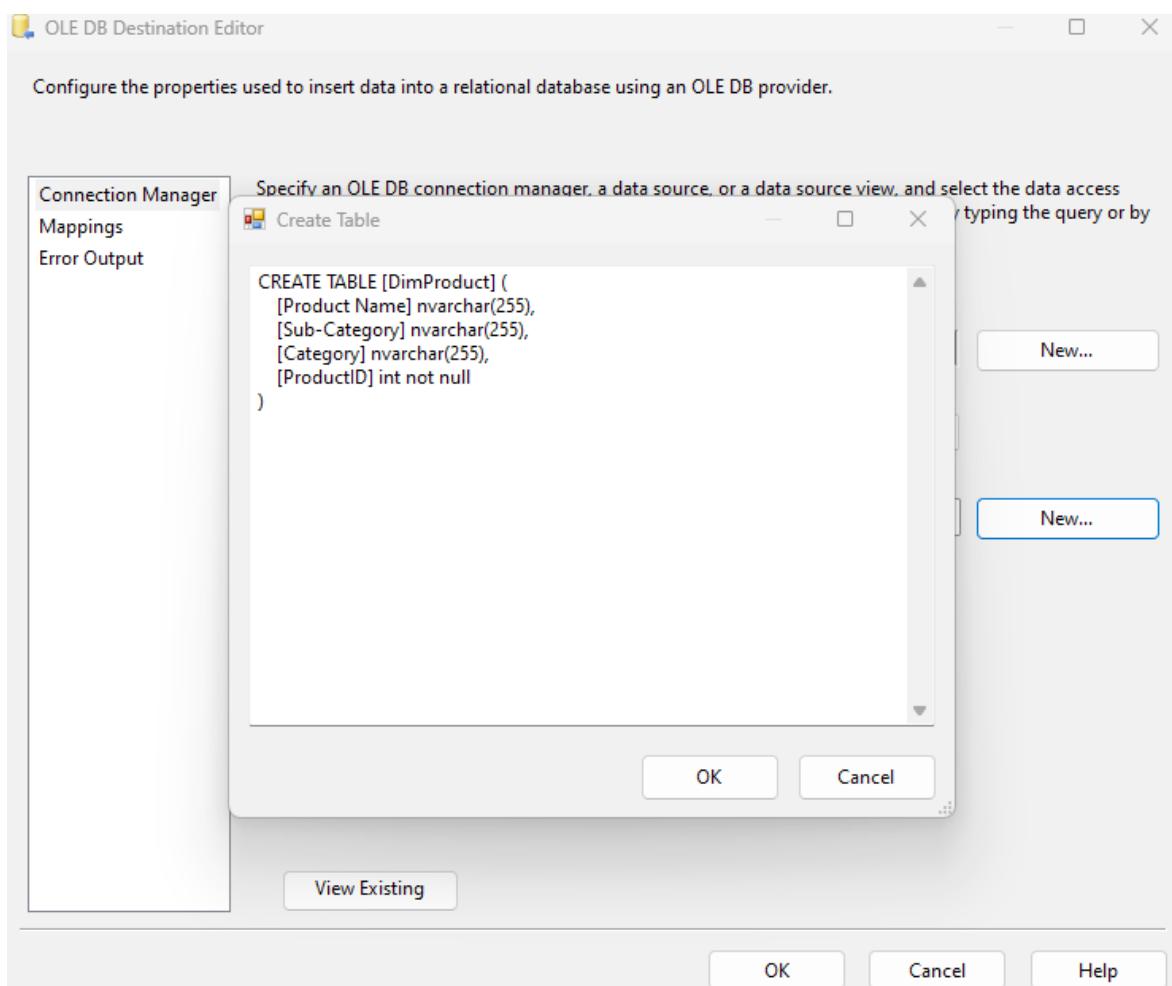
```

102     */
103     int count;
104     /// <summary>
105     /// This method is called once for every row that passes through the component from Input0.
106     ///
107     /// Example of reading a value from a column in the the row:
108     /// string zipCode = Row.ZipCode
109     ///
110     /// Example of writing a value to a column in the row:
111     /// Row.ZipCode = zipCode
112     /// </summary>
113     /// <param name="Row">The row that is currently passing through the component</param>
114
115     public override void Input0_ProcessInputRow(Input0Buffer Row)
116     {
117         /*
118         * Add your code here
119         */
120         count++;
121         Row.ProductID=count;
122     }

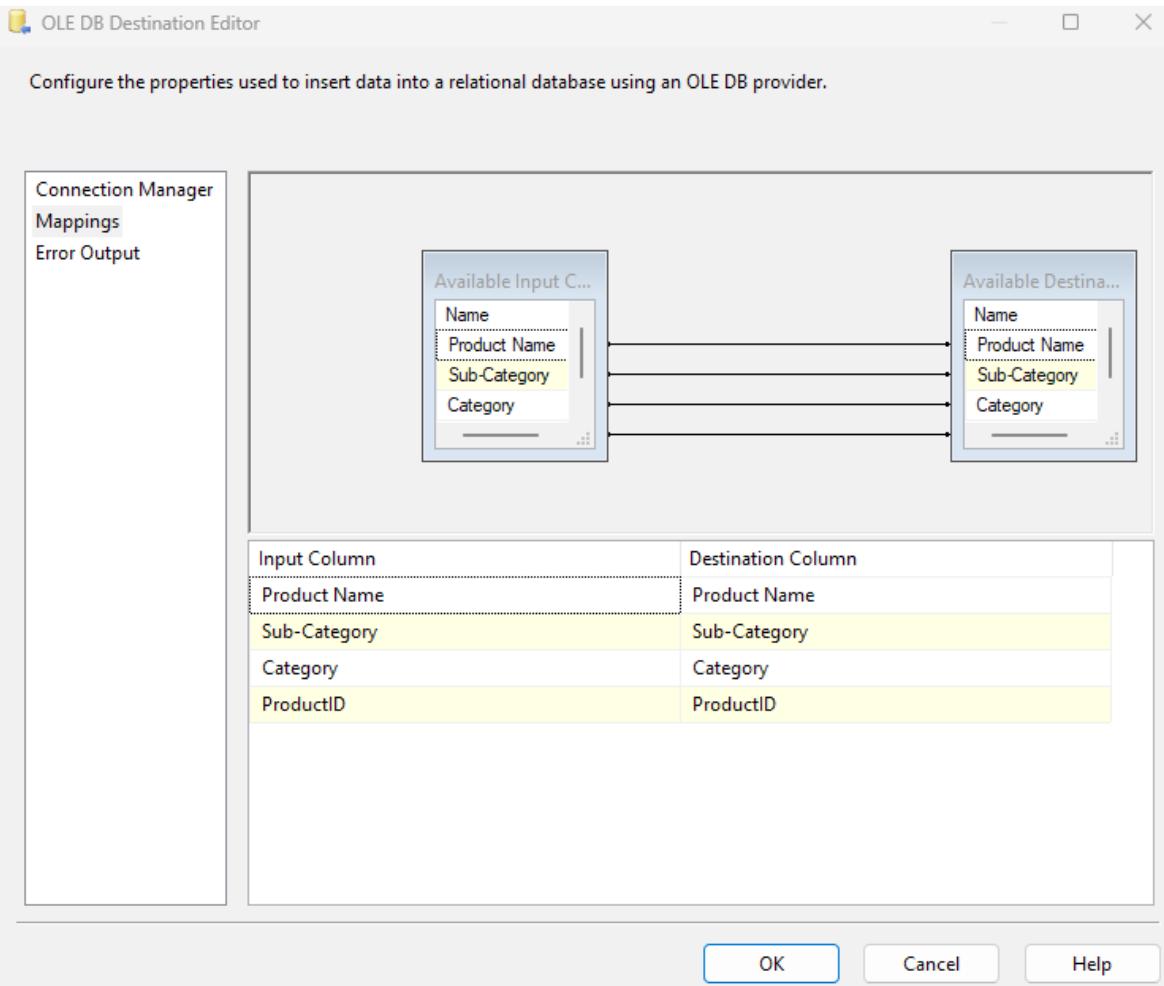
```

Chuột phải vào **OLE DB Destination – Edit**

Tạo bảng **DimProduct** như hình



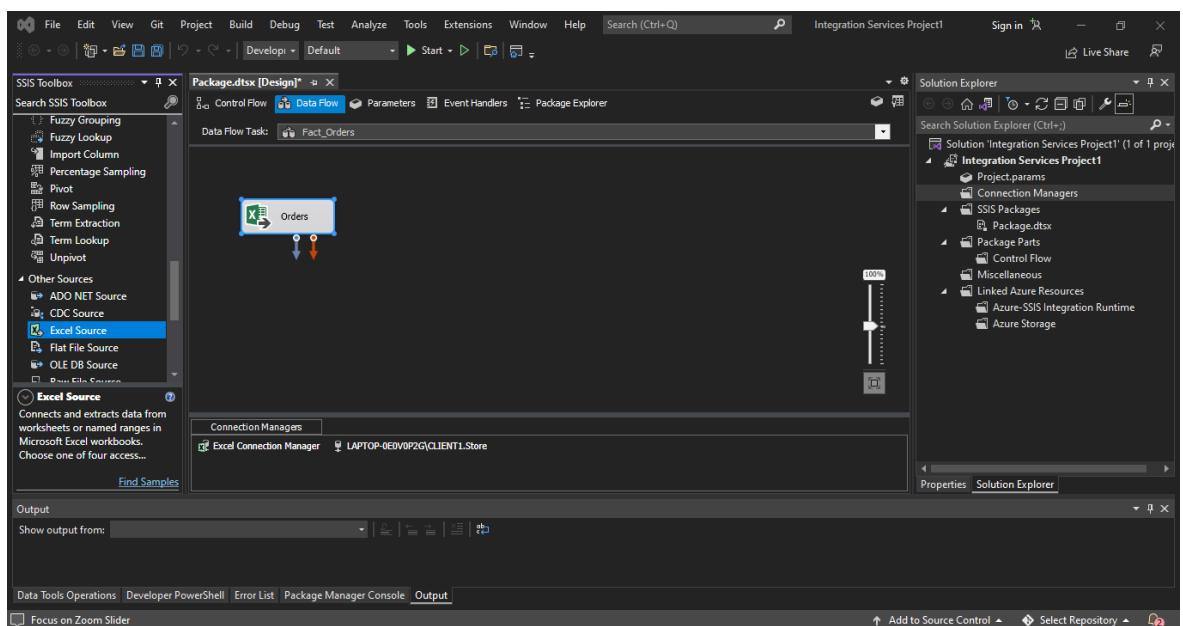
Vào **Mappings**



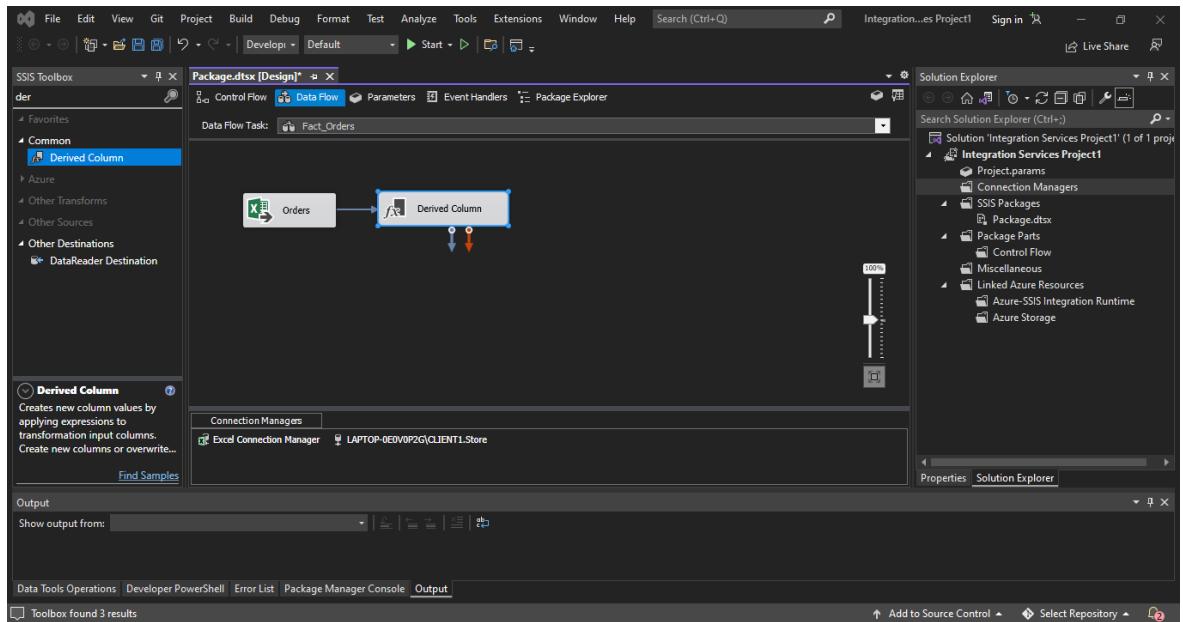
Click OK

2.4.5. Bảng Fact_Orders

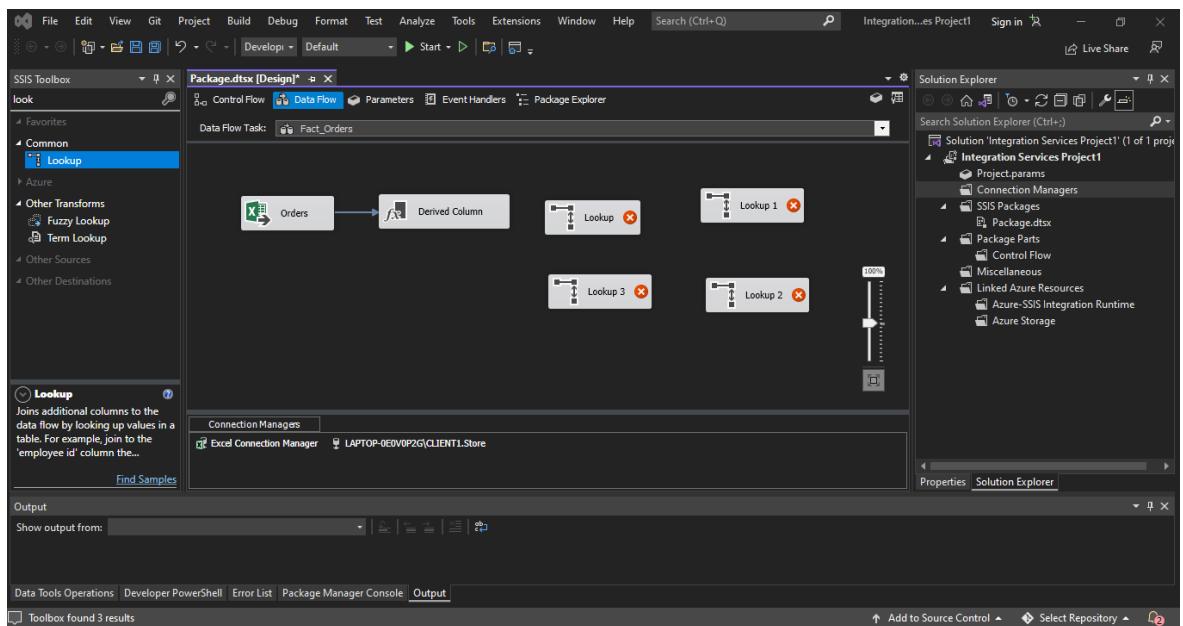
Chọn Excel Source và kéo vào Data Flow



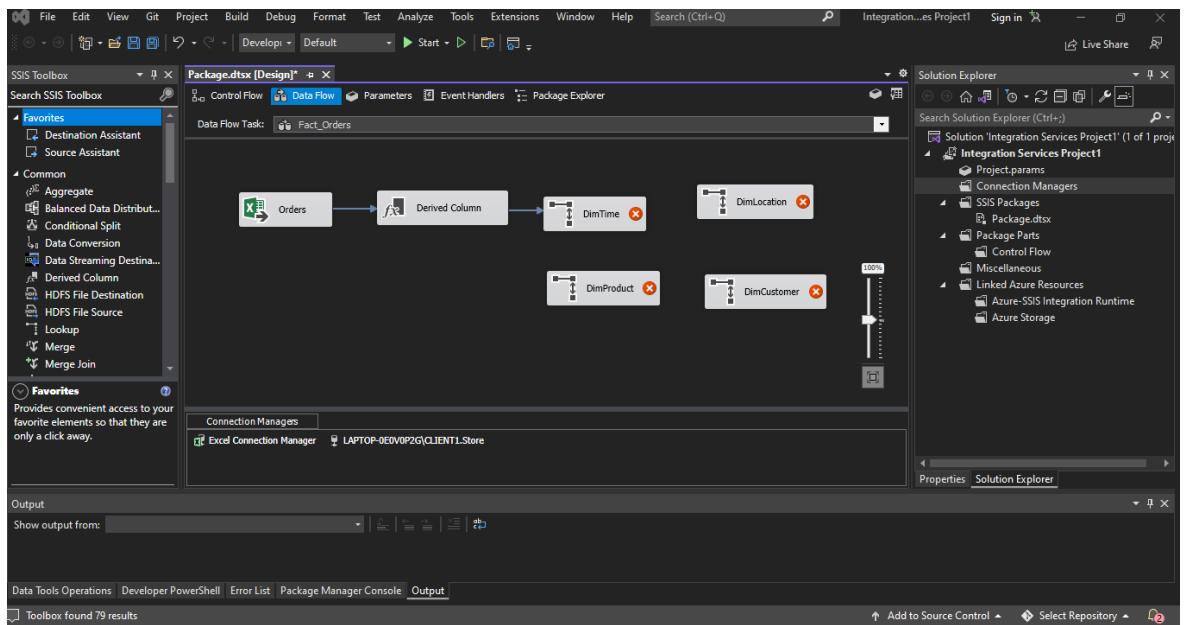
Kéo thả Derived Column như hình



Tạo ra 4 Lookup như hình

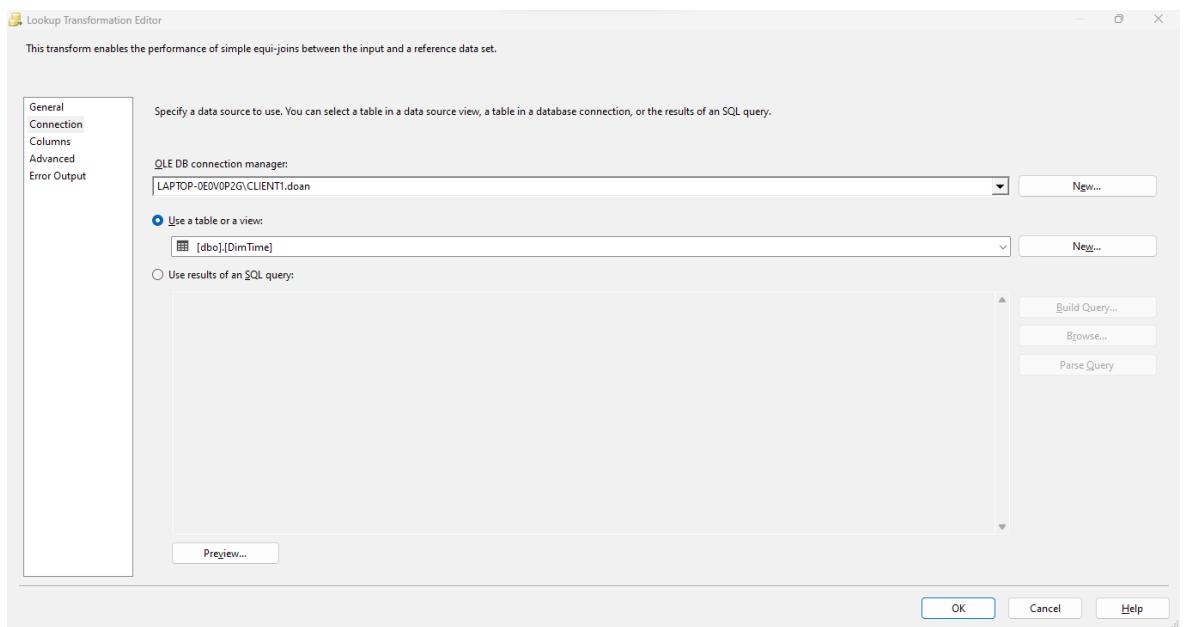


Đổi tên lần lượt lại thành: DimTime, DimLocation, Customer, Product

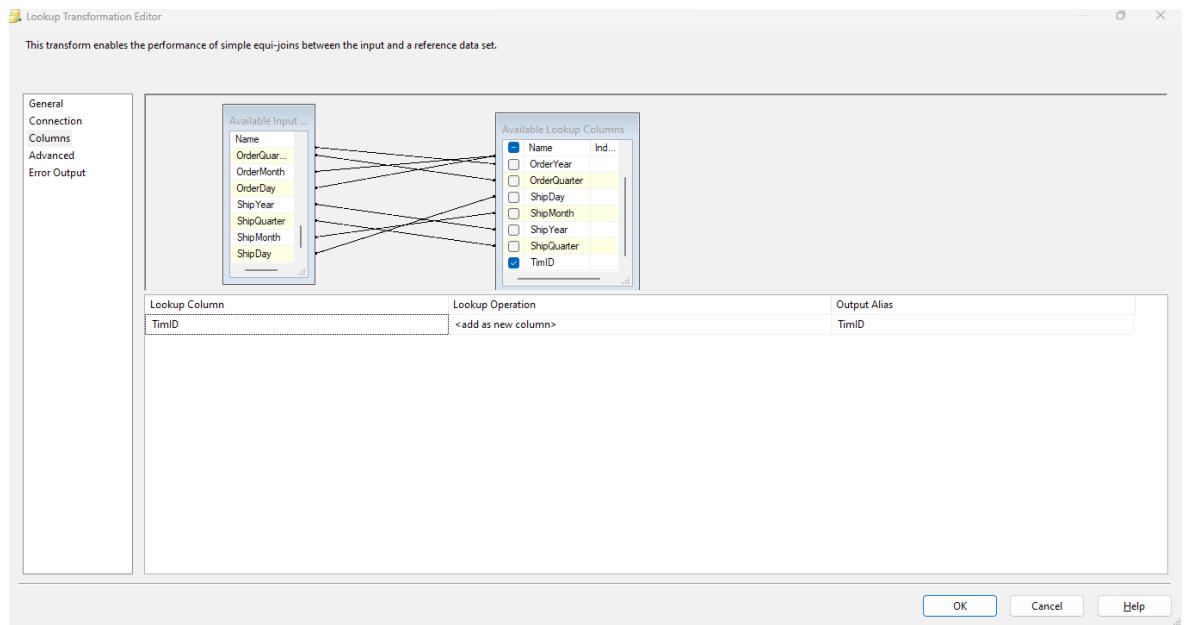


Chuột phải vào **DimTime – Edit**

Ở **Connection** chọn bản **DimTime** đã tạo ra trước đó



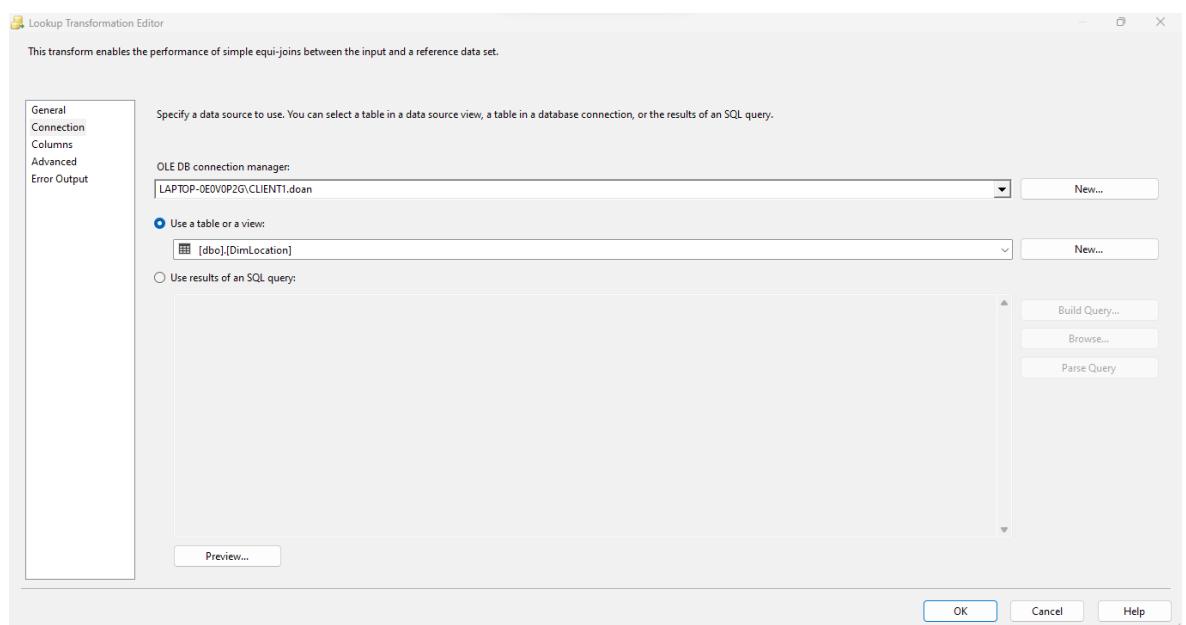
Ở **Columns** chọn cột **TimeID** (vì cột này là cột được tạo ra)



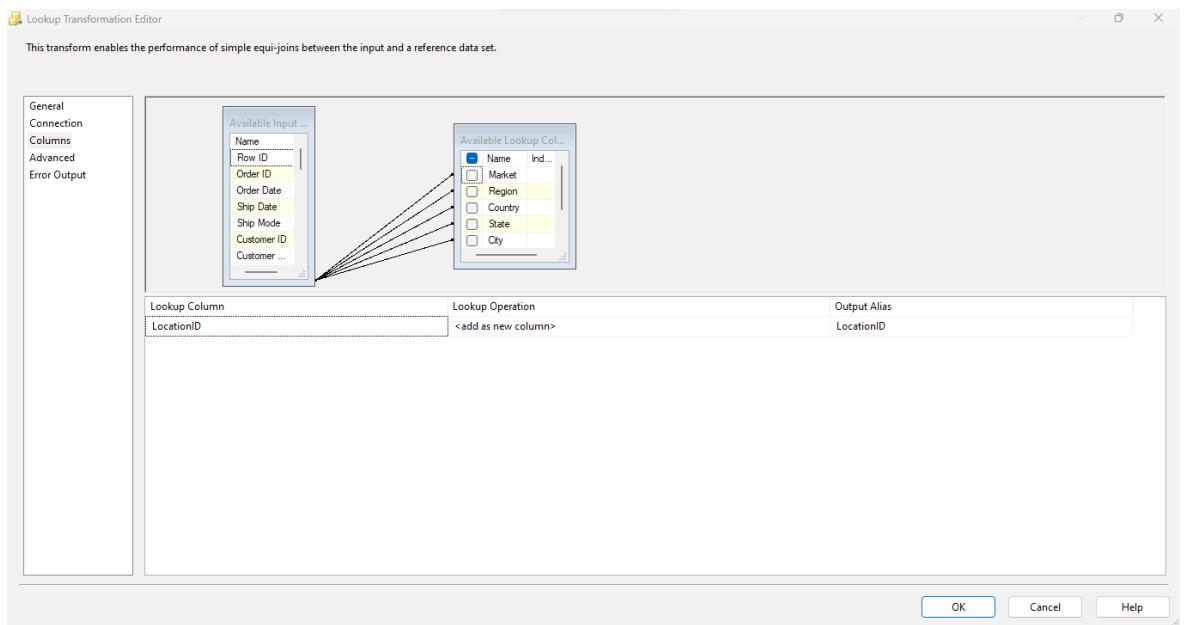
Click OK

Chuột phải vào **DimLocation – Edit**

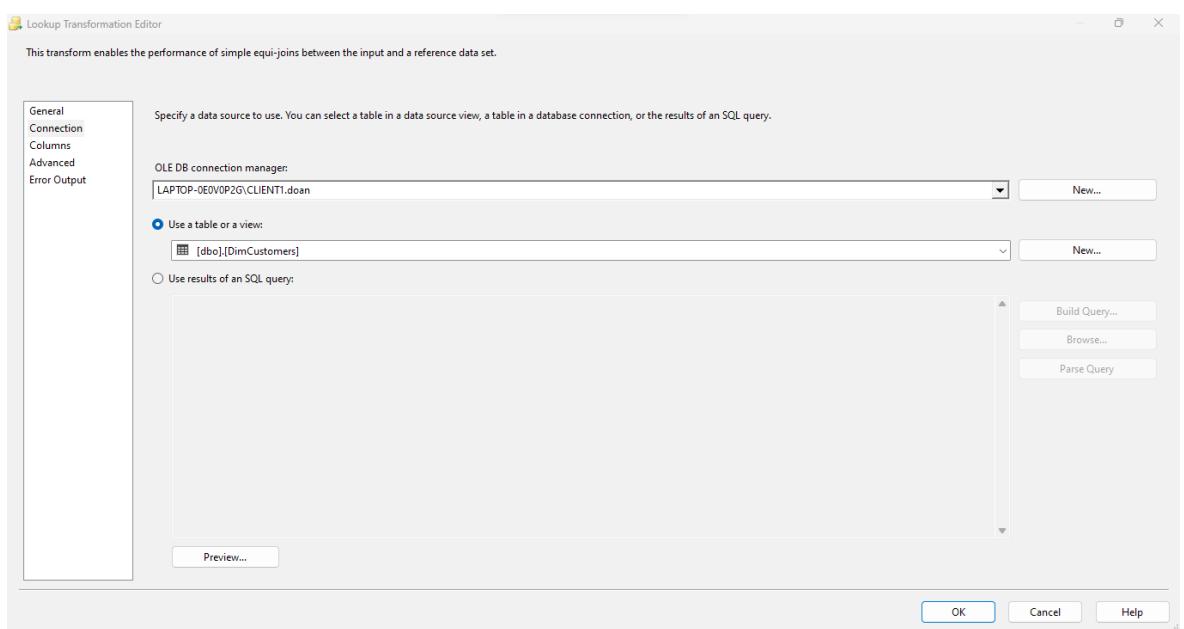
Ở **Connection** chọn bản **DimLocation** đã tạo ra trước đó

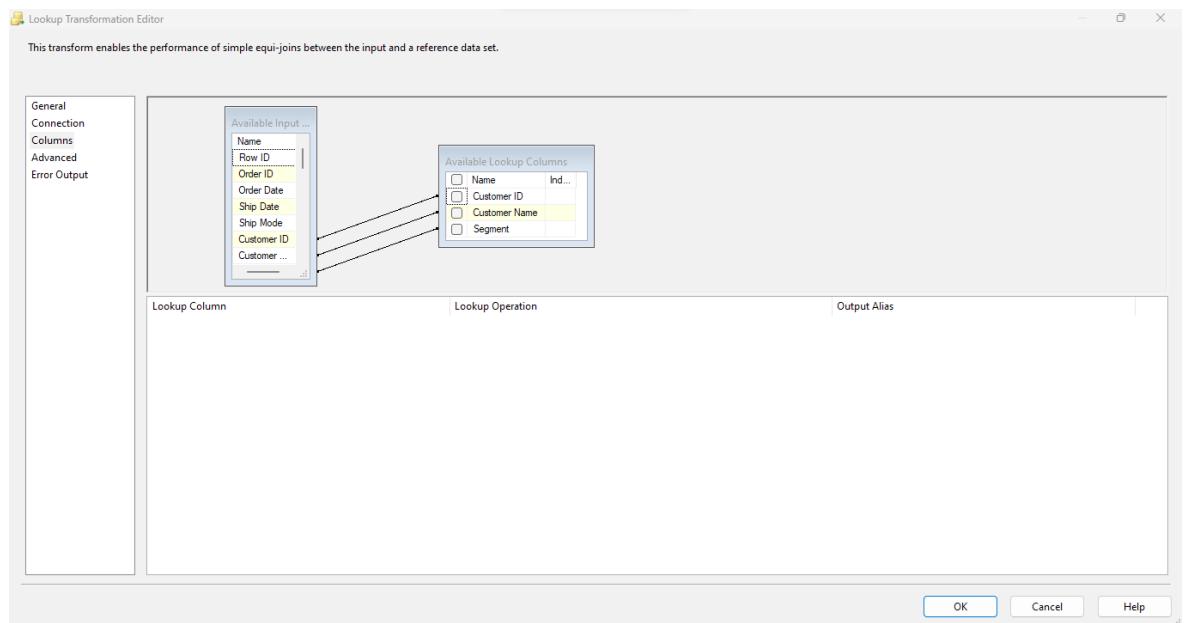


Ở **Column** chọn **LocationID** (vì đây là cột mới được tạo ra)



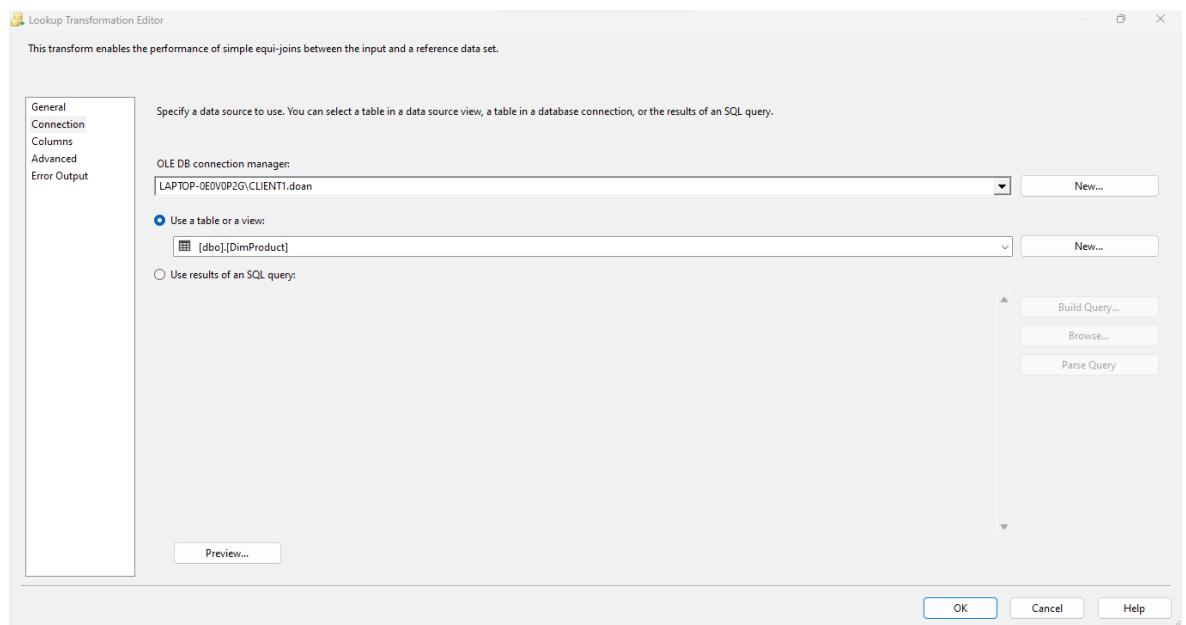
DimCustomer

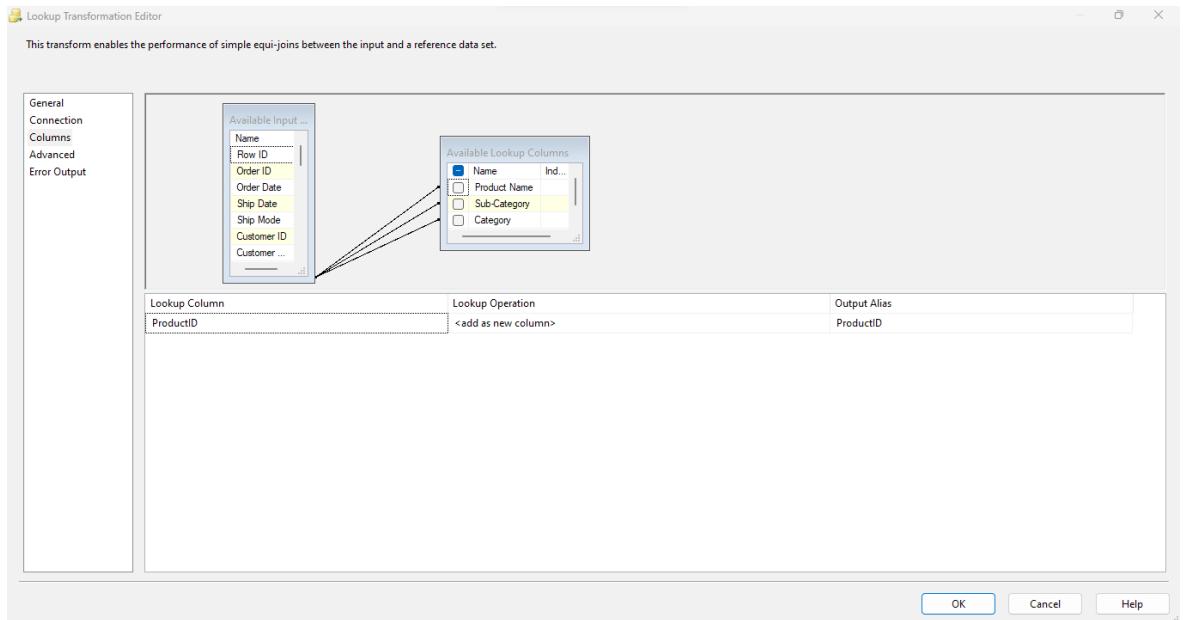




Ở đây không có CustomerID vì cột này đã có sẵn trong Excel

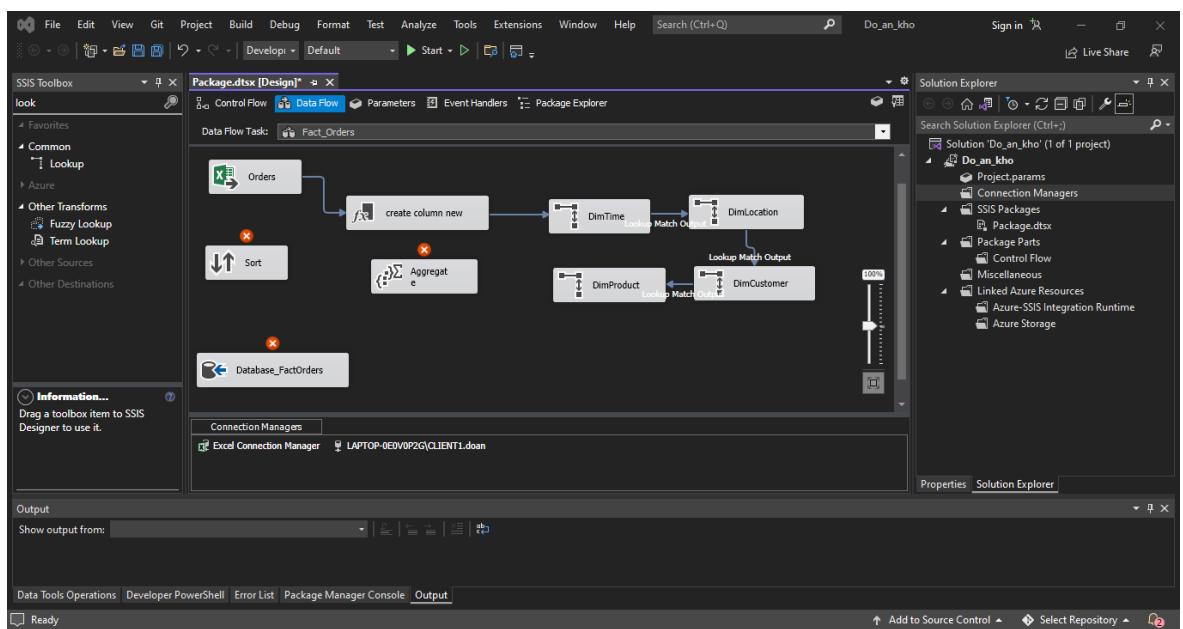
DimProduct



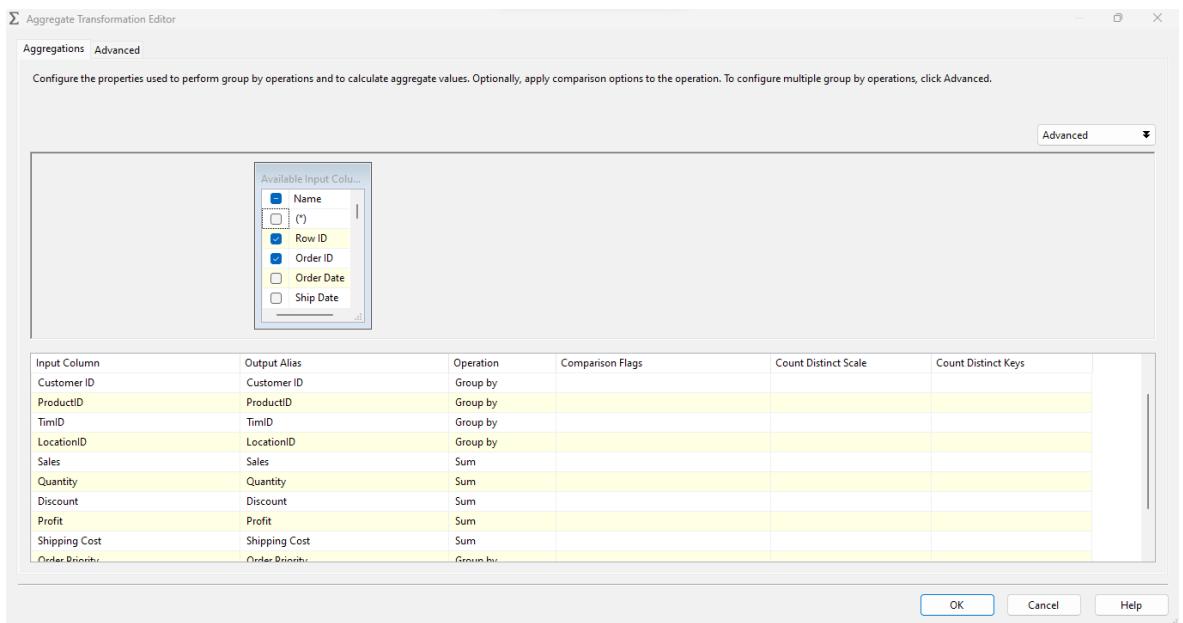


Chọn cột ProductID vì đây là cột mới tạo ra để thay thế các giá trị ProductID cũ (ProductID này khác với ProductID trong Excel)

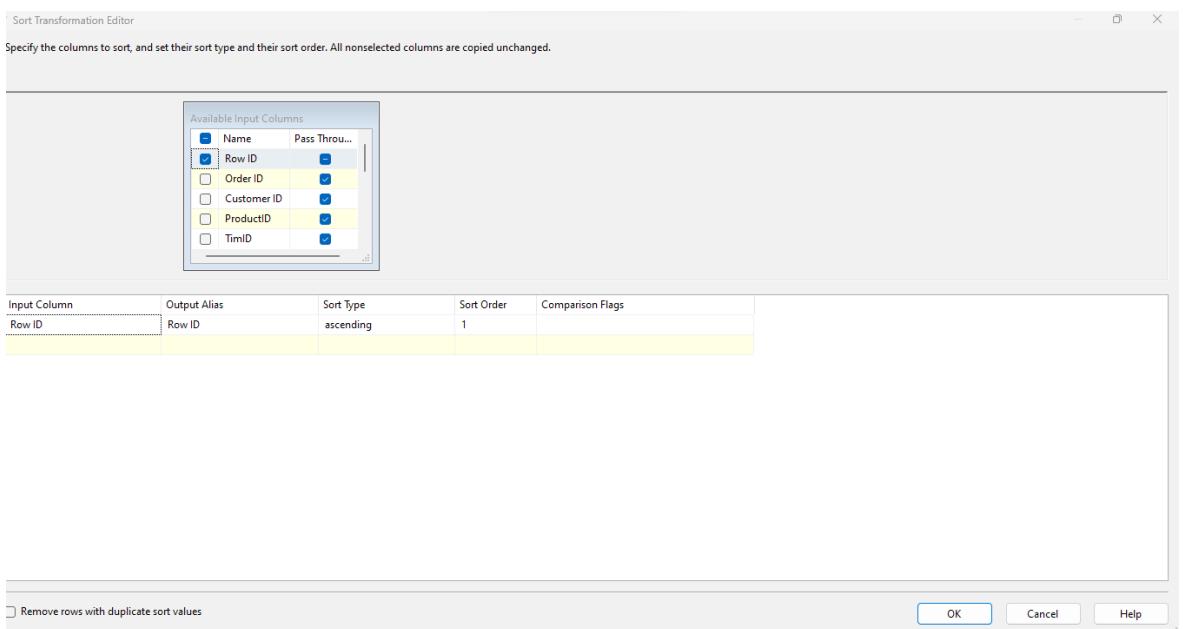
Tạo thêm Aggregate, Sort và OLE DB Destination kết nối tới SQL của bảng Fact như hình



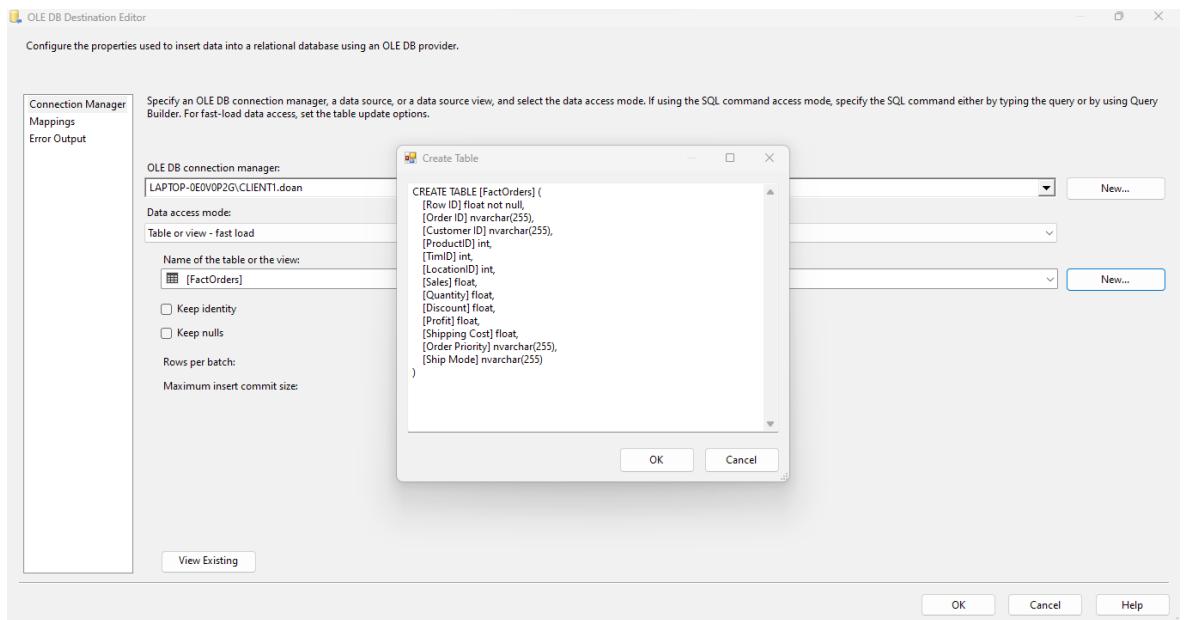
Chuột phải vào **Aggregate – Edit**



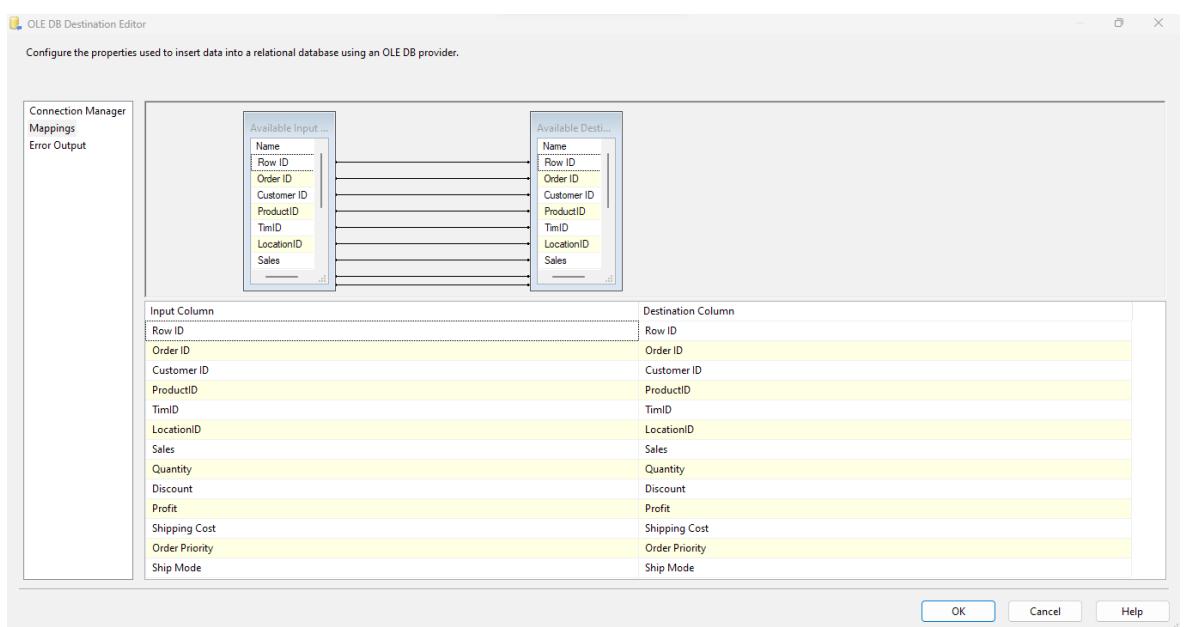
Chuột phải vào Sort – Edit



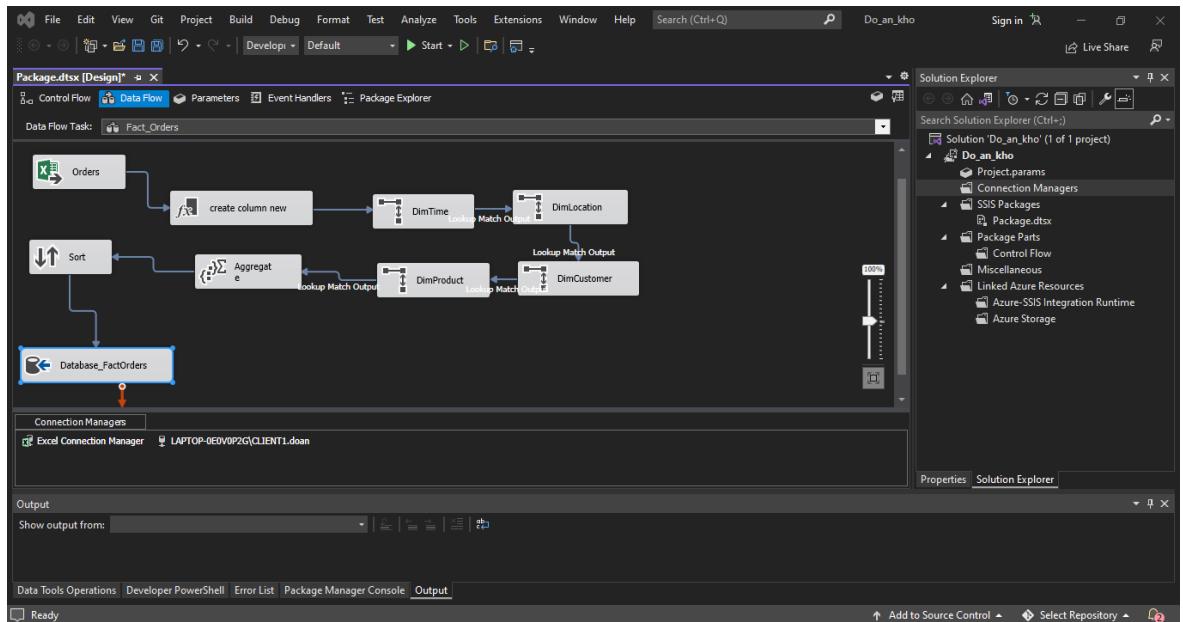
Chuột phải vào OLE DB Destination – Edit



Mappings

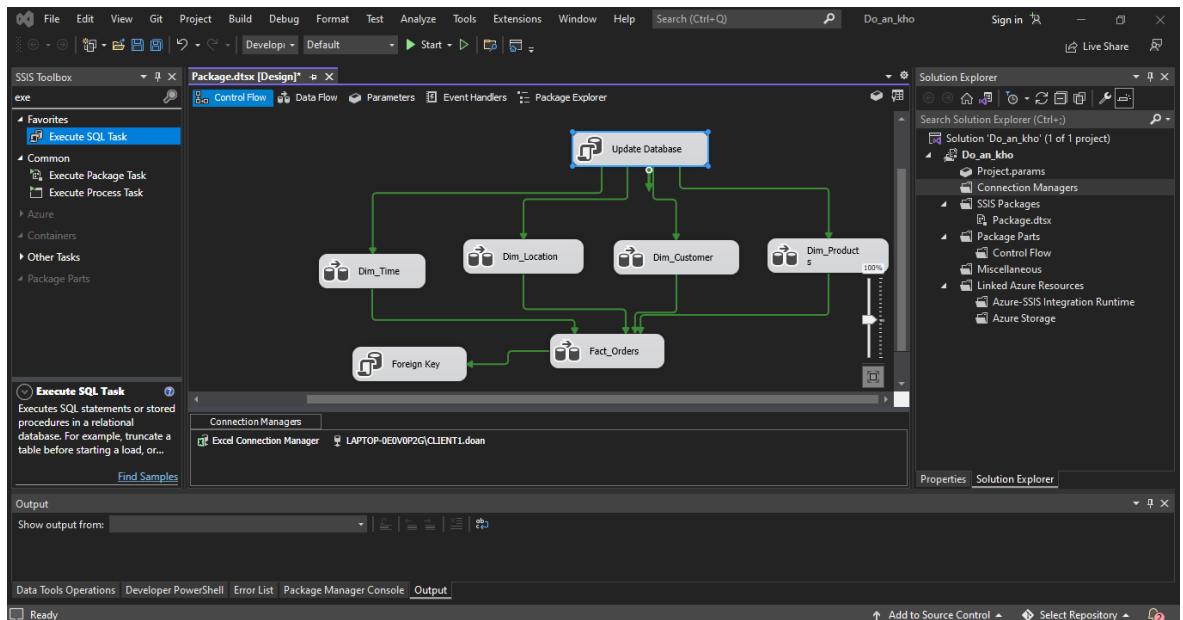


Data Flow của bảng Fact sẽ trông giống như hình:



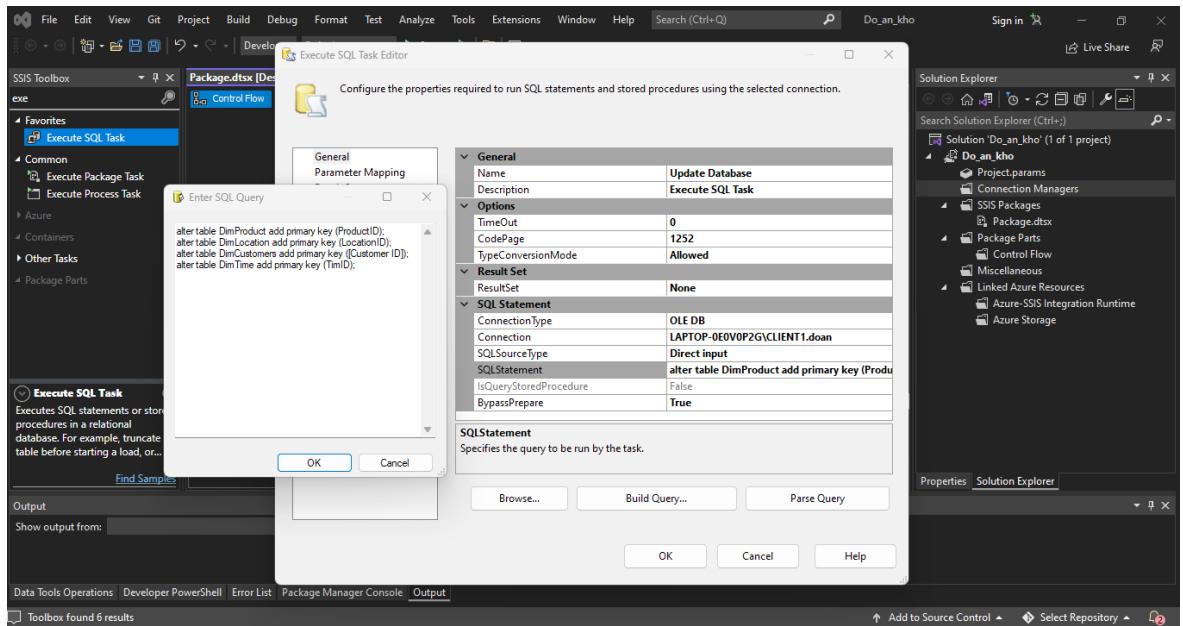
2.5. Quá trình tạo Execute SQL Task

Ở Control Flow kéo thả 2 Execute SQL Task và đặt tên lần lượt là Update Database và Foreign Key



Chuột phải vào **Update DataBase – Edit**

Ở **SQL Statement**, ngay mục **Connection** chọn kết nối với Database, mục **SQLStatement** viết câu query như sau:

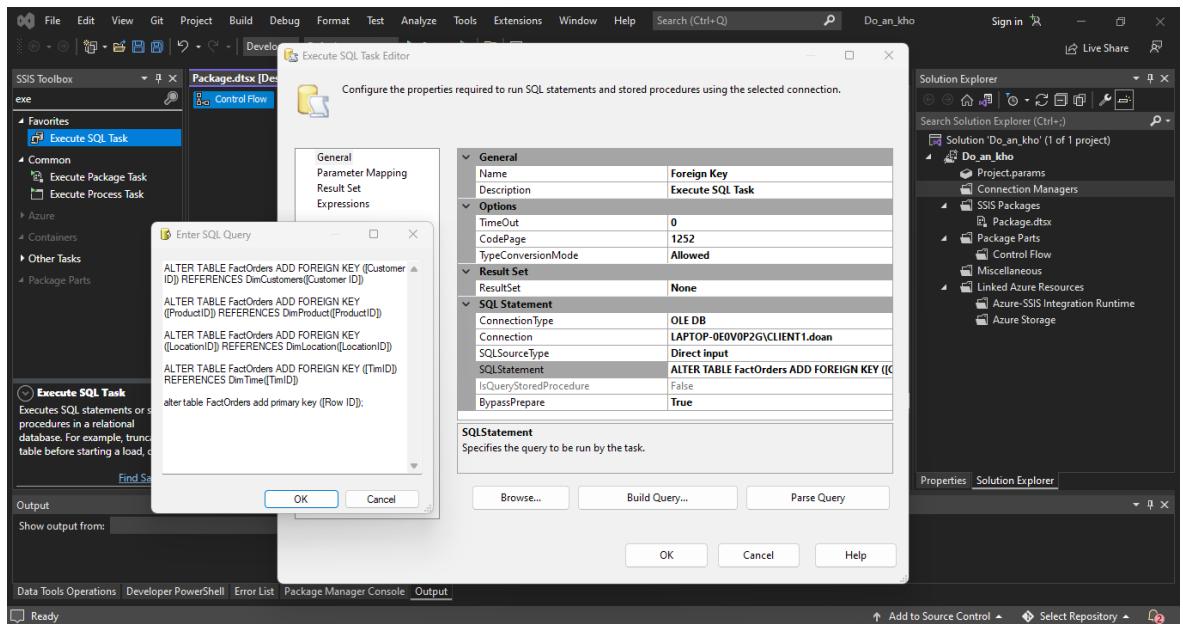


Câu query này nhằm tạo khóa chính cho 4 bảng Dim được tạo ra trong SQL

Click OK.

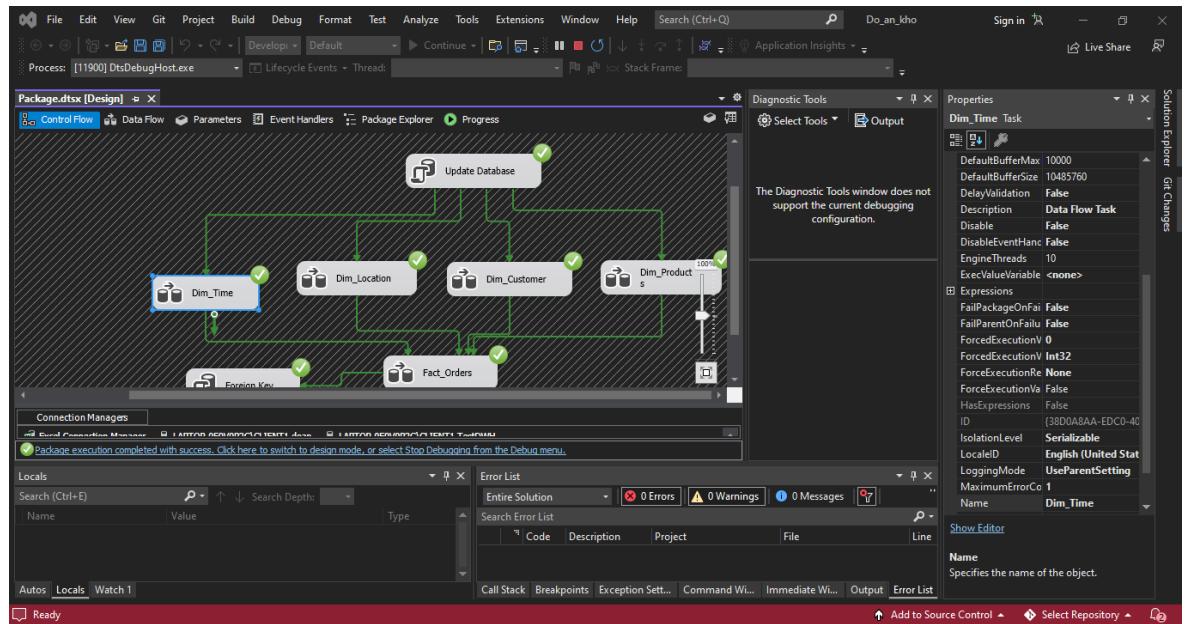
Chuột phải vào Foreign Key – Edit

Ở SQL Statement, ngay mục Connection chọn kết nối với Database, mục SQLStatement viết câu query như sau:



Câu query này nhằm tạo khóa ngoại tham chiếu đến 4 bảng Dim được tạo ra trong SQL

Click OK.



CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU (SSAS)

Chương này, dùng công cụ SSAS xây dựng mô hình và xây dựng khôi – cube, xây dựng các API, phân tích dữ liệu trên Cube. Sử dụng chức năng Pivot table trong Excel thực hiện tương tự như quá trình SSAS và tiến hành đổi chiều kết quả với phân tích bằng tay.

1. Danh sách câu truy vấn

- Có bao nhiêu đơn hàng được giao
- Thống kê tỉ lệ loại hình giao hàng được người dùng chọn ?
- Thống kê doanh số mang lại từ 3 loại hình sản phẩm chính (furniture, office supplies, technology)
- Thống kê chi tiết doanh thu mang lại từ từng loại sản phẩm theo Sub-Category (theo Phone, Tables, ...)
- Thống kê lợi nhuận theo khu vực
- 10 sản phẩm bán chạy nhất
- 10 bang có số lượng mua hàng nhiều nhất (State)
- 5 khu vực có số lượng mua hàng nhiều nhất (Region)
- So sánh tổng doanh thu với chi phí ship
- So sánh số lượng bán hàng giữa các danh mục (Category)
- Thống kê những người hay mua sản phẩm nhất (sắp xếp giảm dần theo lượt mua)
- Xu hướng số lượng đơn hàng theo từng tháng trong năm

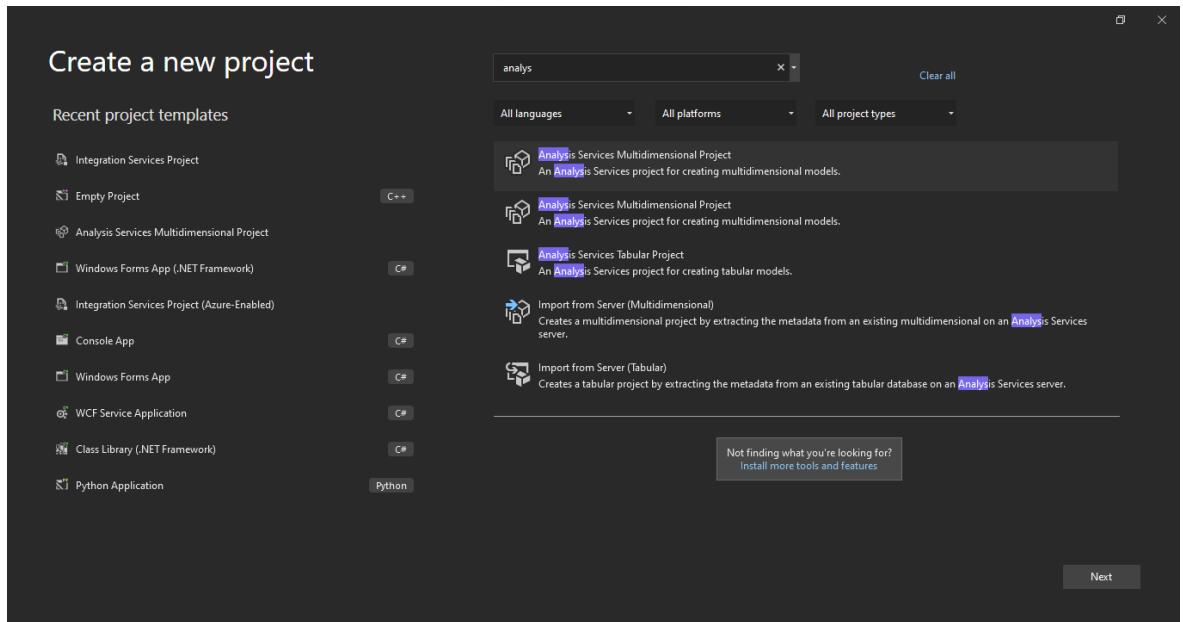
2. Quá trình xây dựng mô hình

2.1. Data Sources

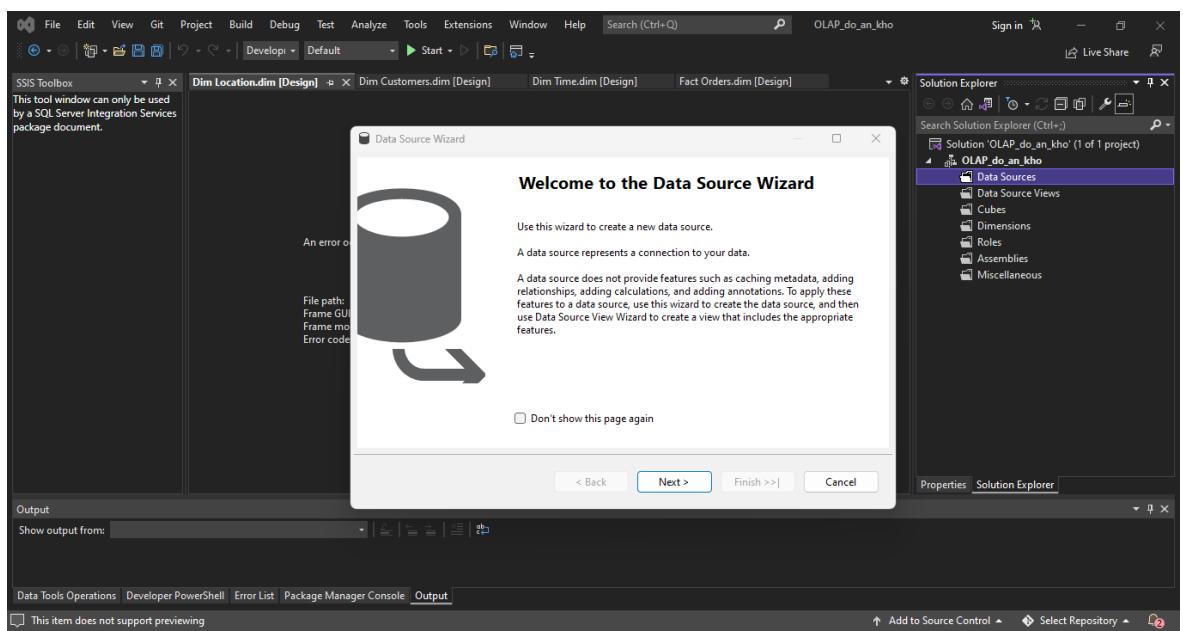
Chọn File – New – Project

Chọn Analysis Services Multidimensional Project

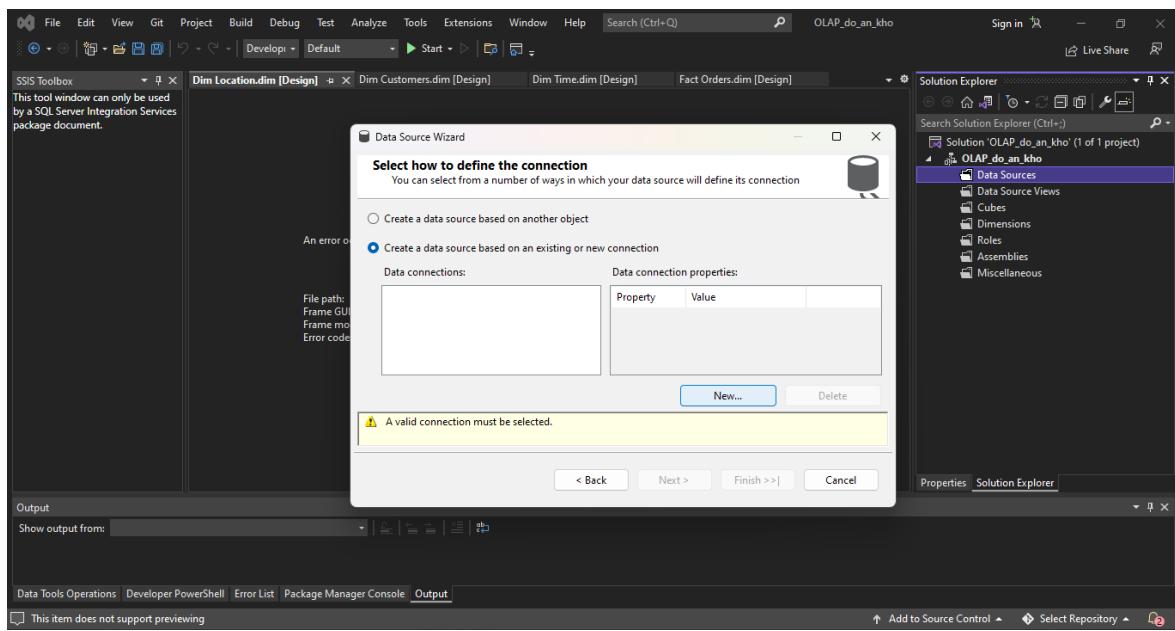
Đặt tên Project và click Create



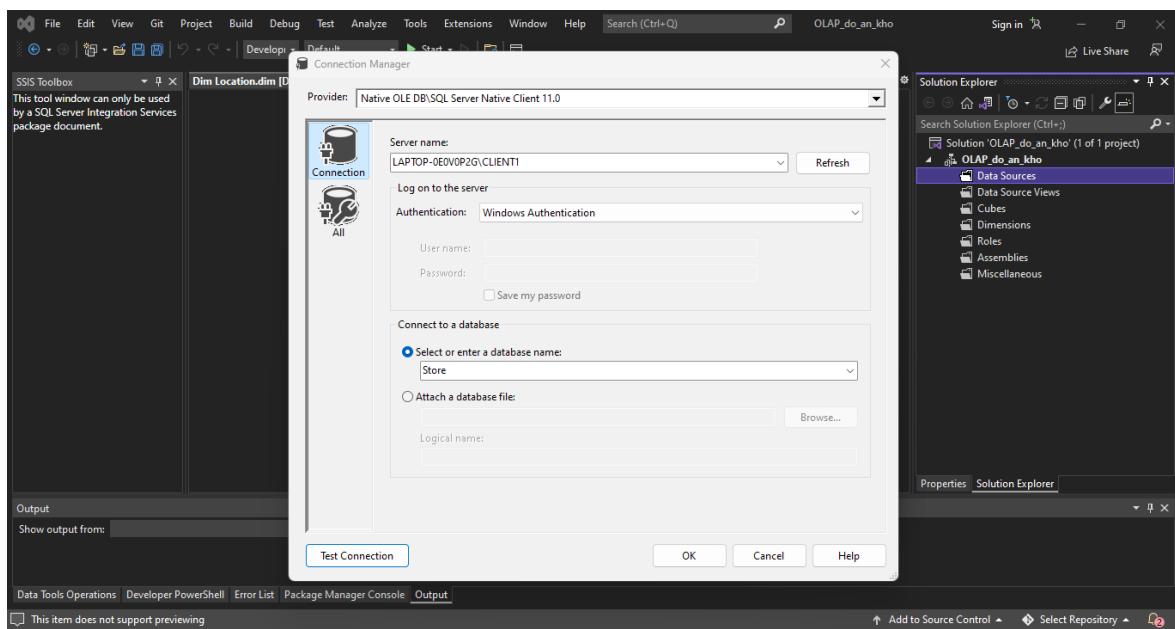
Chọn New Data Source



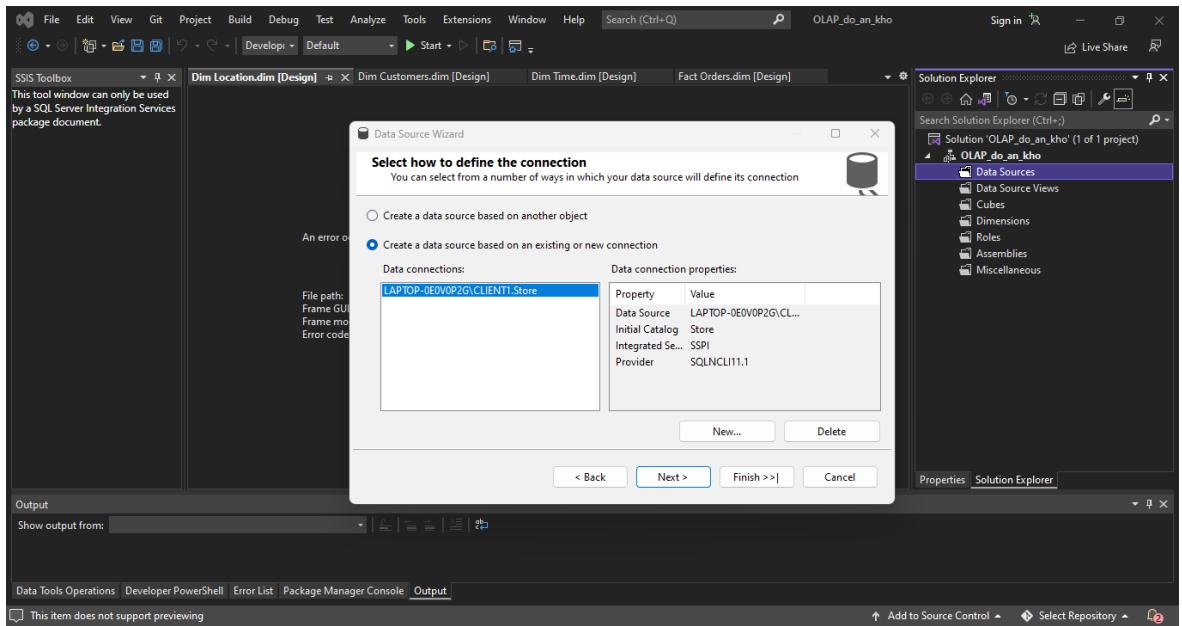
Click Next



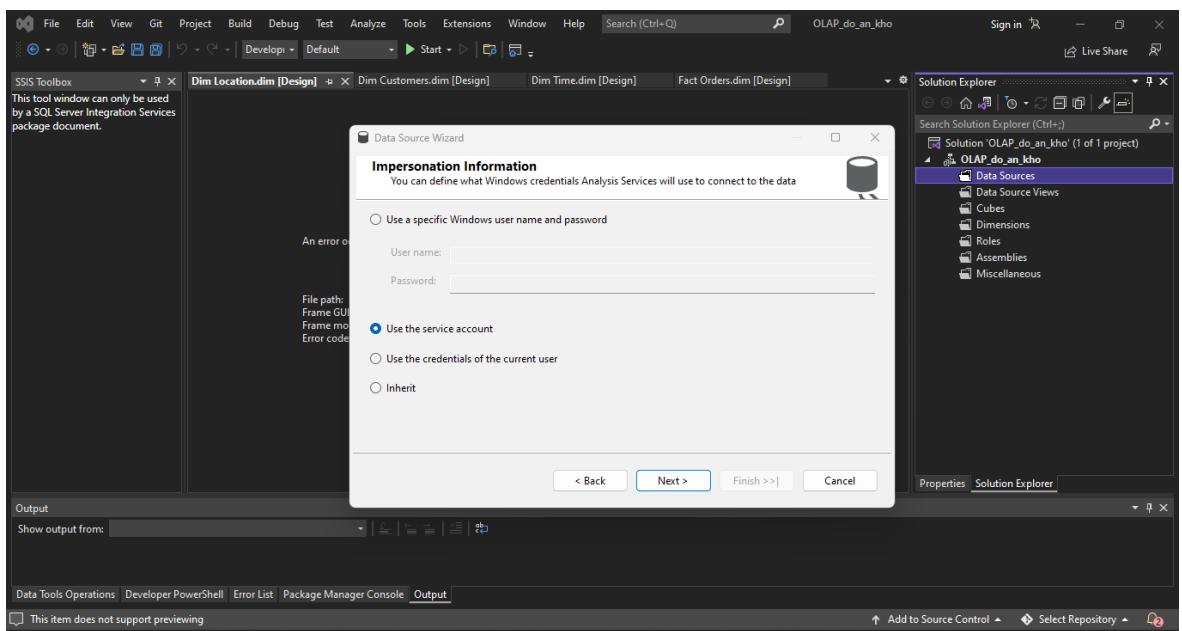
Click New, chọn kết nối đến database



Click OK

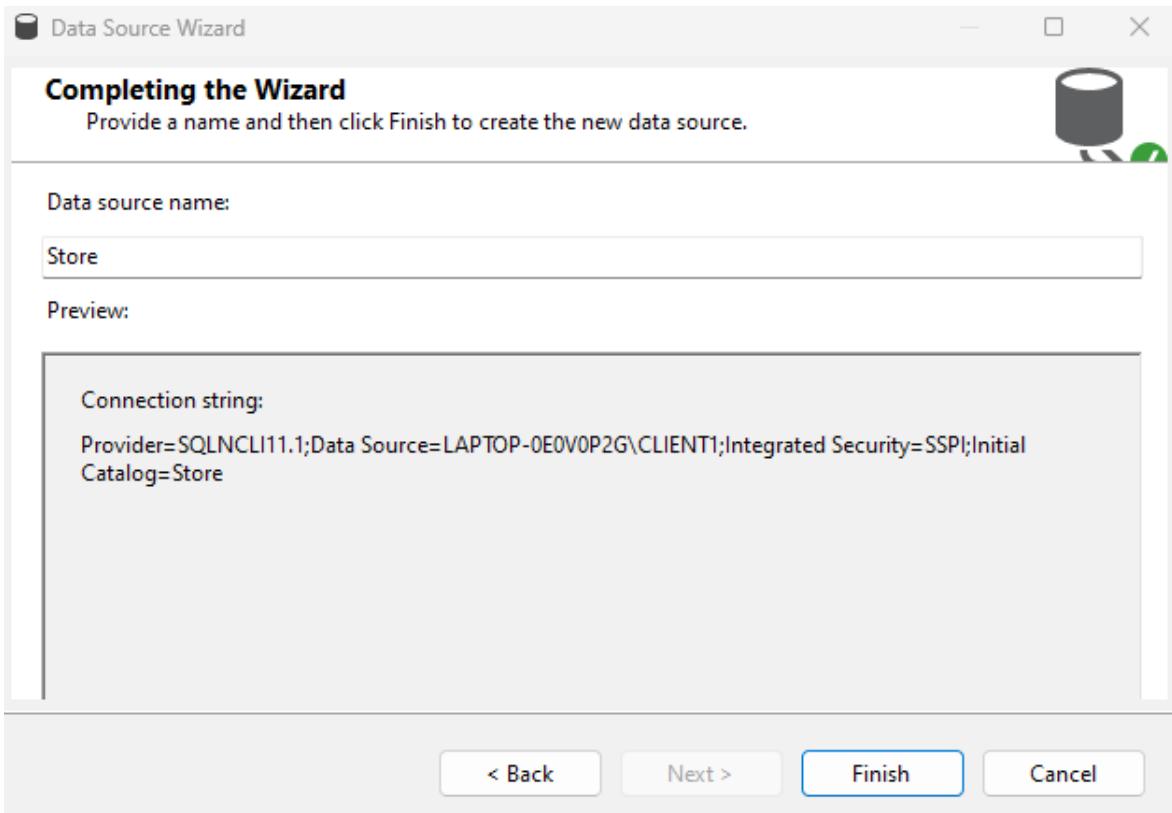


Chọn Next



Tích chọn **Use the service account – Next**

Đặt tên cho Data Source rồi chọn Next

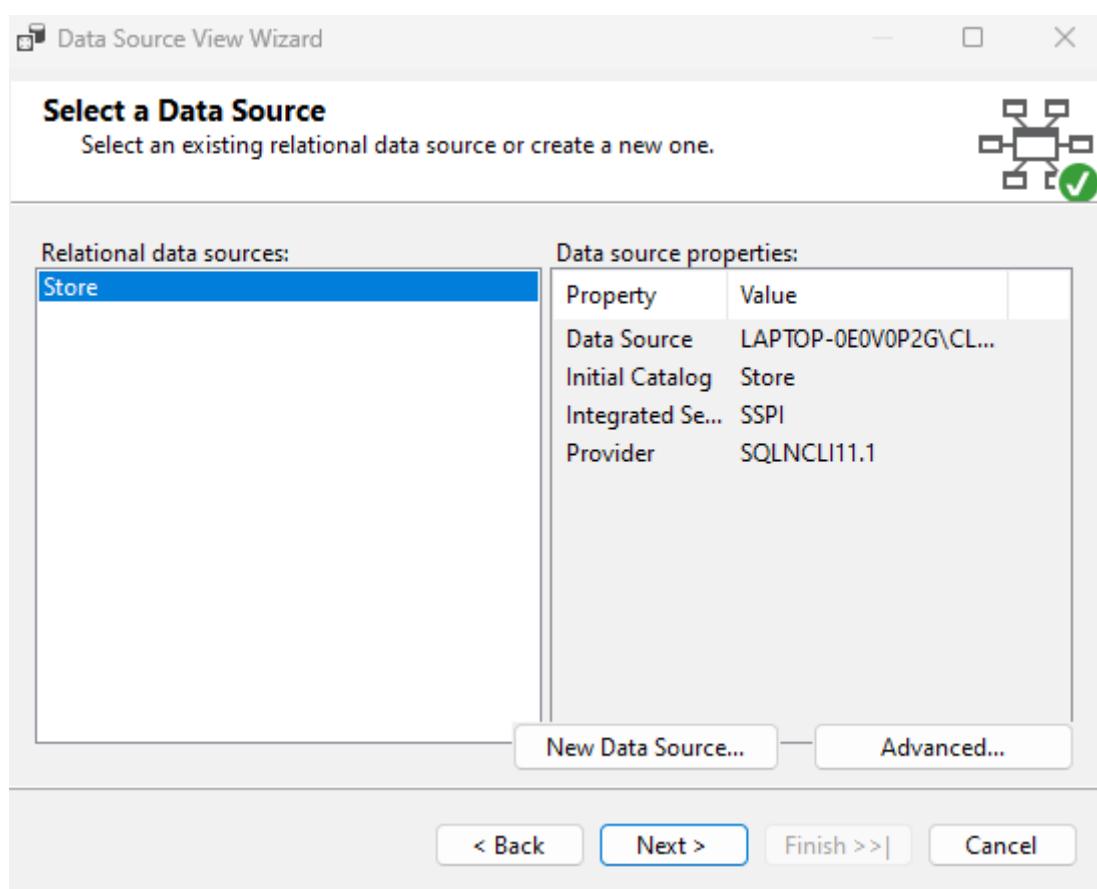


Bấm Finish

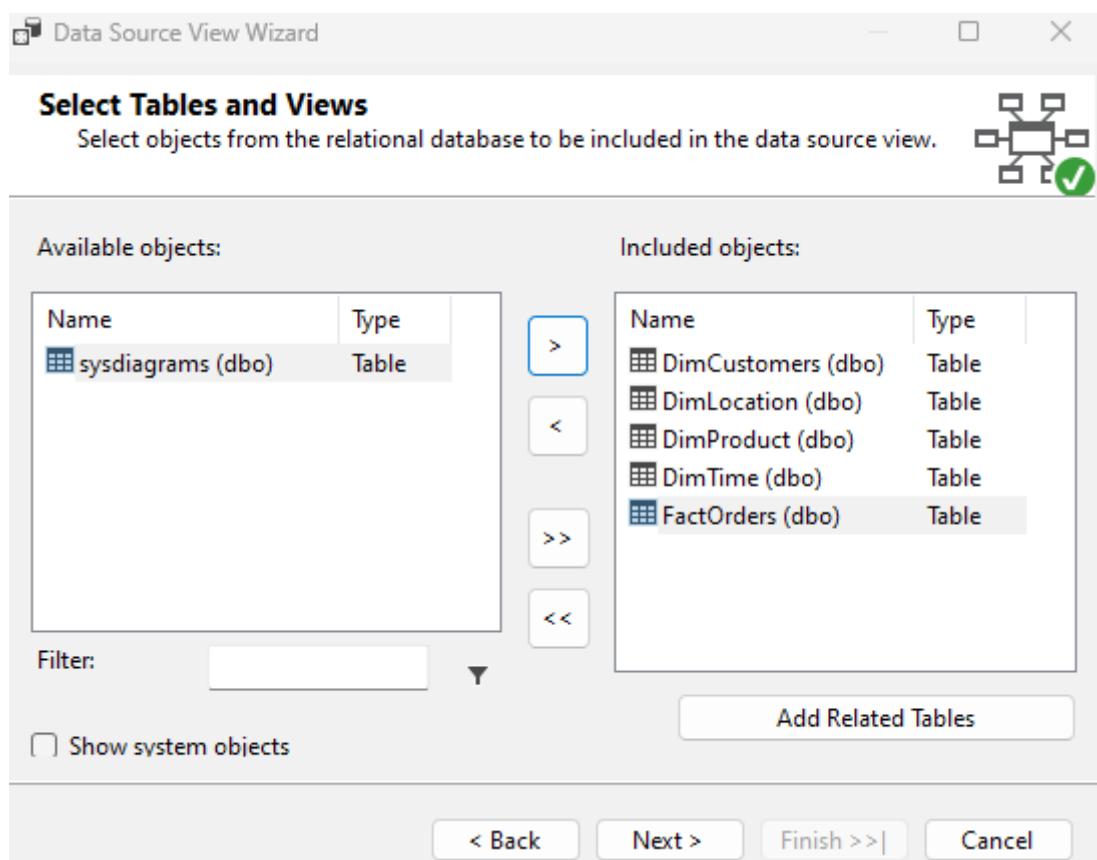
2.2. Data Sources View

Click chuột phải vào **Data Source Views – New Data Source View ...**

Một cửa sổ mới hiện ra , chọn **Next**



Click Next

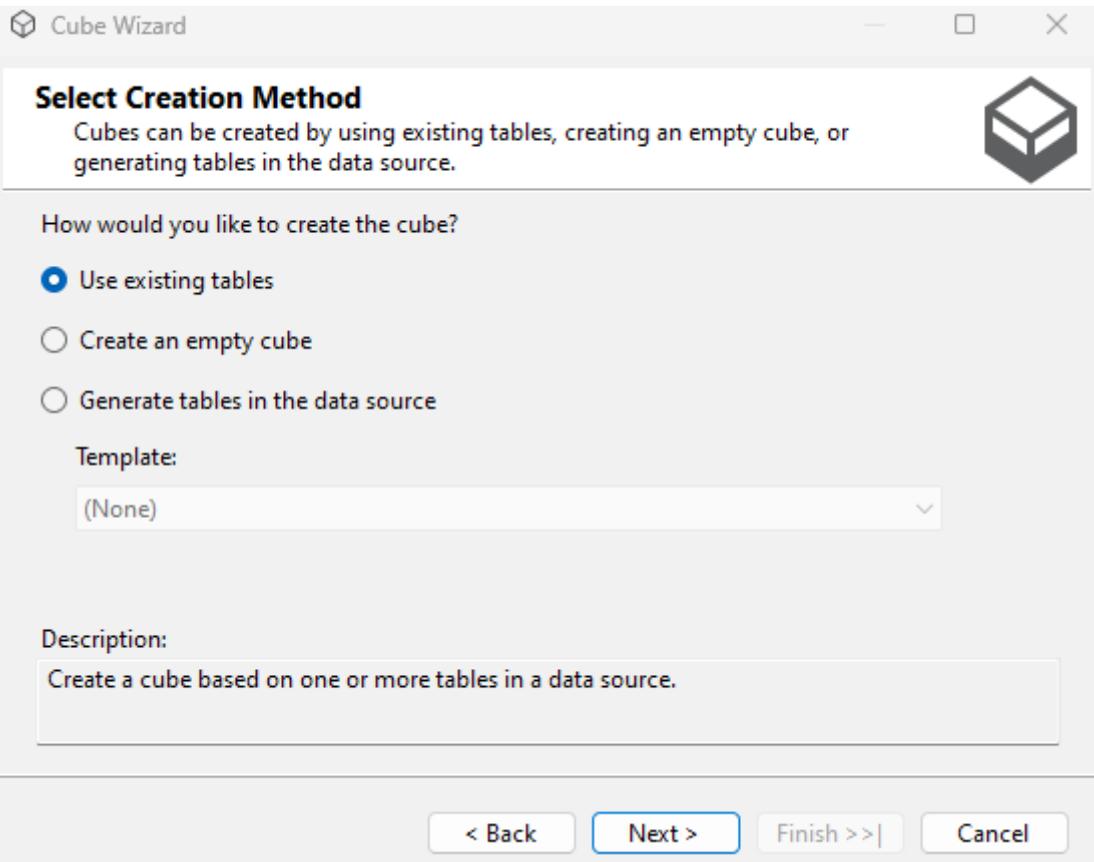


Click Next

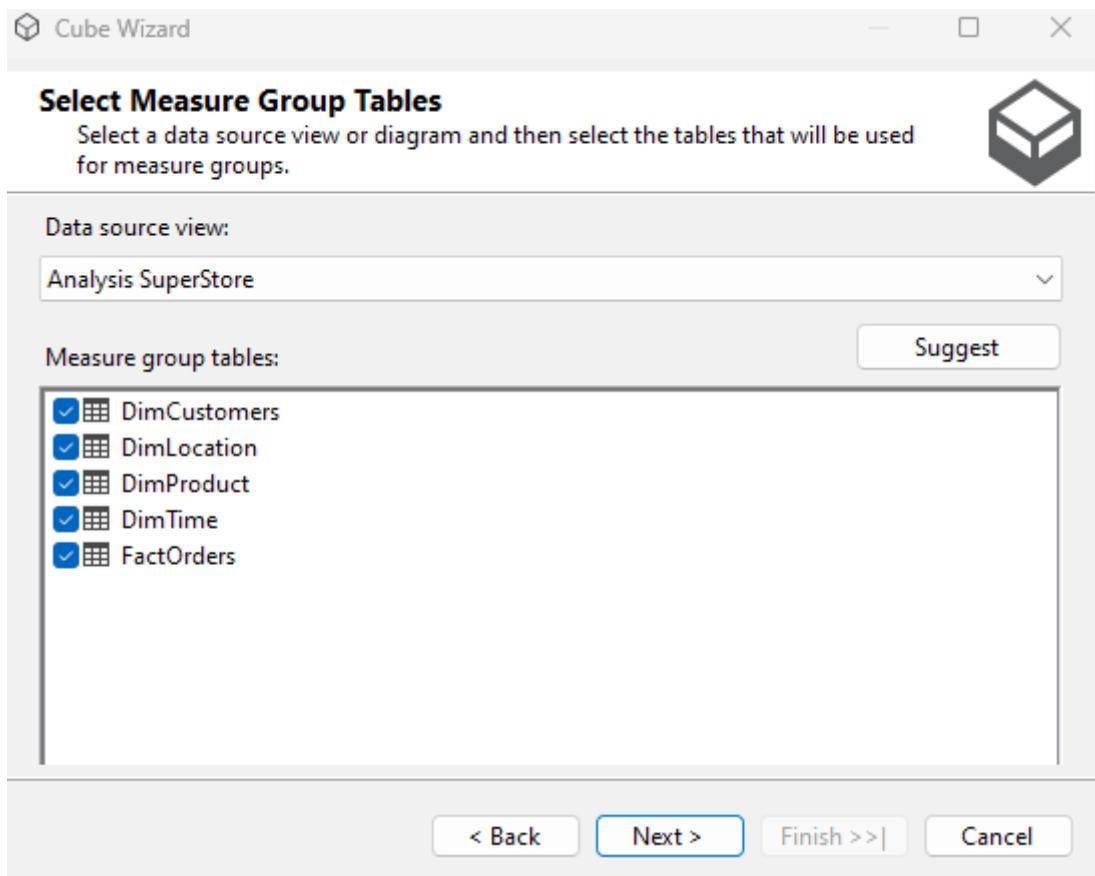
Đặt tên cho **Data Source Views** rồi chọn **Finish**

2.3. Cubes

Click chuột phải vào **Cubes – New Cubes ...**

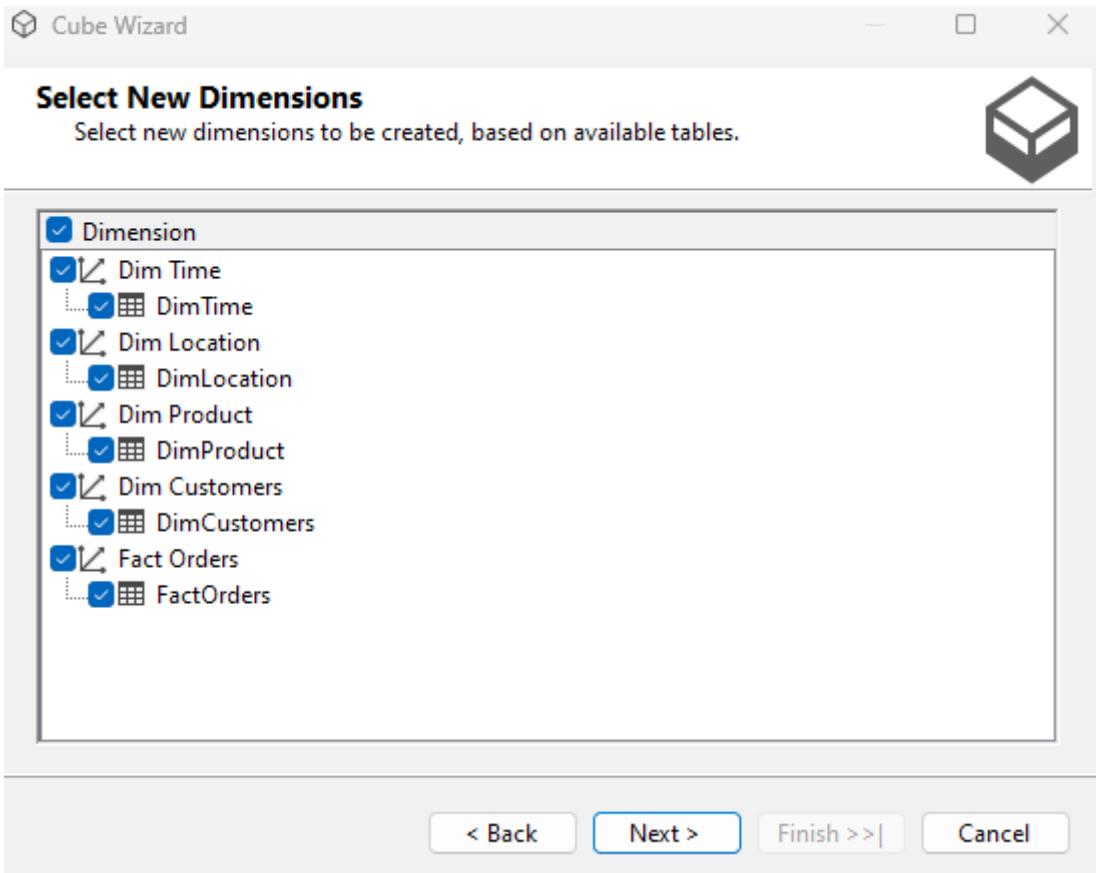


Chọn **Use existing tables** và bấm **Next**

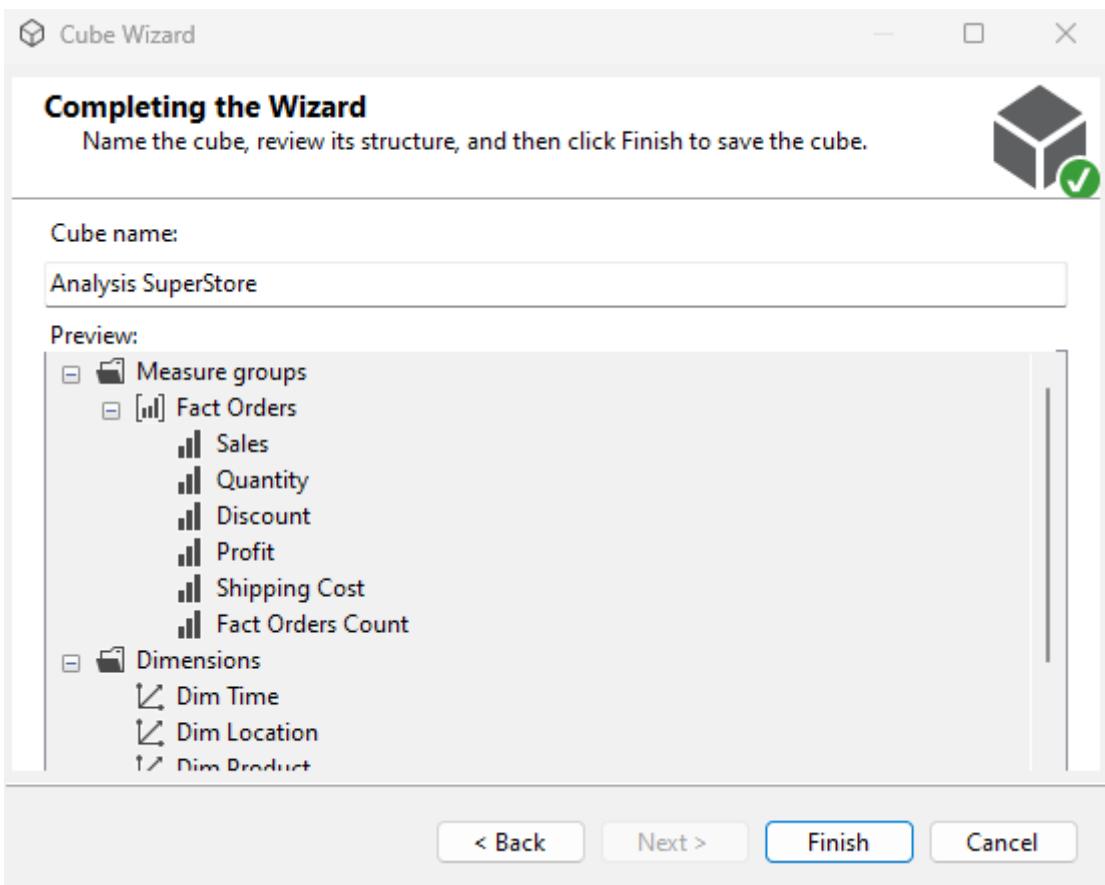


Chọn **FactOrders** và bấm **Next**

Chọn tất cả và click **Next**

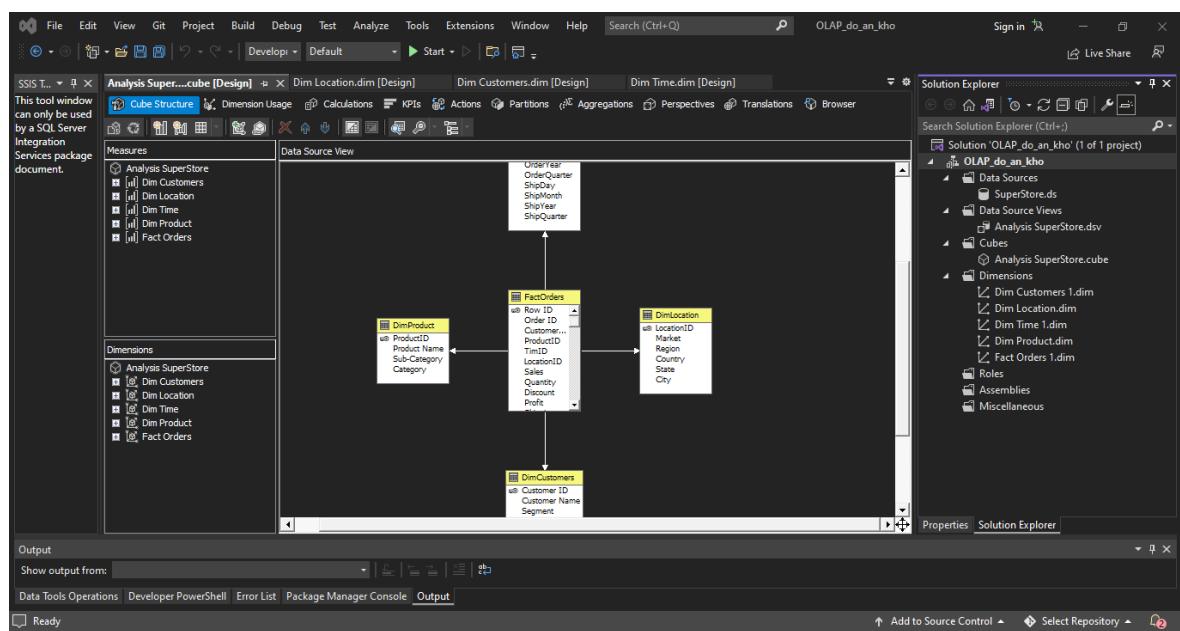


Click Next



Click Finish

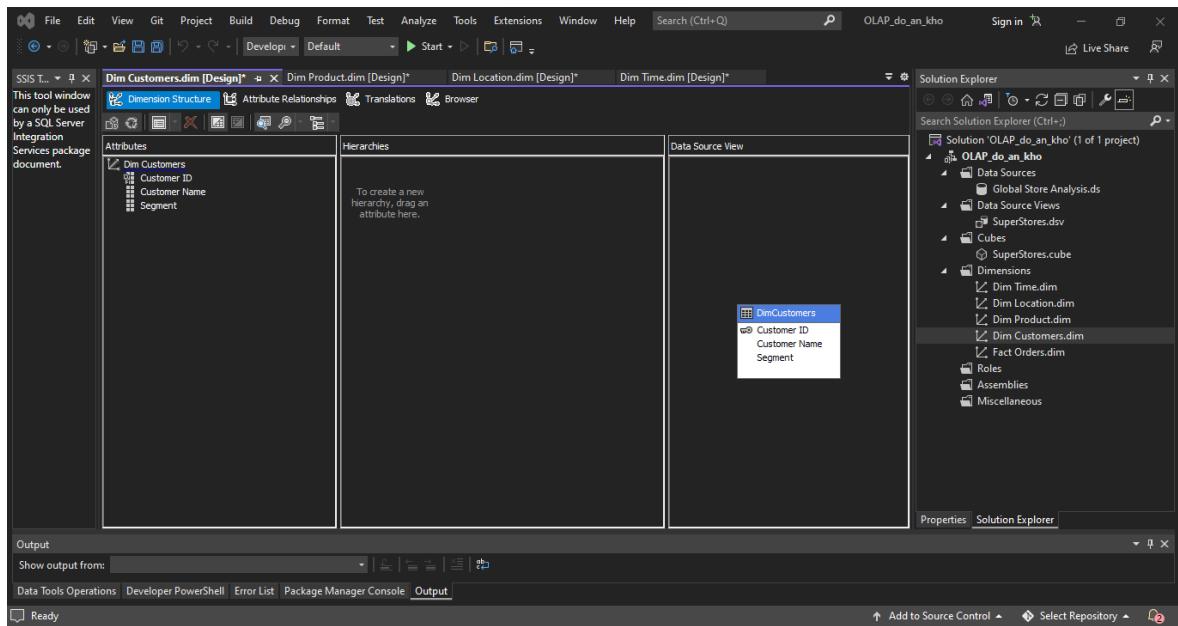
Cube đã tạo



2.4. Dimension

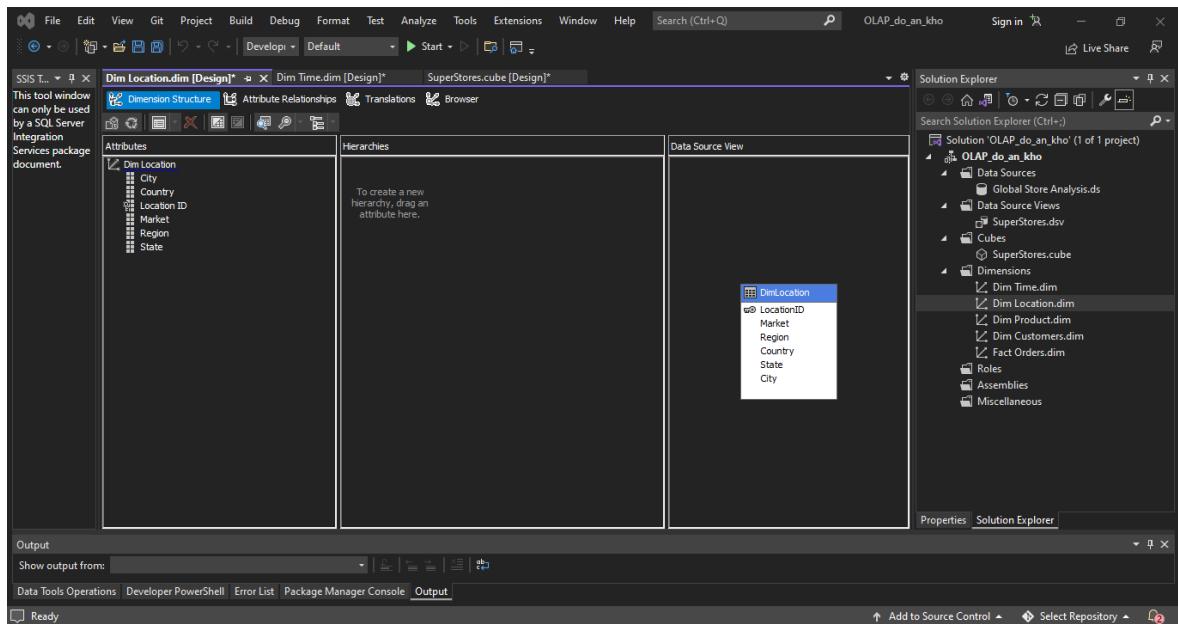
Dim Customer

Chọn các thuộc tính từ bảng DimCustomer sang Attributes như hình



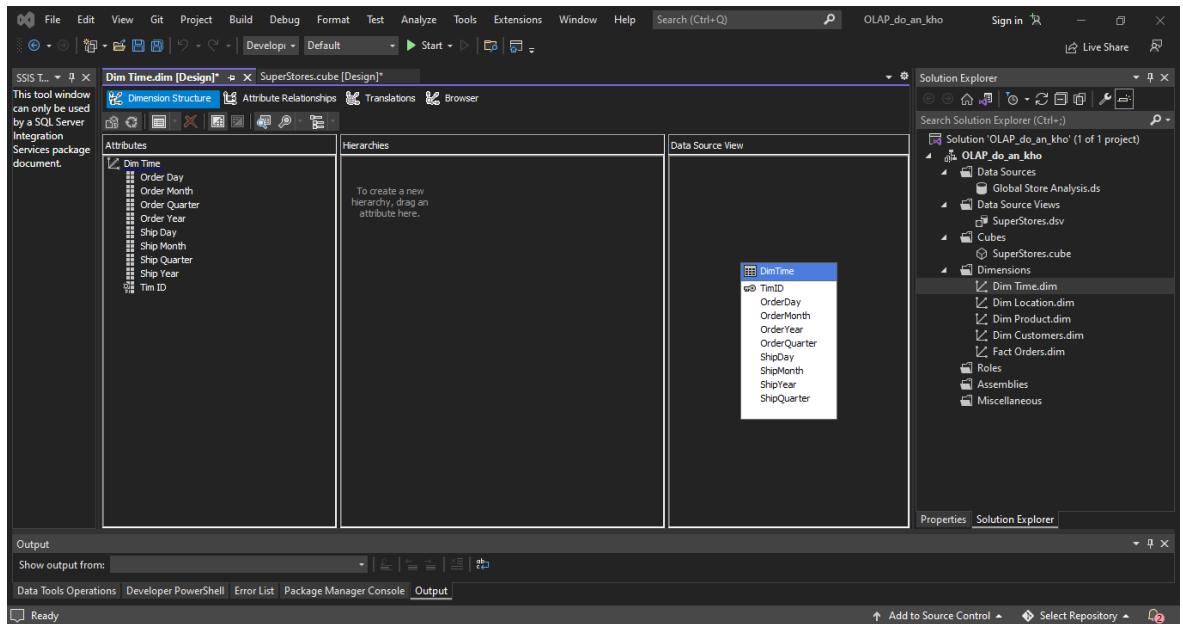
Dim Location

Chọn các thuộc tính từ bảng DimLocation sang Attributes như hình



Dim Time

Chọn các thuộc tính từ bảng DimTime sang Attributes như hình

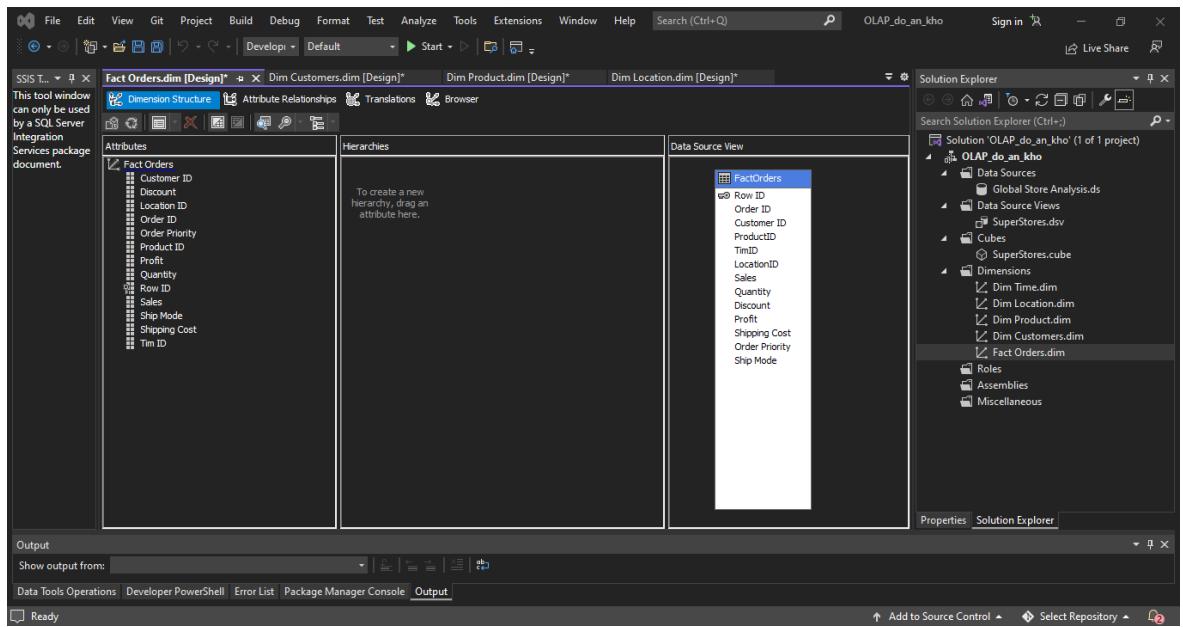


Dim Product

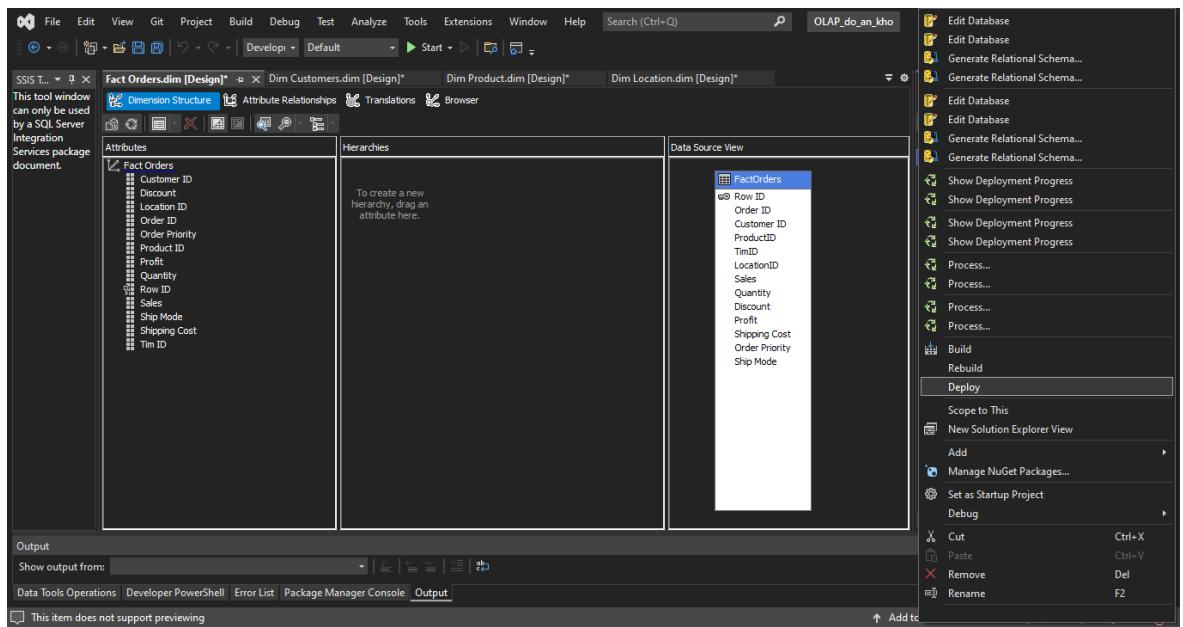
Chọn các thuộc tính từ bảng DimProduct sang Attributes như hình

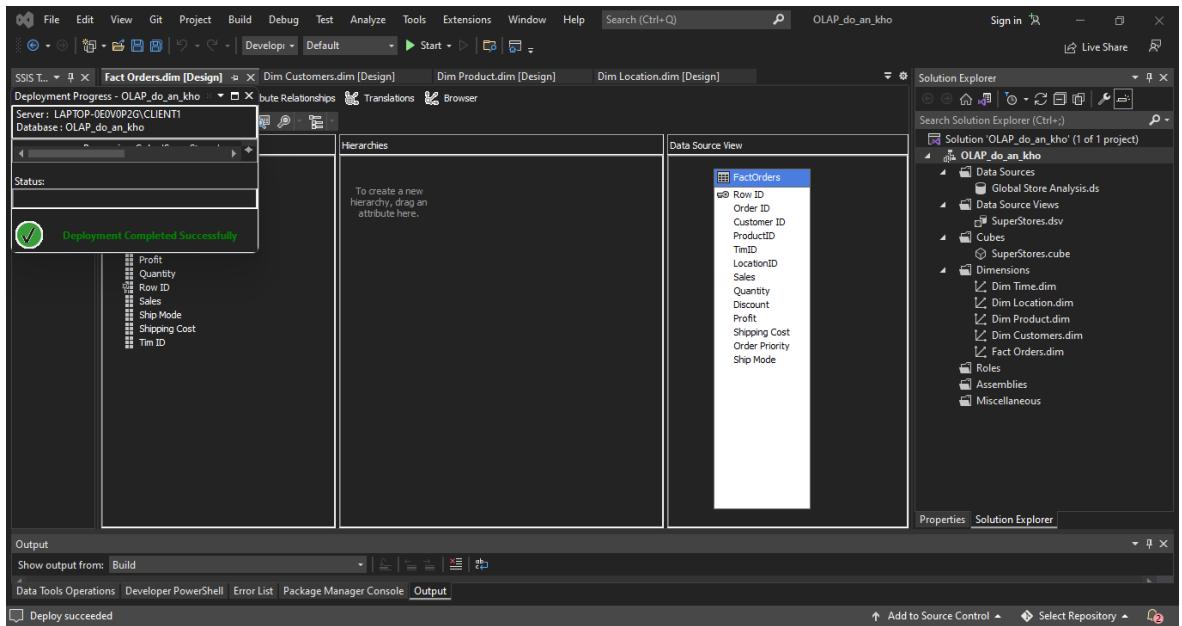
Fact Orders

Chọn các thuộc tính từ bảng FactOrders sang Attributes như hình

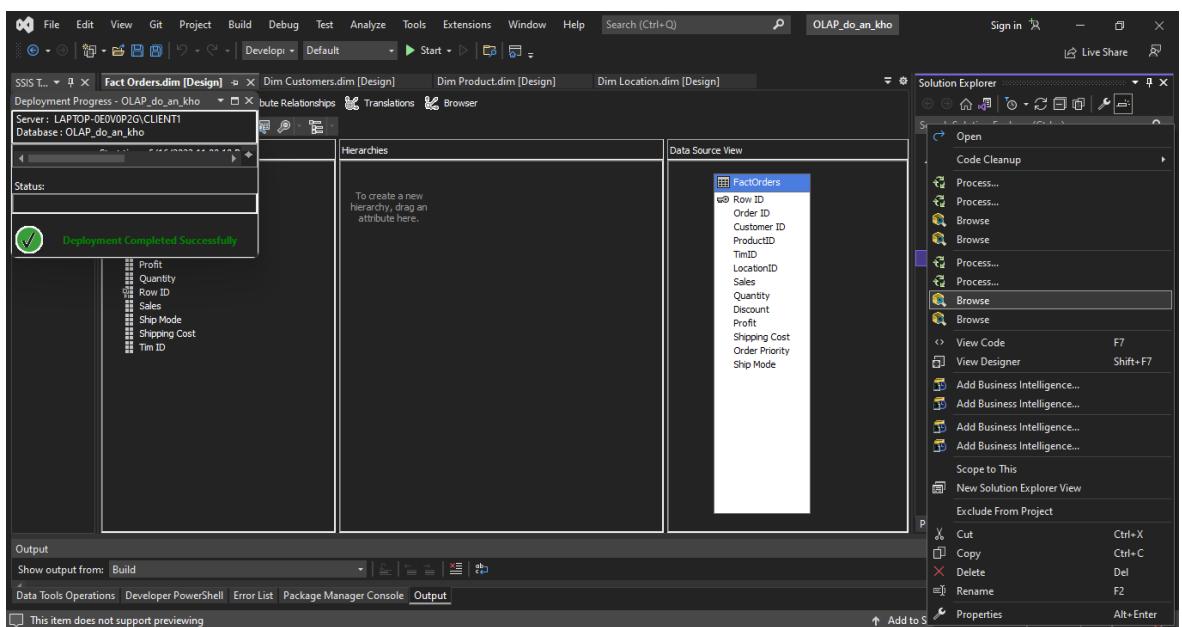


Click chuột phải vào project chọn Deploy





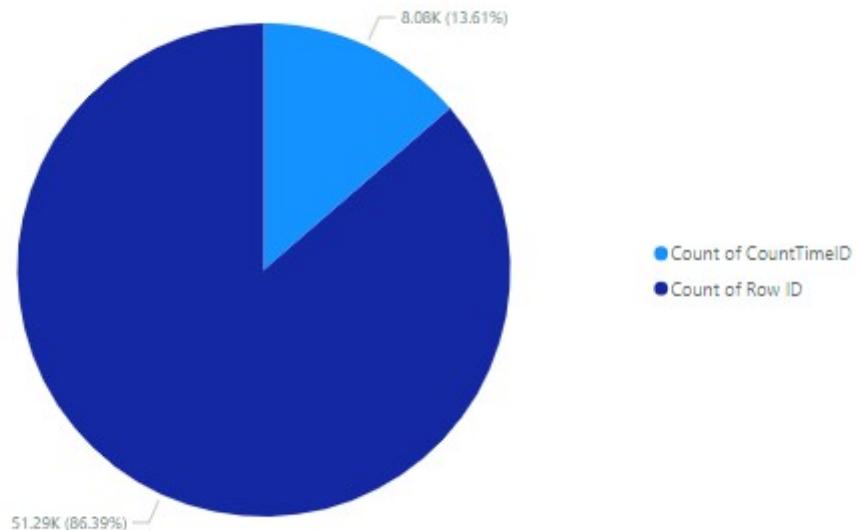
Click chuột phải vào Browse



3. Các câu truy vấn

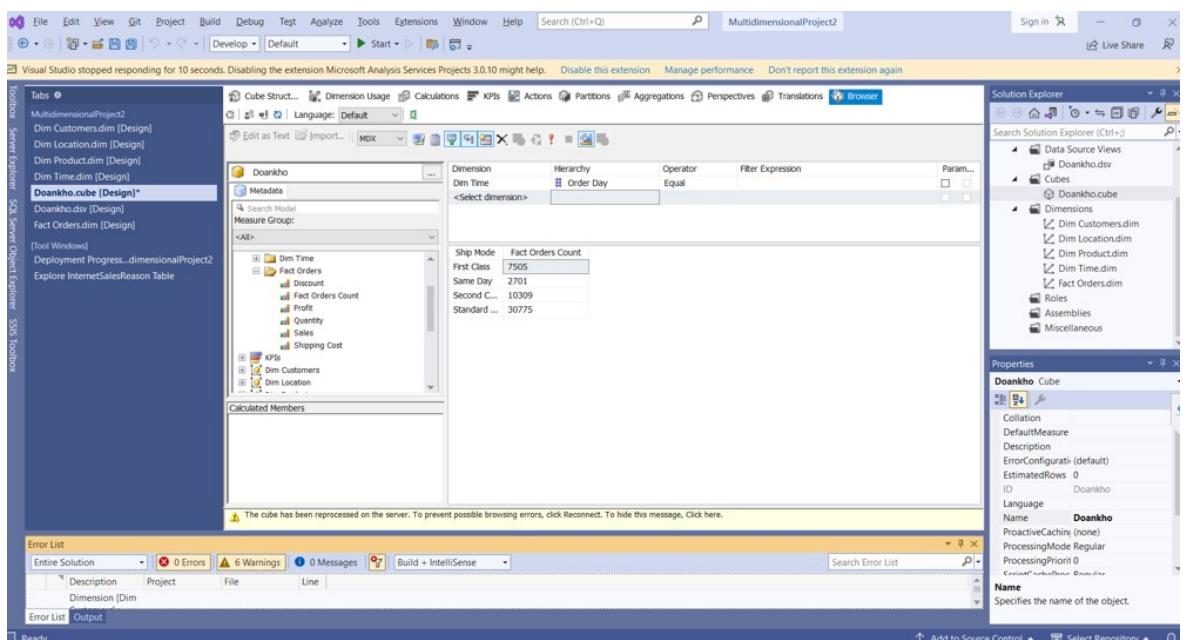
3.1. Tính tỷ lệ đơn hàng được giao ở nước Mỹ so với tổng thẻ

- Sử dụng Power BI



3.2. Thống kê tỉ lệ phương thức giao hàng được người dùng chọn

- Sử dụng công cụ SSAS

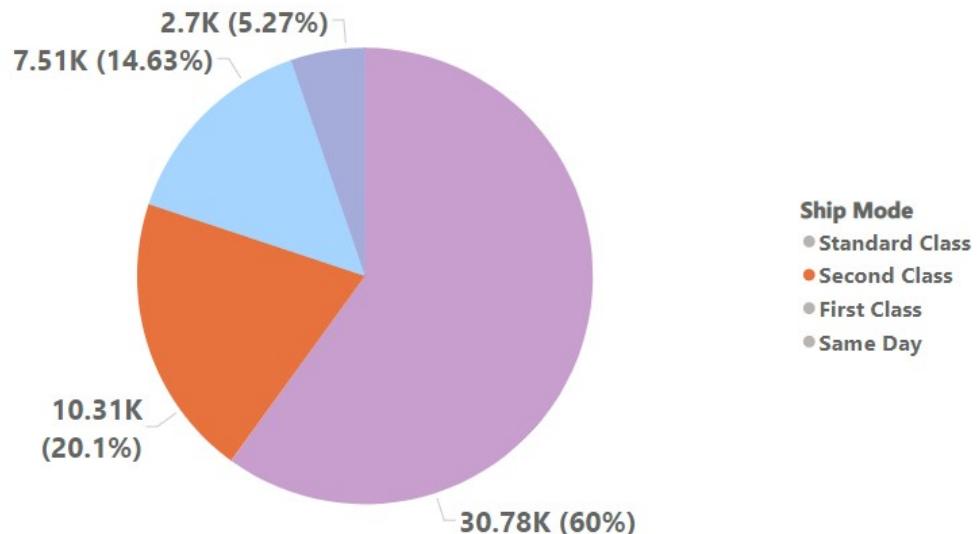


- Sử dụng Pivot table trong Excel

The screenshot shows a Microsoft Excel spreadsheet titled 'Book2 - Excel'. The PivotTable is located in the range A1:C6. The data is as follows:

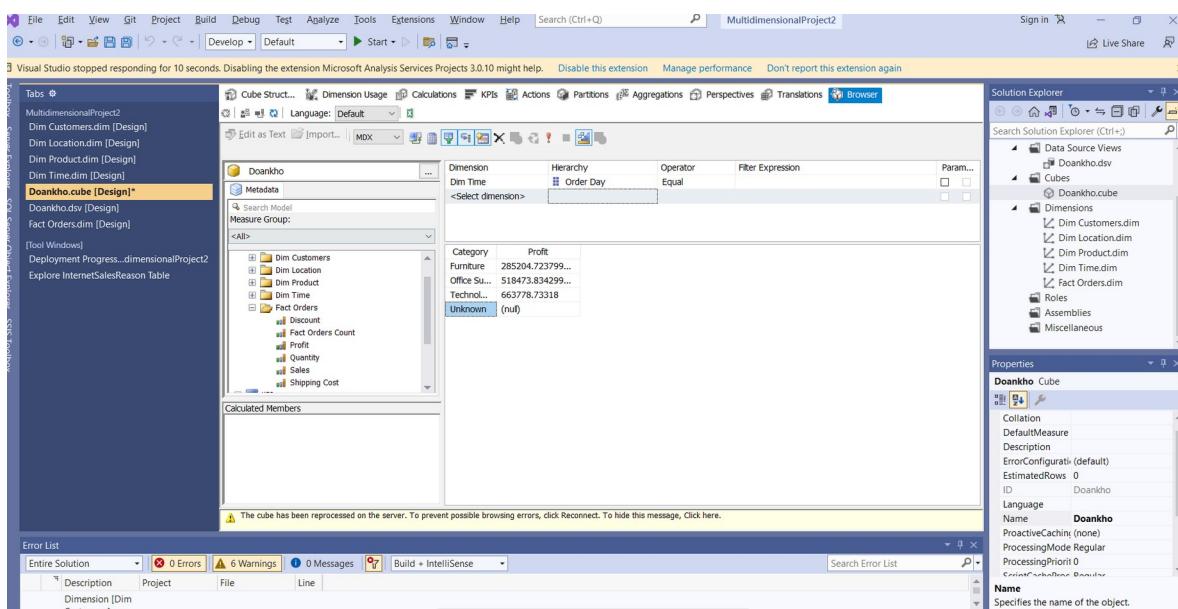
	Row Labels	Fact Orders Count
1	First Class	7505
2	Same Day	2701
3	Second Class	10309
4	Standard Class	30775
5	Grand Total	51290

- Sử dụng Power BI



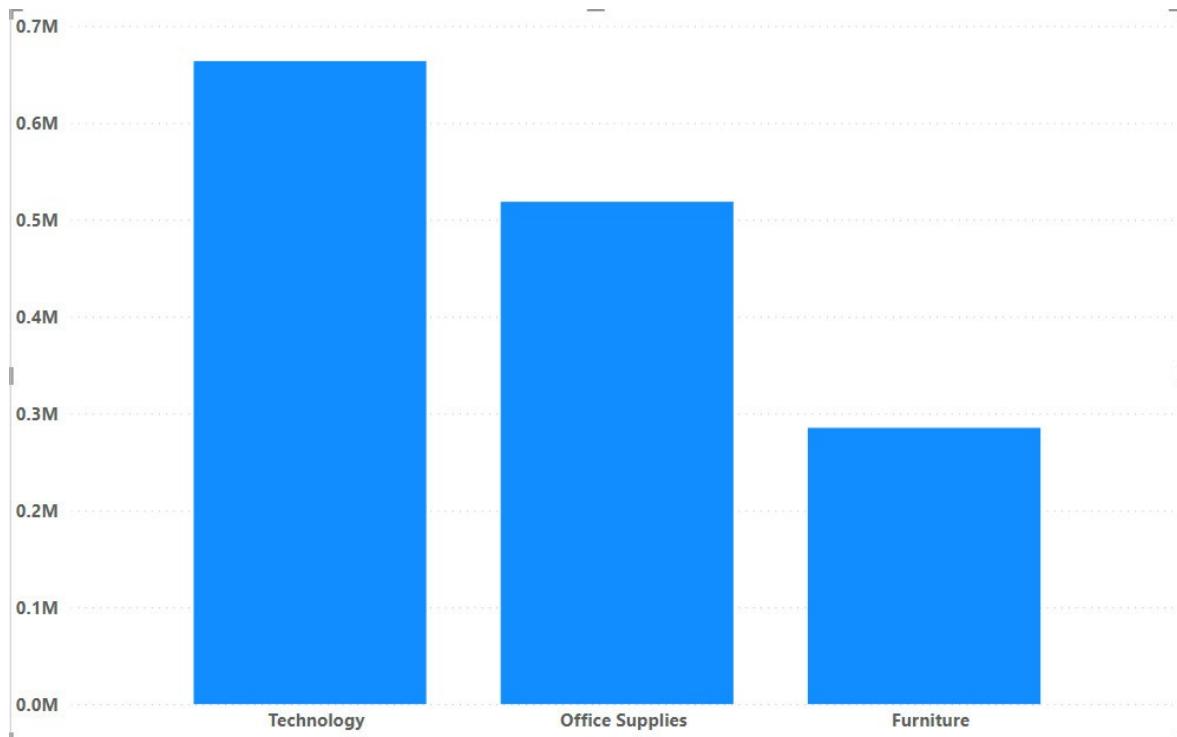
3.3. Thống kê doanh số mang lại từ 3 loại hình sản phẩm

- Sử dụng công cụ SSAS



- Sử dụng Pivot Table trong Excel

- Sử dụng Power BI



3.4. Thống kê chi tiết doanh thu mang lại từ từng loại sản phẩm theo Sub-Category

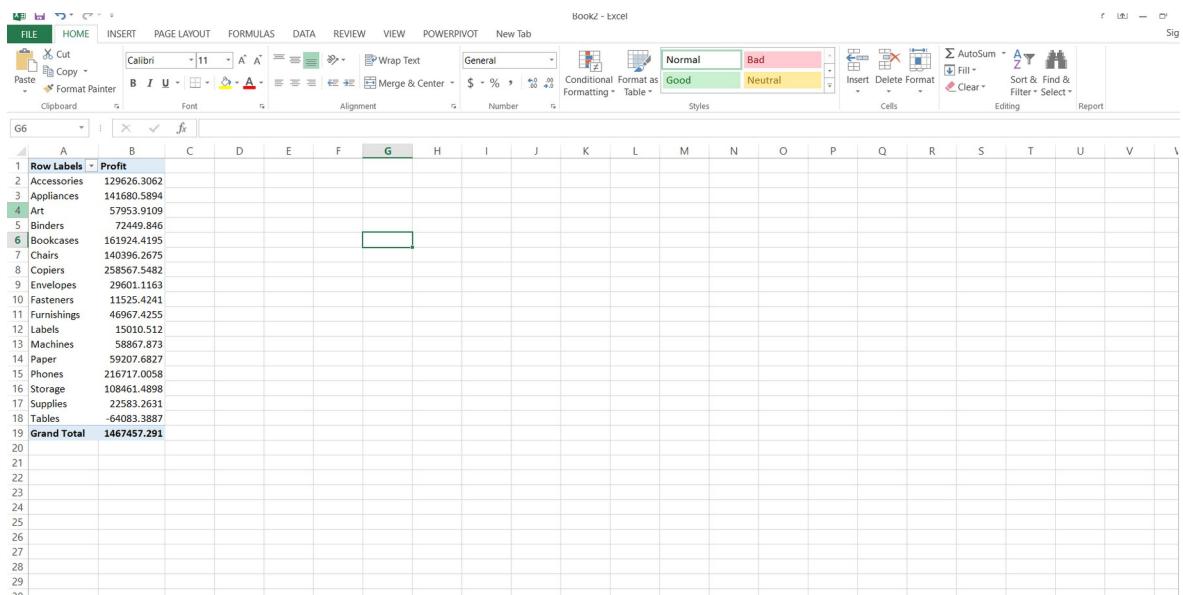
- Sử dụng SSAS

The screenshot shows the Microsoft Analysis Services Projects interface in Visual Studio. The 'Doankho.cube [Design]' is selected in the left pane. The central pane displays a table of profit data by sub-category:

Sub-Category	Profit
Accessories	129626.3062
Appliances	141680.5894
Art	57953.9108999999
Binders	72449.8460000003
Bookcases	161924.4195
Chars	140396.2675
Copiers	258567.54818
Envelopes	29601.1163
Fasteners	11525.4241
Furnishings	46967.4254999999
Labels	15010.512
Machines	58867.873
Paper	59207.6826999999
Phones	216717.1058
Storage	108461.4898
Supplies	22583.2631
Tables	-64083.3886999999

The right pane shows the Solution Explorer with the cube selected. The Properties pane shows the cube's name is 'Doankho'.

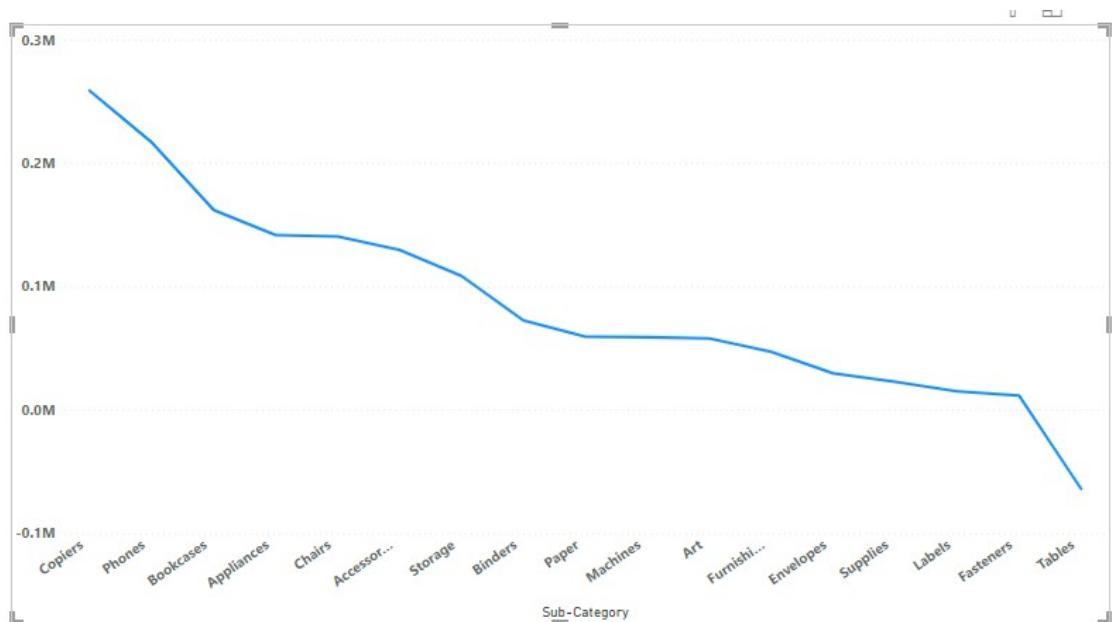
- Sử dụng Pivot Table trong Excel



A screenshot of Microsoft Excel showing a Pivot Table report titled "BookZ - excel". The Pivot Table is displayed in the range A19:G29. The first column contains row labels with category names like "Accessories", "Appliances", "Art", etc., and their corresponding values. The last row shows the "Grand Total" value. The Pivot Table interface includes various ribbon tabs like FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEW, VIEW, and POWERPIVOT. The "HOME" tab is selected, showing tools for Cut, Copy, Paste, Font, Alignment, Number, Styles, Cells, and Editing.

Row Labels	Profit
Accessories	129626.3062
Appliances	141680.5894
Art	57953.9109
Binders	72449.846
Bookcases	161924.4195
Chairs	140396.2675
Copiers	25867.5482
Envelopes	29601.1163
Fasteners	11525.4241
Furnishings	46967.4255
Labels	15010.512
Machines	58867.873
Paper	59207.6827
Phones	216717.0058
Storage	108461.4898
Supplies	22583.2631
Tables	-64083.3887
Grand Total	1467457.291

- Sử dụng Power BI



3.5. Thống kê doanh thu theo khu vực

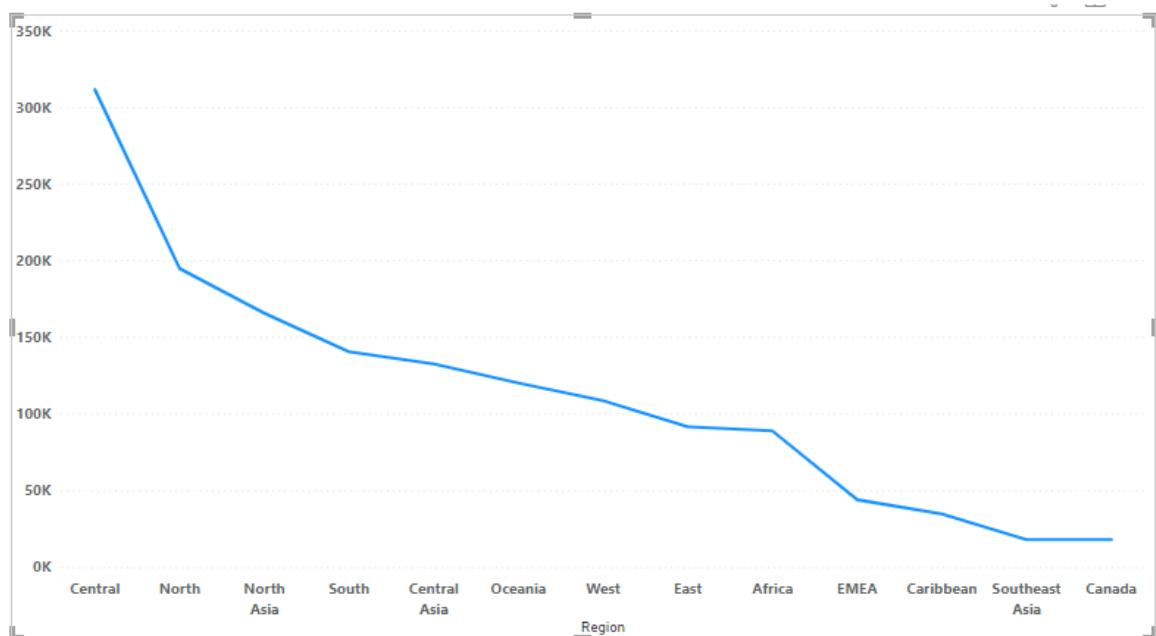
- Sử dụng SSAS

The screenshot shows the Microsoft Analysis Services Project interface in Visual Studio. The 'Doankho.cube [Design]' tab is selected. The main area displays a cube browser with a hierarchy of dimensions and measures. A calculated member named 'West' is highlighted. The properties pane on the right shows the cube's name as 'Doankho'. The error list at the bottom shows 6 warnings.

- Sử dụng Pivot Table trong Excel

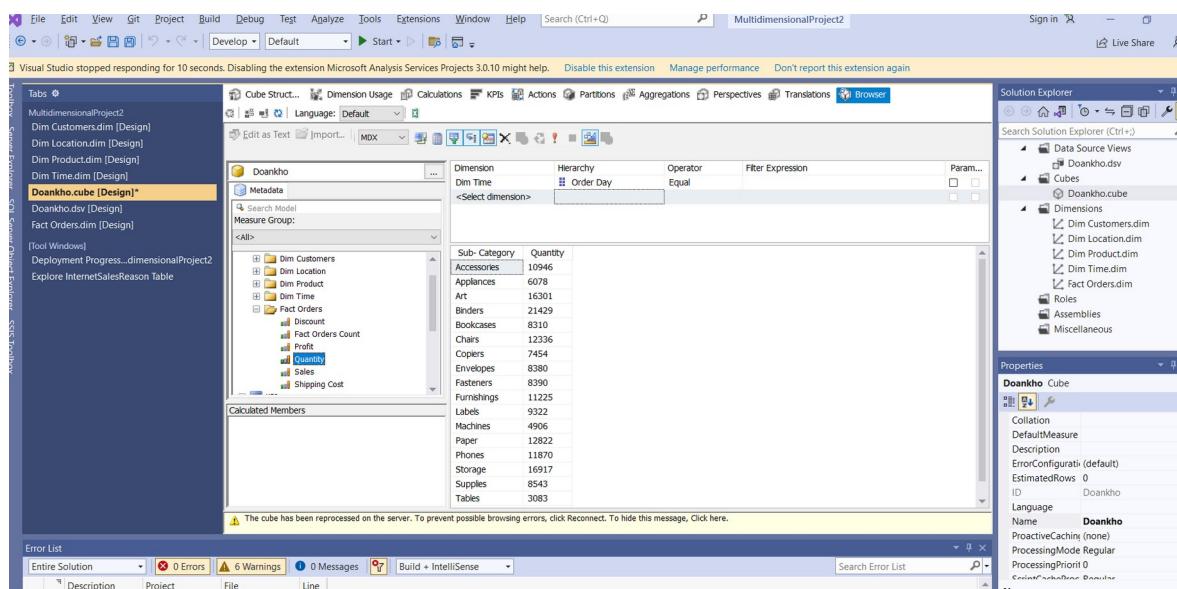
The screenshot shows a Microsoft Excel spreadsheet titled 'Book2 - Excel'. The PivotTable Fields ribbon is open, showing fields from the 'Fact Orders' dimension: Ship Month, Ship Quarter, Ship Year, Profit, Discount, Fact Orders Count, Quantity, Sales, and Shipping Cost. The PivotTable Fields pane also shows the 'Dim Customers' dimension. The main worksheet shows a PivotTable with data grouped by Region and Profit.

- Sử dụng Power BI

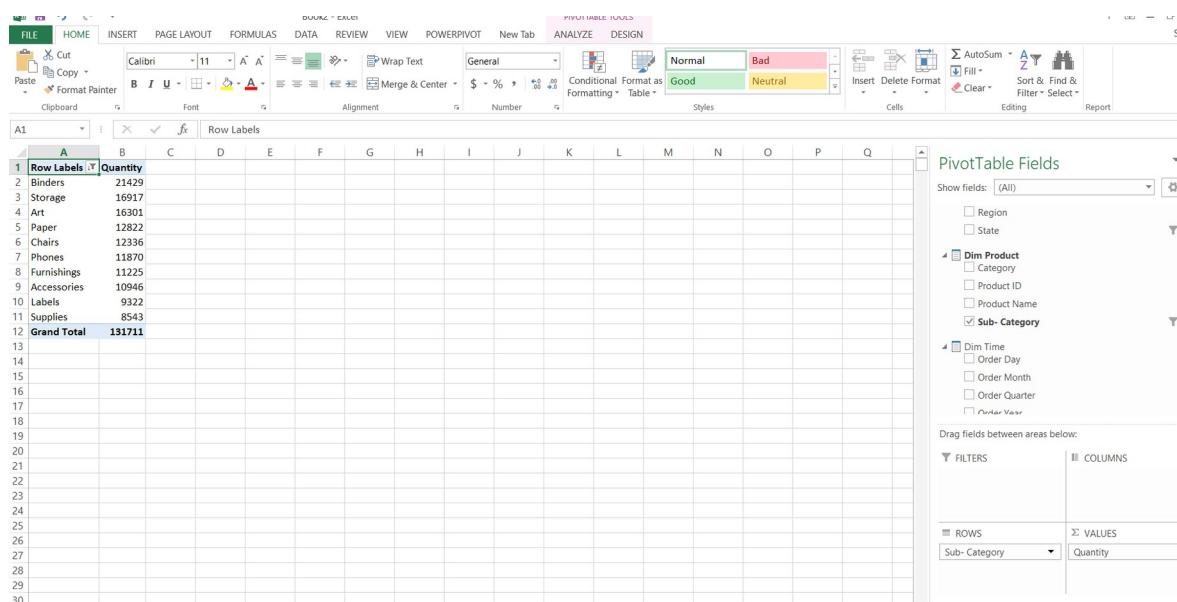


3.6. Liệt kê các sản phẩm bán chạy nhất

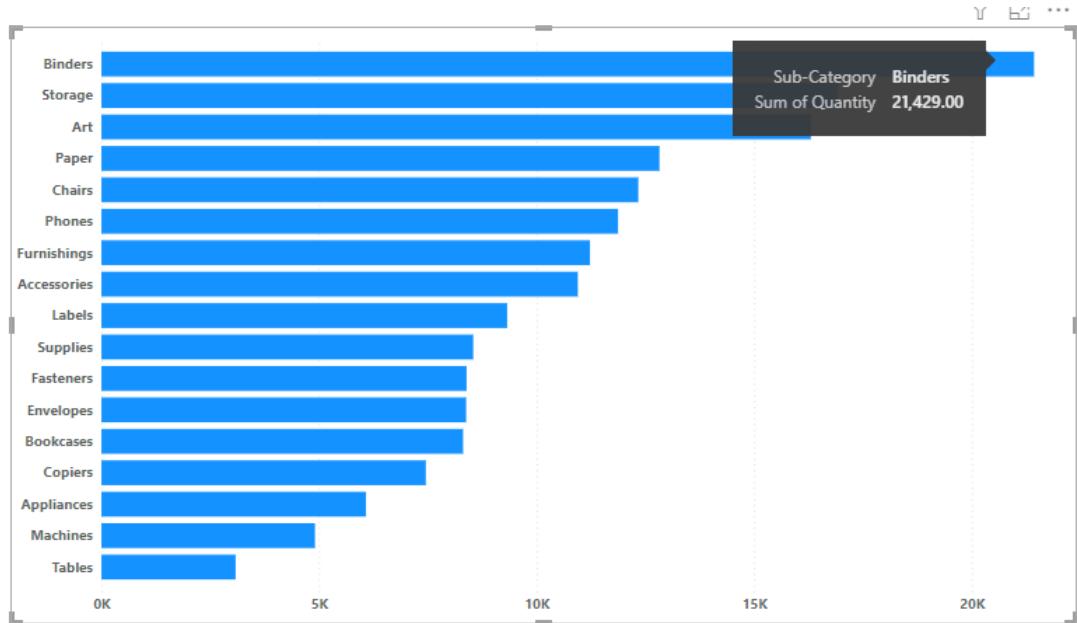
- Sử dụng SSAS



- Sử dụng Pivot Table trong Excel



- Sử dụng Power BI



3.7. Thống kê số sản phẩm bán chạy nhất theo bang

- Sử dụng SSAS

The screenshot shows the Microsoft Visual Studio environment for developing Microsoft Analysis Services (SSAS) cubes. The main window displays a data grid titled "State" with the following data:

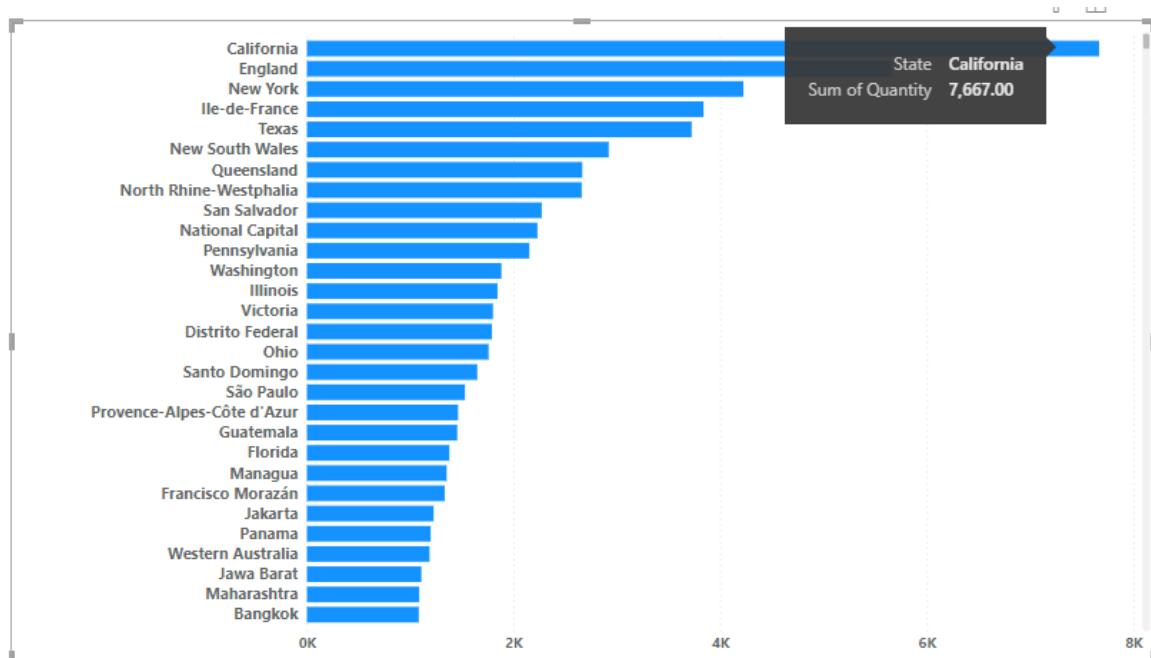
State	Quantity
Abia	52
Abruzzi	75
Abuja Capital Territ...	3
Aceh	24
Acne	18
Ad Dauhah	47
Adamawa	11
Adana	14
Adyaman	128
Afyonkarahisar	14
Aichi	5
'Ajman	44
Aksaray	11
Alava Ibom	4
Al Bahr Al Ahmar	18
Al Bahr Al Ahmar	53

The Solution Explorer on the right shows the project structure, including Data Source Views, Cubes (Doankho.cube), Dimensions, and other components. The Properties window on the far right shows details for the selected "Doankho" cube.

- Sử dụng Pivot Table trong Excel

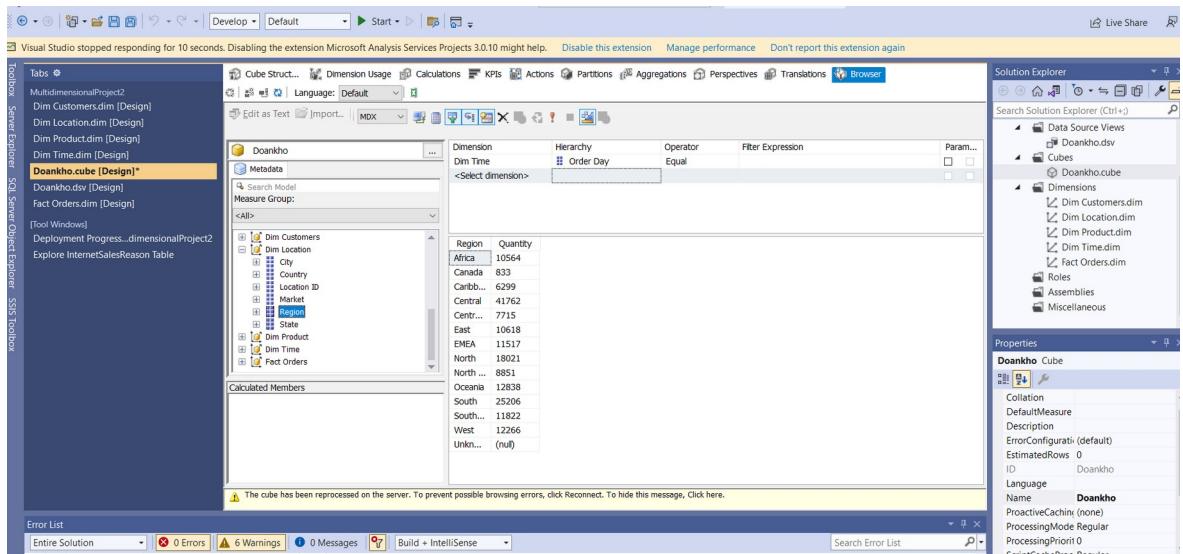
	Quantity
1 Row Labels	
2 California	7667
3 England	5656
4 New York	4224
5 Ile-de-France	3839
6 Texas	3724
7 New South Wales	2921
8 Queensland	2665
9 North Rhine-Westphalia	2660
10 San Salvador	2273
11 National Capital	2231
12 Grand Total	37860

- Sử dụng Power BI



3.8. Liệt kê những khu vực có số lượng mua hàng nhiều nhất

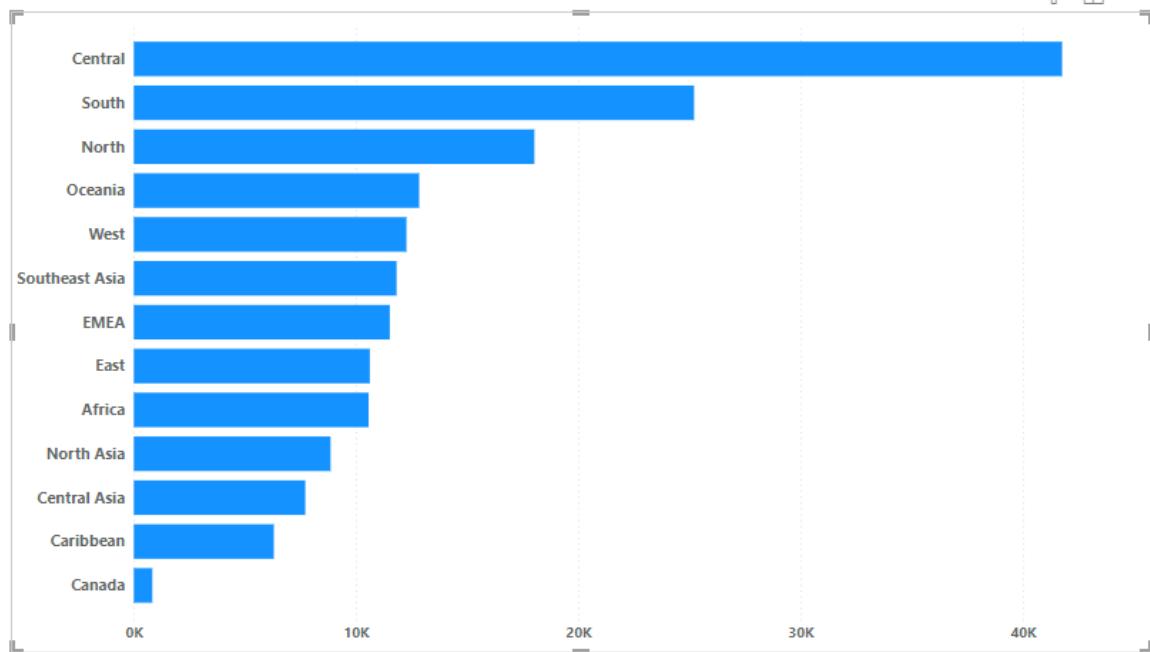
- Sử dụng SSAS



- Sử dụng Pivot Table trong Excel

Row Labels		Quantity
Central		41762
South		25206
North		18021
Oceania		12838
West		12266
Grand Total		110093

- Sử dụng Power BI



3.9. So sánh tổng doanh thu với chi phí ship

- Sử dụng SSAS

The screenshot shows the Microsoft Visual Studio IDE with the Analysis Services Designer open. The main area displays a cube named "Doankho" with a table viewer showing data for Order Year, Profit, and Shipping Cost. The Properties pane on the right shows the cube's name is "Doankho".

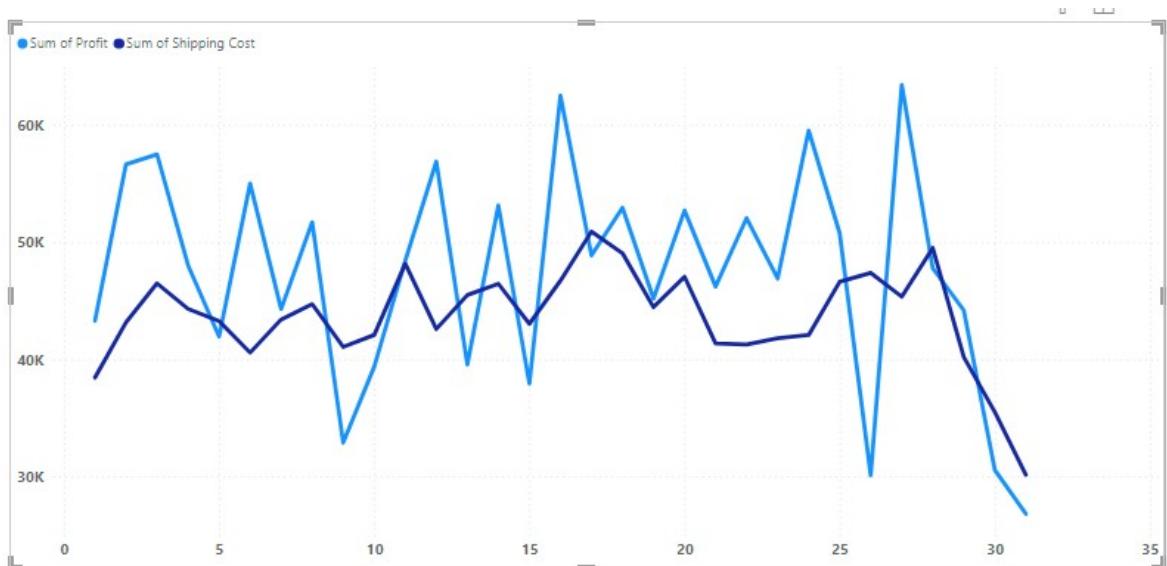
Order Year	Profit	Shipping Cost
2011	248940.811539999	244270.345...
2012	307415.279099999	283490.824...
2013	406935.230180003	364548.744...
2014	504165.970459996	460505.789...
Unknown	(null)	(null)

- Sử dụng Pivot Table trong Excel

The screenshot shows a Microsoft Excel spreadsheet titled "Book2 - Excel". The ribbon menu is visible at the top. A Pivot Table is displayed in the main area, showing data for "Row Labels" (Year) across columns "Profit" and "Shipping Cost". The data rows are as follows:

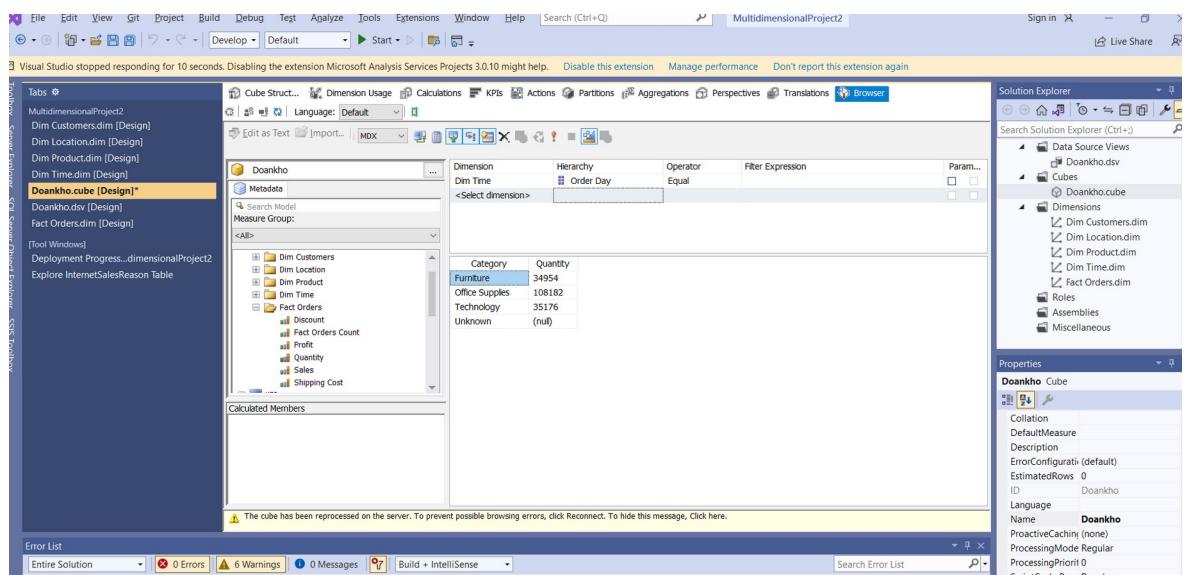
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
Row Labels		Profit	Shipping Cost																			
2011		248940.8115	244270.3455																			
2012		307415.2791	283490.824																			
2013		406935.2302	364548.7444																			
2014		504165.9705	460505.7895																			
Grand Total		1467457.291	1352815.703																			

- Sử dụng Power BI



3.10. So sánh tổng số lượng bán hàng giữa các danh mục

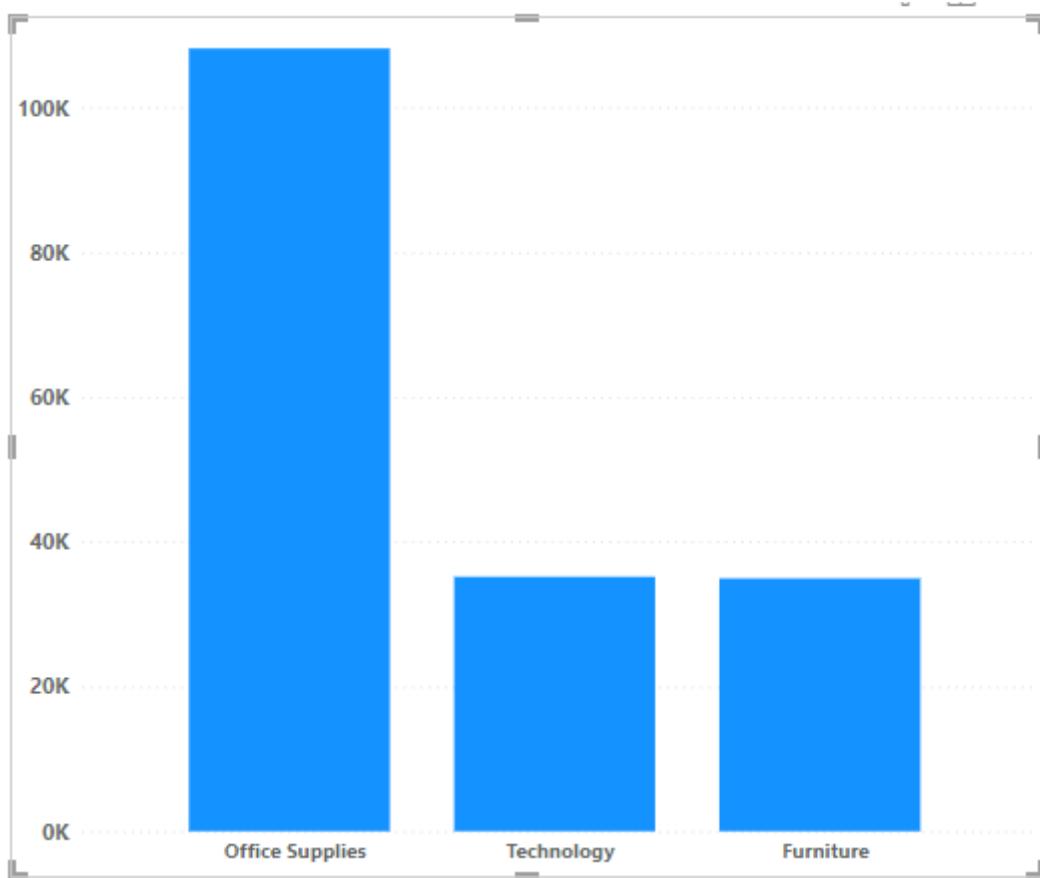
- Sử dụng SSAS



- Sử dụng Pivot Table trong Excel

Row Labels	Quantity
Furniture	34954
Office Supplies	108182
Technology	35176
Grand Total	178312

- Sử dụng Power BI



3.11. Thông kê những khách hàng mua nhiều sản phẩm nhất

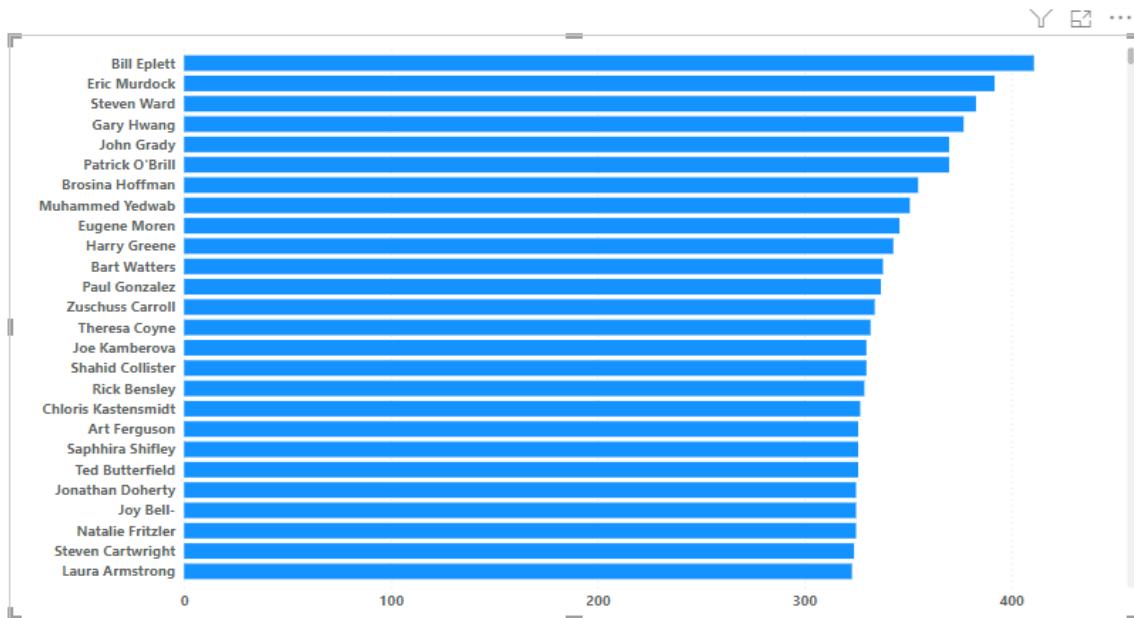
- Sử dụng SSAS

The screenshot displays the Microsoft SQL Server Data Tools (SSDT) environment for building a multidimensional data warehouse. The main workspace shows the 'Doankho' cube structure, with the 'Customer' dimension expanded to reveal its hierarchy and associated fact tables. The properties of the 'Doankho' cube are set to 'Doankho' with an aggregation prefix of 'Doankho'. The solution explorer on the right lists the project components, including the cube and dimension files. The error list at the bottom indicates a clean build.

- Sử dụng Pivot Table trong Excel

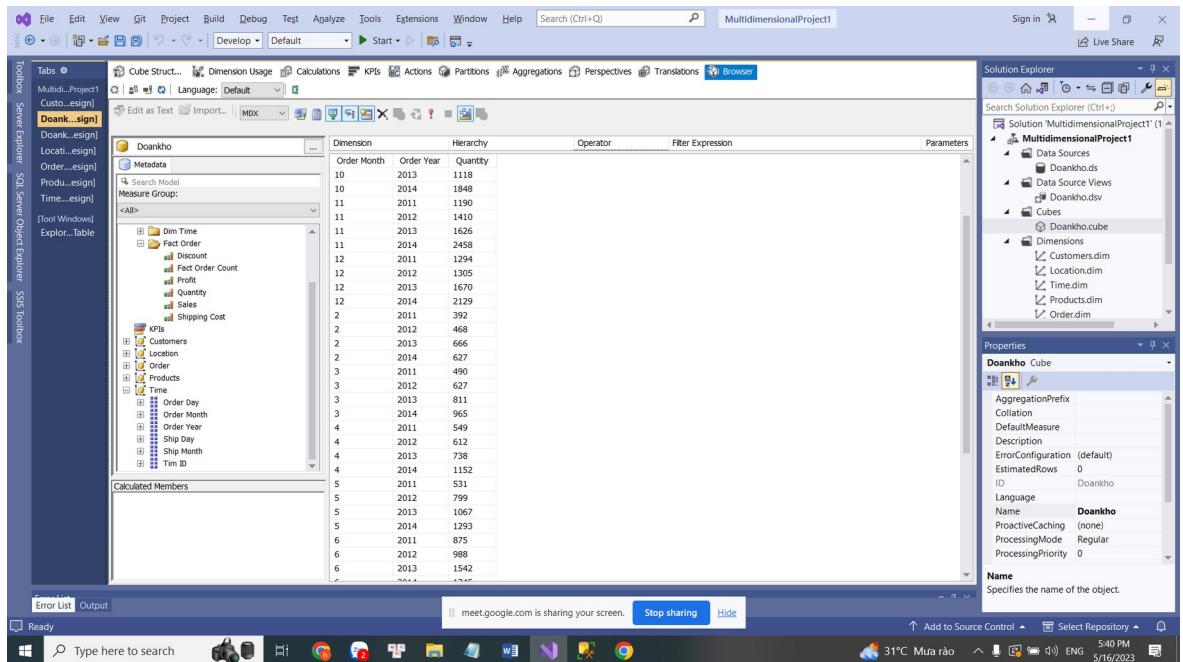
A	B	C	D	E	F	G	H	I	J	K	L	M
Row Labels	Office Supplies	Technology	Furniture	Grand Total								
5 Bill Eplett	256	77	78	411								
6 Eric Murdock	232	82	78	392								
7 Steven Ward	223	72	88	383								
8 Gary Hwang	256	65	56	377								
9 Patrick O'Brill	221	86	63	370								
10 John Grady	232	73	65	370								
11 Brosina Hoffman	202	102	51	355								
12 Muhammed Yedwab	175	69	107	351								
13 Eugene Moren	204	66	76	346								
14 Harry Greene	212	58	73	343								
15 Bart Watters	196	63	79	338								
16 Paul Gonzalez	225	68	44	337								
17 Zuschuss Carroll	202	74	58	334								
18 Theresa Coyne	191	75	66	332								
19 Shahid Collister	217	49	64	330								
20 Joe Kamberova	195	54	81	330								
21 Rick Bensley	170	80	79	329								
22 Chloris Kastensmidt	204	62	61	327								
23 Ted Butterfield	244	35	47	326								
24 Saphhira Shifley	170	82	74	326								
25 Art Ferguson	173	54	99	326								
26 Natalie Fritzler	182	75	68	325								
27 Jonathan Doherty	195	74	56	325								
28 Joy Bell-	218	69	38	325								
29 Steven Cartwright	220	45	59	324								
30 Laura Armstrong	200	66	57	323								
31 Phillip Breyer	189	92	38	319								
32 Joe Elijah	204	50	65	319								
33 Noel Staavos	191	56	71	318								
34 Irene Maddox	176	72	70	318								
35 Seth Vernon	156	79	82	317								
36 Joseph Holt	206	61	50	317								
37 Laurel Beltran	174	68	74	316								
38 Ben Ferrer	185	75	56	316								
39 Tracy Blumstein	185	84	45	314								
40 Michael Raino	191	59	61	314								

- Sử dụng Power BI



3.12. Tìm xu hướng số lượng đơn hàng trong năm

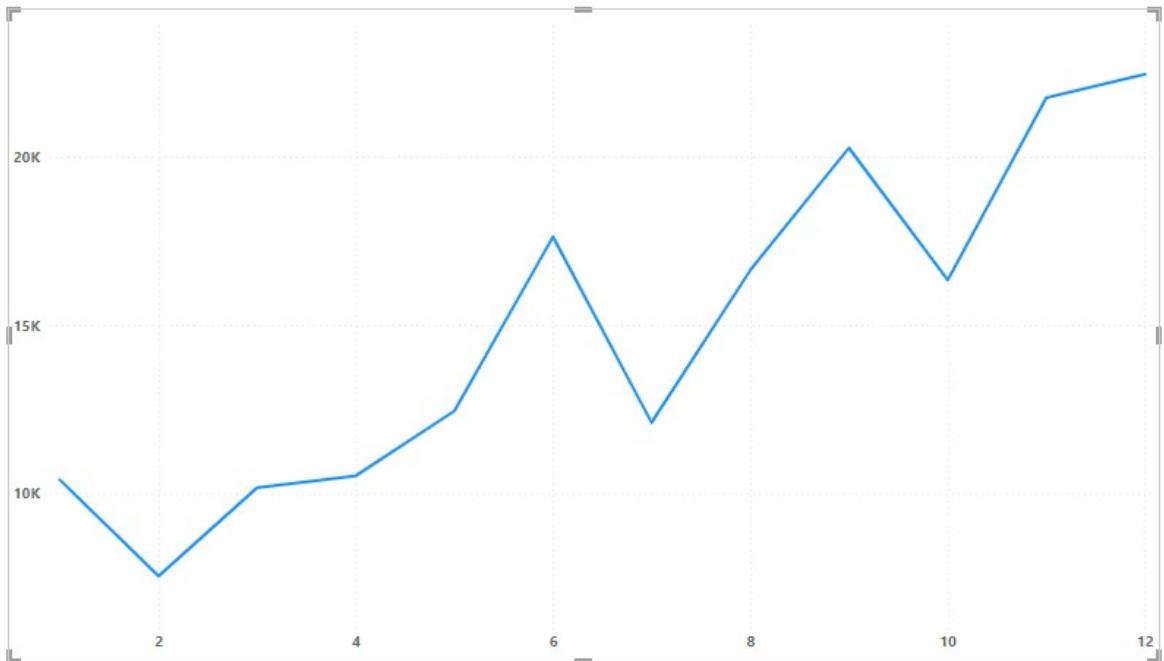
- Sử dụng SSAS



- Sử dụng Pivot Table trong Excel

	2011	2012	2013	2014	Grand Total
1	402	16	713	867	2498
2	802	995	1118	1848	4763
3	1190	1410	1626	2458	6684
4	1294	1305	1670	2129	6398
5	392	468	666	627	2153
6	490	627	811	965	2893
7	549	612	738	1152	3051
8	531	799	1067	1293	3690
9	875	988	1542	1745	5150
10	493	709	934	1013	3149
11	906	1157	1393	1664	5120
12	1000	1126	1571	1839	5536
13	Grand Total	8924	10712	13849	17600
14					\$1085

- Sử dụng Power BI



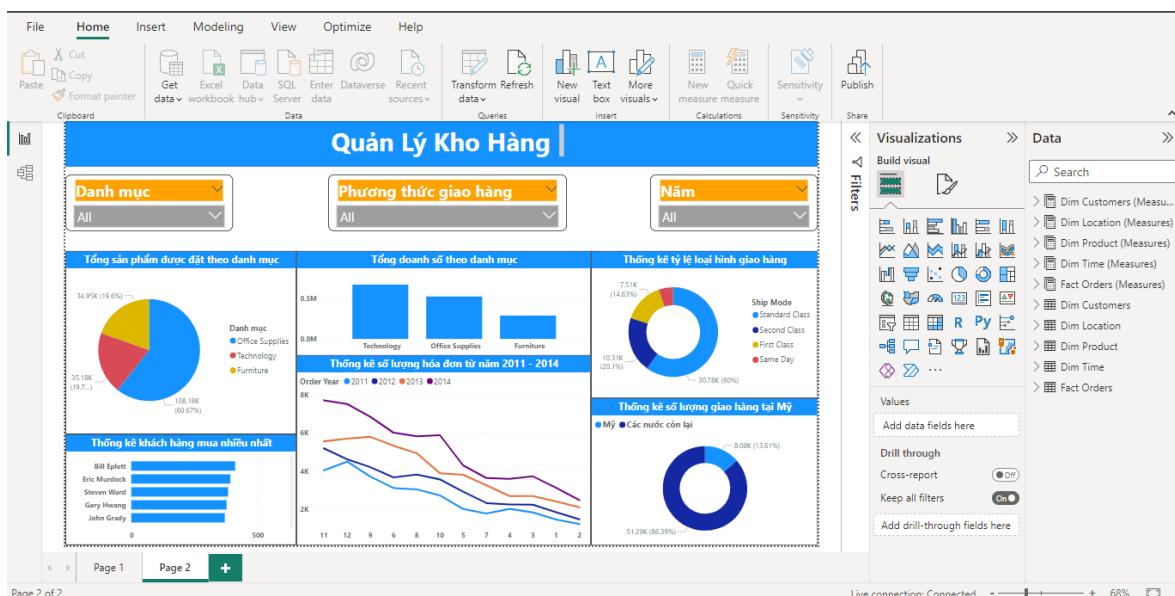
4. Dashboard đã thực hiện

4.1. Dashboard thứ nhất

- Phương thức giao hàng: Tìm kiếm theo phương thức giao hàng
- Bang: Tìm kiếm theo bang

- Danh mục: Tìm kiếm theo danh mục
- Quý: Tìm kiếm theo quý
- 2 thẻ thể hiện doanh thu và lợi nhuận của tất cả sản phẩm
- Biểu đồ đường thể hiện chi phí giao hàng và lợi nhuận
- Biểu đồ thể giới thống kê lợi nhuận theo bang
- Biểu đồ tròn thể hiện tỉ lệ lợi nhuận theo khu vực và tỉ lệ doanh thu theo khu vực
- Biểu đồ cột ngang thống kê top những sản phẩm bán chạy nhất

4.2. Dashboard thứ 2



- Danh mục: Tìm kiếm theo danh mục
- Phương thức giao hàng: Tìm kiếm theo phương thức giao hàng
- Năm: Tìm kiếm theo năm (2011 – 2014)
- Biểu đồ tròn thể hiện tổng sản phẩm được đặt theo danh mục
- Biểu đồ cột ngang thể hiện top khách hàng mua nhiều nhất
- Biểu đồ đường thể hiện xu hướng mua hàng từ năm 2011 – 2014
- Biểu đồ cột đứng thể hiện tổng doanh thu theo danh mục
- Biểu đồ tròn thể hiện loại hình giao hàng được chọn và số lượng đơn hàng tại Mỹ.

BẢNG THÔNG TIN CHI TIẾT NHÓM 11

STT	HỌ VÀ TÊN	MSSV
1	Nguyễn Hiếu Gia Cường	20133027
2	Huỳnh Công Hậu	20133039

BẢNG PHÂN CÔNG CÔNG VIỆC

Công việc	Người thực hiện	
	Nguyễn Hiếu Gia Cường	Huỳnh Công Hậu
Chọn dataset	50%	50%
Nghiên cứu tính cần thiết để xây dựng kho dữ liệu	50%	50%
Khảo sát nghiên cứu, phân tích báo cáo nghiệp vụ	60%	40%
Xây dựng lượt đồ hình sao	30%	70%
Tích hợp dữ liệu vào kho (SSIS)	30%	70%
Quá trình SSAS – Xây dựng Cube, Phân chiều	50%	50%
Thực hiện 15 câu truy vấn bằng SSAS, Pivot table	50%	50%
Trực quan với Power BI	70%	30%
Viết báo cáo	90%	10%

TÀI LIỆU THAM KHẢO

1. Microsoft, 22/04/2023, SSIS Projects extension for VS2022 troubleshooting guide, <https://learn.microsoft.com/en-us/sql/ssdt/release-notes-integration-services-2022?view=sql-server-ver16>, truy cập ngày: 14/05/2023.
2. Leslie Richardson, 12/07/2022, Analysis Services and Reporting Services extensions for Visual Studio 2022 are here, <https://devblogs.microsoft.com/visualstudio/analysis-services-and-reporting-services-extensions-for-visual-studio-2022-are-here/>, truy cập ngày: 12/05/2023
3. Microsoft, 14/03/2023, Tạo báo cáo Power BI để phân tích dự án, <https://learn.microsoft.com/vi-vn/power-apps/maker/canvas-apps/sharepoint-scenario-build-report>, truy cập ngày 12/05/2023
4. File bài giảng của thầy Nguyễn Văn Thành.