

ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
BÁO CÁO ĐỒ ÁN
THỊ GIÁC MÁY TÍNH NÂNG CAO
CS331.O11.KHCL



ĐỀ TÀI

BÁO CÁO ĐỒ ÁN MÔN HỌC KHO DỮ LIỆU VÀ OLAP

Giảng viên hướng dẫn: **Nguyễn Thị Kim Phụng**

Sinh viên thực hiện:

Lê Hoàng Long - 20521563

Phạm Văn Nghĩa – 20521656

Nguyễn Tú Luân - 20521583

TP.Hồ Chí Minh, ngày 11, tháng 1, năm 2024

MỤC LỤC

MỤC LỤC	2
LỜI CẢM ƠN	5
NHẬN XÉT CỦA GIÁNG VIÊN	6
CHƯƠNG 1. GIỚI THIỆU KHO DỮ LIỆU	7
1.1. Mô tả dữ liệu	7
1.1.1. Nguồn dữ liệu	7
1.1.2. Mô tả thuộc tính	7
1.1.3. Mô tả chi tiết thuộc tính	8
1.2. Xây dựng kho dữ liệu	9
1.2.1. Snowflake Schema	9
1.2.2. Bảng Fact	9
1.2.3. Dim_Area	10
1.2.4. Dim_SaleTime	10
1.2.5. Dim_BuiltTime	11
1.2.6. Dim_Property	11
1.2.7. Dim_Type	11
1.2.8. Dim_Utility	11
1.2.9. Dim_Street	11
1.2.10. Dim_QualityScore	11
1.2.11. Dim_Condition	12
CHƯƠNG 2: XÂY DỰNG KHO DỮ LIỆU - QUÁ TRÌNH SSIS	13
2.1. Tạo mới hai database	13
2.1.1. Database chứa dữ liệu gốc	13
2.1.2. Data warehouse	15
2.2. Tạo project mới trong công cụ SQL Data Tool	15
2.3. Mô hình SSIS	17
2.4. Load Dimension Tables	17
2.4.1. Load Dim_Area	17

2.4.2. Load Dim_SaleTime	22
2.4.3. Load Dim_BuiltTime	27
2.4.4. Load Dim_Property	32
2.4.5. Load Dim_Type	36
2.4.6. Load Dim_Utility	40
2.4.7. Load Dim_Street	44
2.4.8. Load Dim_QualityScore	48
2.4.9. Load Dim_Condition	52
2.5. Load Fact Table	56
2.6. Viết Execute SQL Task:	85
CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN - QUÁ TRÌNH SSAS	88
3.1. Cấu hình Project	88
3.1.1. Connect đến Data Sources	89
3.1.2. Tạo datasource Views	91
3.1.3. Tạo Cubes	92
3.2. Thực thi truy vấn trên Visual Studio và Power BI và Excel	94
3.2.1. Cuộn lên (Roll up)	94
3.2.2. Truy xuống (Drill down)	95
3.2.3. Chọn và chiểu (Slice and Dice)	98
3.2.4. Xoay chiều (Pivot)	108
3.3. Ngôn ngữ MDX	113
3.3.1. Câu 1:	113
3.3.2. Câu 2:	114
3.3.3. Câu 3:	115
3.3.4. Câu 4:	116
3.3.5. Câu 5:	117
3.3.6. Câu 6:	117
3.3.7. Câu 7:	118

3.3.8. Câu 8:	119
3.3.9. Câu 9:	120
3.3.10. Câu 10:	120
CHƯƠNG 4: KHAI PHÁ DỮ LIỆU - DATA MINING	122
4.1. Nhập thư viện	122
4.2. Tiền xử lý dữ liệu	122
4.3. Huấn luyện mô hình	126
4.4. Xuất ra tập luật	130
4.5. Đánh giá mô hình	131

LỜI CẢM ƠN

Lời đầu tiên chúng em xin chân thành cảm ơn quý Thầy Cô trường Đại học Công nghệ Thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh, đặc biệt là cô Nguyễn Thị Kim Phụng – giảng viên môn Kho dữ liệu và OLAP, đã hỗ trợ, hướng dẫn và giải đáp những thắc mắc của em trong suốt quá trình thực hiện để em có thể hoàn thành tốt đồ án của mình. Trong suốt gần một học kỳ thực hiện đồ án, em đã vận dụng những kiến thức được giảng dạy đồng thời tìm hiểu thêm từ những nguồn thông tin bên ngoài để có thể phát triển hơn đồ án của bản thân. Dù vậy, em vẫn không thể tránh khỏi những sai sót xảy ra trong quá trình tìm hiểu và thực hiện đồ án. Do đó, em mong nhận được sự góp ý từ cô để hiểu rõ hơn về kiến thức liên quan đến môn học, cũng như có được kinh nghiệm để tiếp tục thực hiện các đề tài, dự án tiếp theo trong thời gian tới. Em xin chân thành cảm ơn!

Thành phố Hồ Chí Minh, tháng 4 năm 2023 Nhóm
thực hiện

NHẬN XÉT CỦA GIẢNG VIÊN

....., ngày.....tháng.....năm 2023

Người nhận xét

(Ký tên và ghi rõ họ tên)

CHƯƠNG 1. GIỚI THIỆU KHO DỮ LIỆU

1.1. Mô tả dữ liệu

1.1.1. Nguồn dữ liệu

- Bộ dữ liệu Chennai Housing Sales Price là bộ dữ liệu về giá bất động sản tại Chennai, Ấn Độ. Bộ dữ liệu này chứa thông tin về giá bán bất động sản và các thuộc tính của bất động sản như: Khu vực, loại bất động sản, tình trạng của bất động sản, các tiện ích công cộng và giá bán của bất động sản (Từ năm 2004 đến năm 2015).
- Dataset bao gồm 7109 dòng với 22 thuộc tính mô tả.
- Link: <https://www.kaggle.com/datasets/kunwarakash/chennai-housing-sales-price>

1.1.2. Mô tả thuộc tính

STT	Tên thuộc tính	Kiểu dữ liệu	Mô tả
1	PRT_ID	int	Mã bất động sản
2	AREA	nvarchar(50)	Khu vực
3	INT_SQFT	int	Diện tích bất động sản (Đơn vị m ²)
4	DATE_SALE	datetime	Ngày bán bất động sản
5	DIST_MAINROAD	int	Khoảng cách từ bất động sản đến đường chính (Đơn vị m)
6	N_BEDROOM	int	Số lượng phòng ngủ
7	N_BATHROOM	int	Số lượng phòng tắm
8	N_ROOM	int	Tổng số phòng
9	SALE_COND	nvarchar(50)	Tình trạng của bất động sản khi bán
10	PARK_FACIL	bit	Có bãi đỗ xe hay không
11	DATE_BUILD	datetime	Ngày bất động sản được xây dựng
12	BUILDTYPE	nvarchar(50)	Loại bất động sản
13	UTILITY_AVAIL	nvarchar(50)	Tiện ích của bất động sản

14	STREET	nvarchar(50)	Tên đường
15	MZZONE	nvarchar(50)	Mã định danh khu vực
16	QS_ROOMS	decimal(2, 1)	Chất lượng tất cả các phòng
17	QS_BATHROOM	decimal(2, 1)	Chất lượng phòng tắm
18	QS_BEDROOM	decimal(2, 1)	Chất lượng phòng ngủ
19	QS_OVERALL	decimal(2, 1)	Tổng thể chất lượng bất động sản
20	REG_FEE	int	Phí đăng ký
21	COMMIS	int	Phí cho người môi giới
22	SALES_PRICE	int	Giá bán bất động sản

1.1.3. Mô tả chi tiết thuộc tính

Dictionary thuộc tính SALE_COND

Thuộc tính SALE_COND thể hiện tình trạng của bất động sản khi bán

SALE_COND	
Tên	Mô tả
AbNormal	Bất động sản có các vấn đề như: tình trạng pháp lý, tình trạng hư hỏng, thiếu cơ sở hạ tầng,...
Family	Bất động sản được xây dựng phù hợp với nhu cầu của một gia đình, bao gồm các tiện ích như: phòng ngủ, phòng khách, phòng bếp, phòng tắm,...
Partial	Bất động sản được chia thành các căn hộ riêng lẻ và bán theo từng căn hộ,...
AdjLand	Bất động sản là đất trống hoặc đất có một số cơ sở hạ tầng nhưng chưa được xây dựng.
Normal sale	Bất động sản thông thường.

Dictionary thuộc tính UTILITY_AVAIL

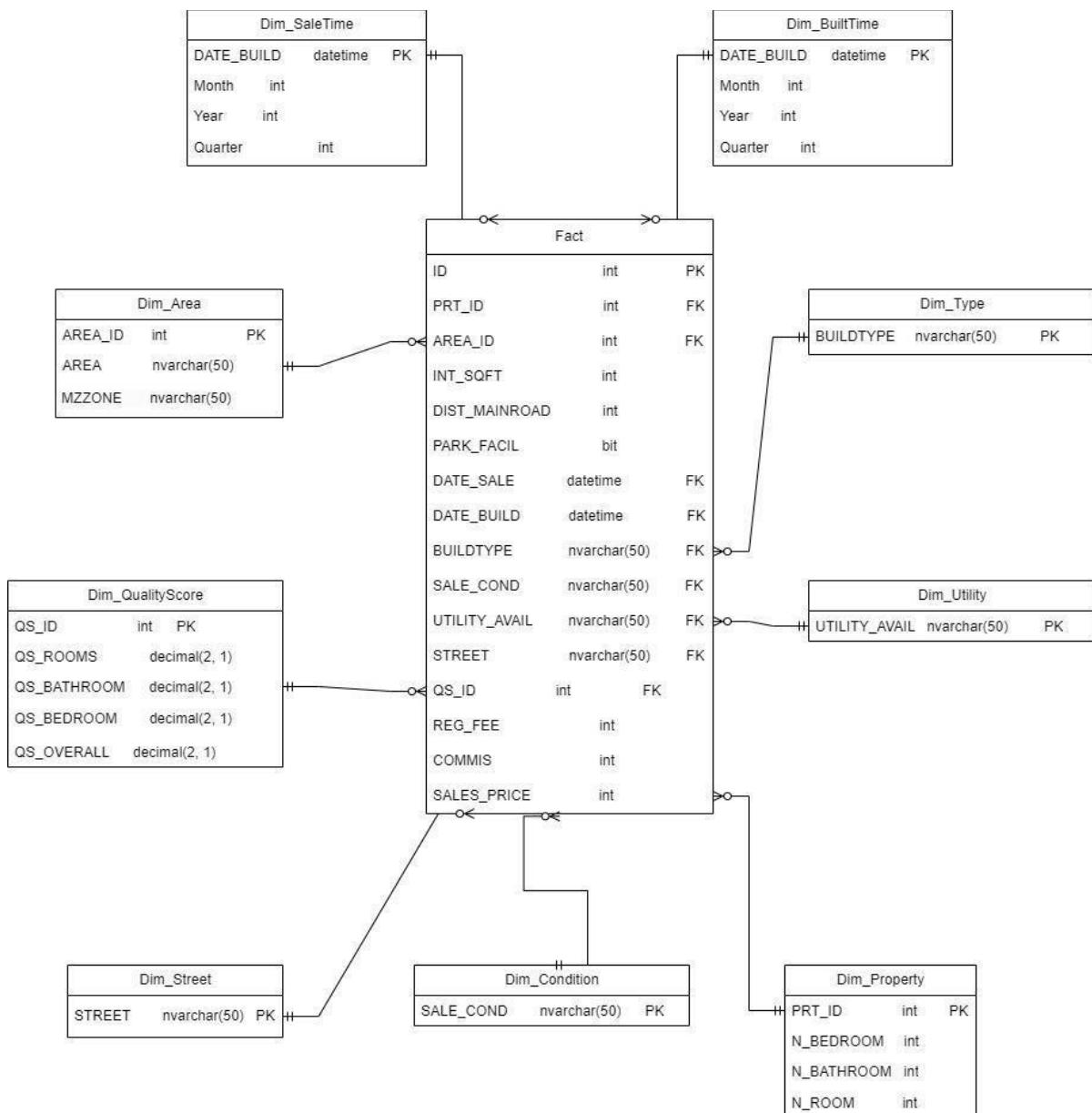
Thuộc tính UTILITY_AVAIL thể hiện tiện ích của bất động sản

UTILITY_AVAIL

Tên	Mô tả
AllPub	Bất động sản đầy đủ tất cả các tiện ích công cộng.
ELO	Bất động sản chỉ được cung cấp điện áp thấp.
No Sewer	Bất động sản không có hệ thống thoát nước chính trong khu vực.
NoSeWa	Bất động sản không có hệ thống cung cấp nước chính trong khu vực.

1.2. Xây dựng kho dữ liệu

1.2.1. Snowflake Schema



1.2.2. Bảng Fact

Tên thuộc tính	Kiểu dữ liệu	Mô tả
ID	int	Khóa chính
PRT_ID	int	Mã bất động sản
AREA_ID	int	Mã khu vực
INT_SQFT	int	Diện tích bất động sản (Đơn vị m ²)
DIST_MAINROAD	int	Khoảng cách từ bất động sản đến đường chính (Đơn vị m)
PARK_FACIL	int	Có bãi đỗ xe hay không
DATE_SALE	datetime	Ngày bán bất động sản
DATE_BUILD	datetime	Ngày bất động sản được xây dựng
BUILDTYPE	nvarchar(50)	Loại bất động sản
SALE_COND	nvarchar(50)	Tình trạng của bất động sản khi bán
UTILITY_AVAIL	nvarchar(50)	Tiện ích của bất động sản
STREET	nvarchar(50)	Tên đường
QualityScore_ID	int	Mã đánh giá chất lượng
REG_FEE	int	Phí đăng ký
COMMIS	int	Phí cho người môi giới
SALES_PRICE	int	Giá bán bất động sản

1.2.3. Dim_Area

Tên thuộc tính	Kiểu dữ liệu	Mô tả
AREA_ID	int	Mã khu vực (Khóa chính)
AREA	nvarchar(50)	Khu vực
MZZONE	nvarchar(50)	Mã định danh khu vực

1.2.4. Dim_SaleTime

Tên thuộc tính	Kiểu dữ liệu	Mô tả
DATE_SALE	datetime	Ngày bán bất động sản (Khóa chính)
Month	int	Tháng
Year	int	Năm
Quarter	int	Quý

1.2.5. Dim_BuiltTime

Tên thuộc tính	Kiểu dữ liệu	Mô tả
DATE_BUILD	datetime	Ngày bất động sản được xây dựng (Khóa chính)
Month	int	Tháng
Year	int	Năm
Quarter	int	Quý

1.2.6. Dim_Property

Tên thuộc tính	Kiểu dữ liệu	Mô tả
PRT_ID	int	Mã bất động sản (Khóa chính)
N_BEDROOM	int	Số lượng phòng ngủ
N_BATHROOM	int	Số lượng phòng tắm
N_ROOM	int	Tổng số phòng

1.2.7. Dim_Type

Tên thuộc tính	Kiểu dữ liệu	Mô tả
BUILDTYPE	nvarchar(50)	Loại bất động sản (Khóa chính)

1.2.8. Dim_Utility

Tên thuộc tính	Kiểu dữ liệu	Mô tả
UTILITY_AVAIL	nvarchar(50)	Tiện ích của bất động sản (Khóa chính)

1.2.9. Dim_Street

Tên thuộc tính	Kiểu dữ liệu	Mô tả
STREET	nvarchar(50)	Tên đường (Khóa chính)

1.2.10. Dim_QualityScore

Tên thuộc tính	Kiểu dữ liệu	Mô tả
QS_ID	int	Mã đánh giá chất lượng (Khóa chính)
QS_ROOMS	decimal(2, 1)	Chất lượng tất cả các phòng
QS_BATHROOM	decimal(2, 1)	Chất lượng phòng tắm
QS_BEDROOM	decimal(2, 1)	Chất lượng phòng ngủ
QS_OVERALL	decimal(2, 1)	Tổng thể chất lượng bất động sản

1.2.11. Dim_Condition

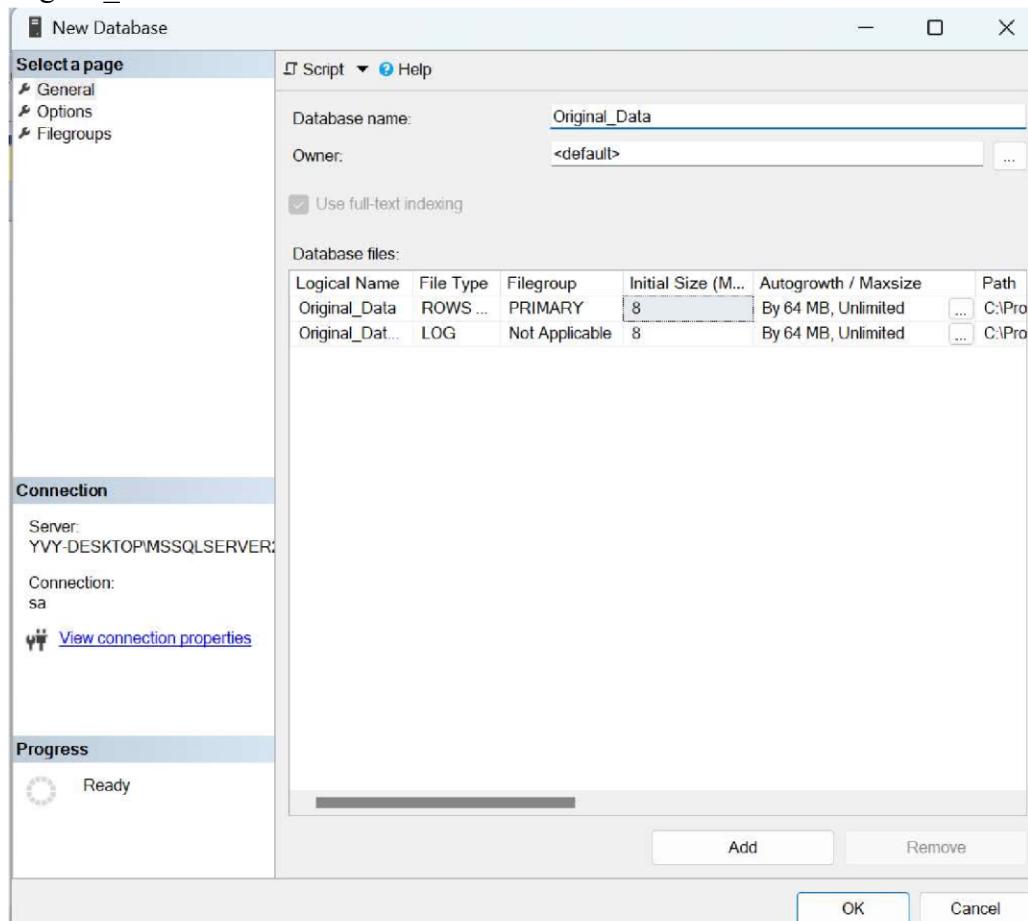
Tên thuộc tính	Kiểu dữ liệu	Mô tả
SALE_COND	nvarchar(50)	Tình trạng của bất động sản khi bán (Khóa chính)

CHƯƠNG 2: XÂY DỰNG KHO DỮ LIỆU - QUÁ TRÌNH SSIS

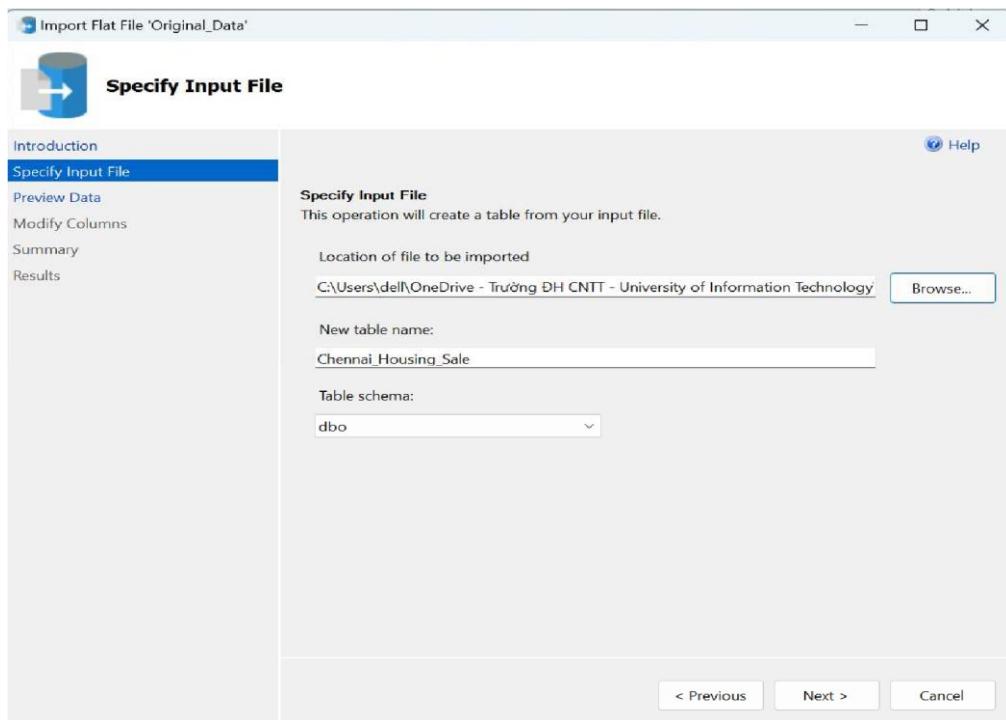
2.1. Tạo mới hai database

2.1.1. Database chứa dữ liệu gốc

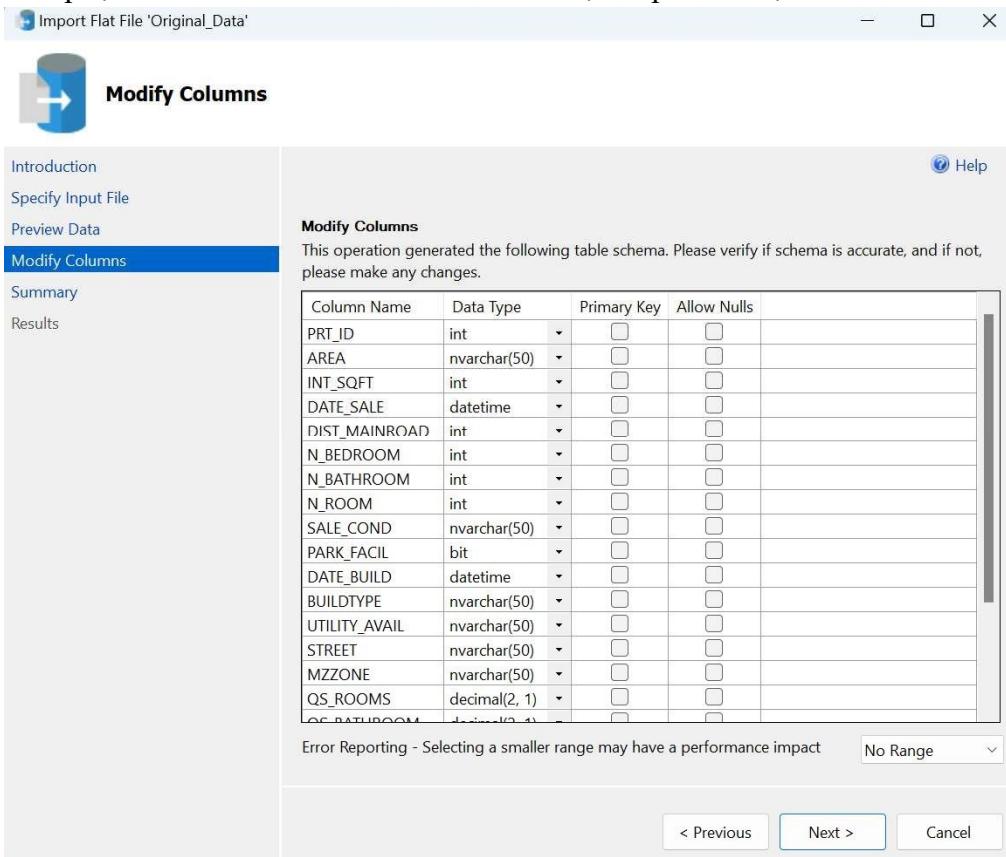
- Mở SQL Server ta thực hiện tạo một database để lưu data gốc ban đầu với tên là Original_Data.



- Nhấn chuột phải vào database Original_Data → Task → Import Flat File.

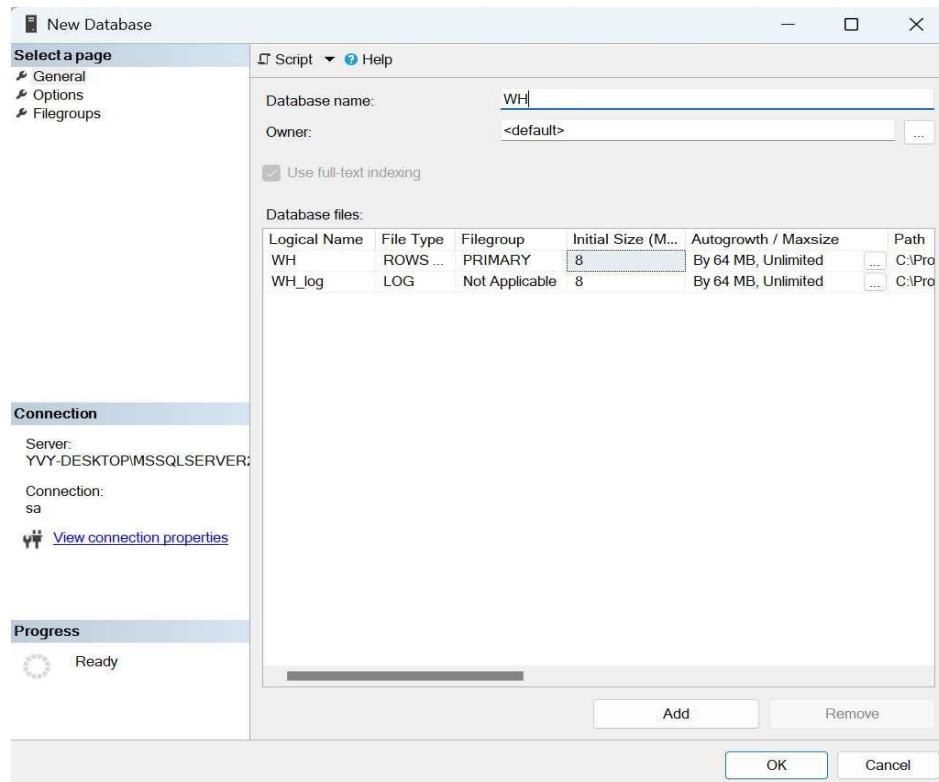


- Nhấn Next để chuyển sang tab Modify Columns sau đó thay đổi kiểu dữ liệu phù hợp và tiếp tục nhấn Next → Finish để hoàn tất việc import dữ liệu



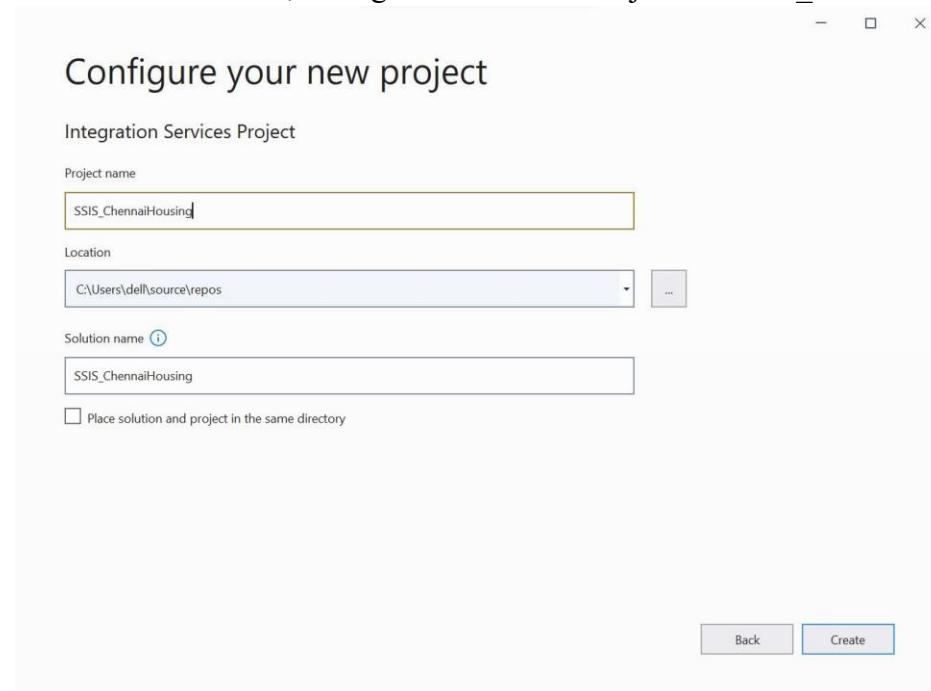
2.1.2. Data warehouse

- Mở SQL Server ta thực hiện tạo một database để lưu data sau khi xử lý dữ liệu với tên là WH.

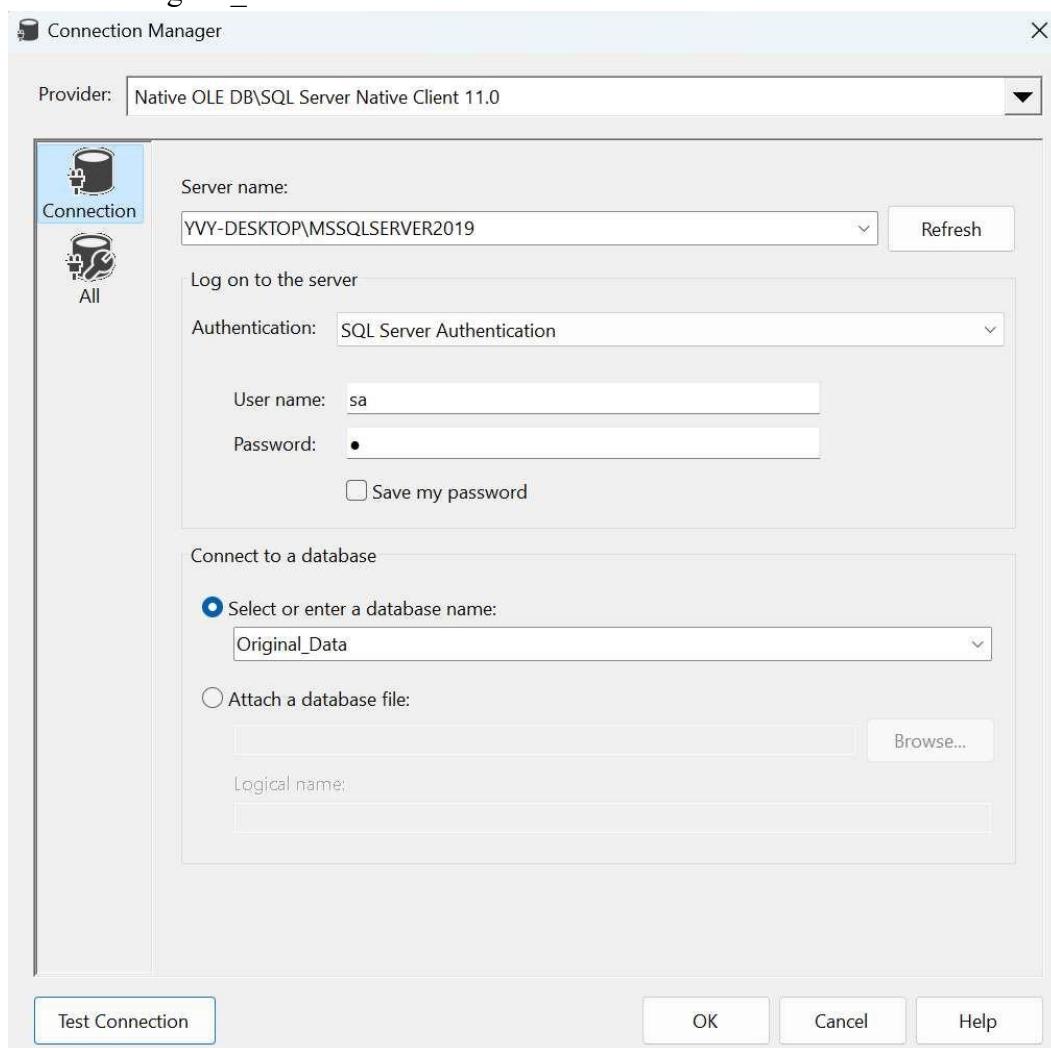


2.2. Tạo project mới trong công cụ SQL Data Tool

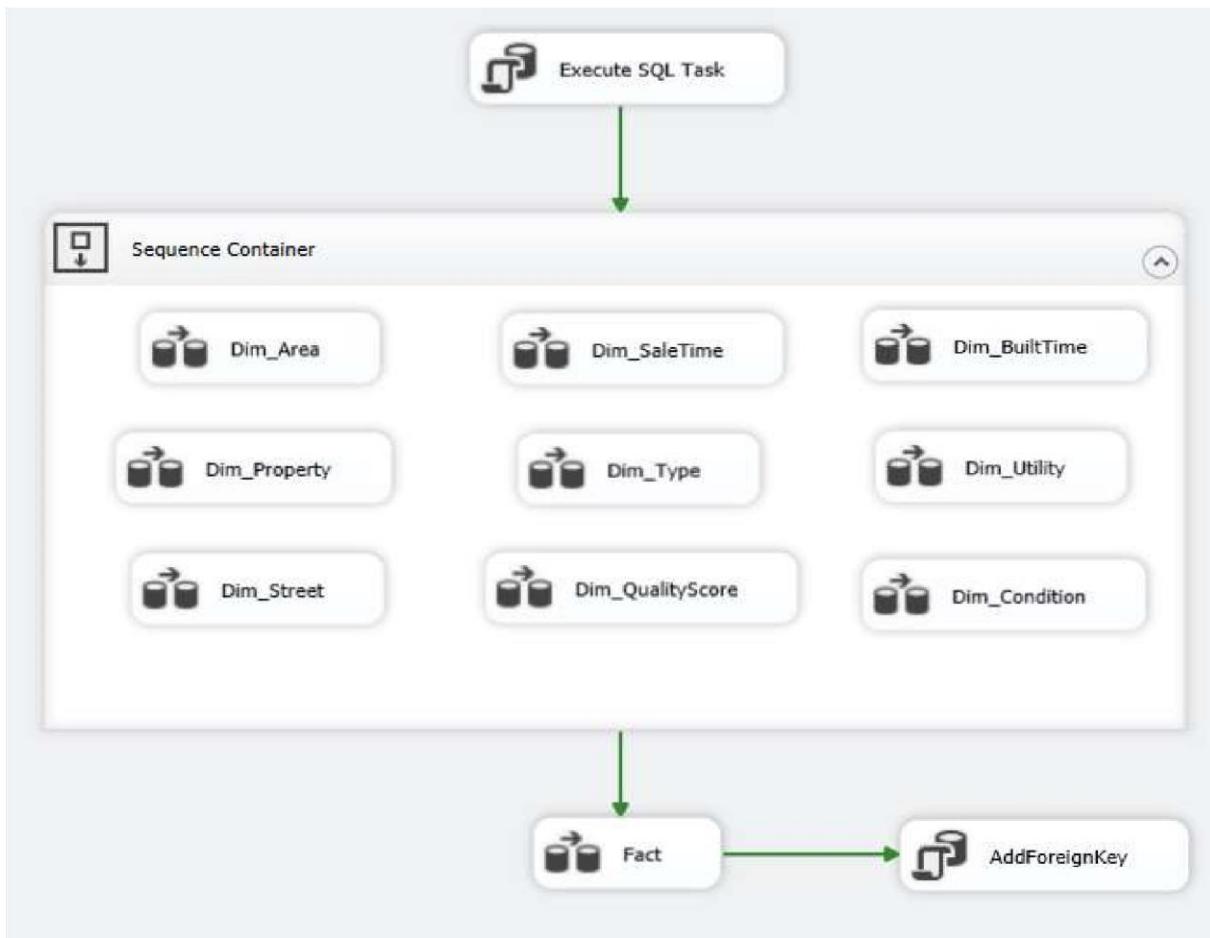
- Vào Visual Studio 2019: Tạo Integration Services Project tên SSIS_ChennaiHousing



- Thêm connection: Thêm tên Server để chọn database. Thực hiện lần lượt chọn 2 database: Original_Data và WH



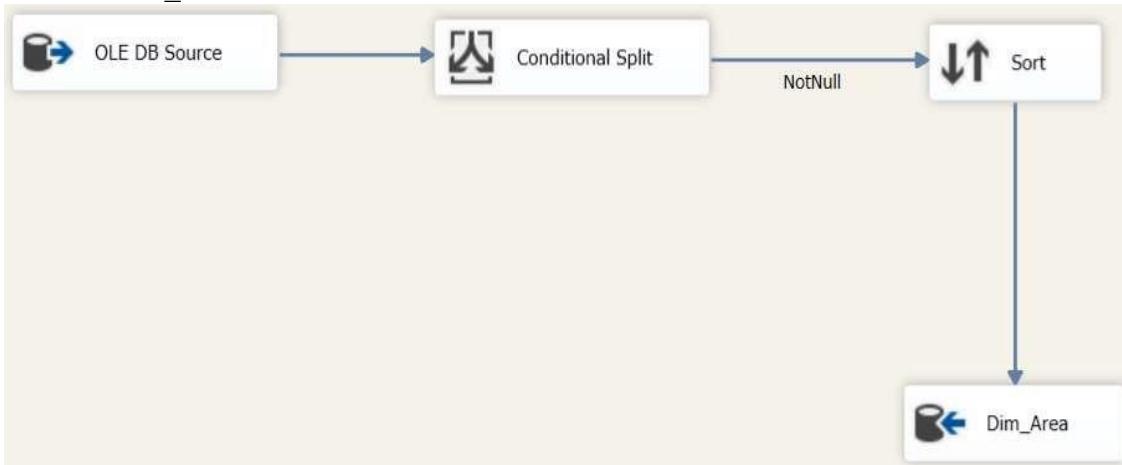
2.3. Mô hình SSIS



- Bước 1: Load Dimension Tables (Tạo các bảng dimension)
- Bước 2: Load Fact Table (Tạo bảng fact)
- Bước 3: Viết Execute SQL Task

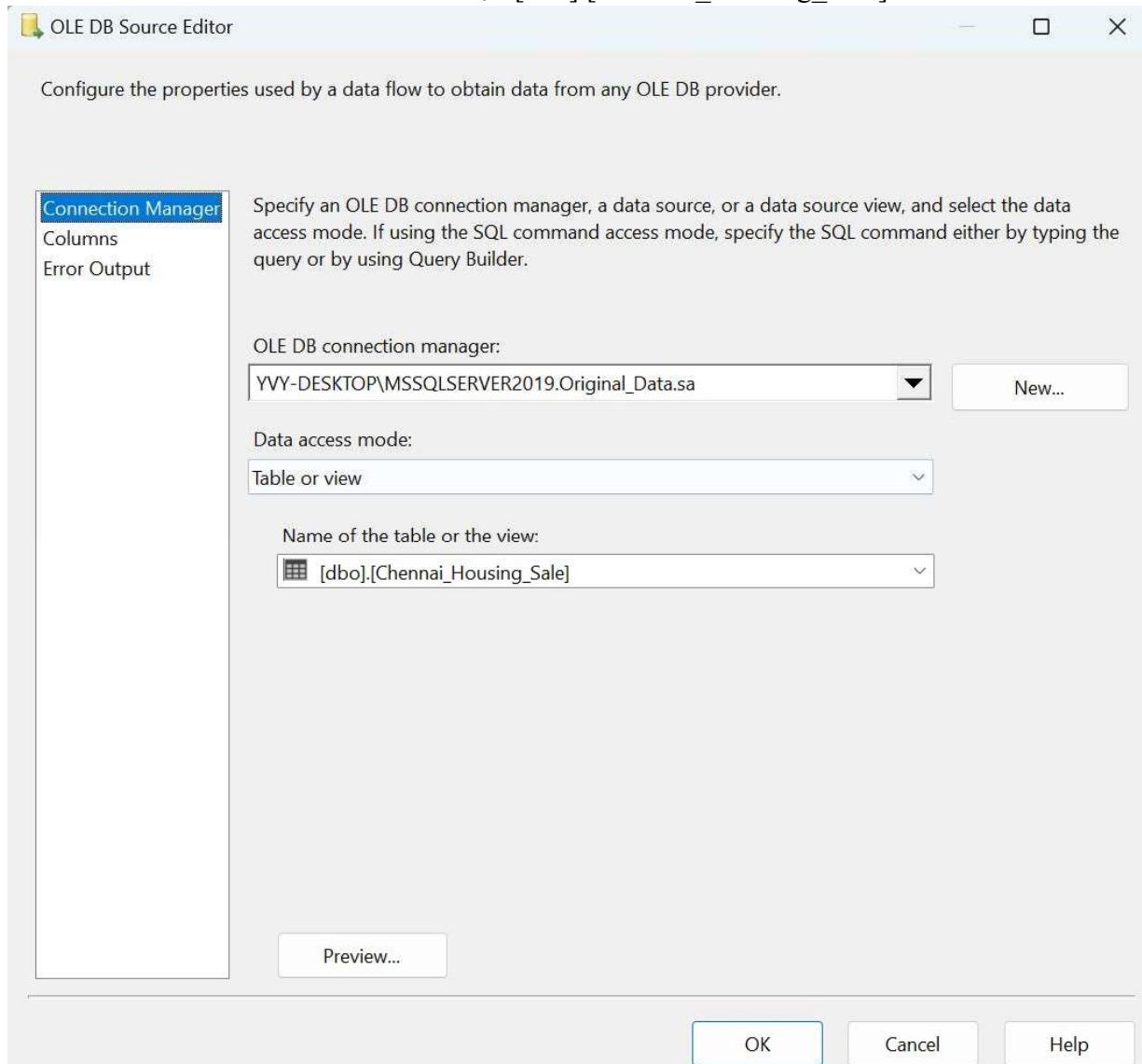
2.4. Load Dimension Tables

2.4.1. Load Dim_Area



OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull

Conditional Split Transformation Editor

Specify the conditions used to direct input rows to specific outputs. If an input row matches no condition, the row is directed to a default output.

Order	Output Name	Condition
1	Null	ISNULL(AREA) ISNULL(MZONE)

Default output name: NotNull

Variables and Parameters

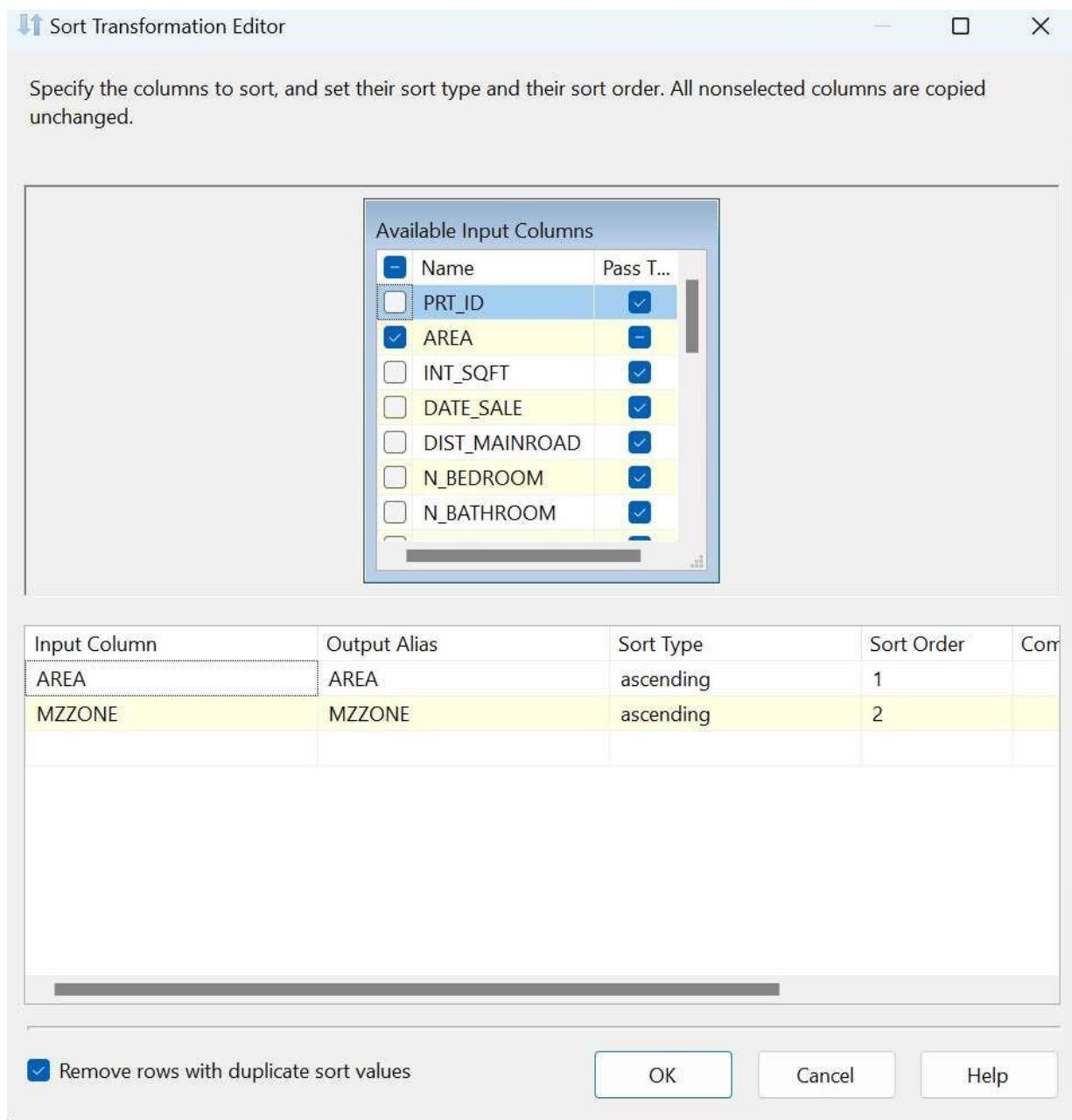
Mathematical Functions
String Functions
Date/Time Functions
NULL Functions
Type Casts
Operators

Description:

OK Cancel Help

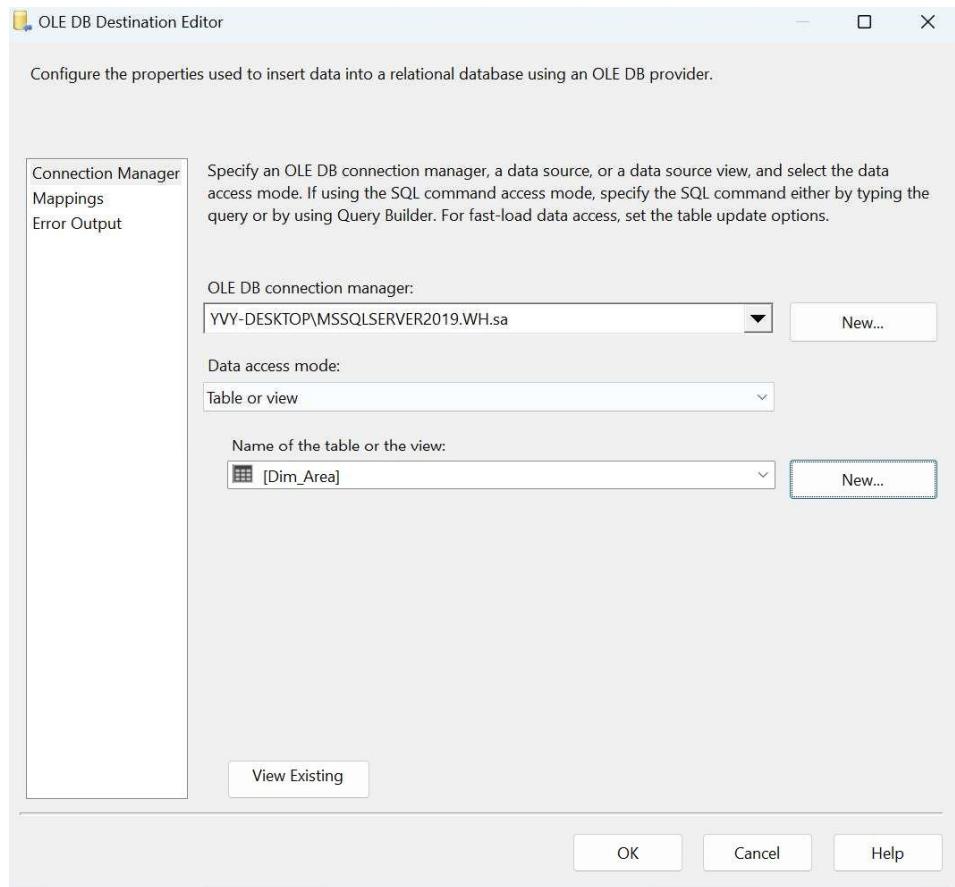
Sort:

- AREA và MZONE là khóa kết
- Tích Remove rows with duplicate sort values

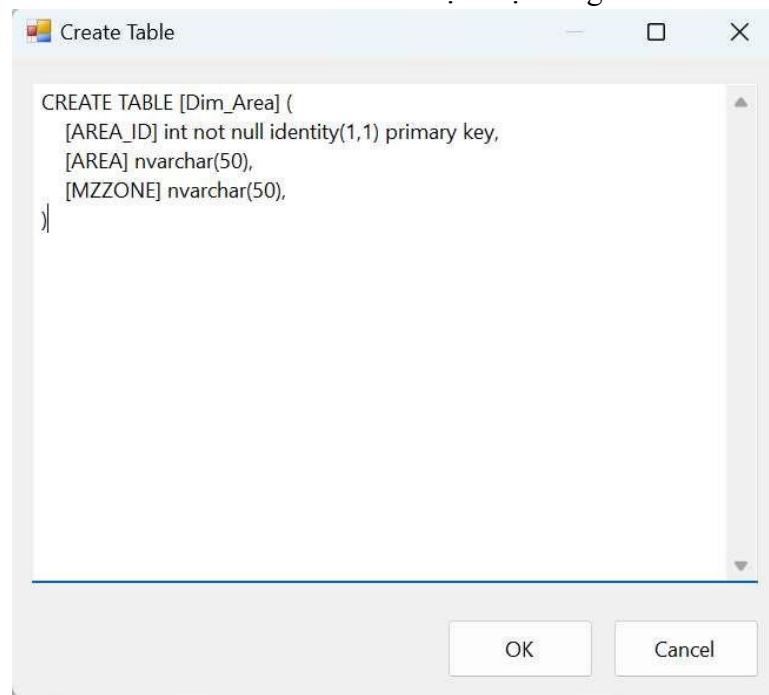


OLE DB Destination:

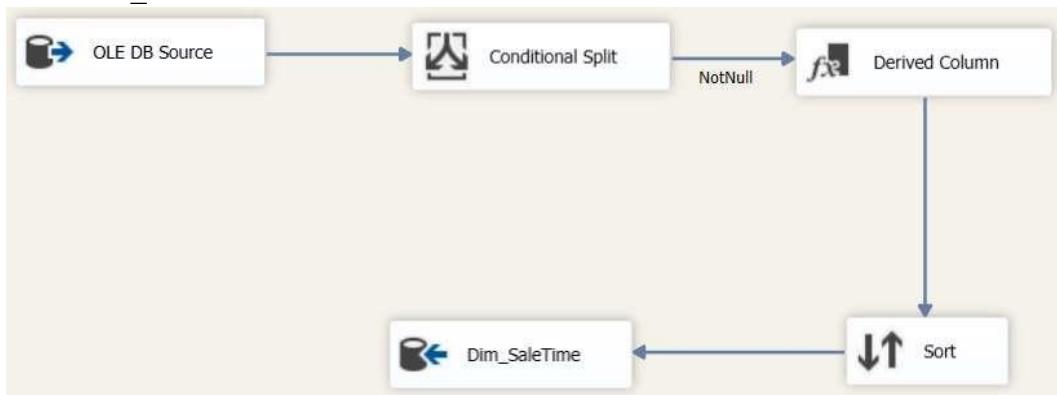
- OLE DB connection manager chọn WH



- Name of the table or the view nhấp New để tạo một bảng mới:

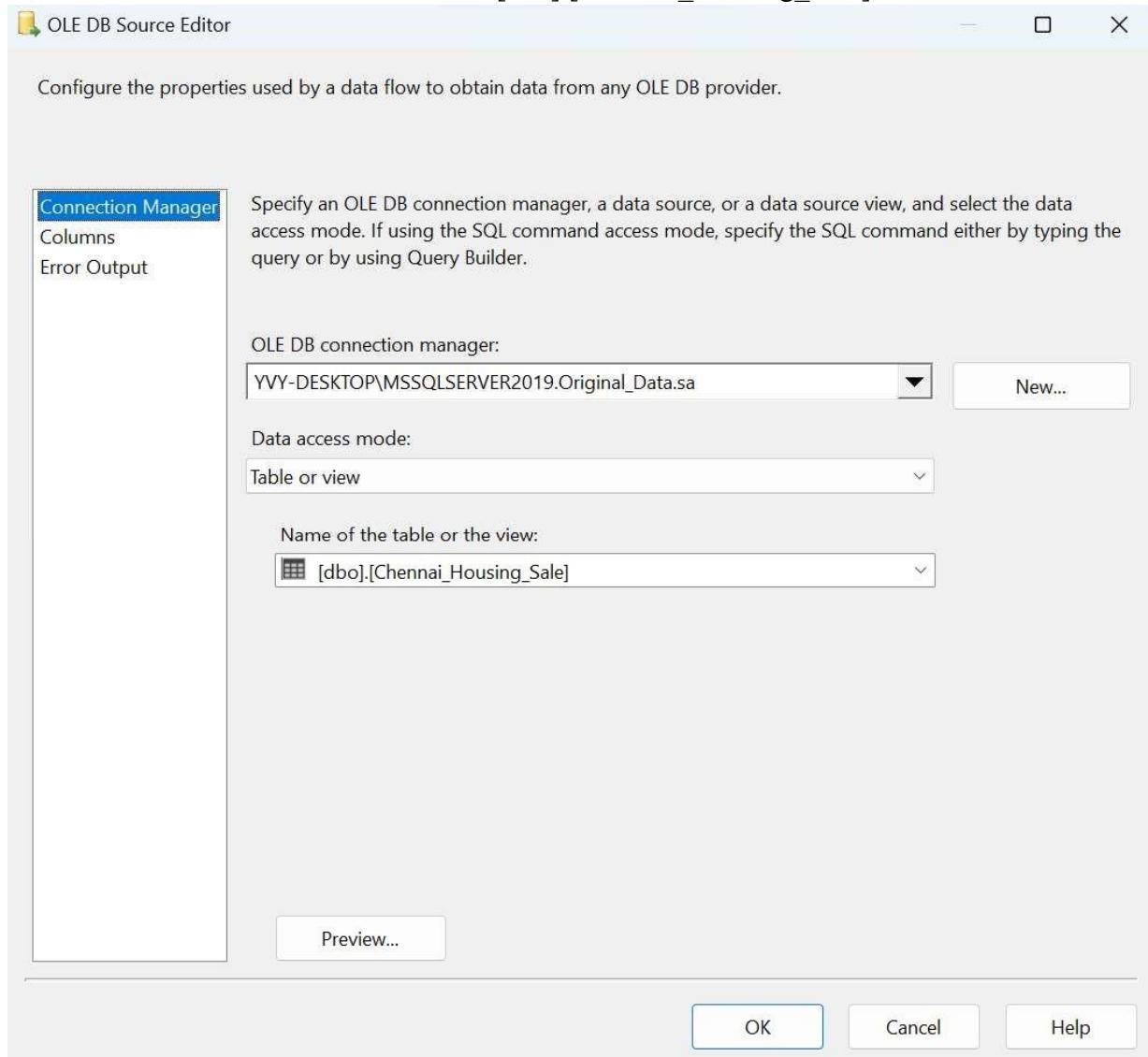


2.4.2. Load Dim_SaleTime



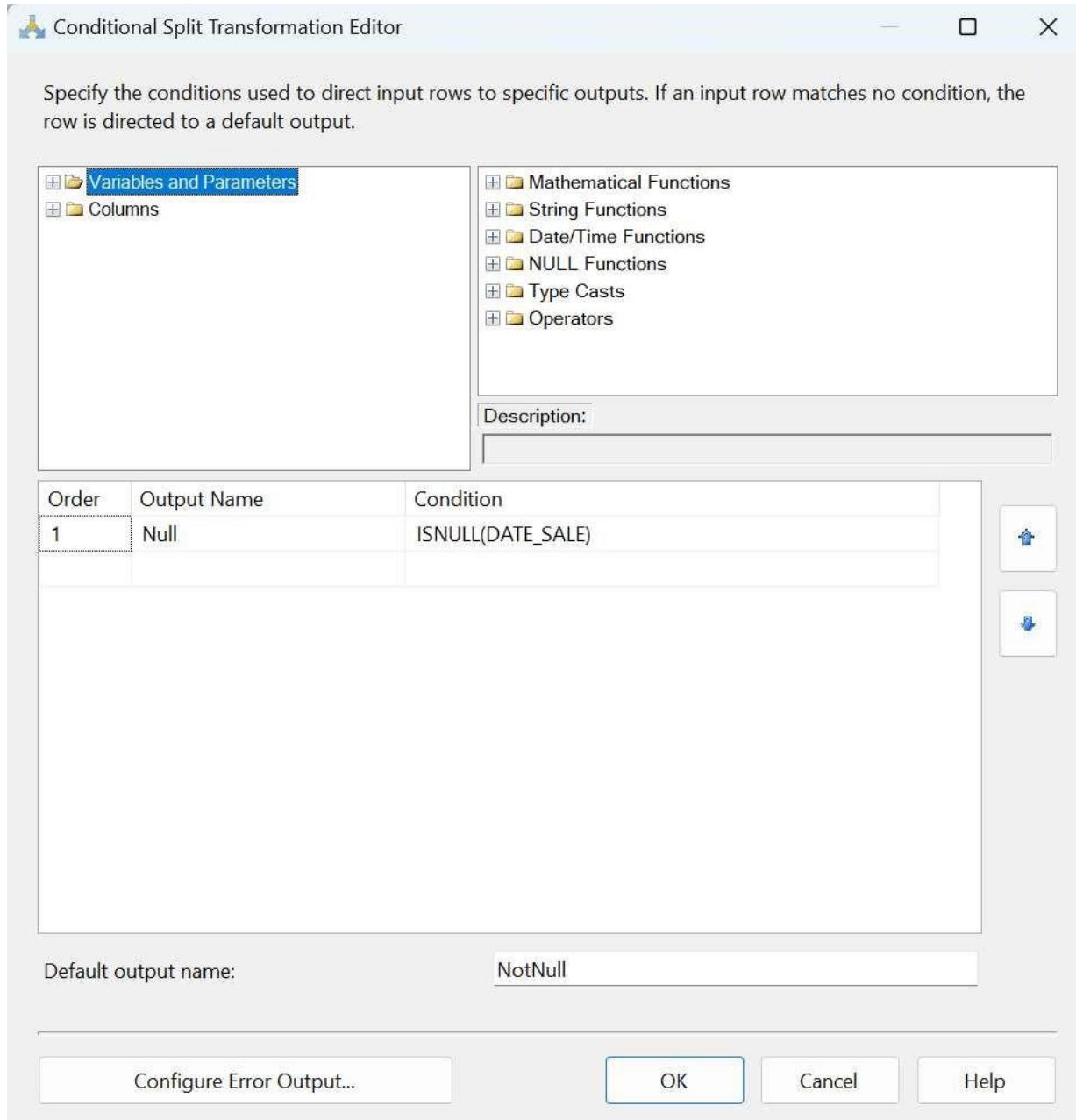
OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



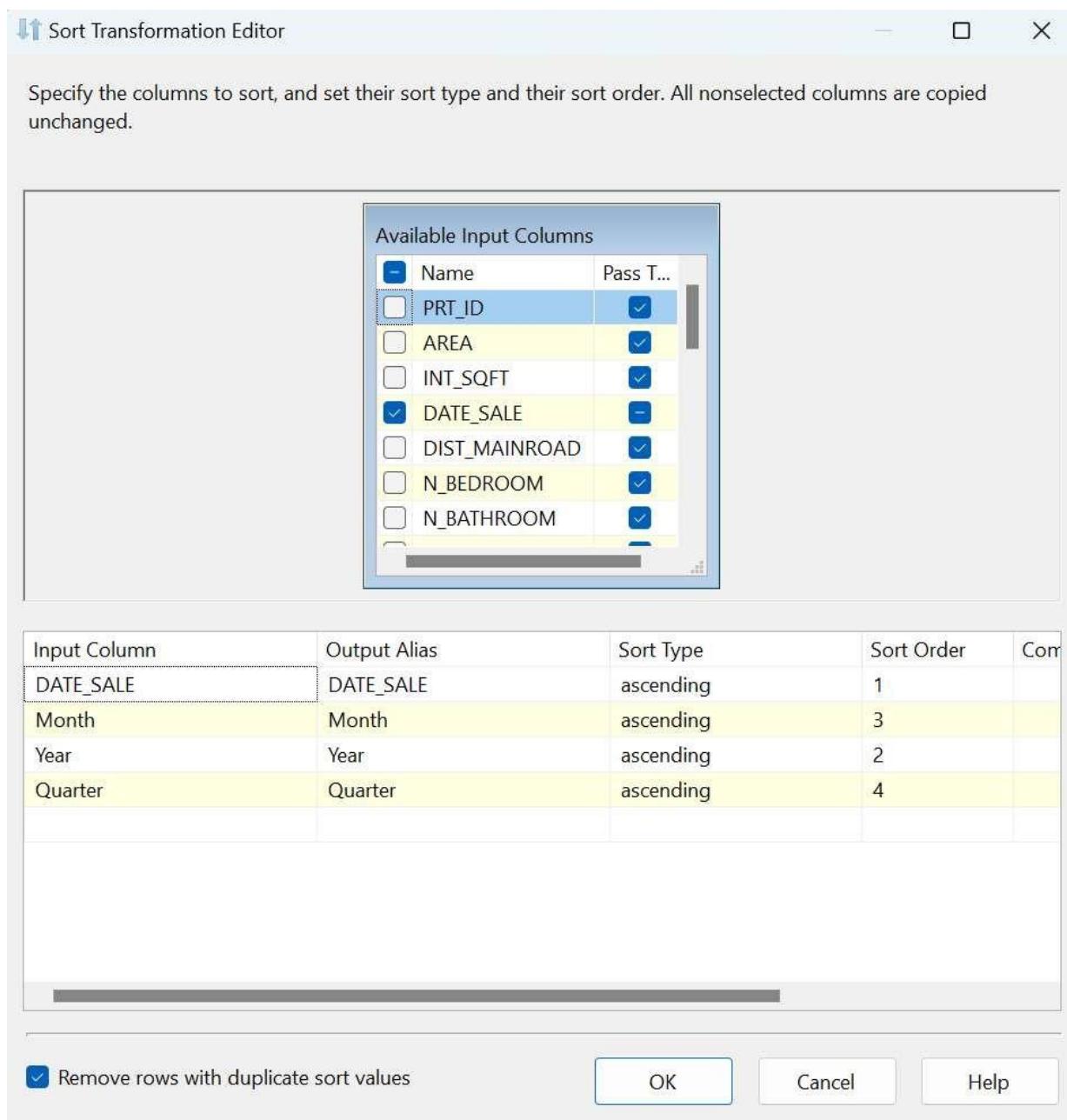
Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull

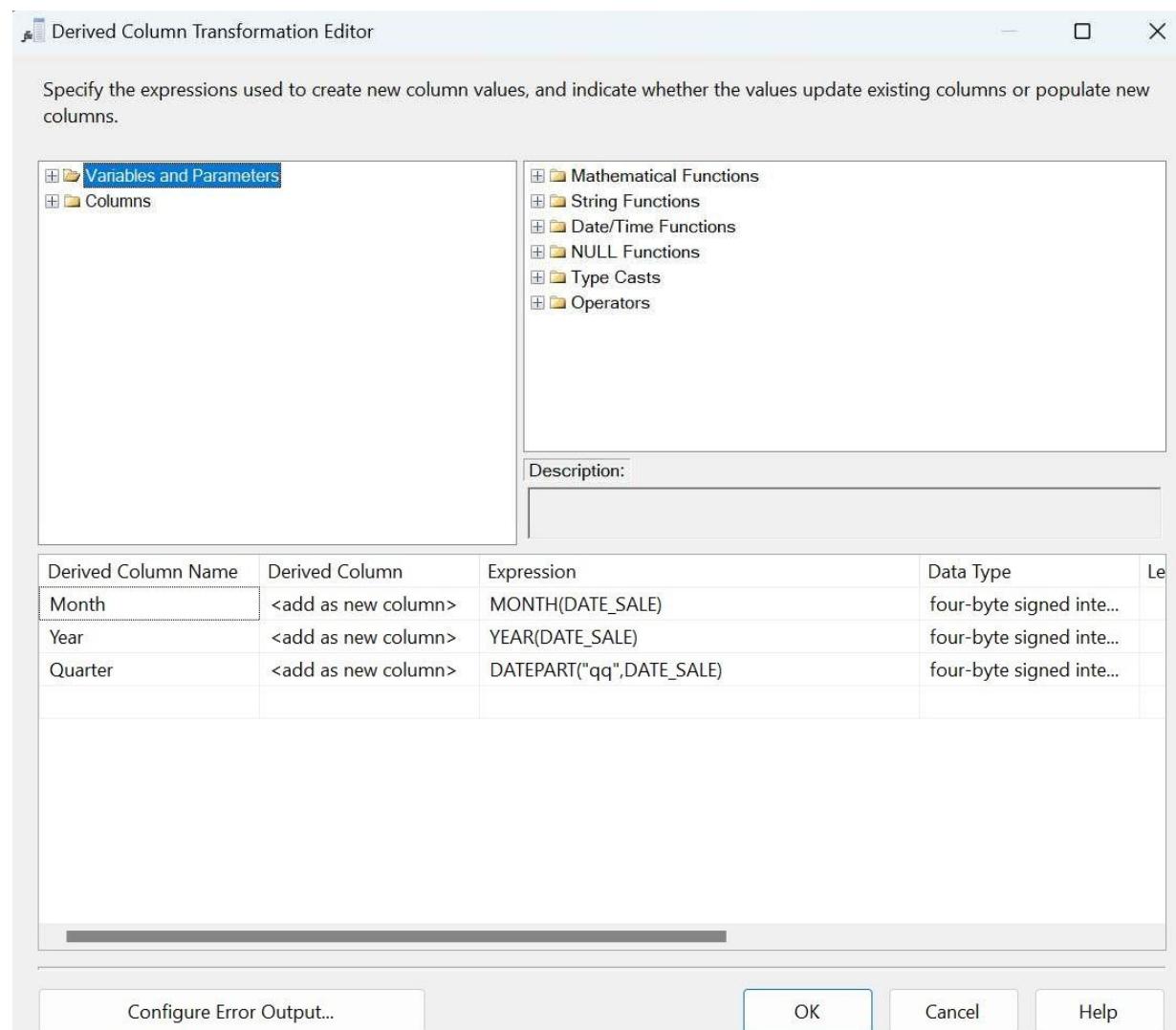


Sort:

- DATE_SALE, Month, Year, Quarter là khóa kết
- Tích Remove rows with duplicate sort values

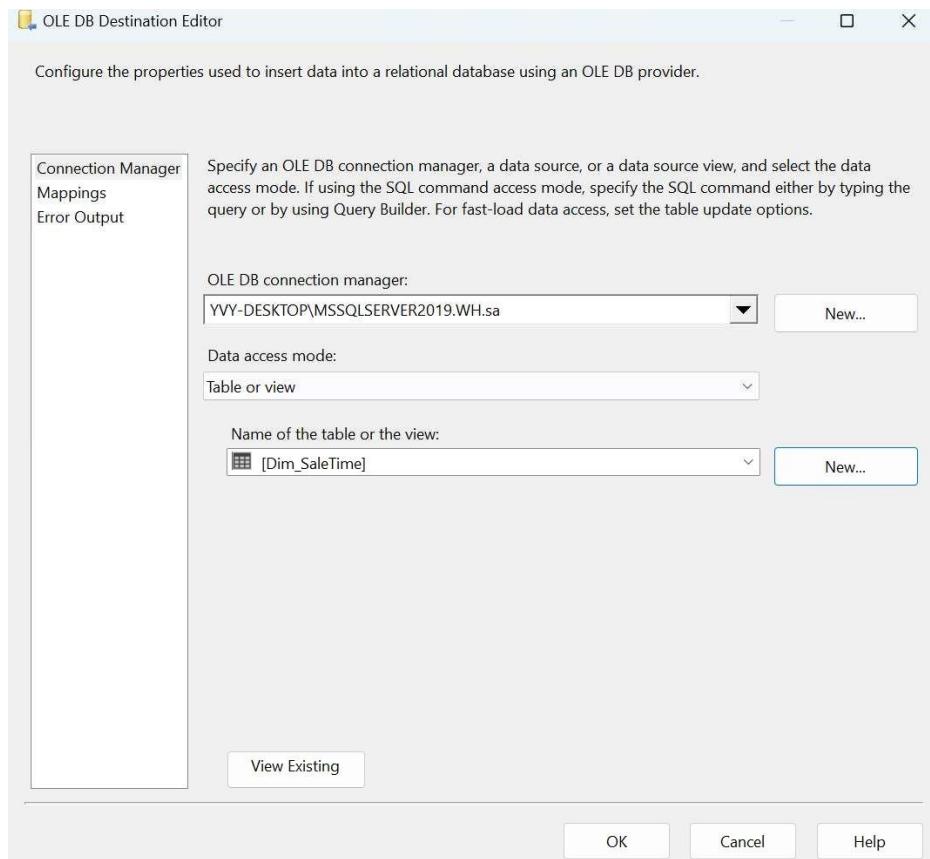
**Derived Column:**

- Tại đây, ta kéo các Date/Time Functions mà ta muốn tách từ thuộc tính “DATE_SALE” và đặt tên cho cột được tách như hình.

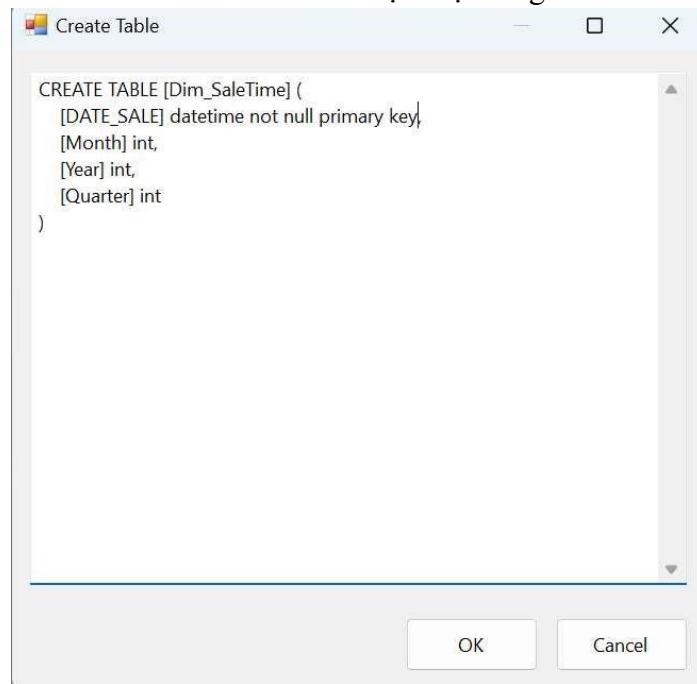


OLE DB Destination:

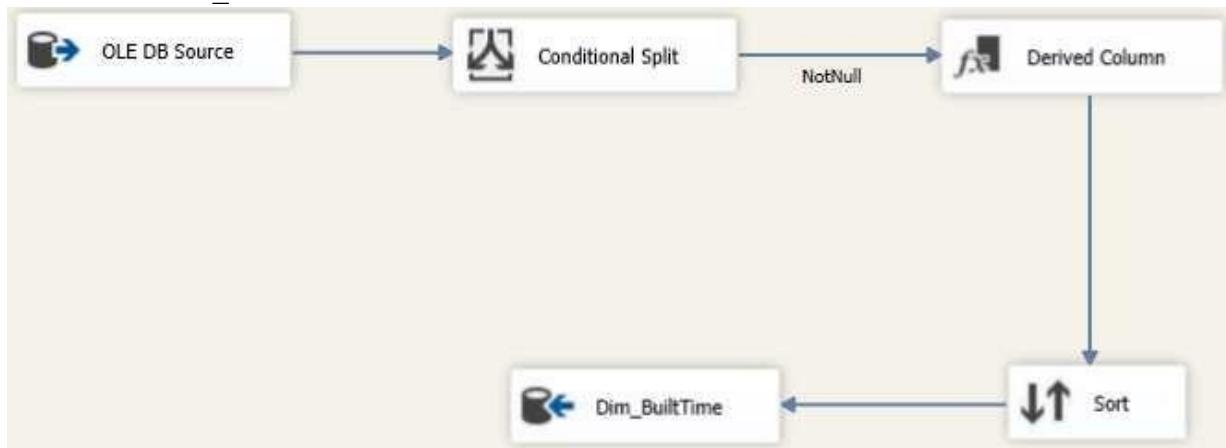
- OLE DB connection manager chọn WH



- Name of the table or the view nhấn New để tạo một bảng mới:

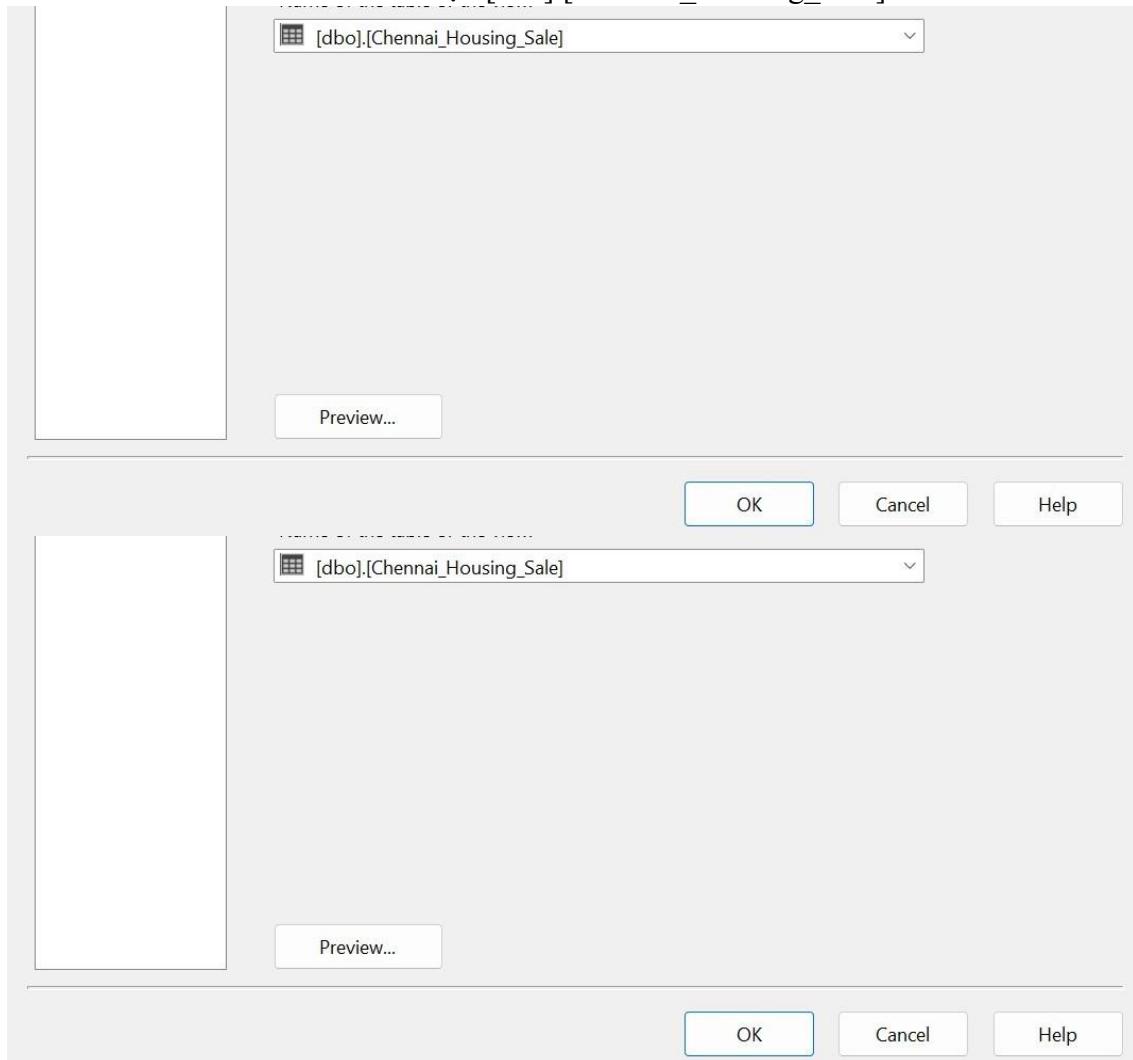


2.4.3. Load Dim_BuiltTime



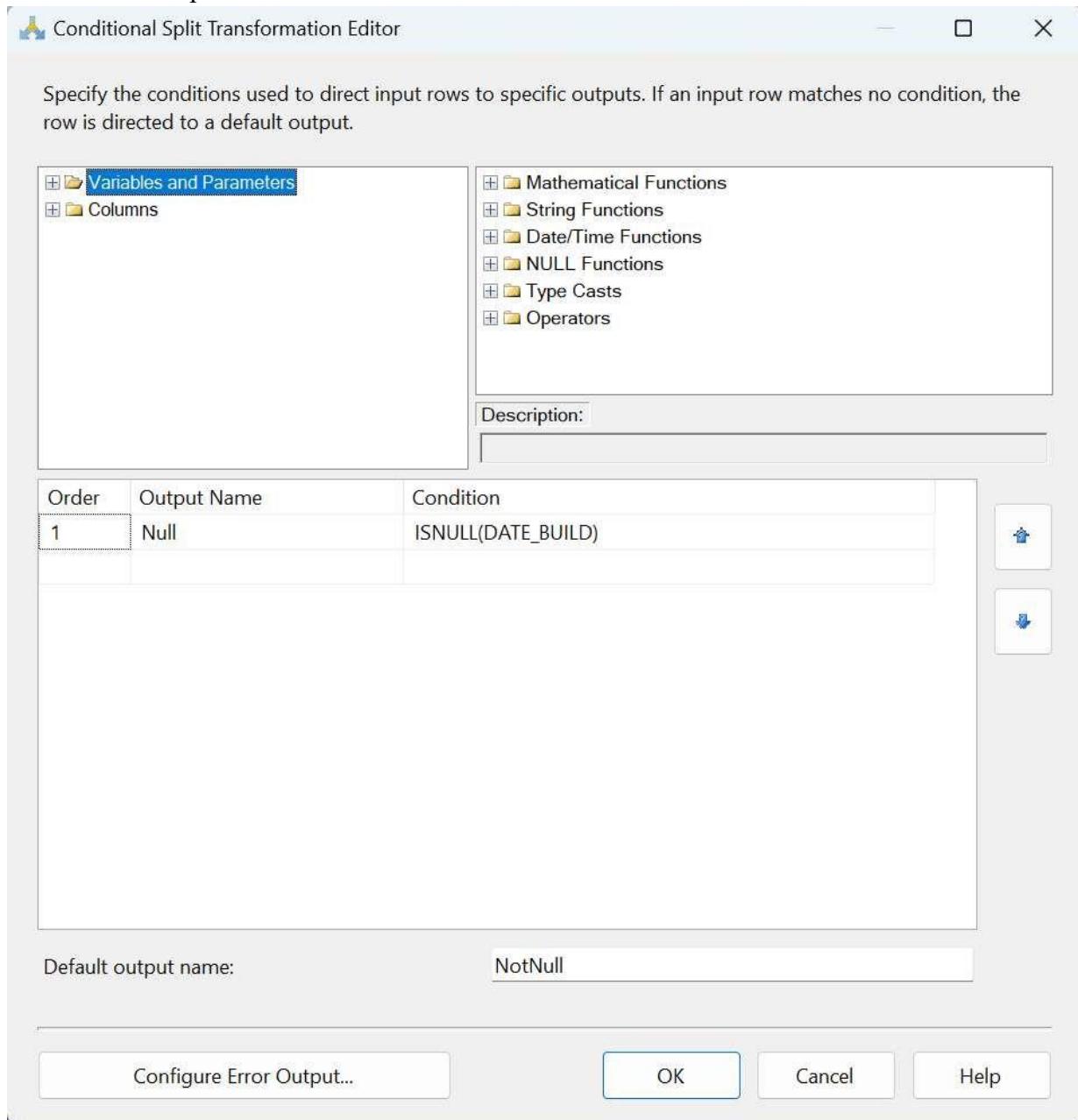
OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



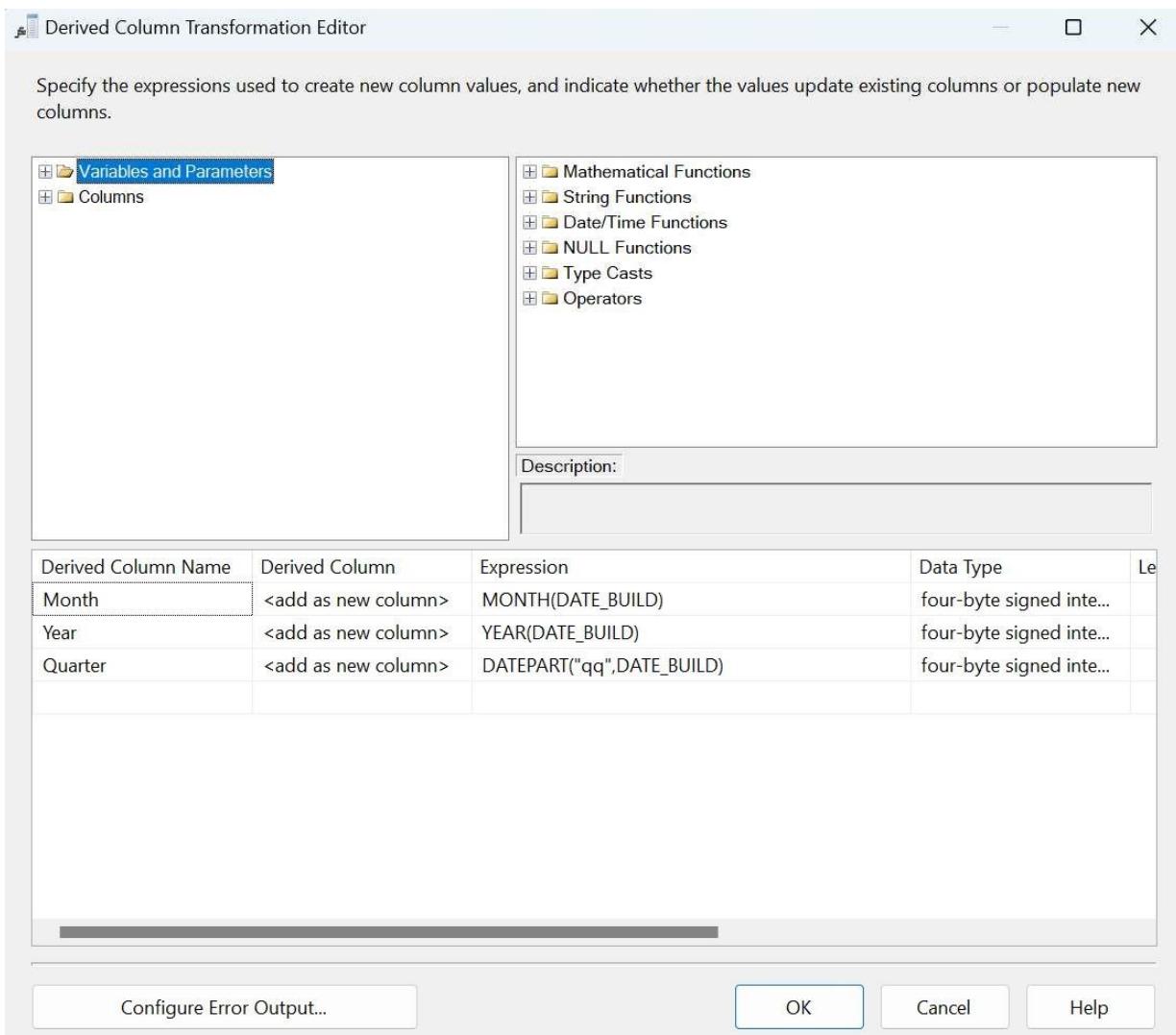
Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull



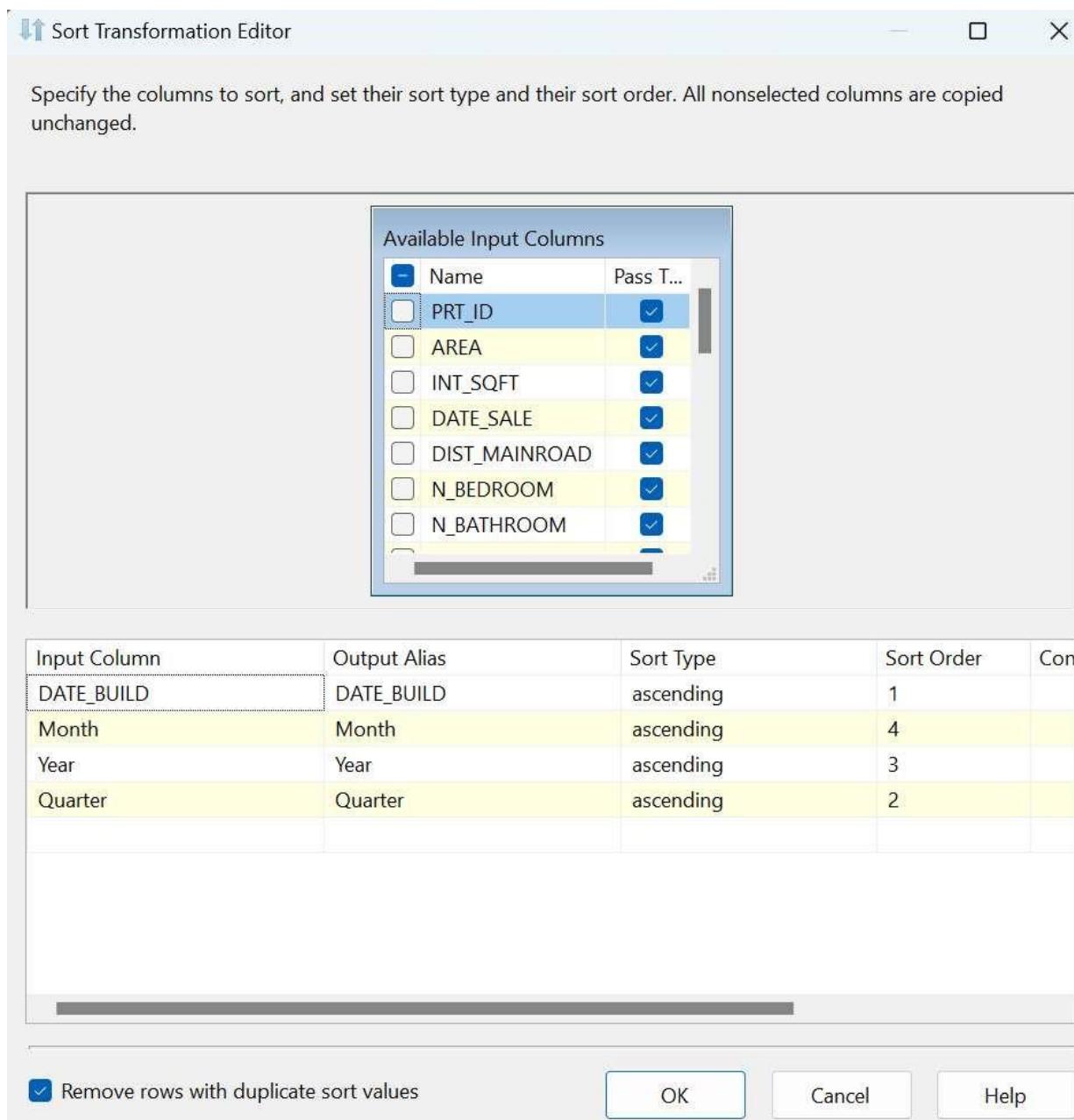
Derived Column:

- Tại đây, ta kéo các Date/Time Functions mà ta muốn tách từ thuộc tính “DATE_BUILD” và đặt tên cho cột được tách như hình.



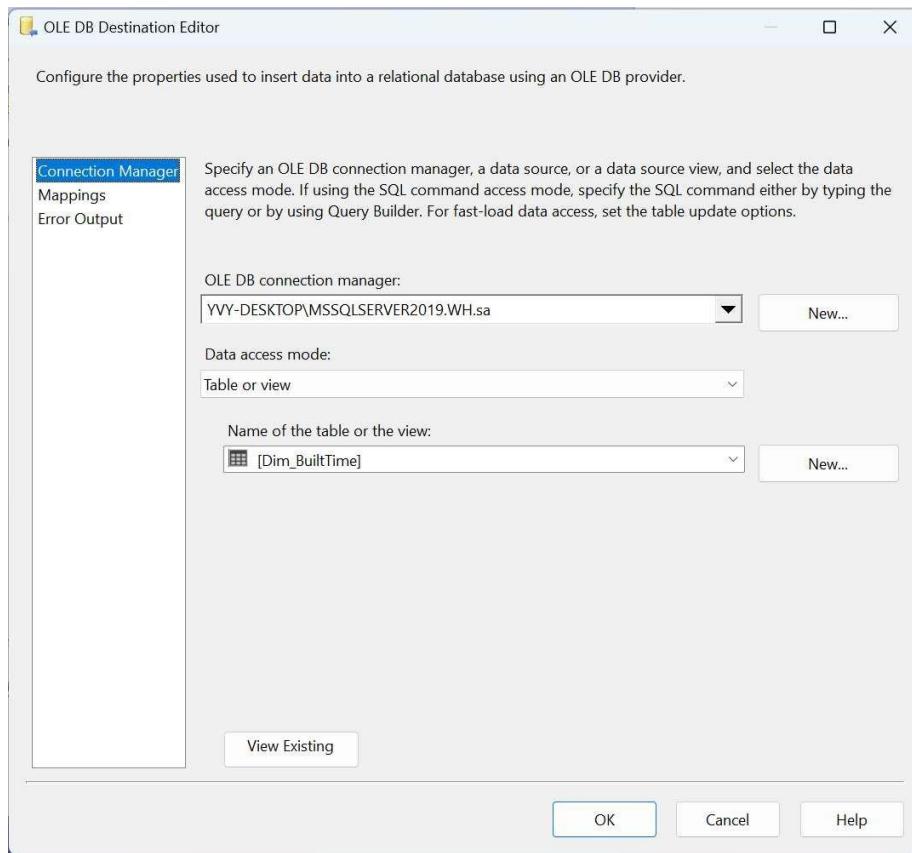
Sort:

- DATE_BUILD, Month, Year, Quarter là khóa két
- Tích Remove rows with duplicate sort values

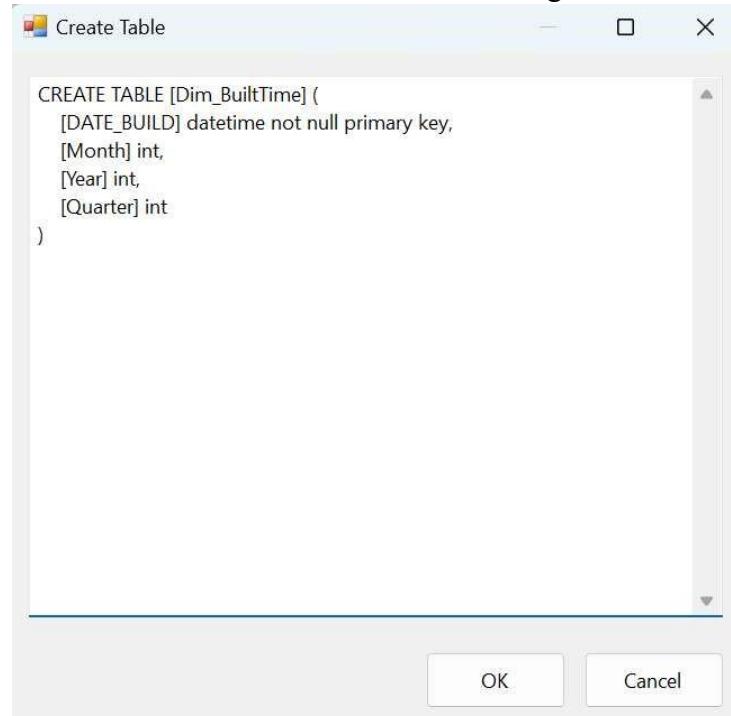


OLE DB Destination:

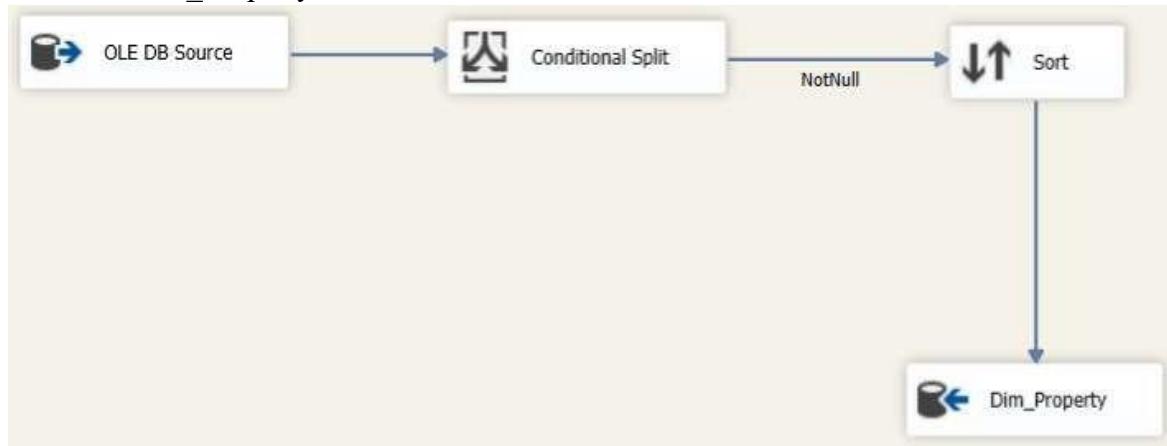
- OLE DB connection manager chọn WH



- Name of the table or the view nhấn New để tạo một bảng mới:

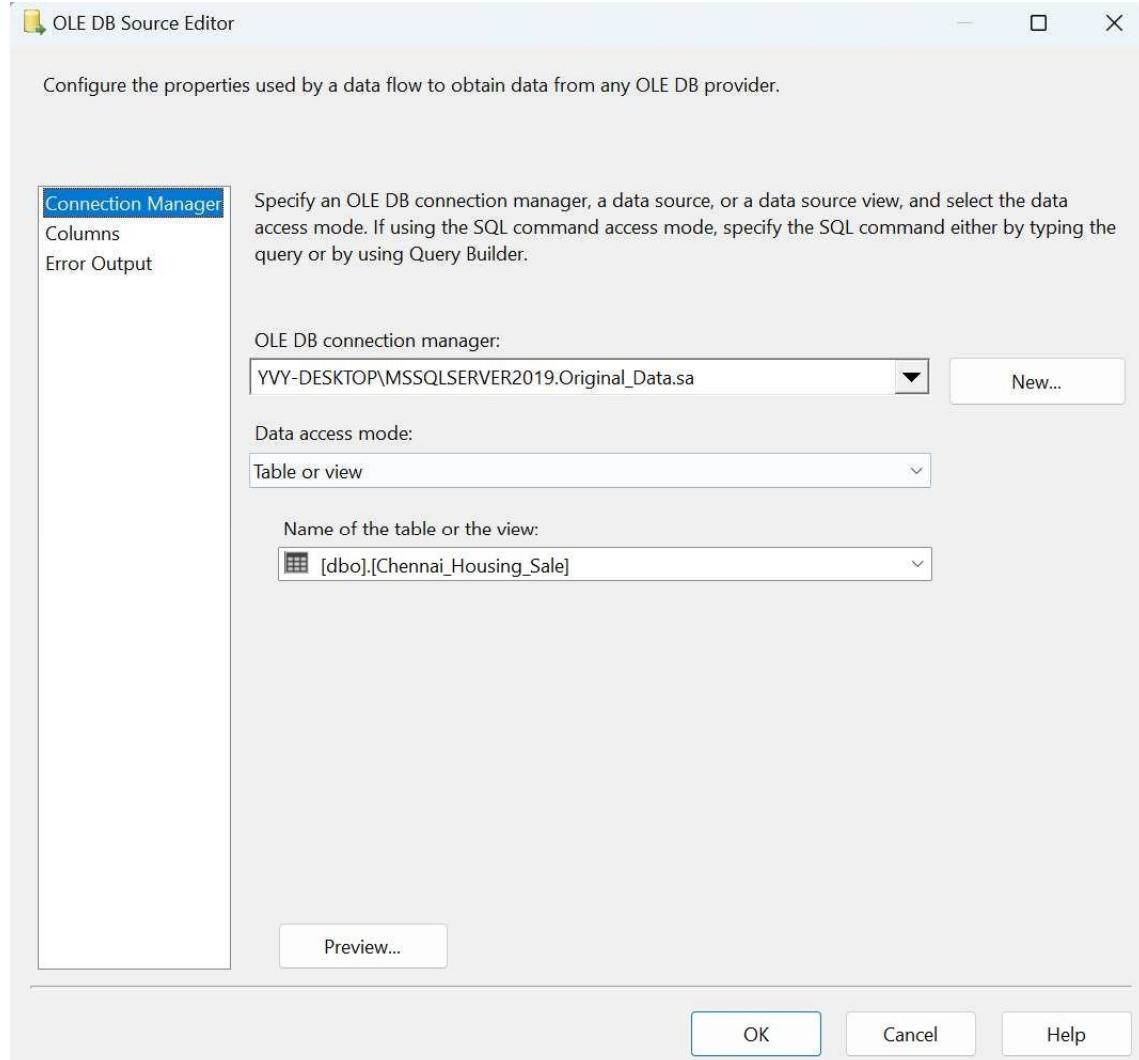


2.4.4. Load Dim_Property



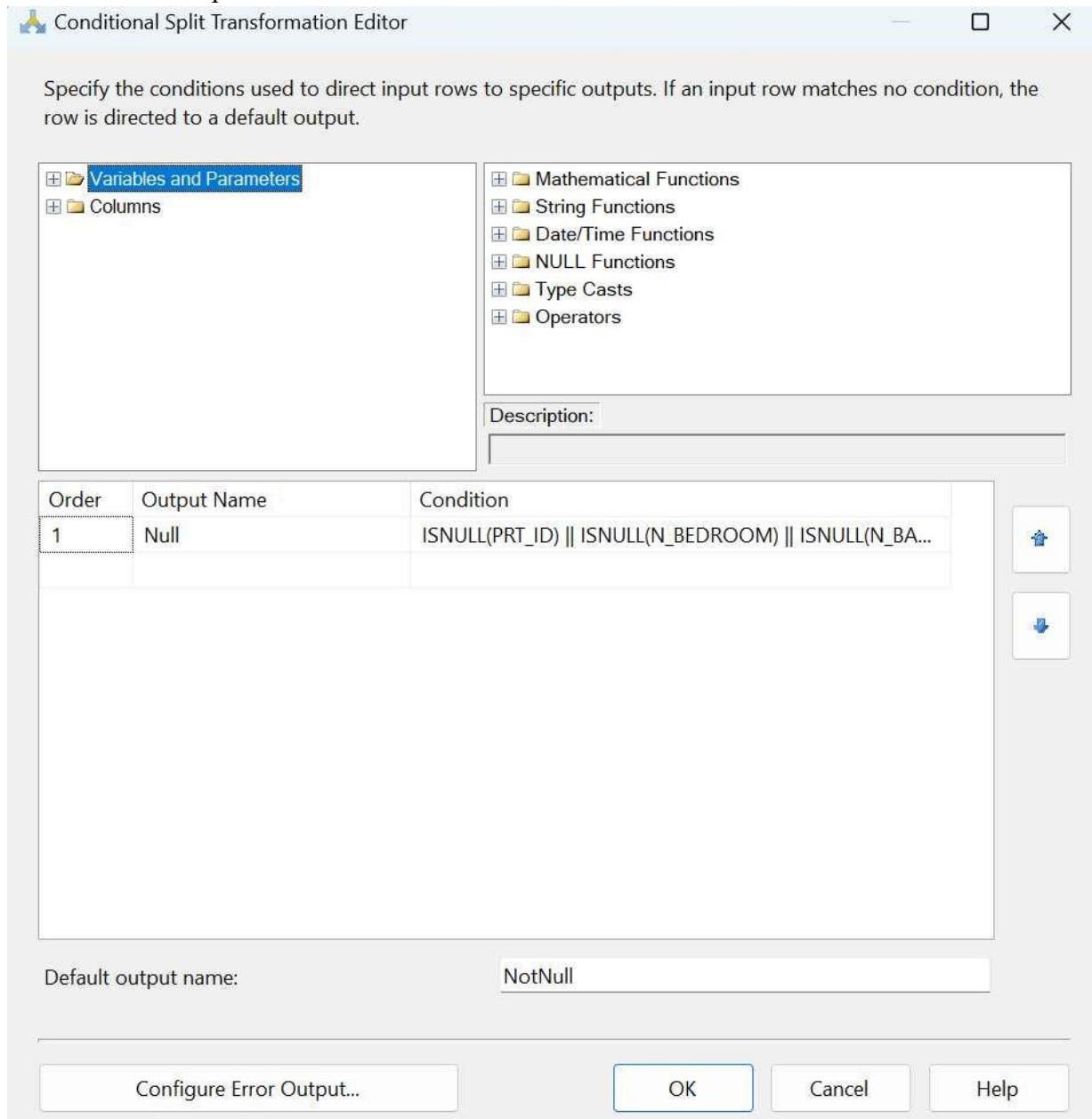
OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



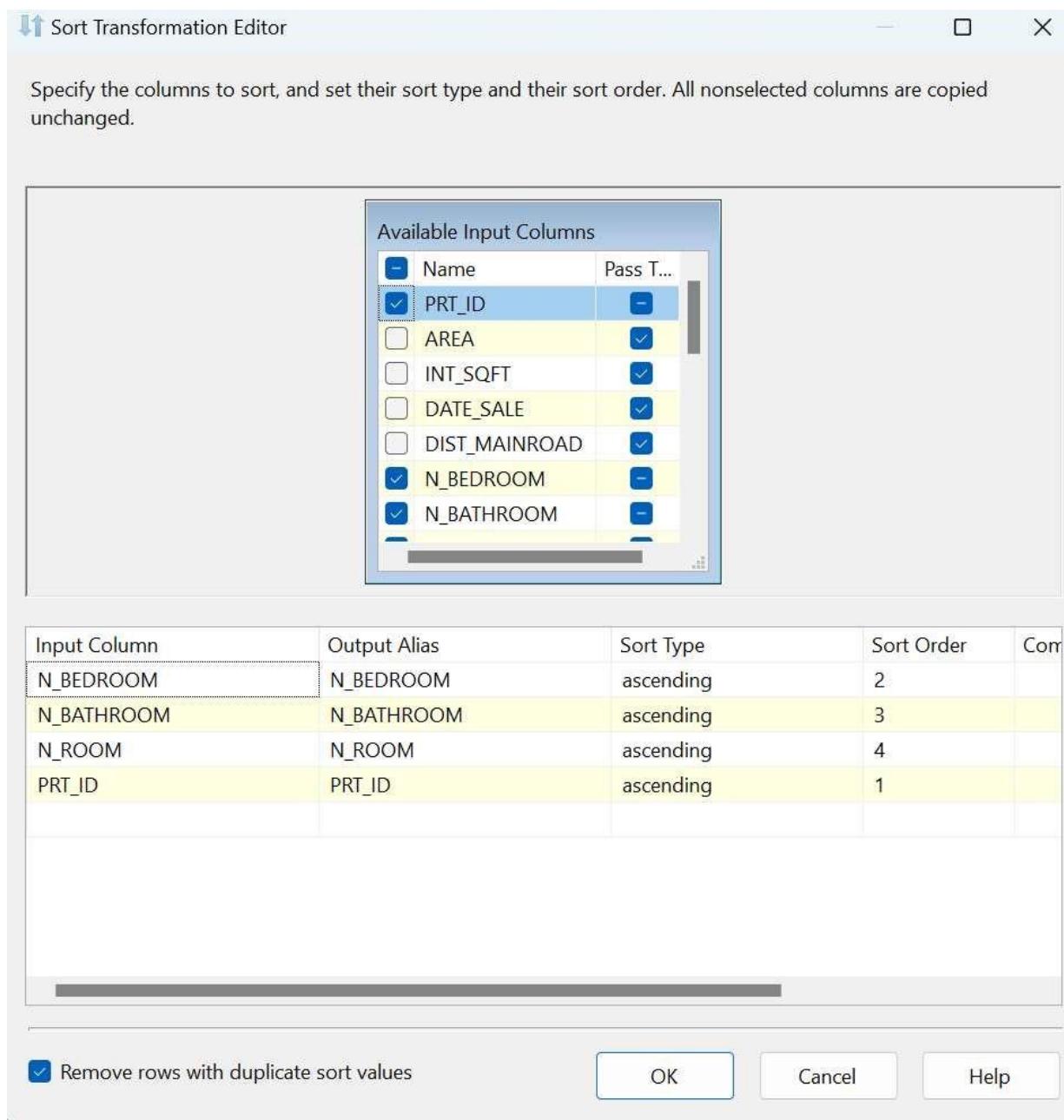
Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull



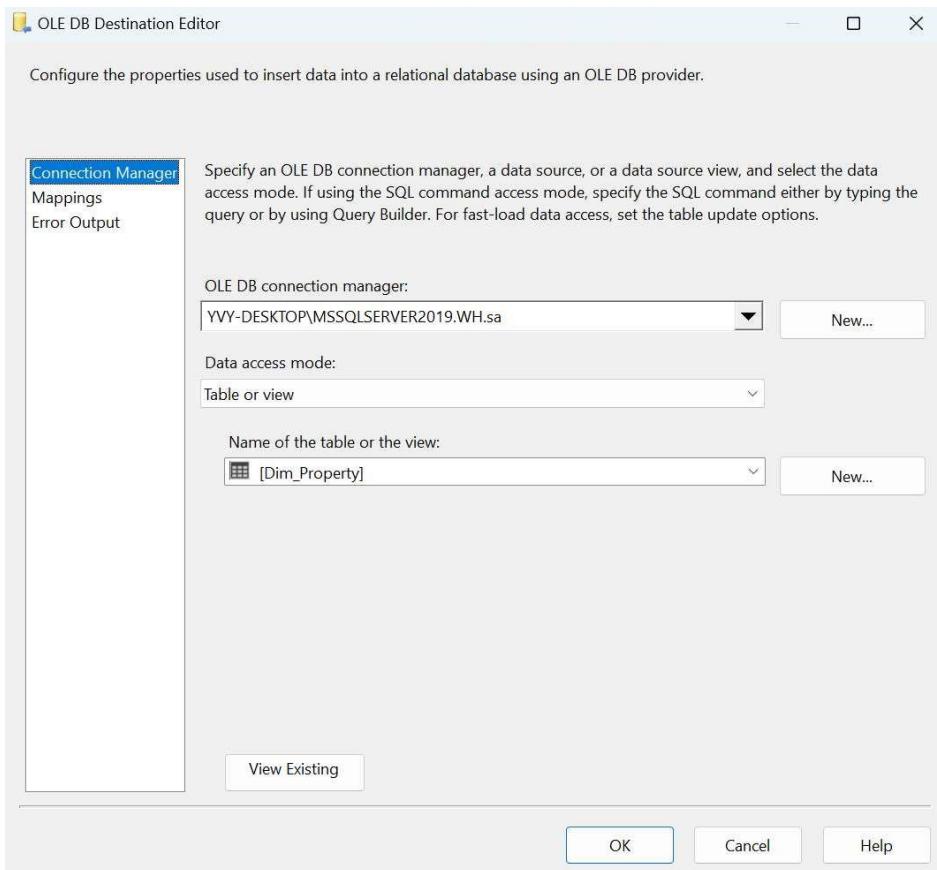
Sort:

- PRT_ID, N_BEDROOM, N_BATHROOM, N_ROOM là khóa kết
- Tích Remove rows with duplicate sort values

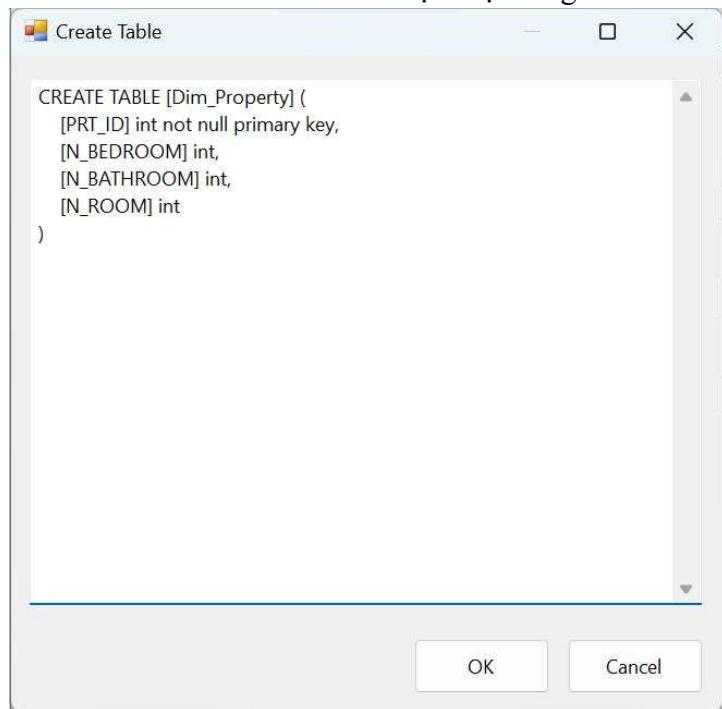


OLE DB Destination:

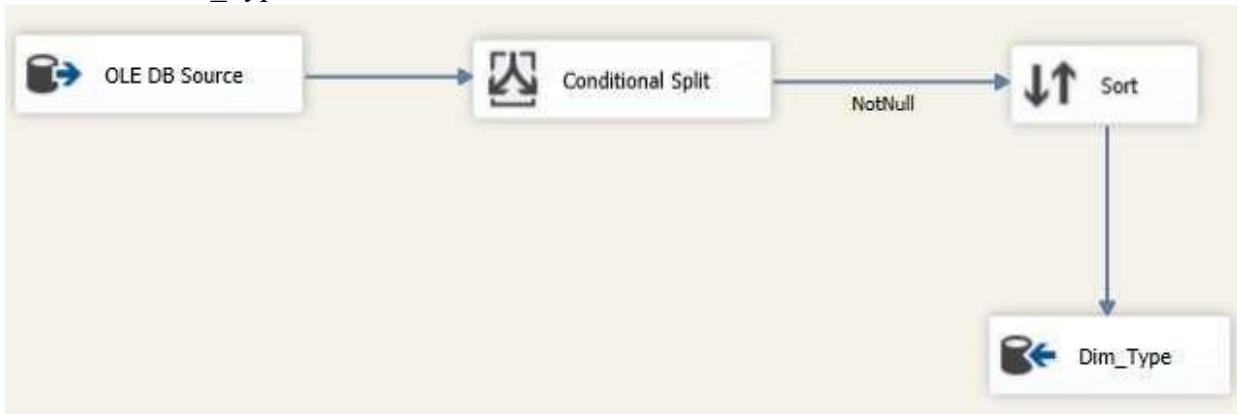
- OLE DB connection manager chọn WH



- Name of the table or the view nhấn New để tạo một bảng mới:

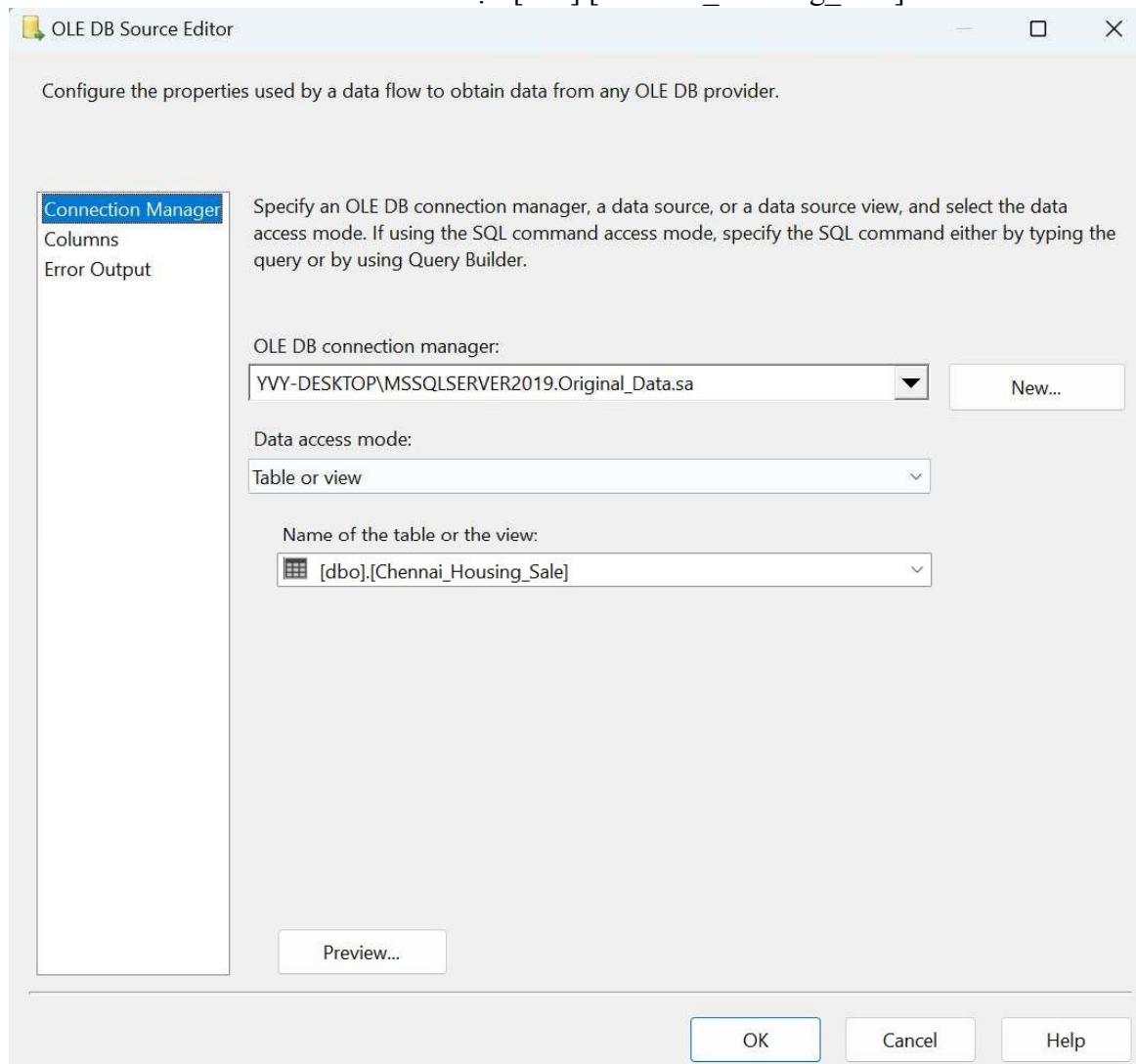


2.4.5. Load Dim_Type



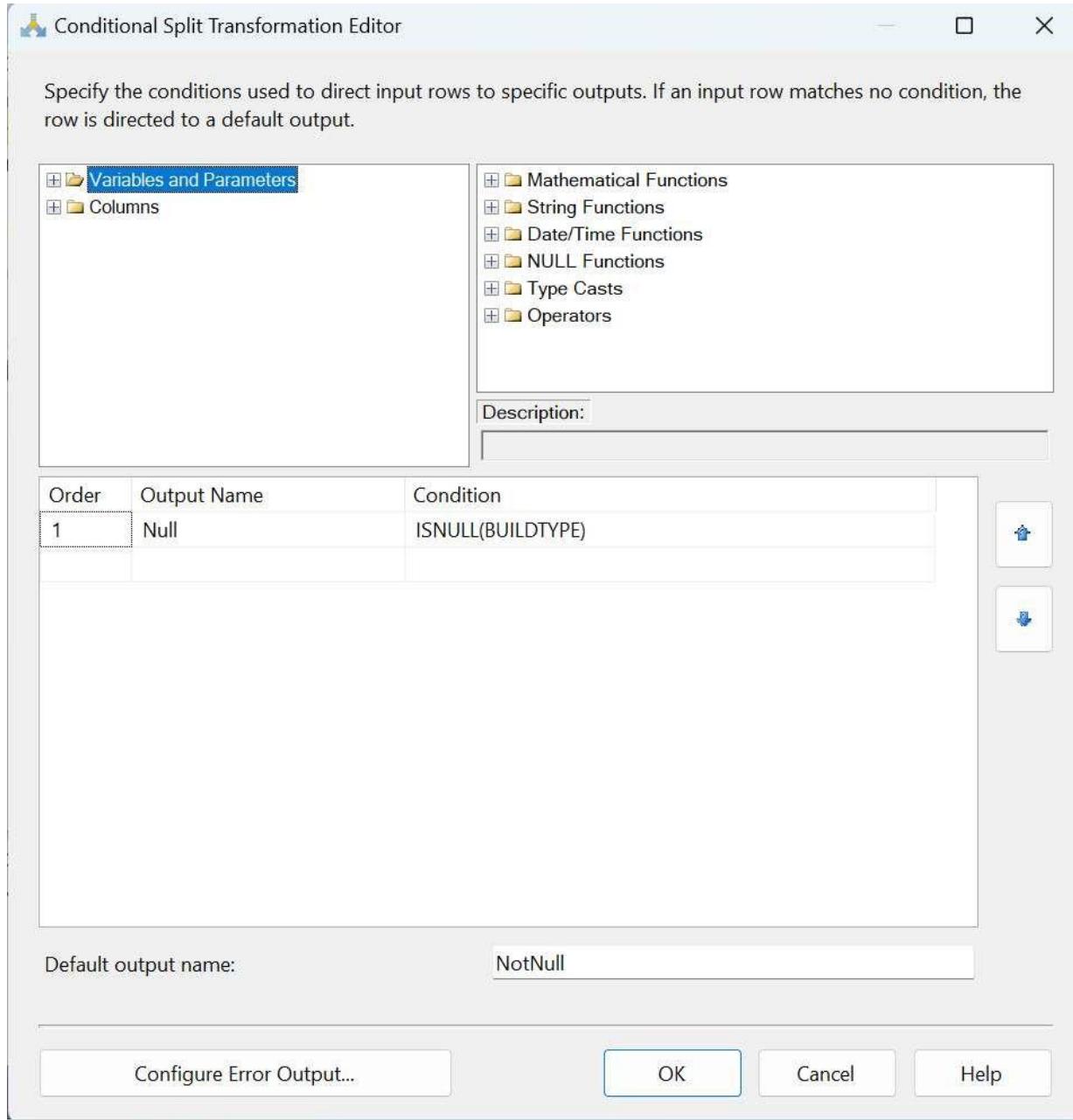
OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



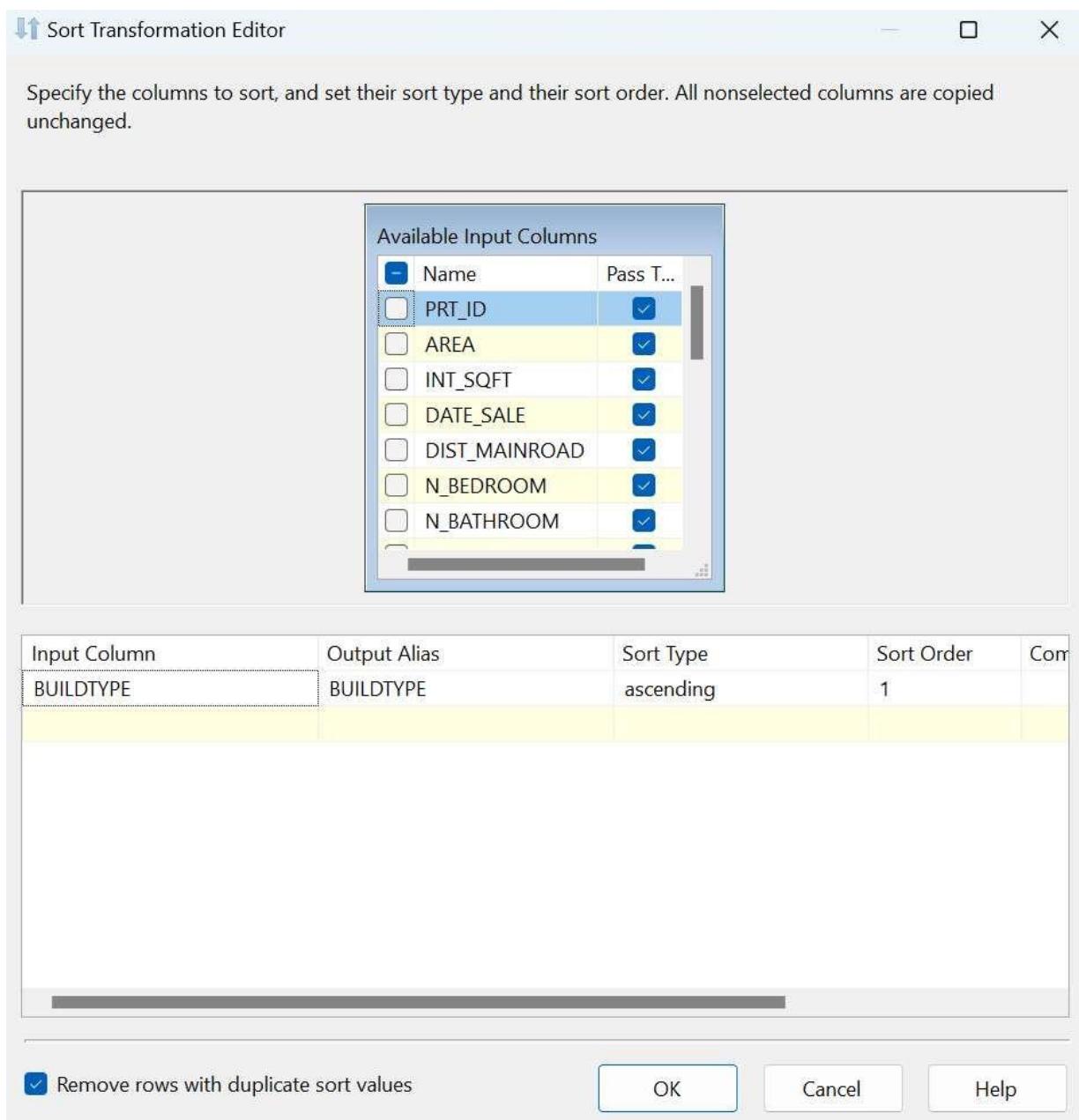
Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull



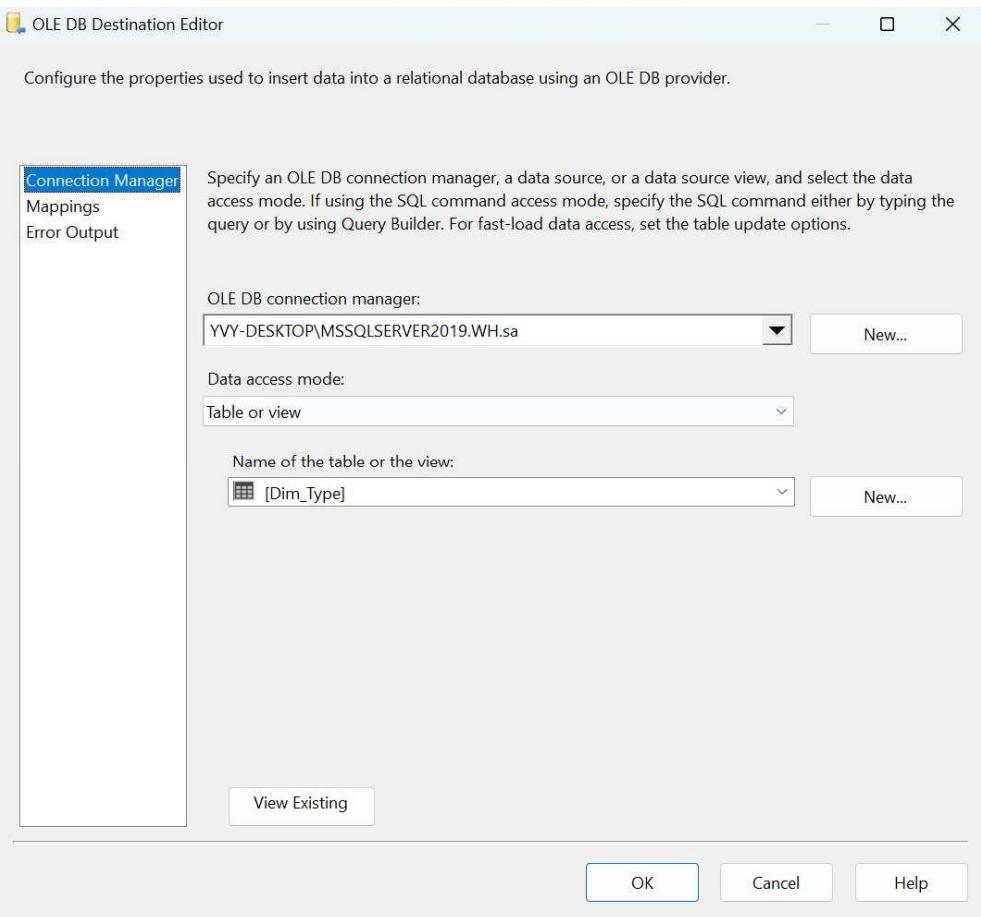
Sort:

- BUILDTYPE là khóa kết
- Tích Remove rows with duplicate sort values

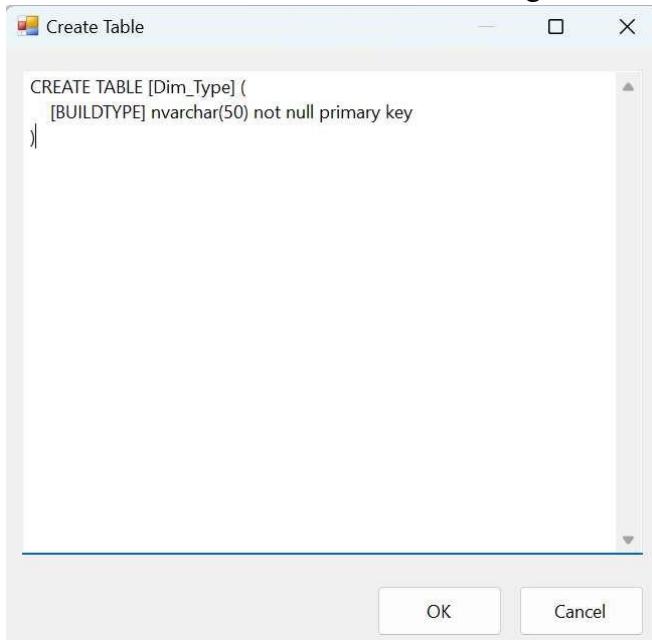


OLE DB Destination:

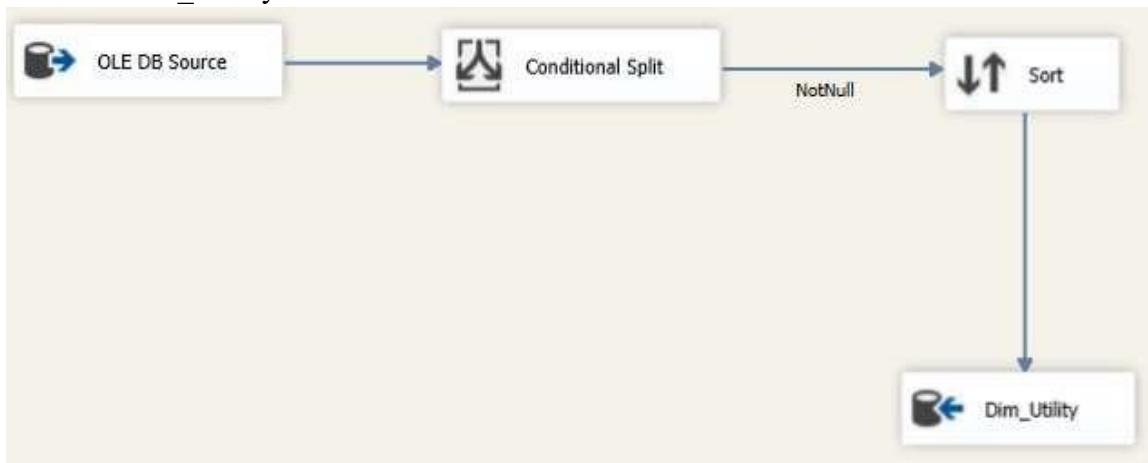
- OLE DB connection manager chọn WH



- Name of the table or the view nhận New để tạo một bảng mới:

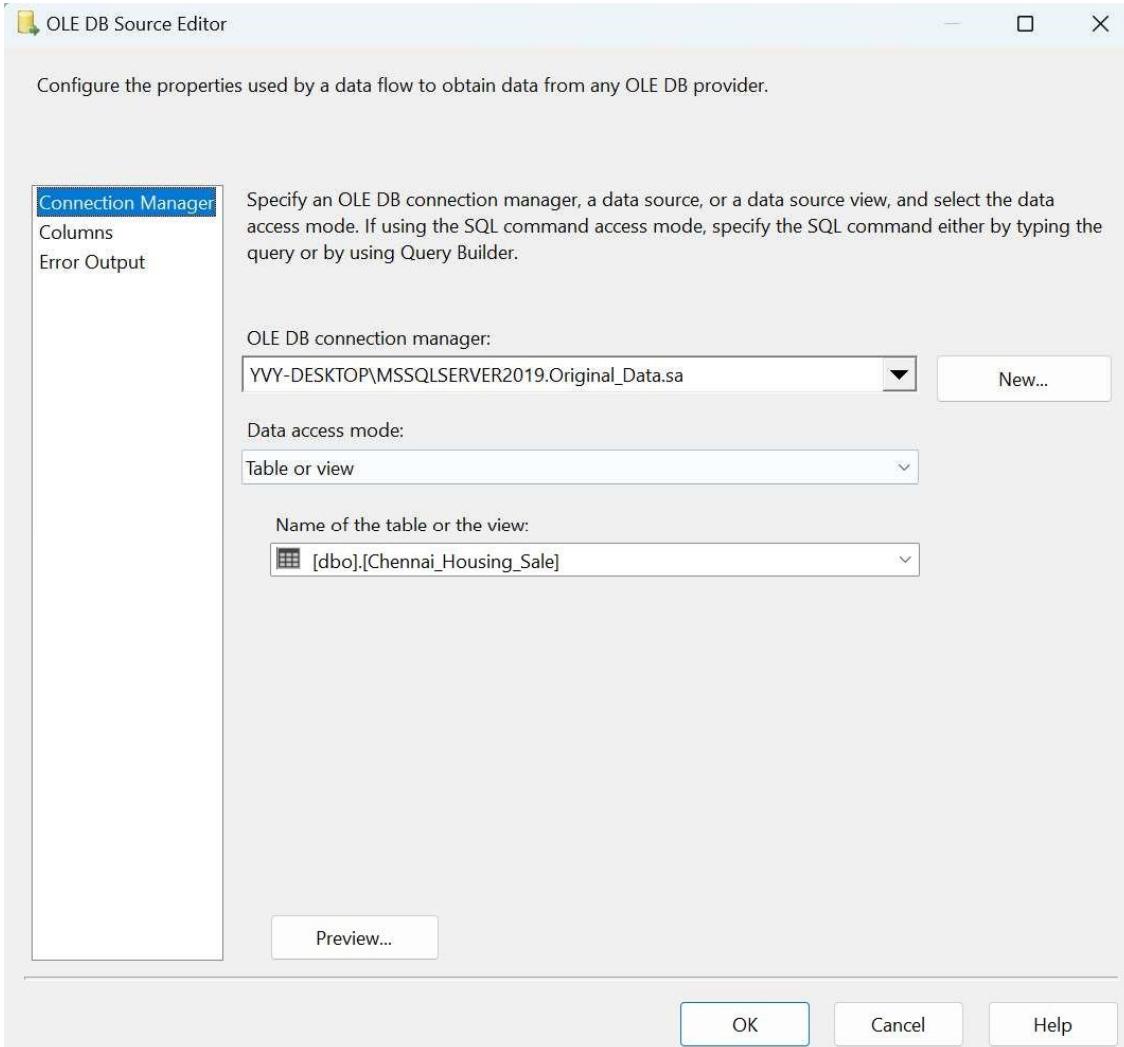


2.4.6. Load Dim_Utility



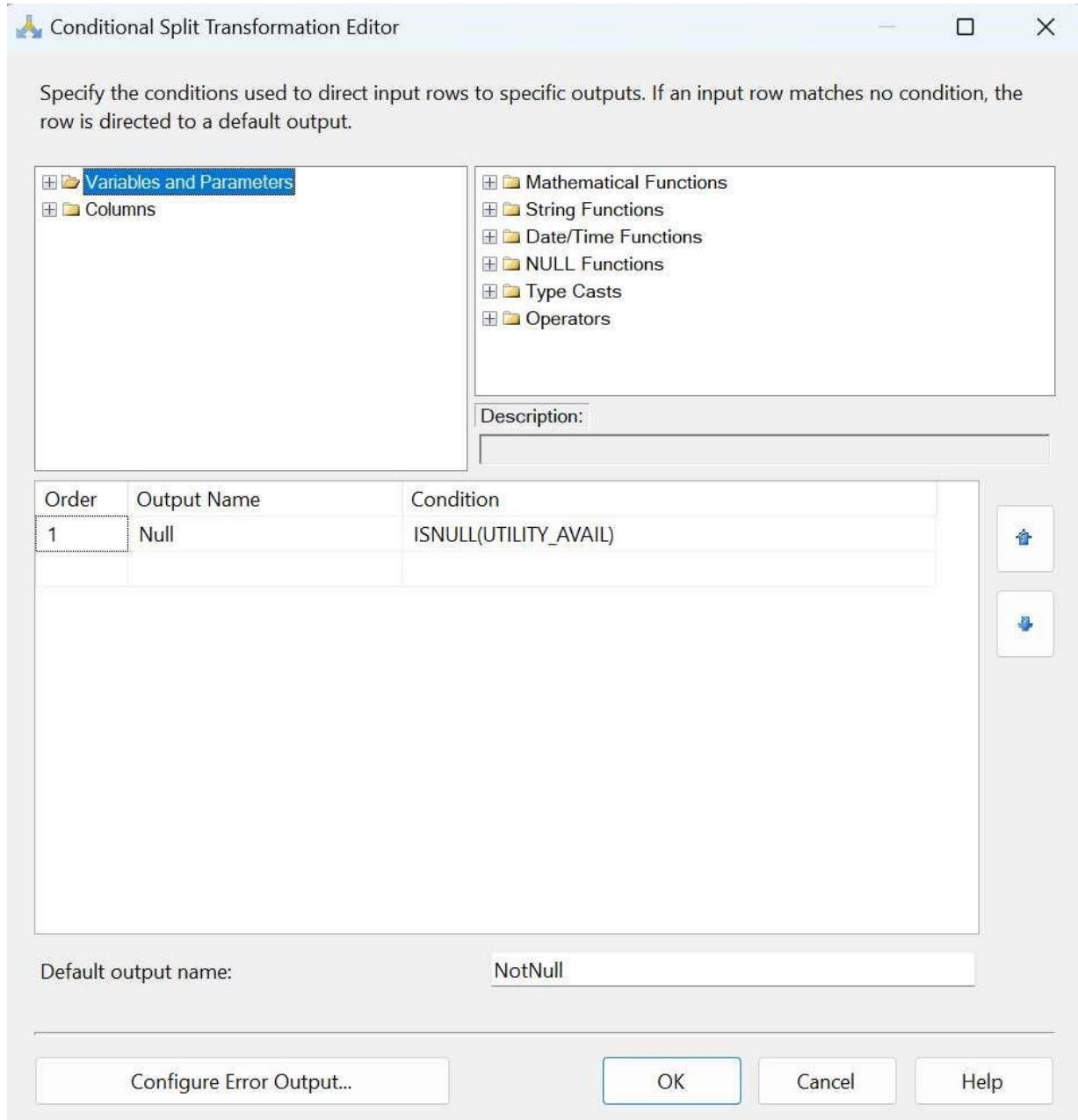
OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



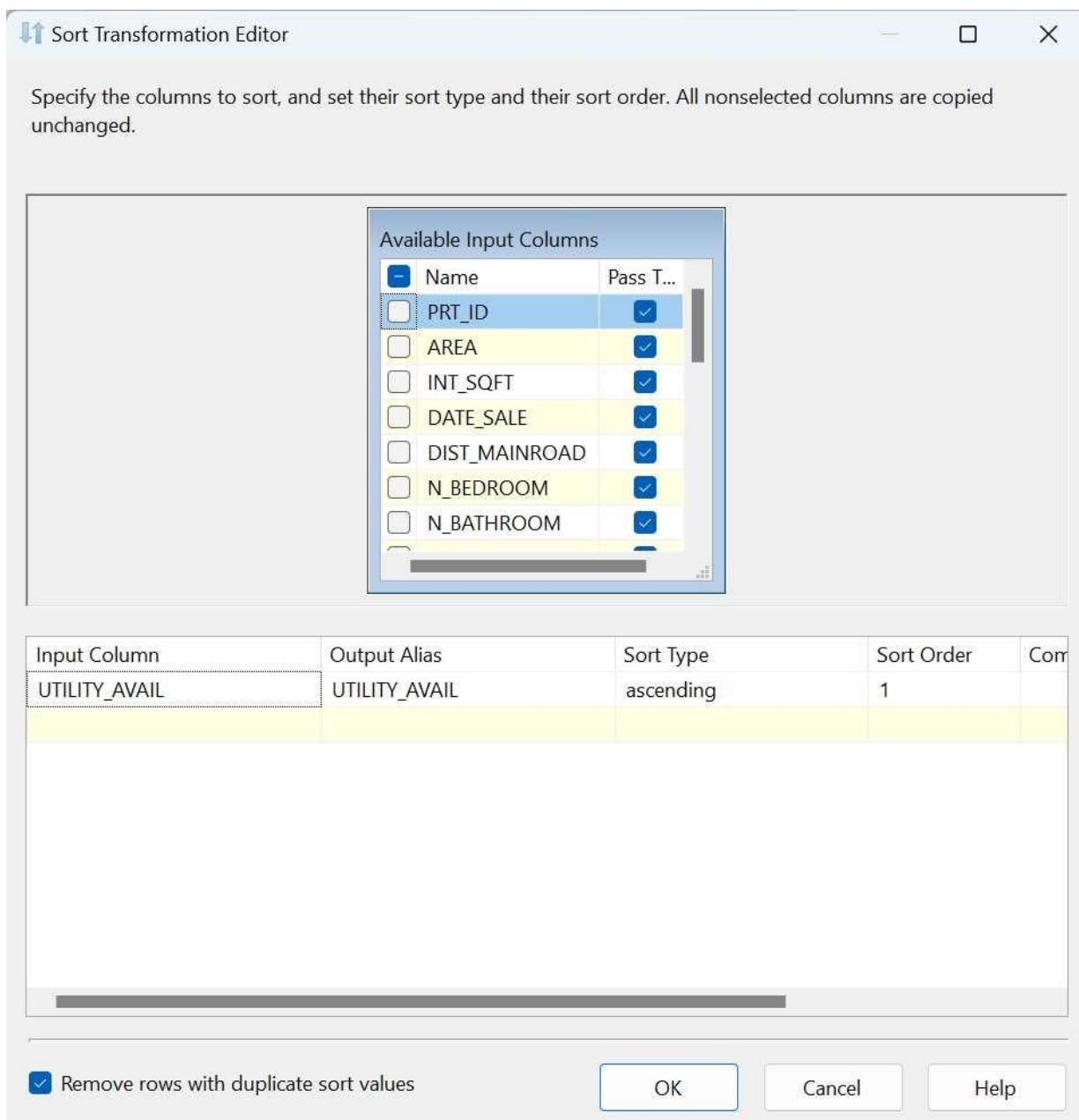
Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull



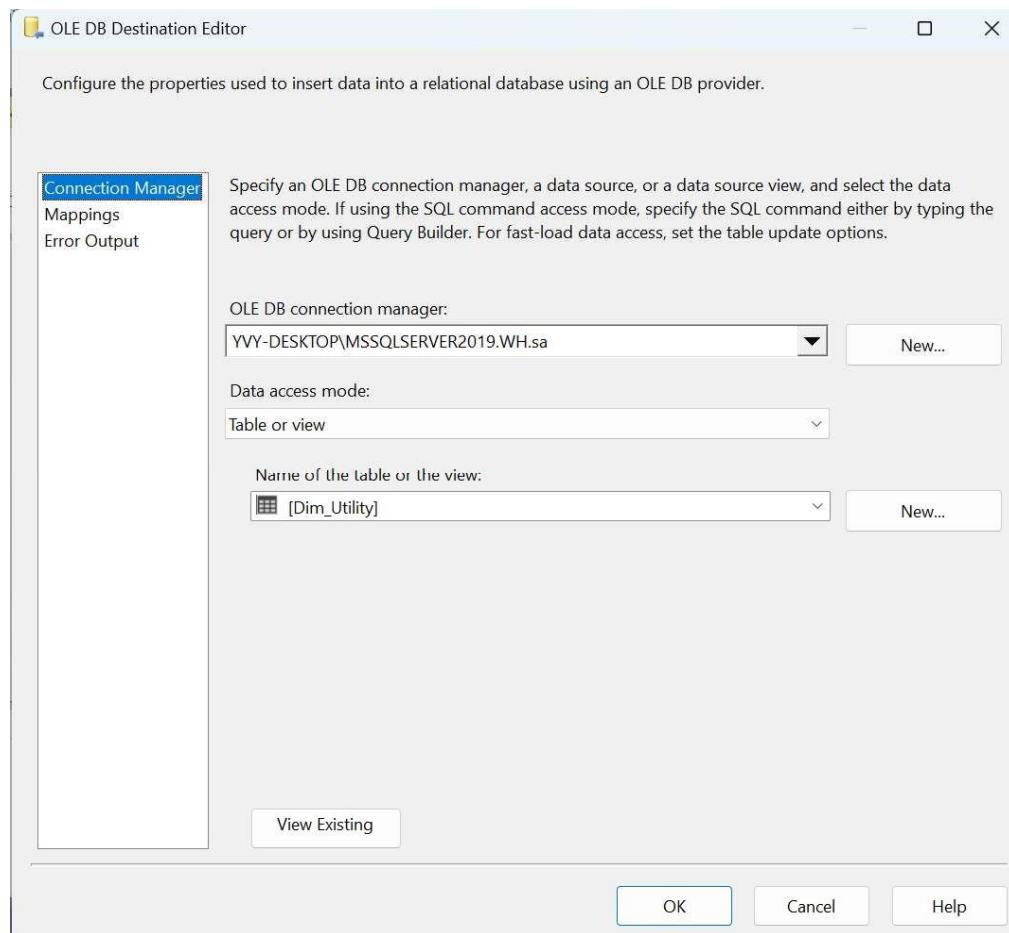
Sort:

- UTILITY_AVAIL là khóa kêt
- Tích Remove rows with duplicate sort values

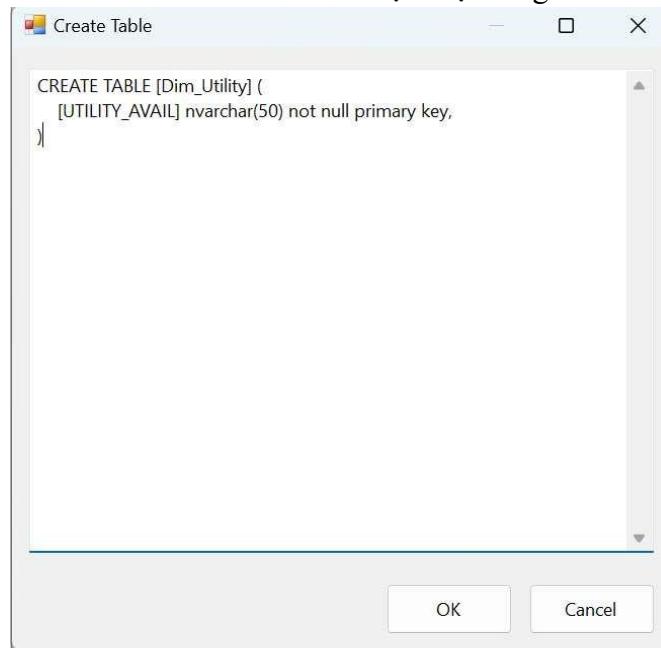


OLE DB Destination:

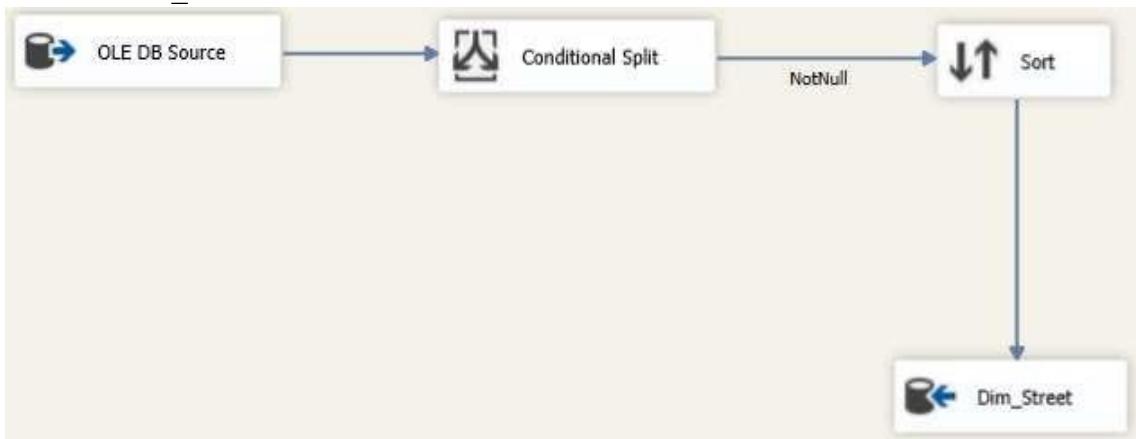
- OLE DB connection manager chọn WH



- Name of the table or the view nhấp New để tạo một bảng mới:

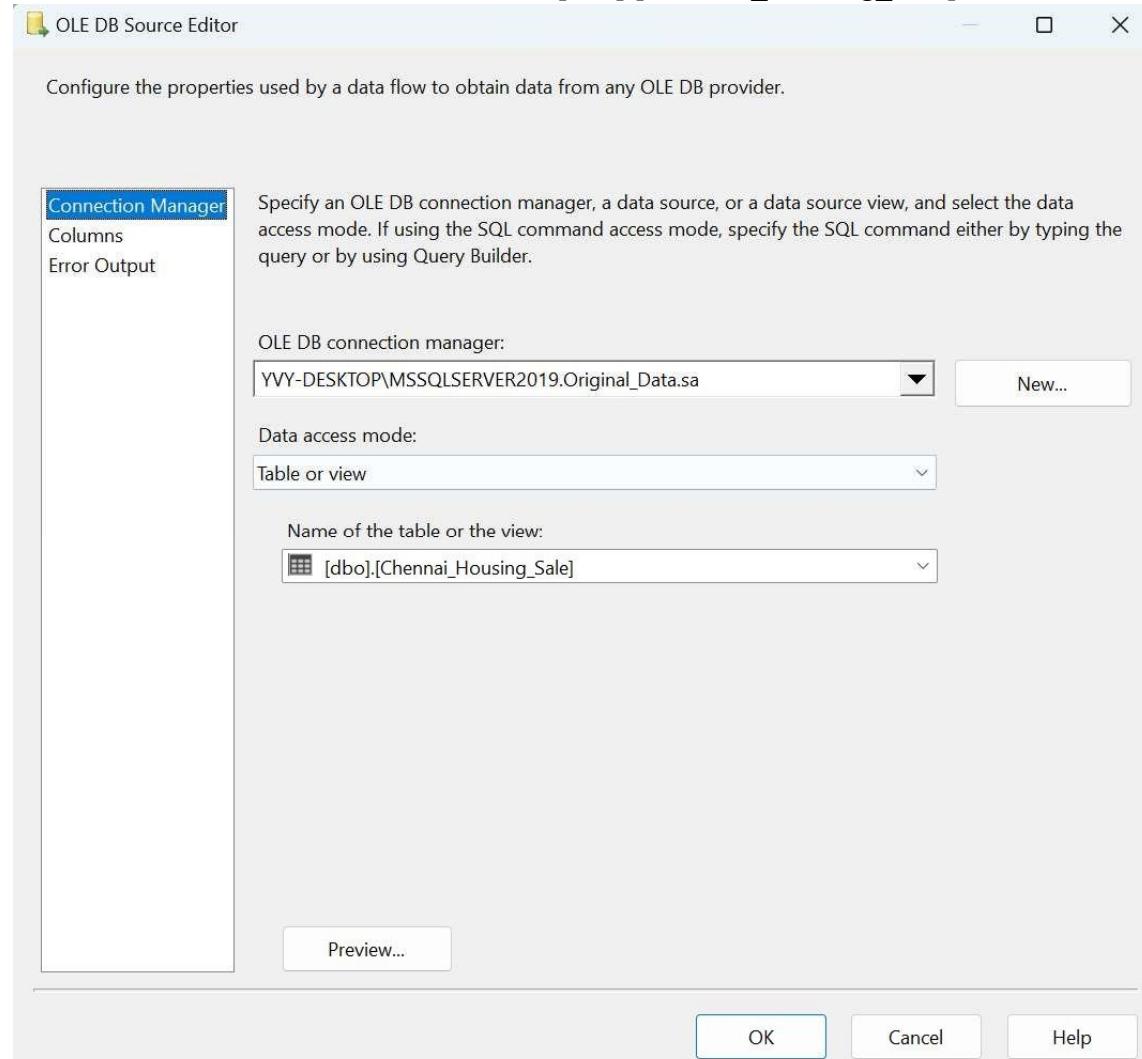


2.4.7. Load Dim_Street



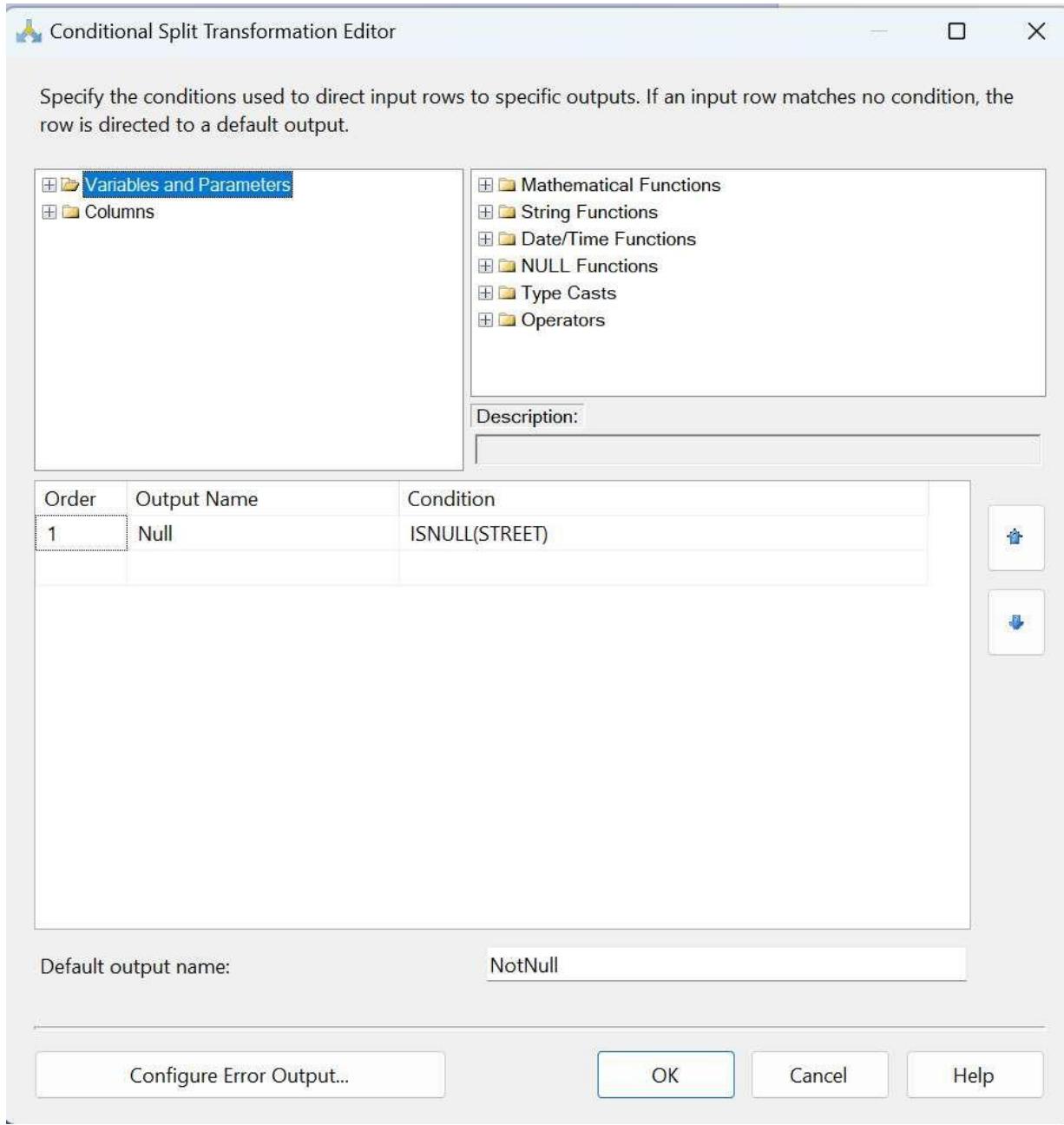
OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



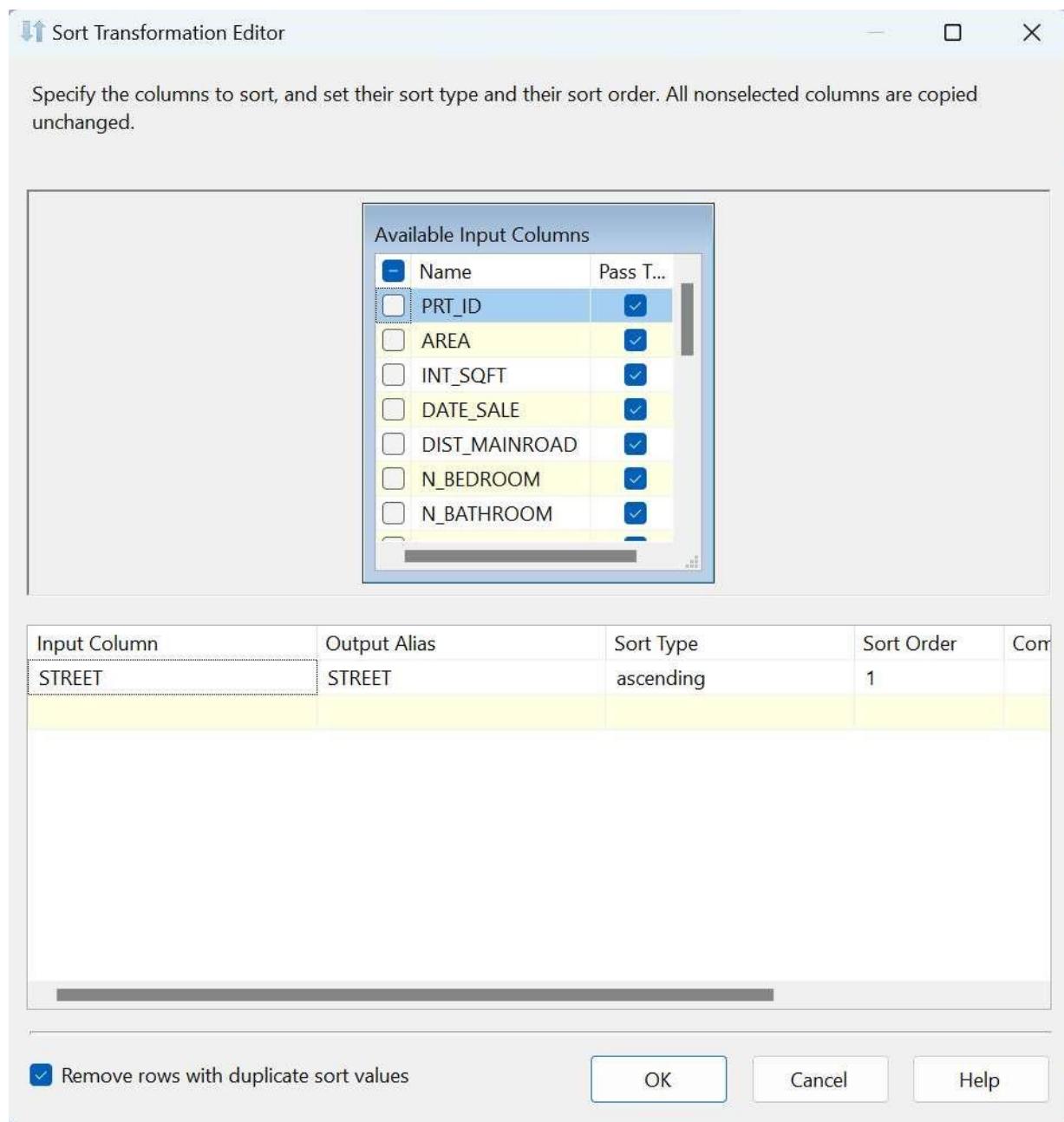
Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull



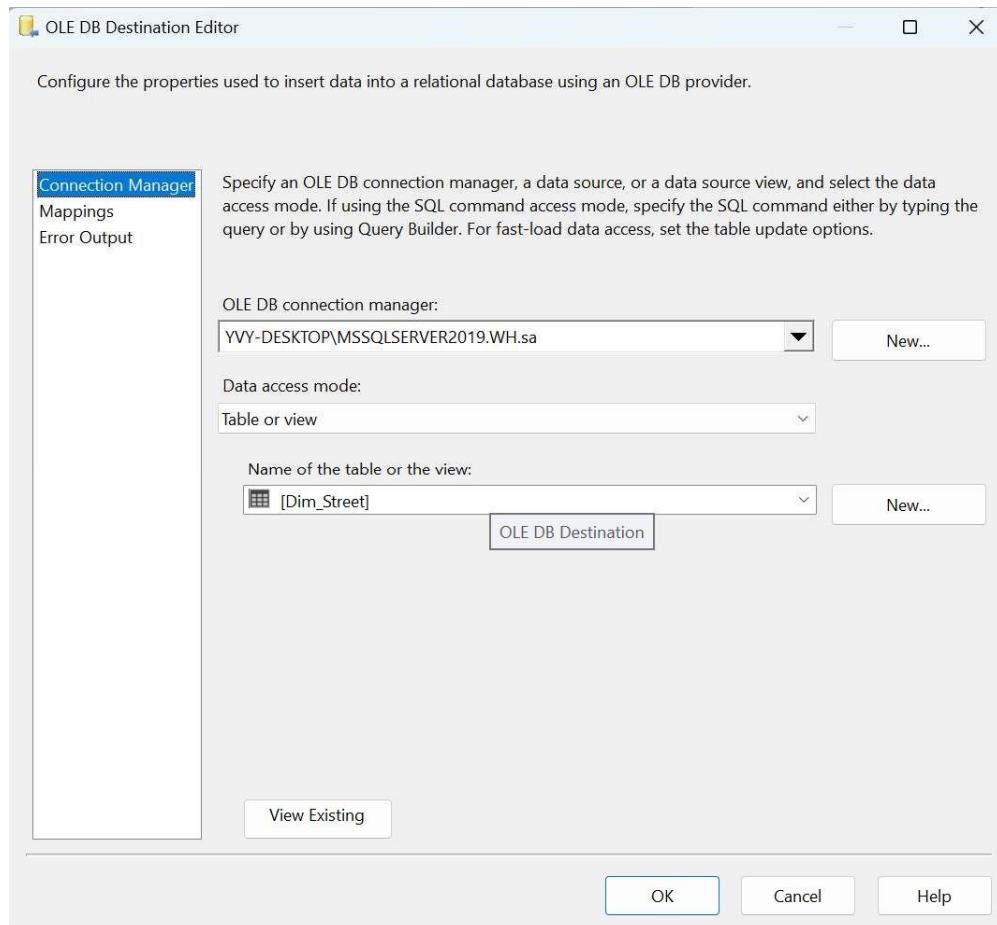
Sort:

- STREET là khóa kết
- Tích Remove rows with duplicate sort values

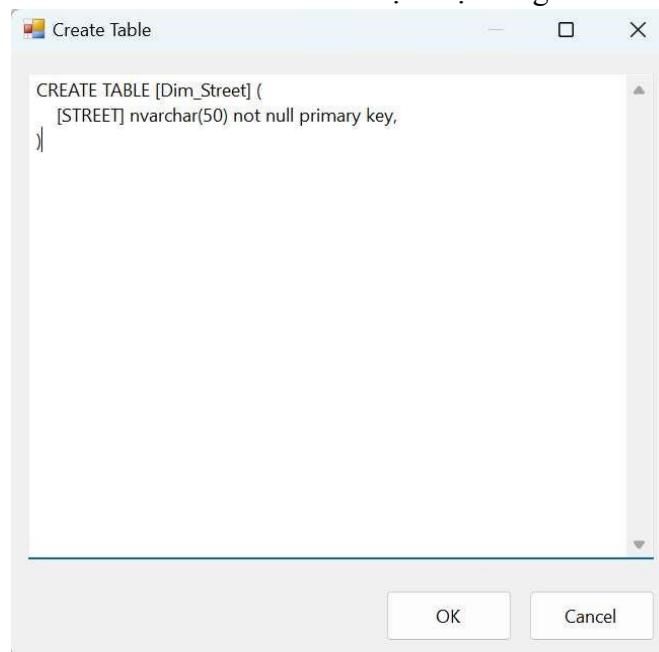


OLE DB Destination:

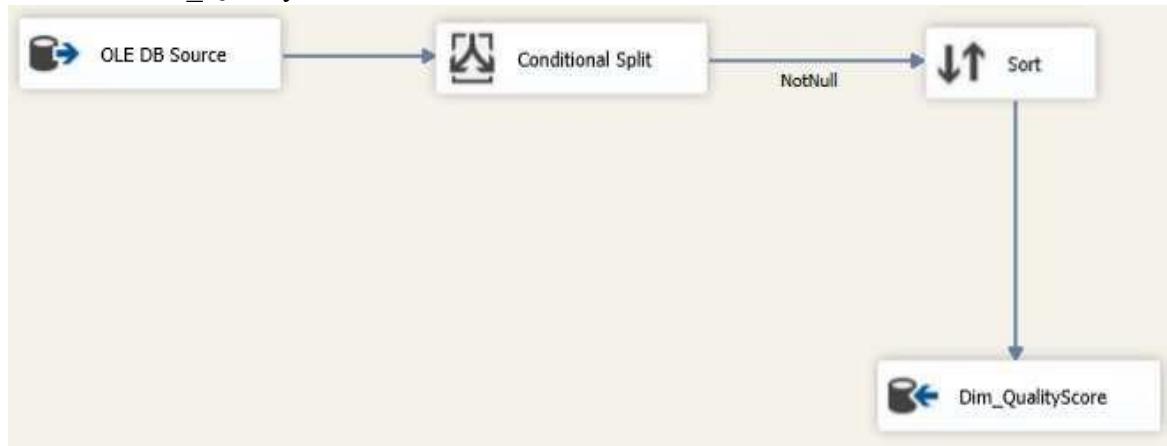
- OLE DB connection manager chọn WH



- Name of the table or the view nhấp New để tạo một bảng mới:

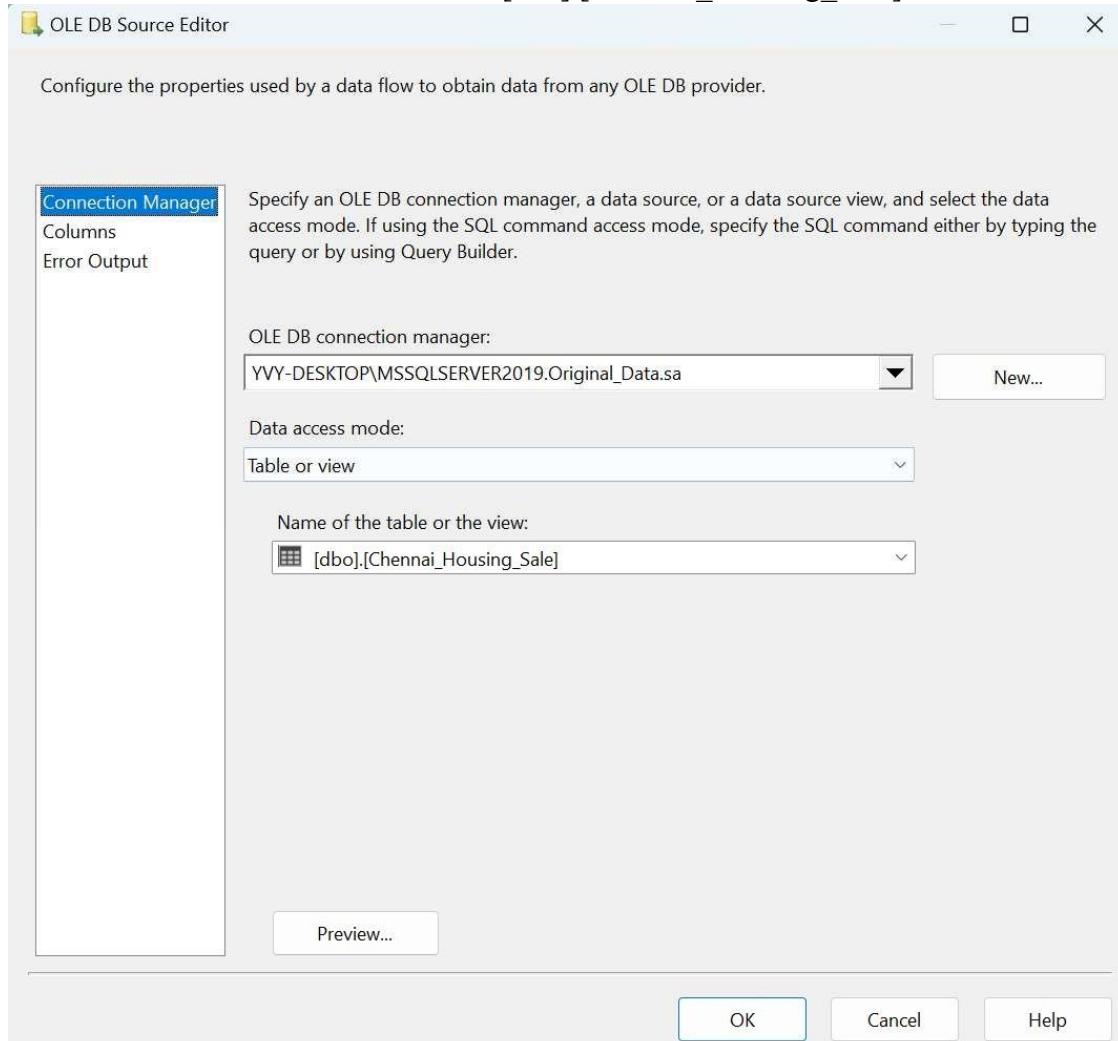


2.4.8. Load Dim_QualityScore



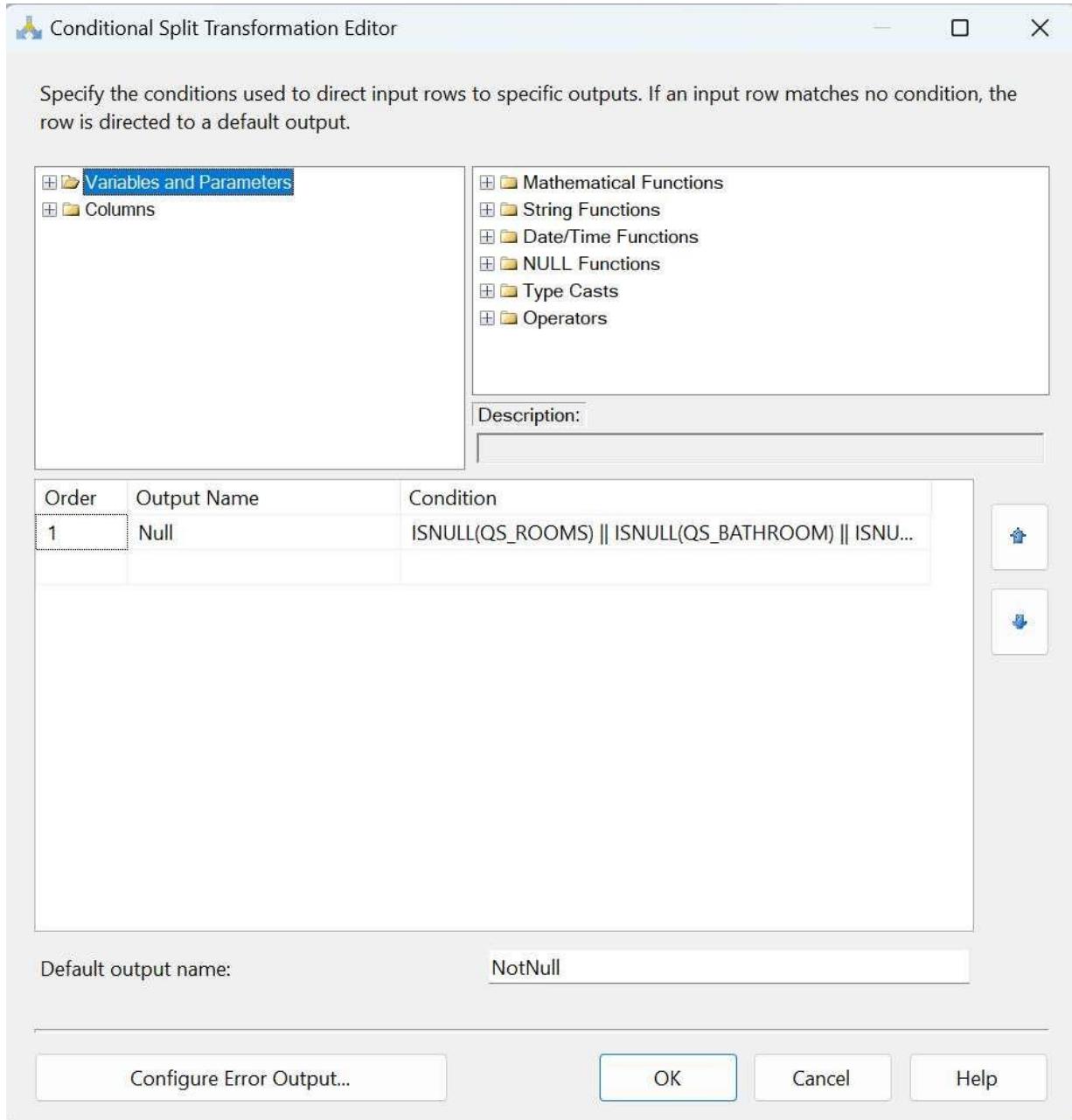
OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



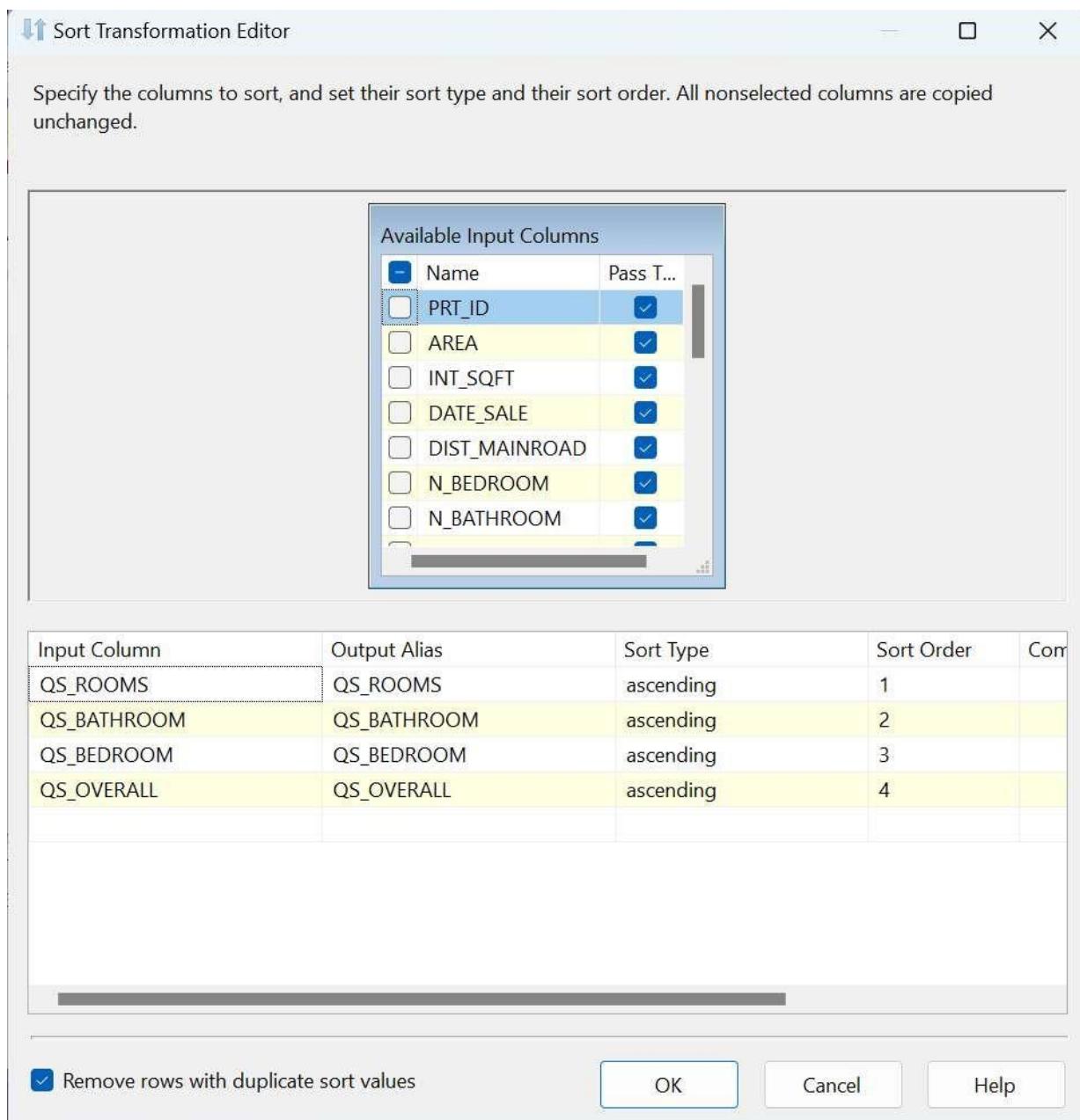
Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull



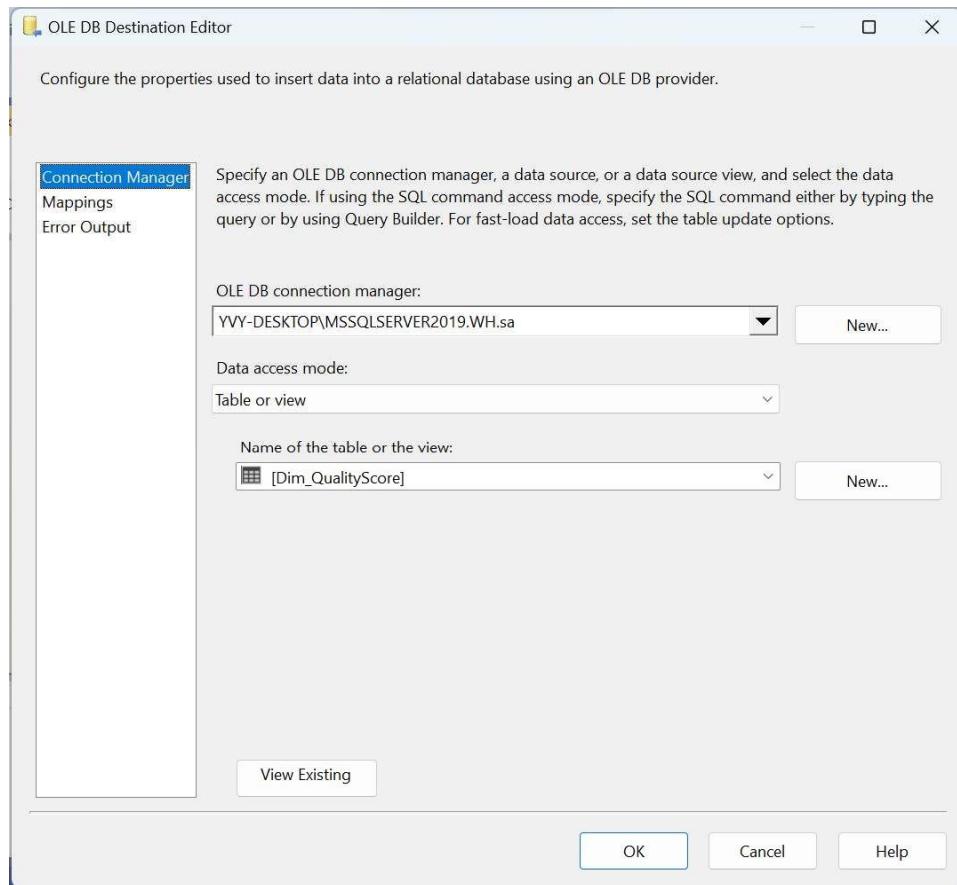
Sort:

- QS_BEDROOM, QS_BATHROOM, QS_ROOMS, QS_OVERALL là khóa kết
- Tích Remove rows with duplicate sort values

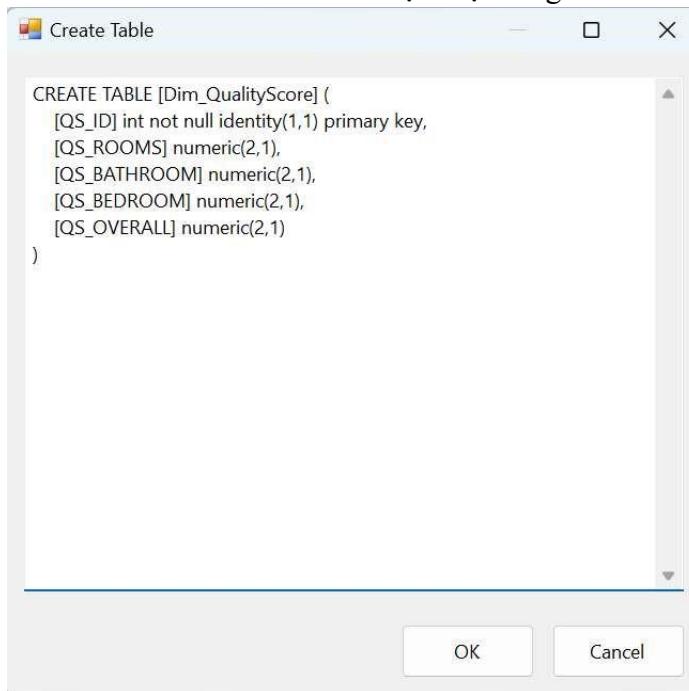


OLE DB Destination:

- OLE DB connection manager chọn WH



- Name of the table or the view nhấn New để tạo một bảng mới:

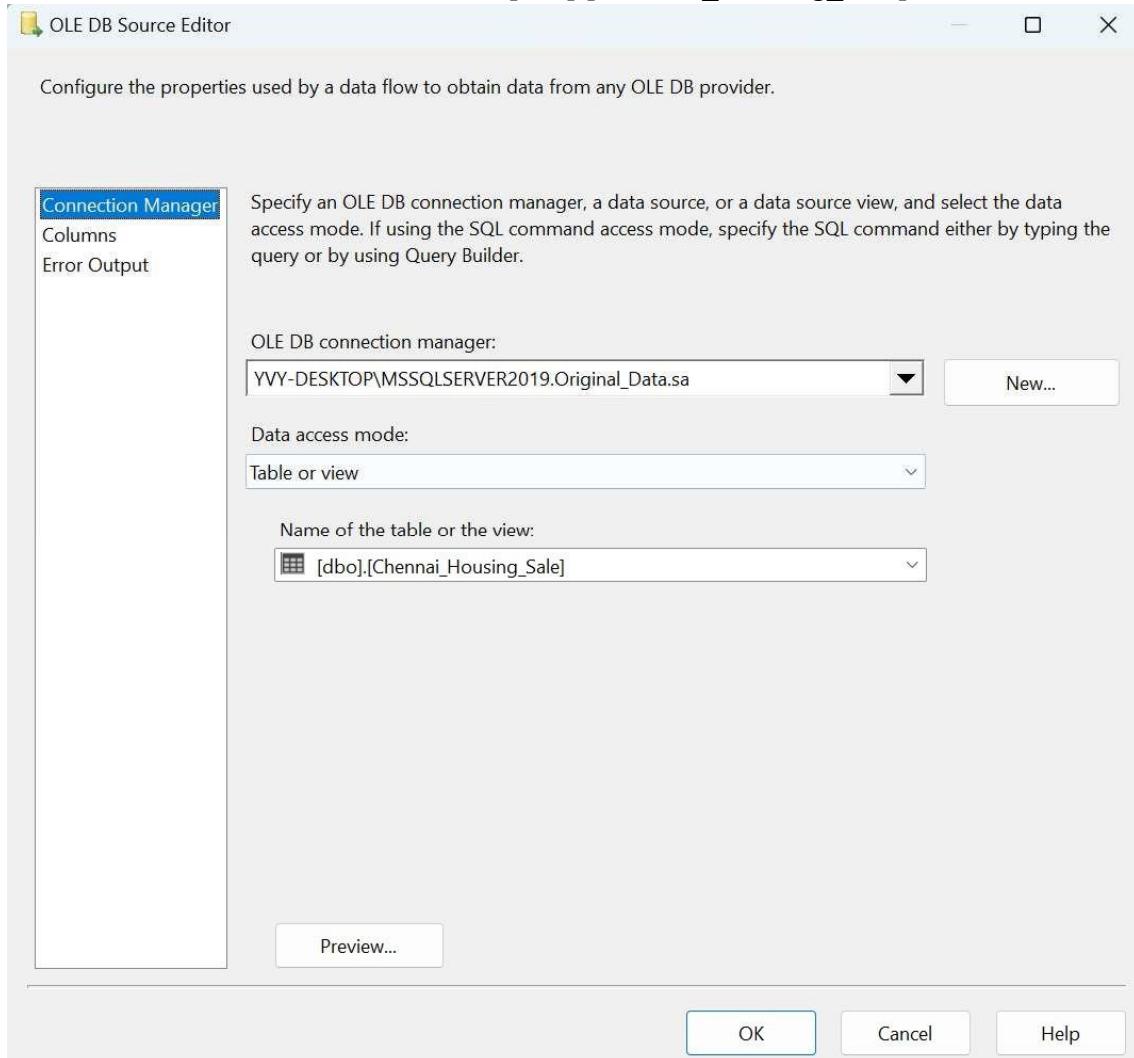


2.4.9. Load Dim_Condition



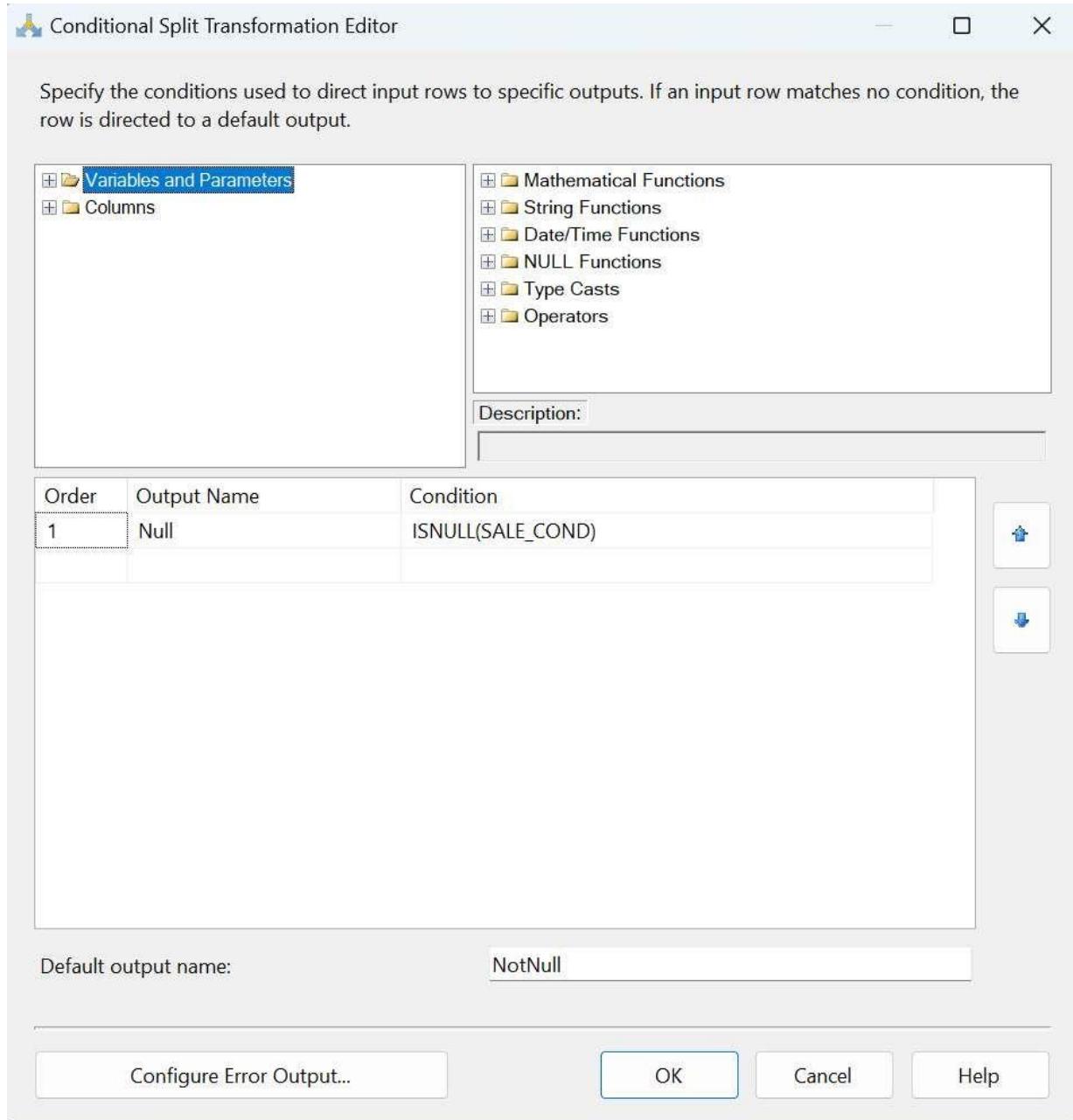
OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].



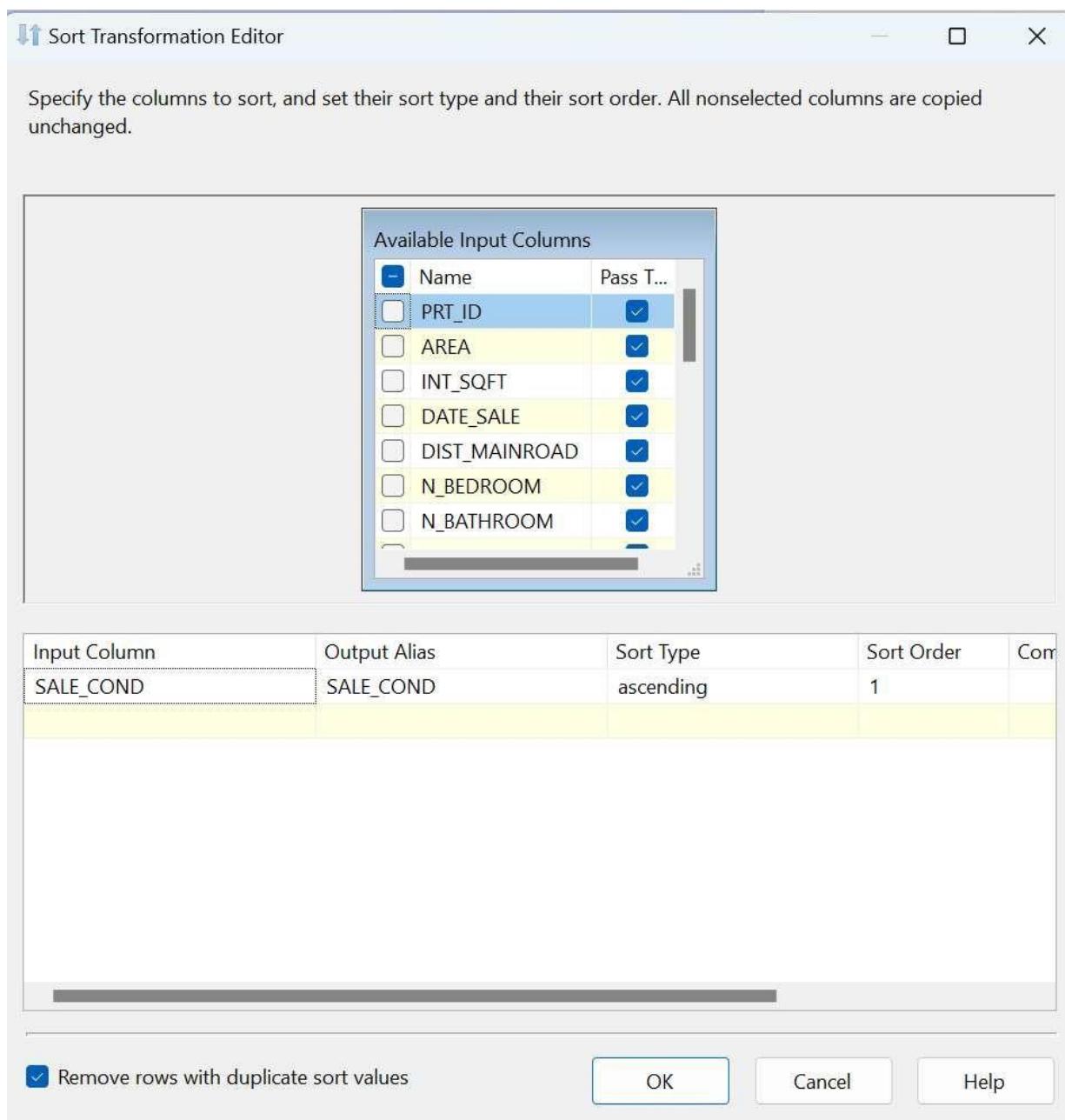
Conditional Split:

- Output Name: Null
- Condition: Dùng ISNULL để lọc ra những dòng dữ liệu NULL
- Default output name: NotNull



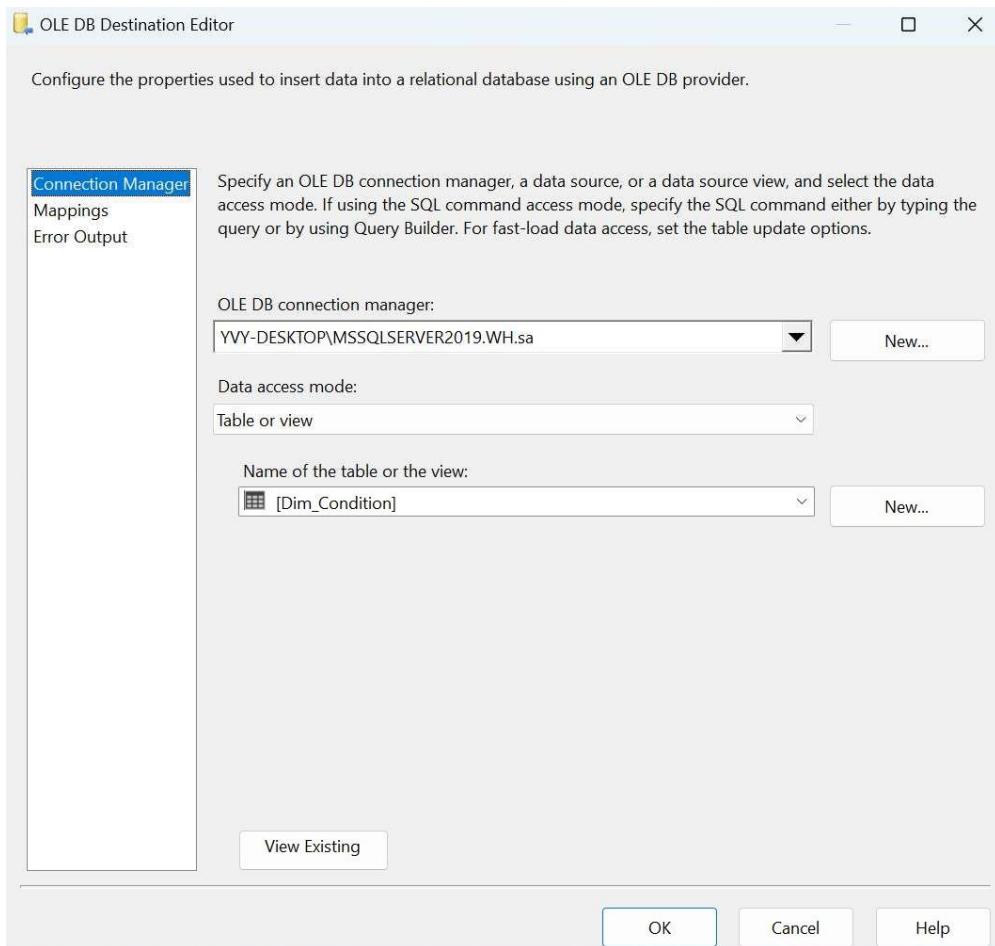
Sort:

- SALE_COND là khóa kết
- Tích Remove rows with duplicate sort values

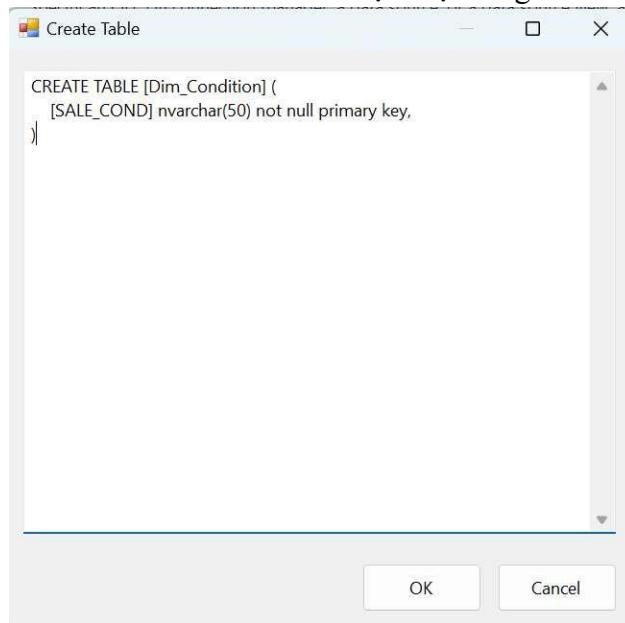


OLE DB Destination:

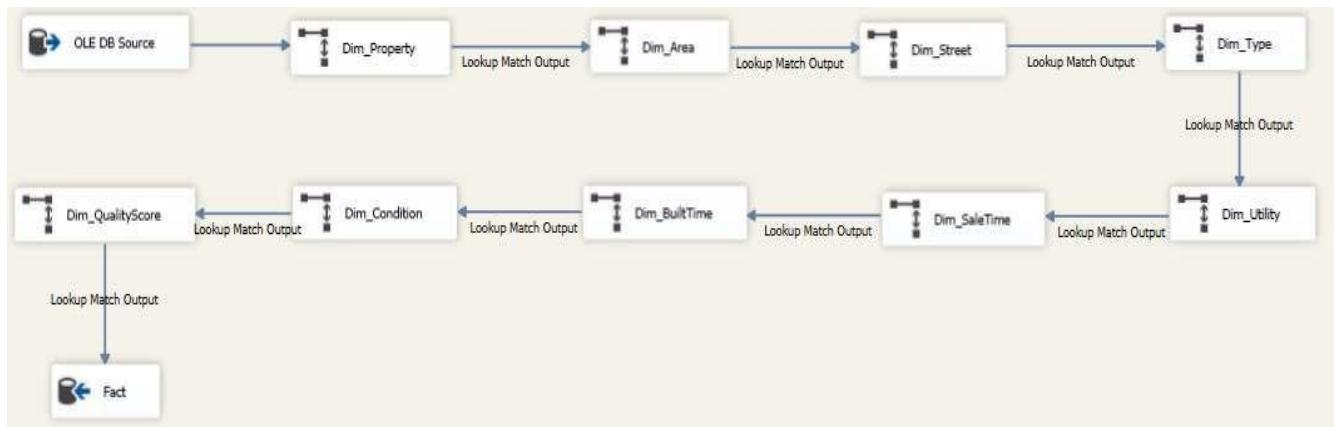
- OLE DB connection manager chọn WH



- Name of the table or the view nhận New để tạo một bảng mới:

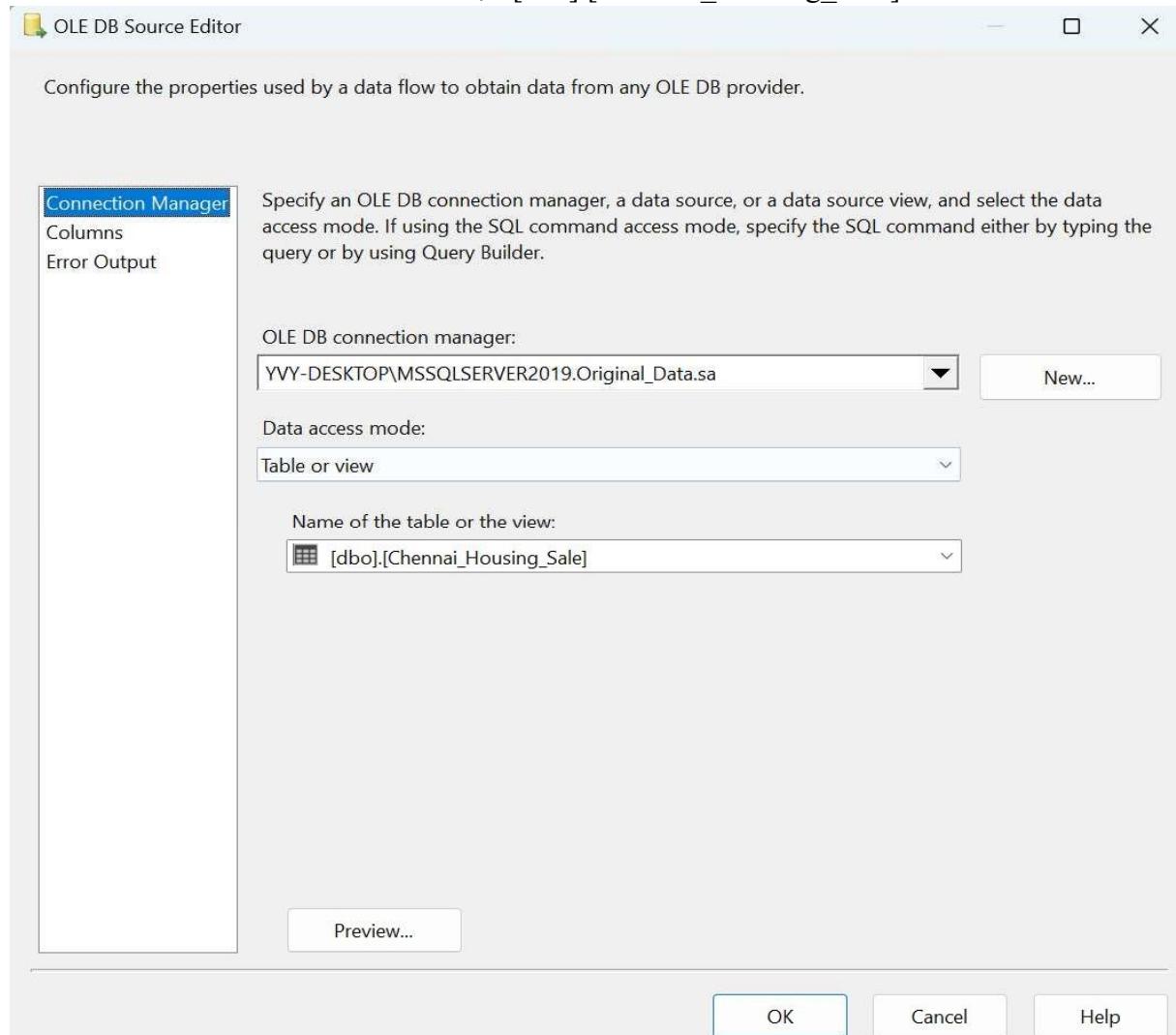


2.5. Load Fact Table



OLE DB Source:

- OLE DB connection manager chọn Original_Data.
- Name of the table or the view chọn [dbo].[Chennai_Housing_Sale].

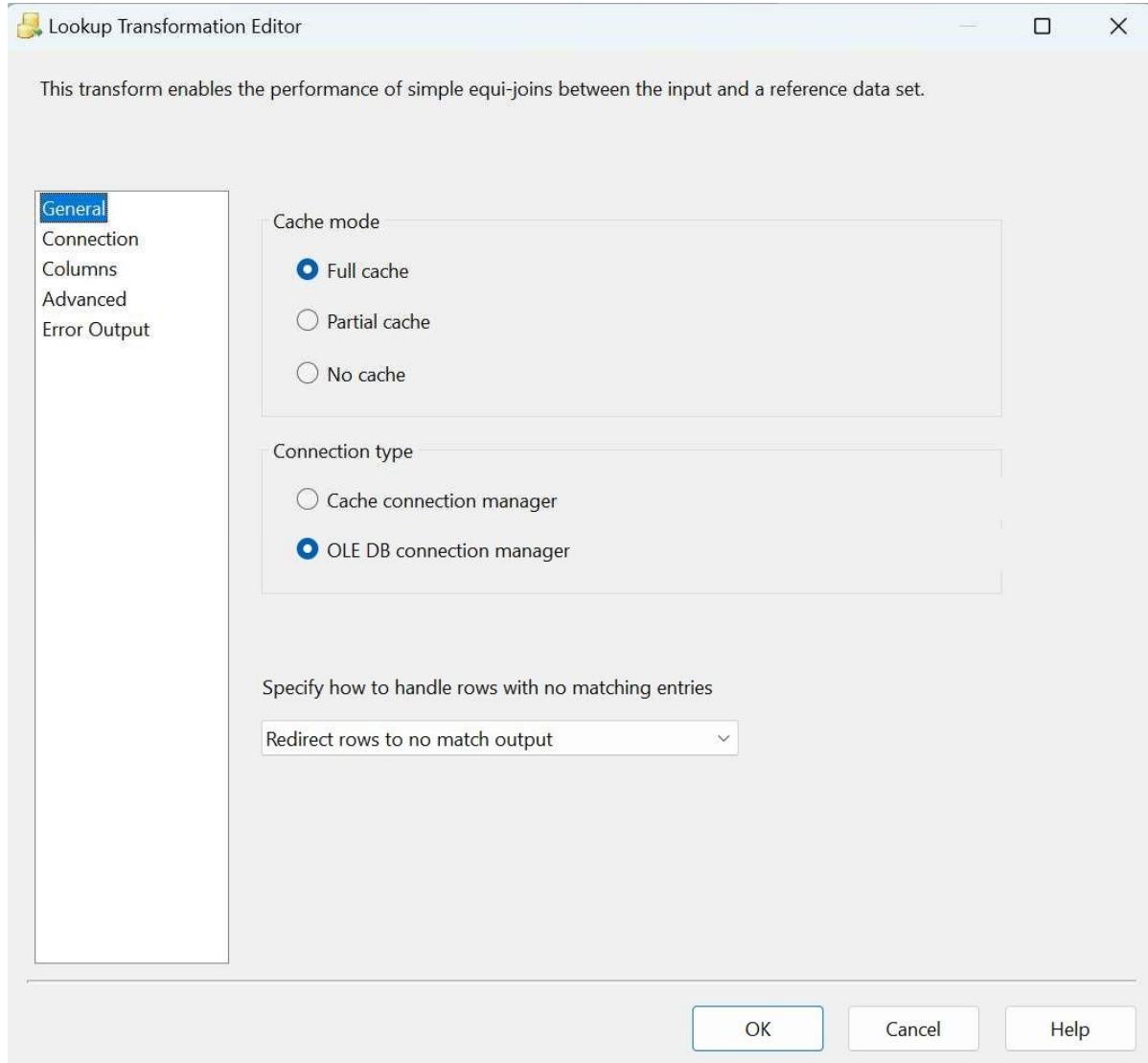


Lookup Transformation Editor:

❖ Dim_Property - Tại

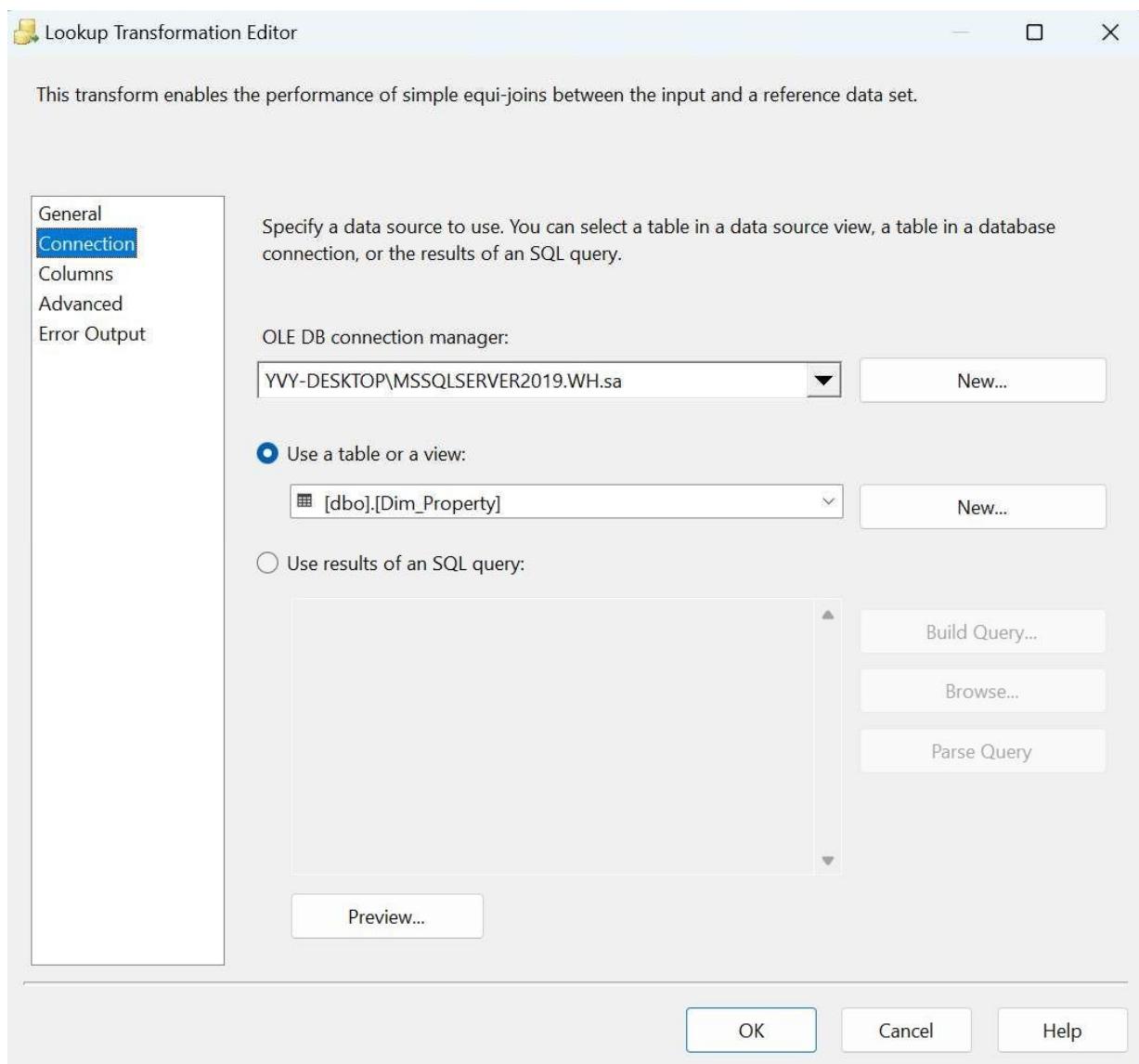
tab General:

- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.

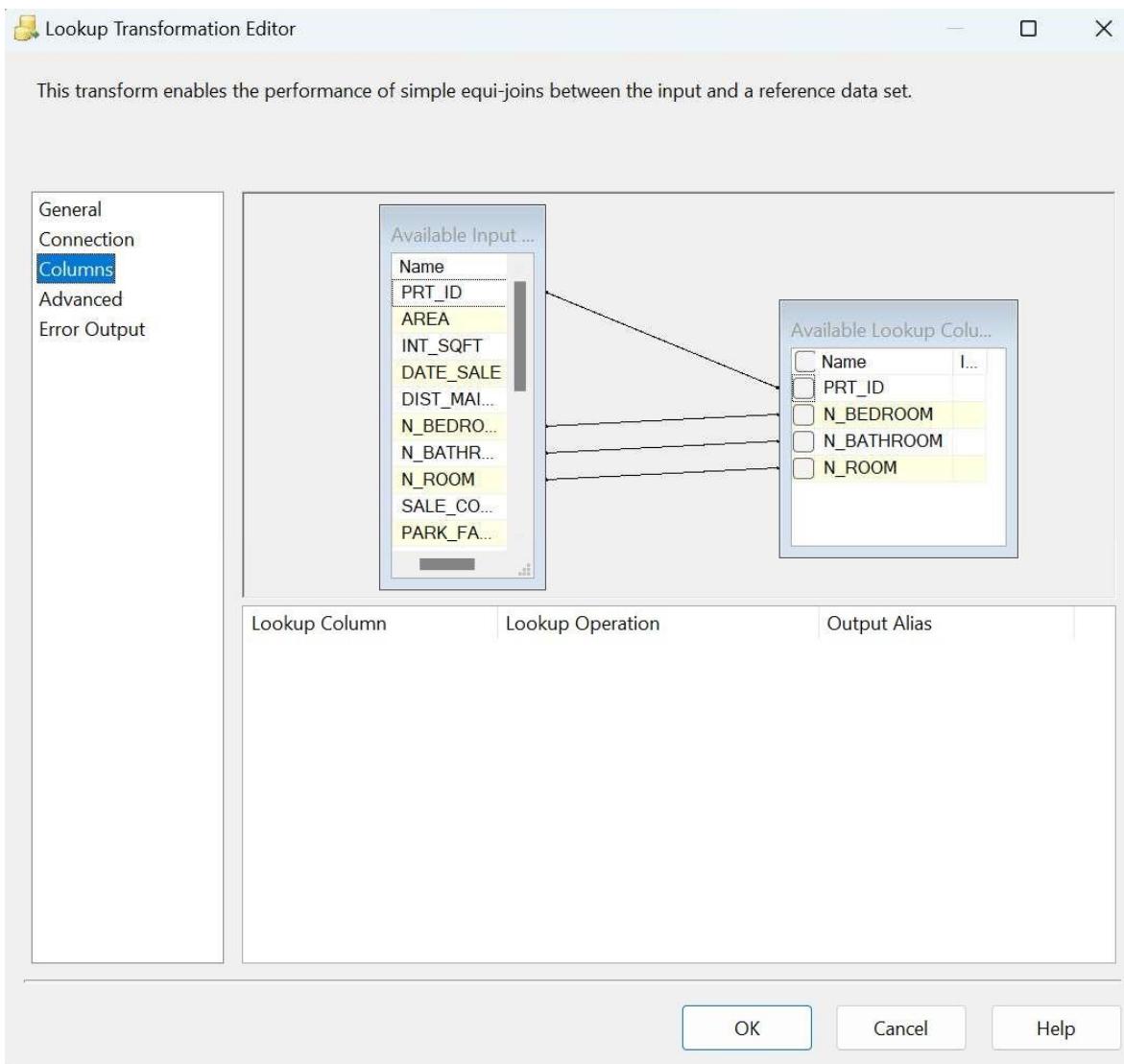


- Tại tab Connection:

- OLE DB connection manager chọn WH.
- Name of the table or the view chọn [dbo].[Dim_Property]



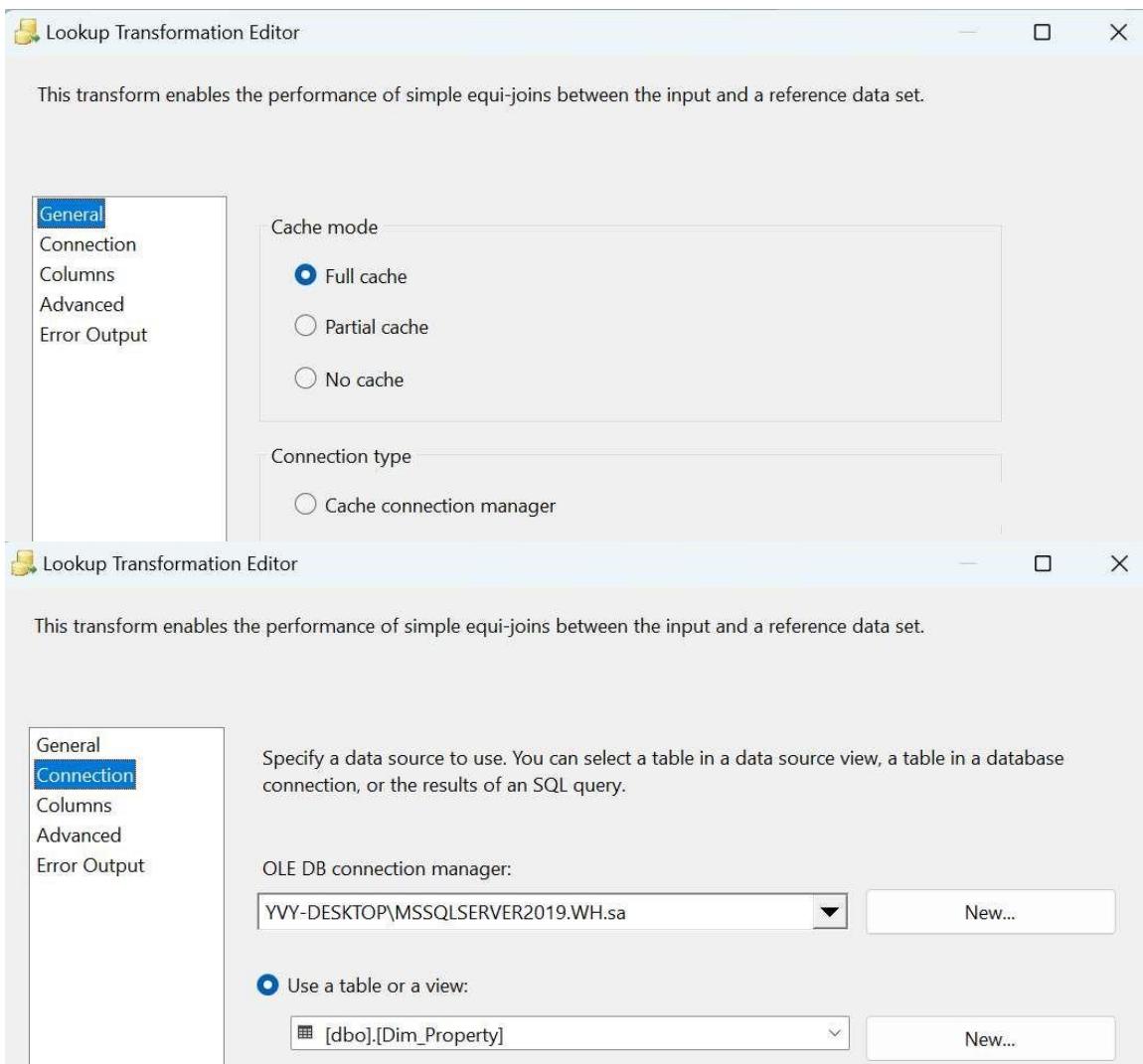
- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.



❖ Dim_Area - Tại tab

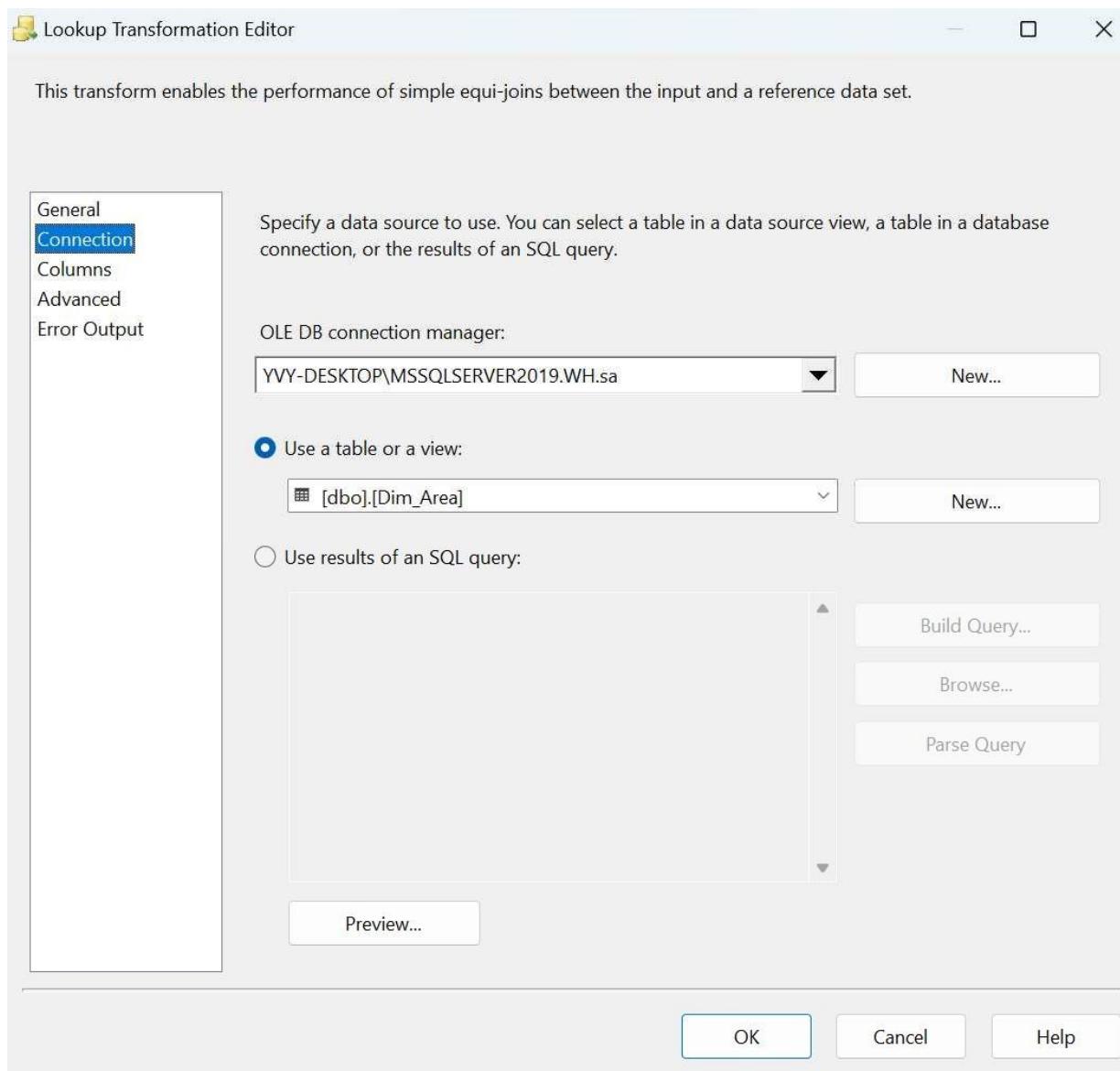
General:

- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.

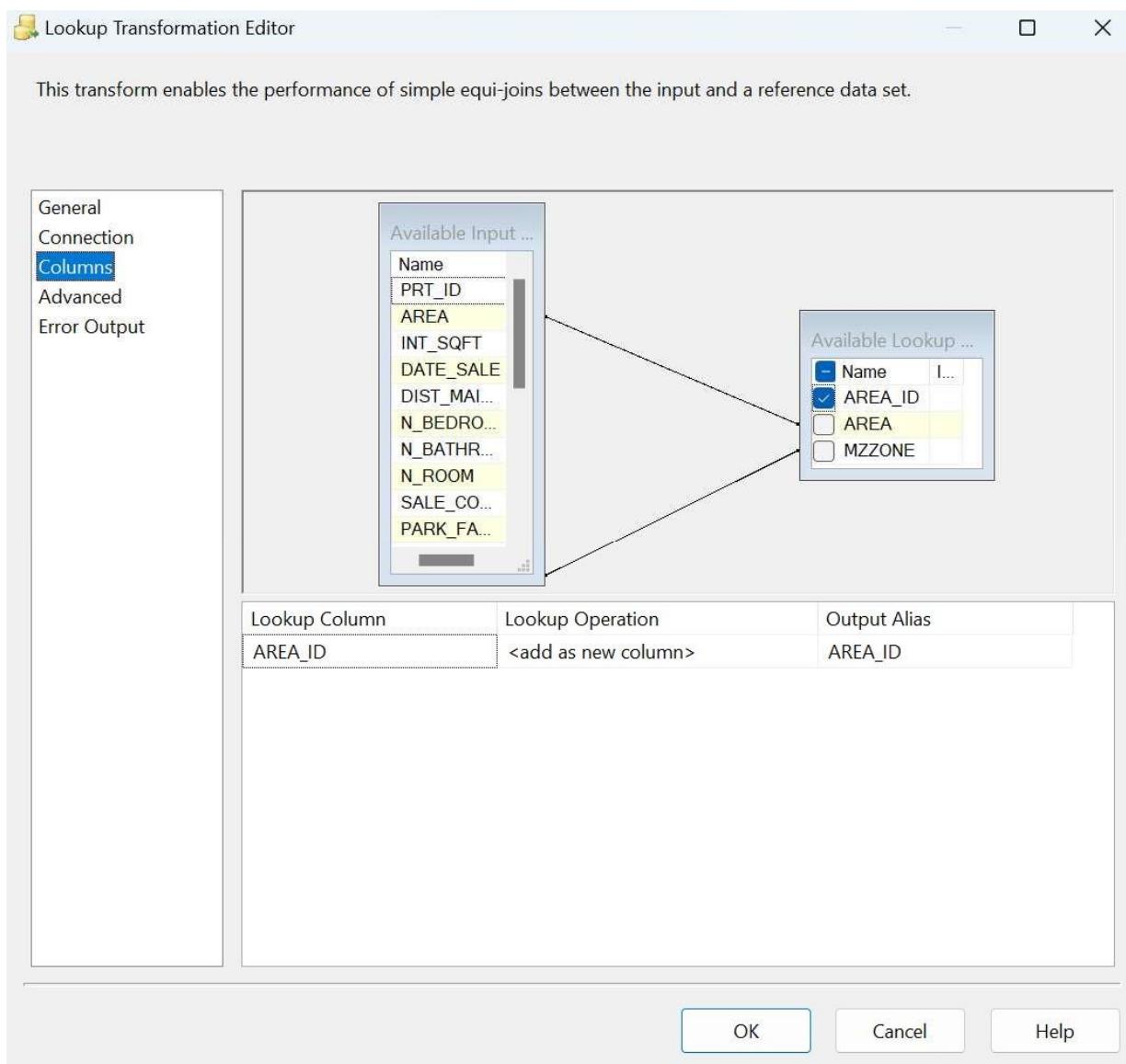


- Tại tab Connection:

- OLE DB connection manager chọn WH.
- Name of the table or the view chọn [dbo].[Dim_Area]



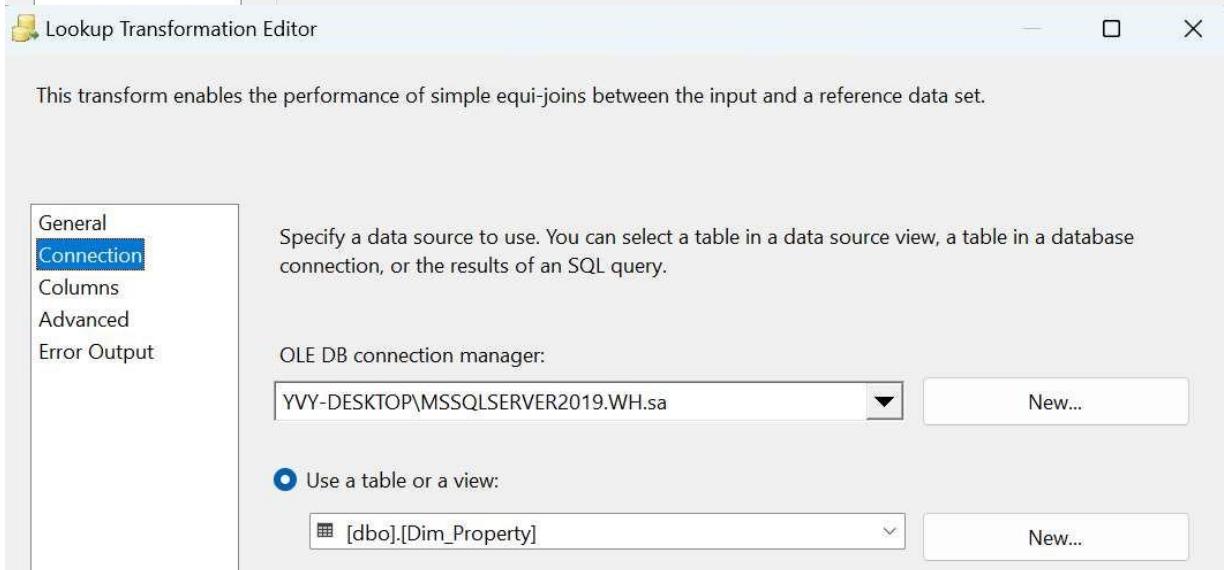
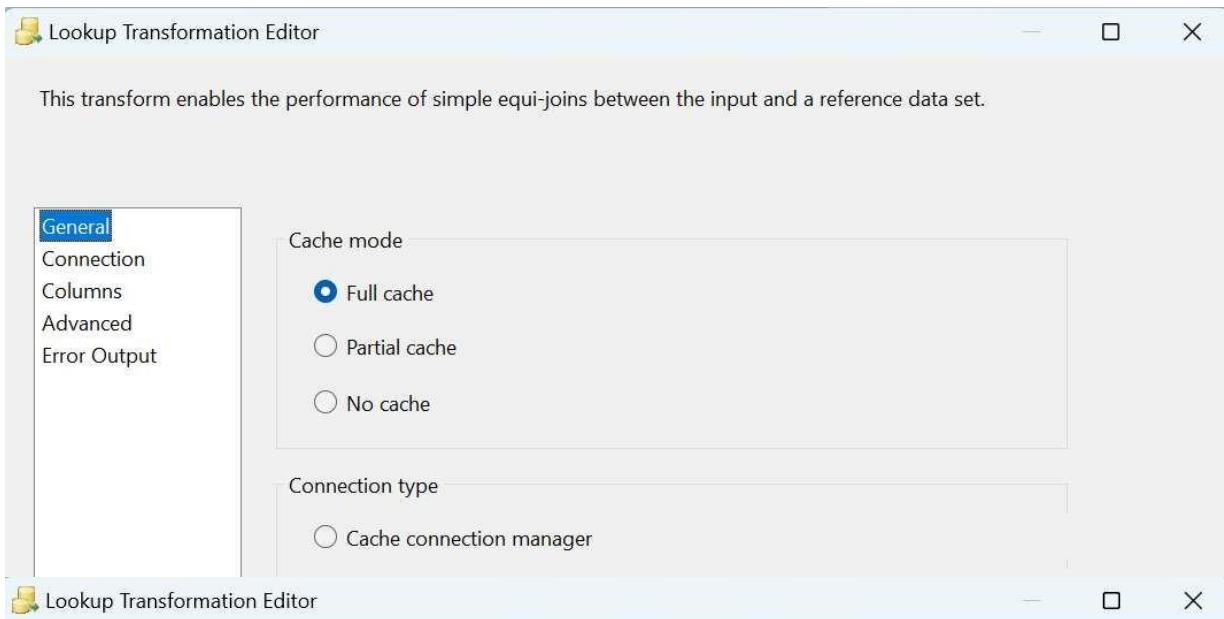
- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.



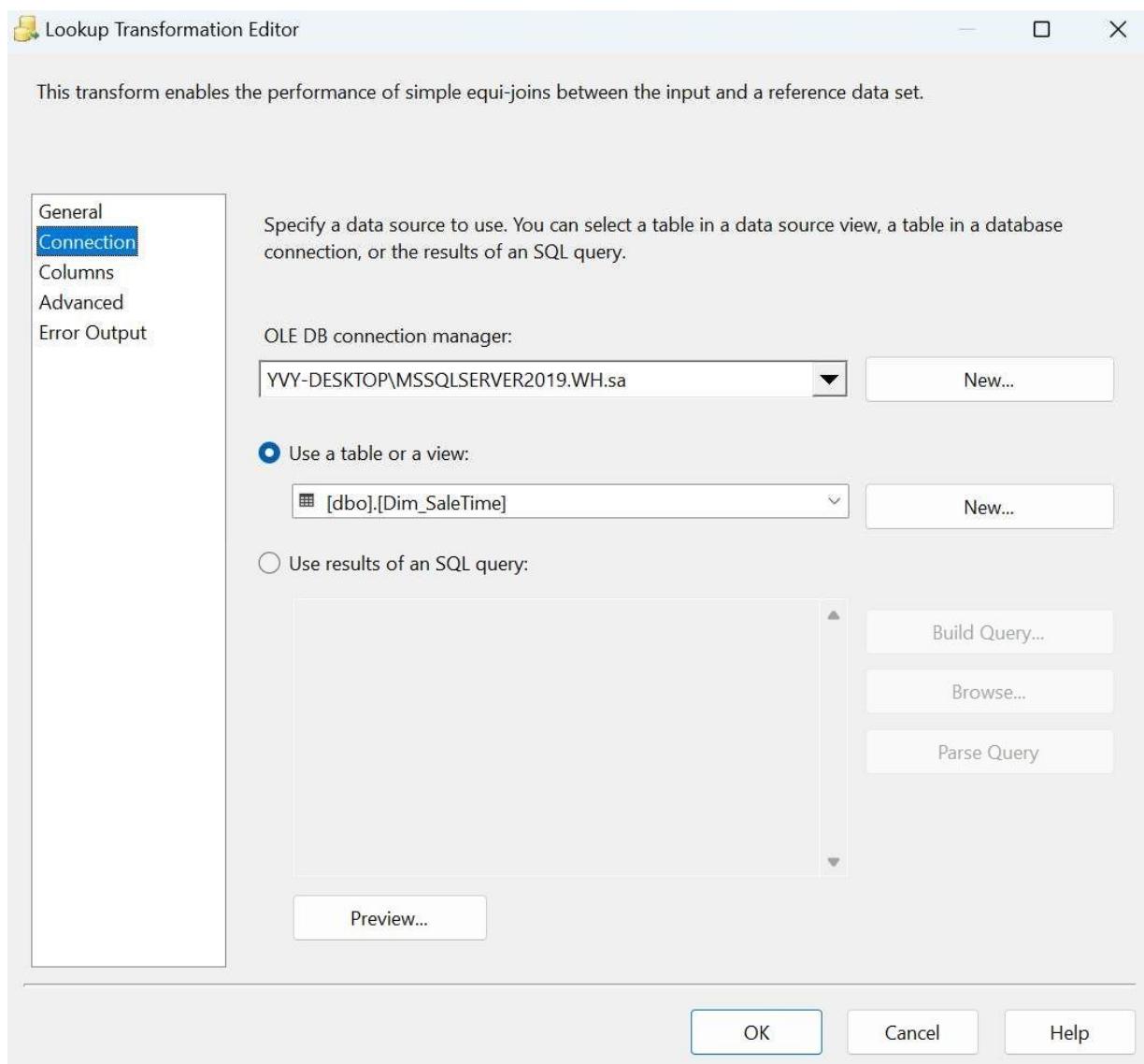
❖ Dim_SaleTime - Tại

tab General:

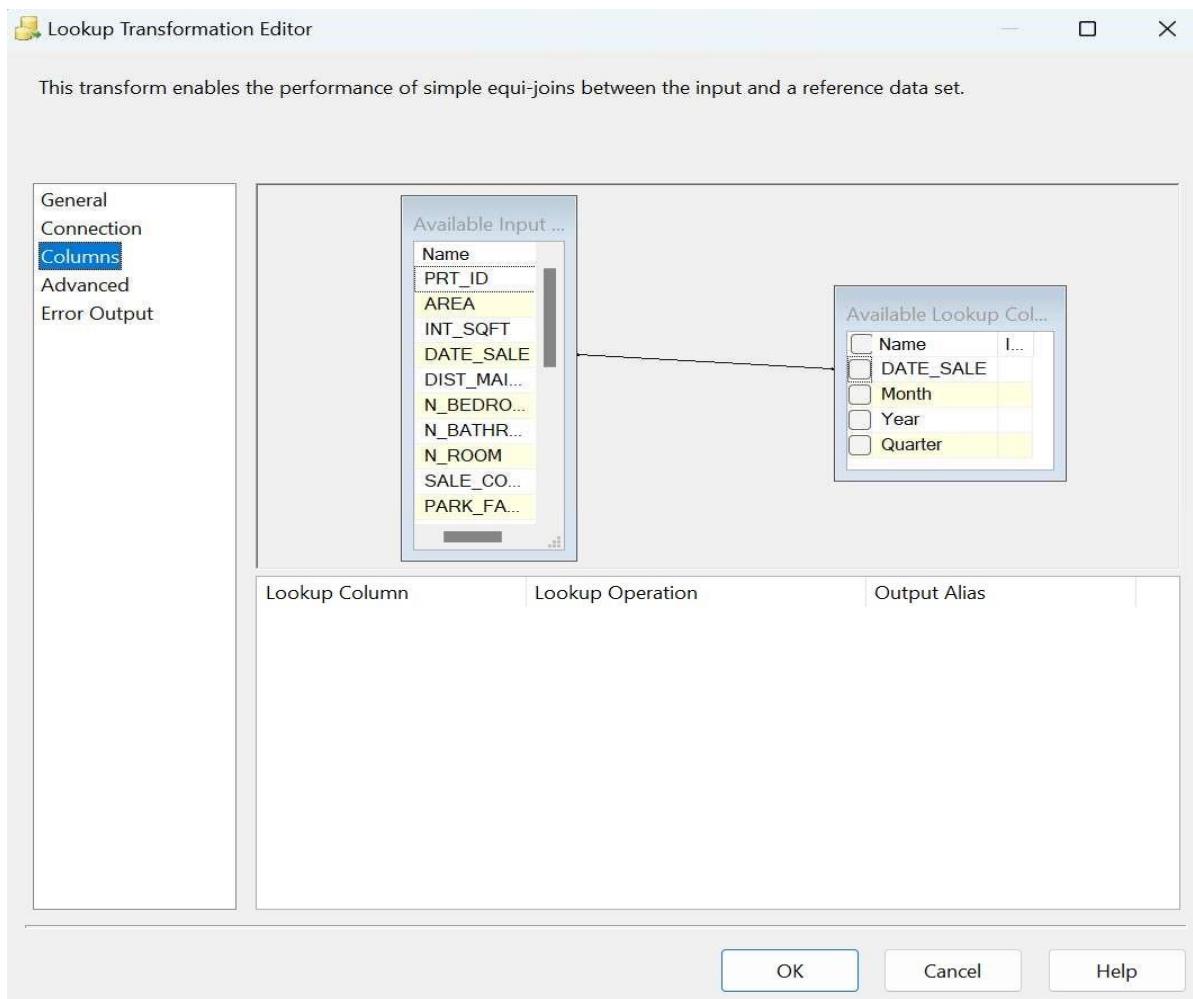
- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.



- Tại tab Connection:
 - OLE DB connection manager chọn WH.
 - Name of the table or the view chọn [dbo].[Dim_SaleTime]



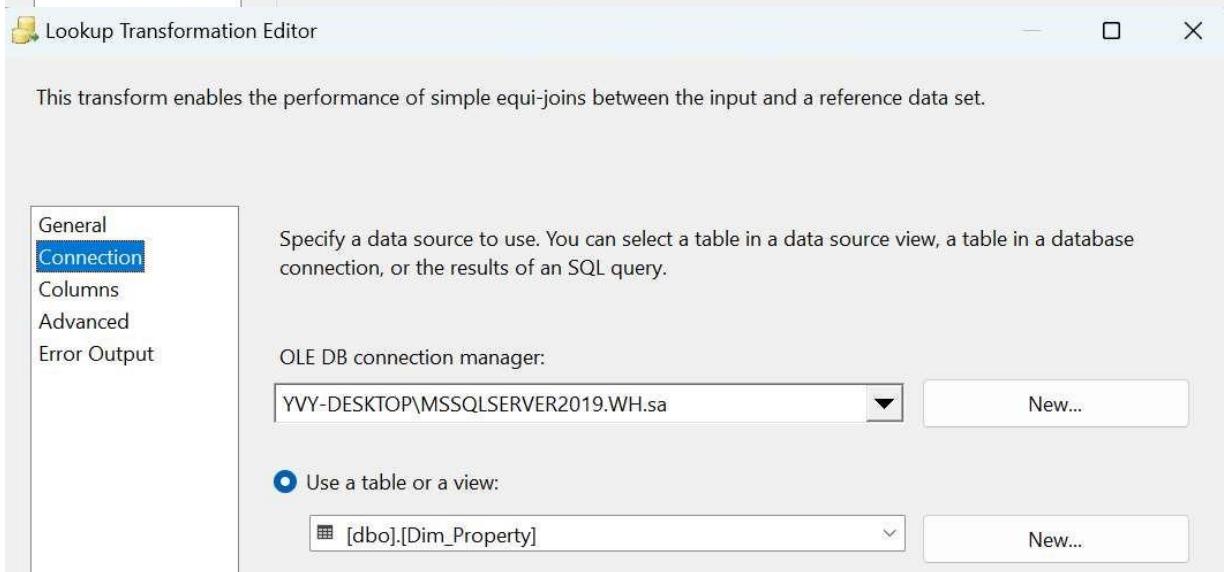
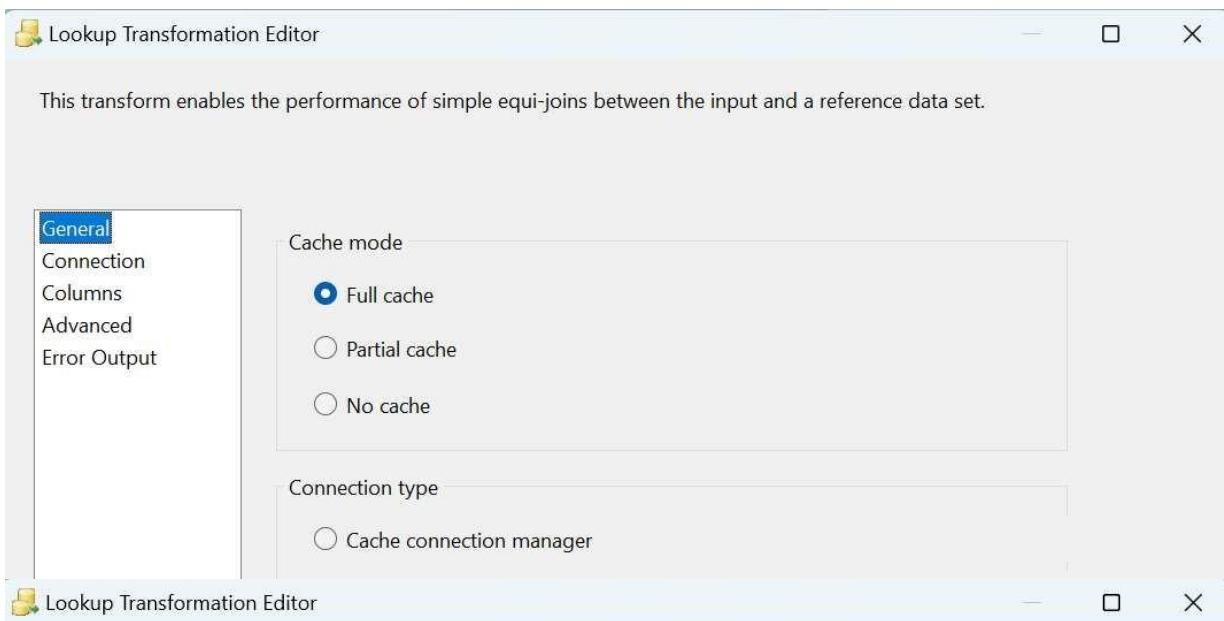
- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.



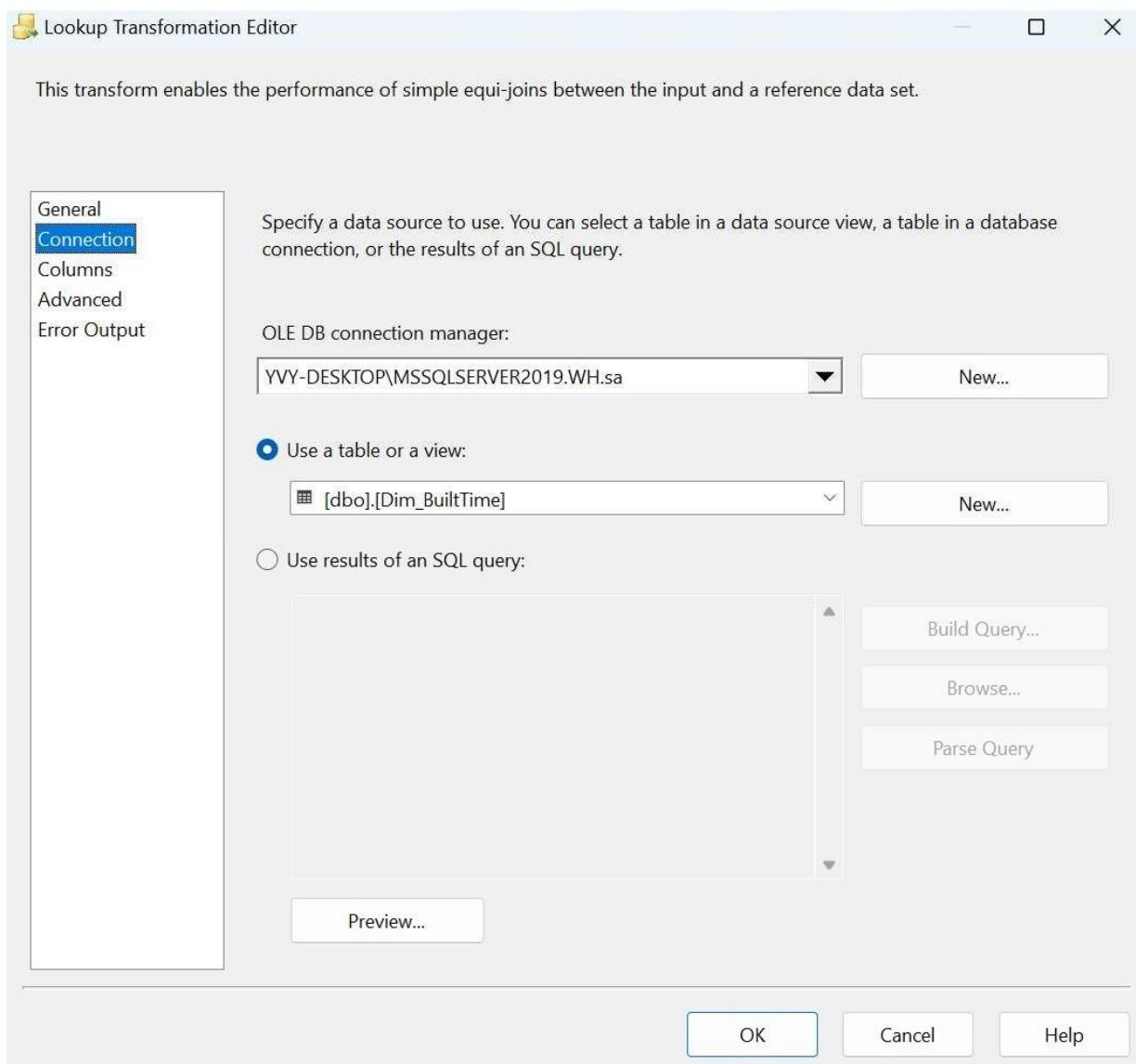
❖ Dim_BuiltTime -

Tại tab General:

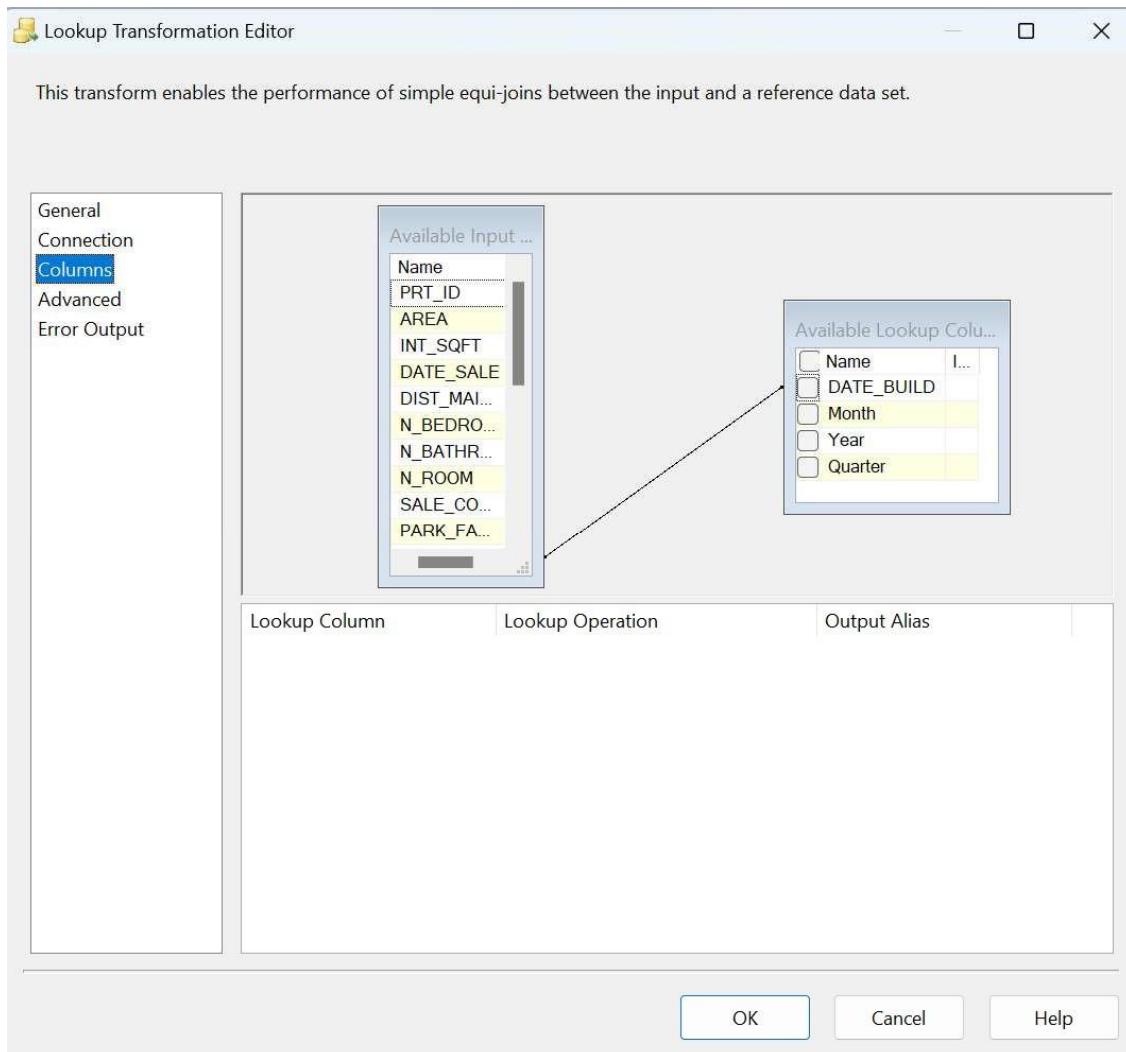
- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.



- Tại tab Connection:
 - OLE DB connection manager chọn WH.
 - Name of the table or the view chọn [dbo].[Dim_BuiltTime]



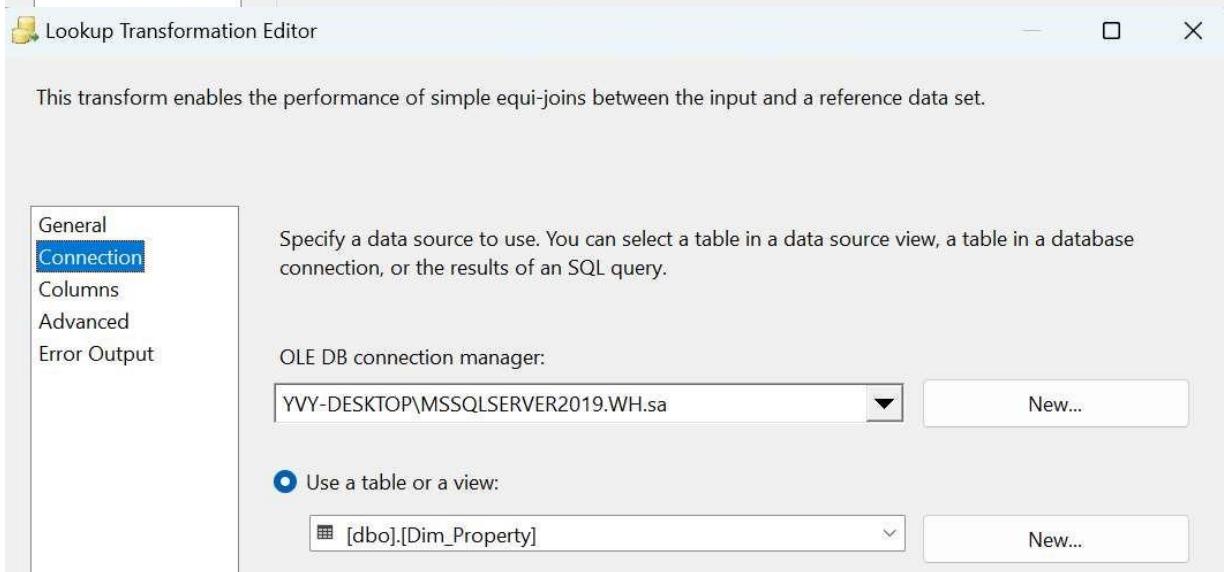
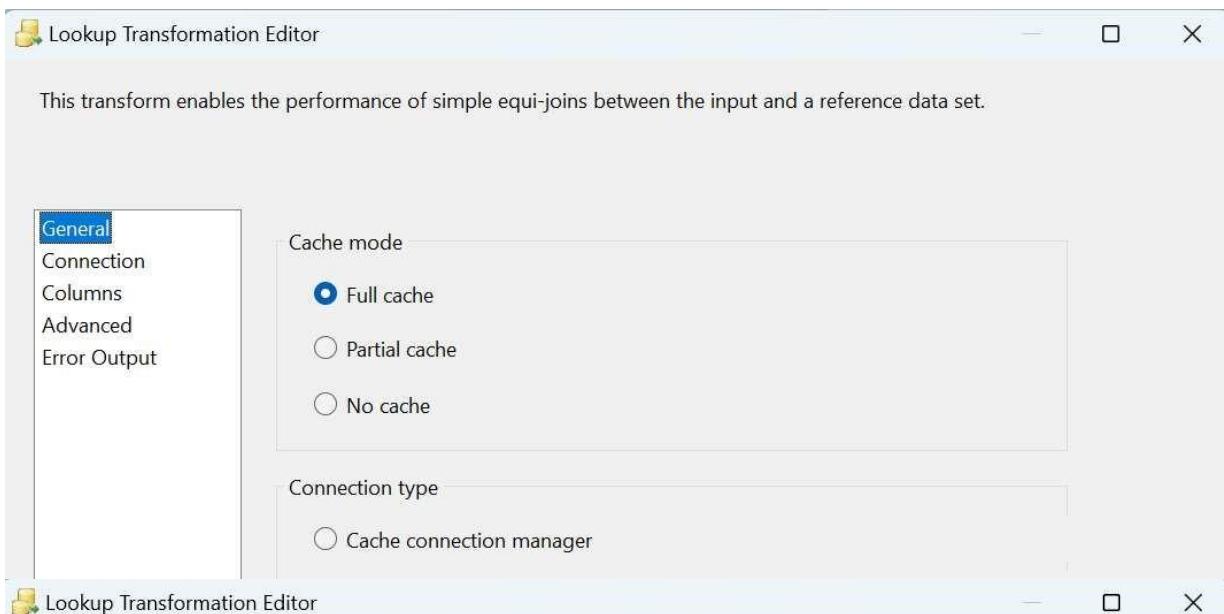
- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.



❖ Dim_Type - Tại tab

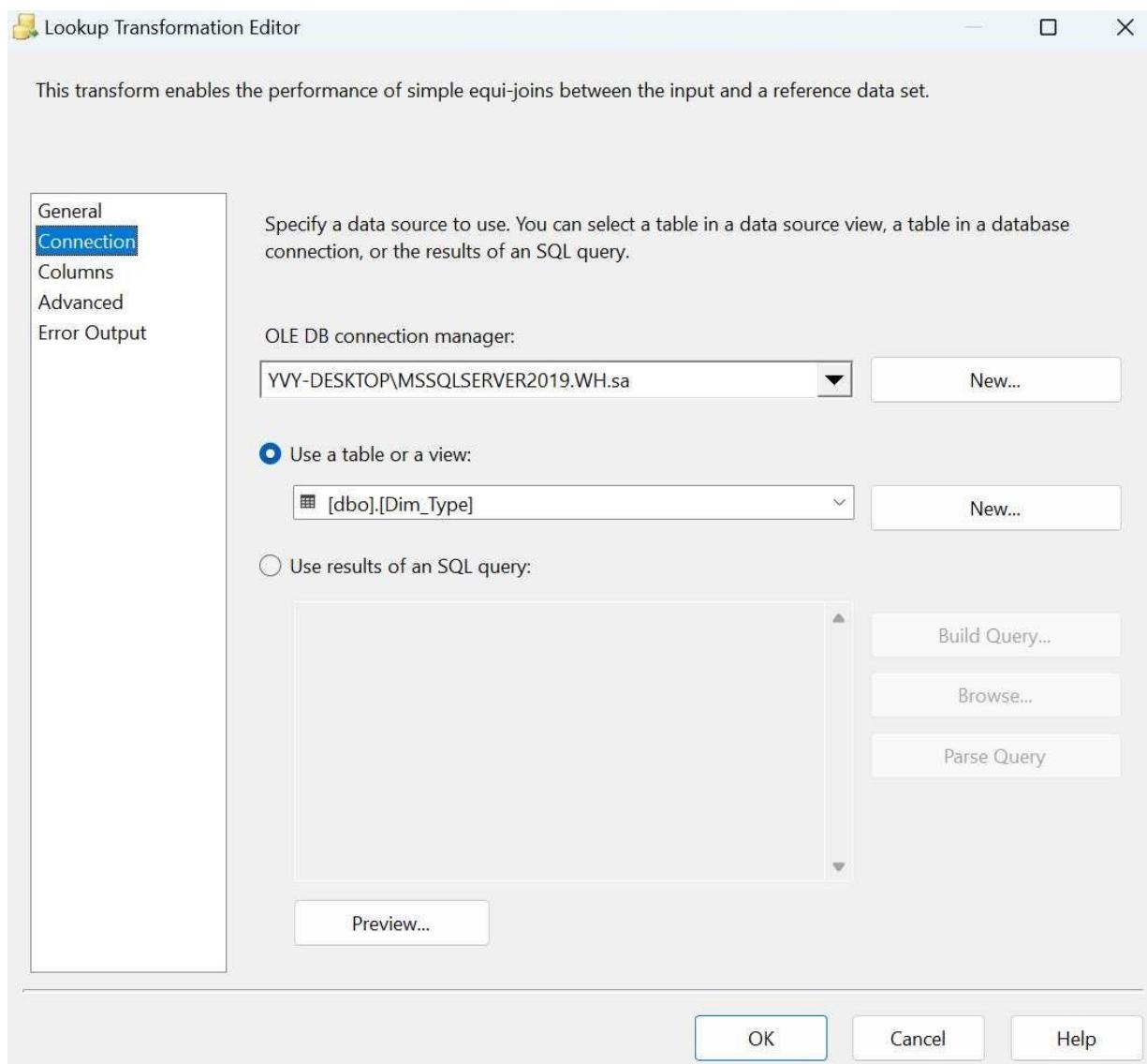
General:

- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.

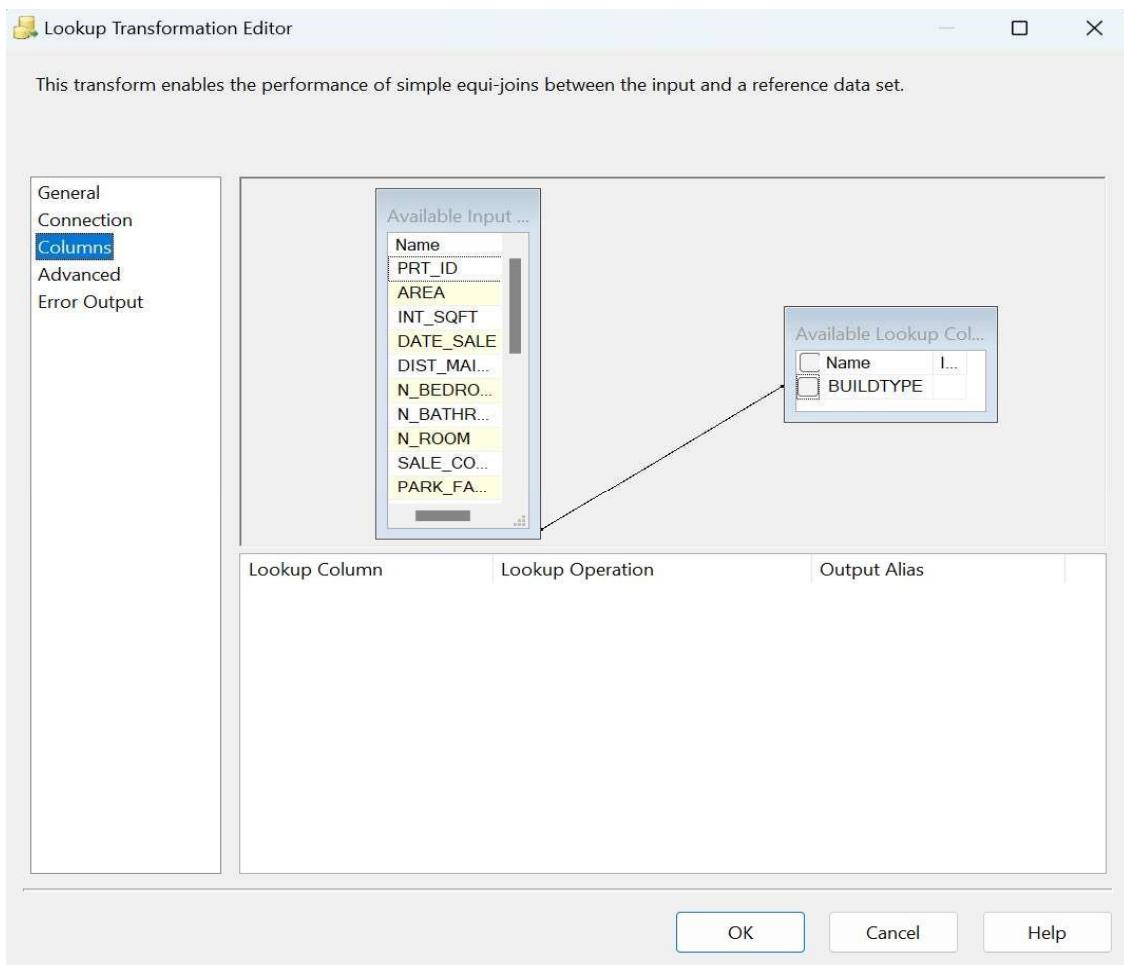


- Tại tab Connection:

- OLE DB connection manager chọn WH.
- Name of the table or the view chọn [dbo].[Dim_Type]



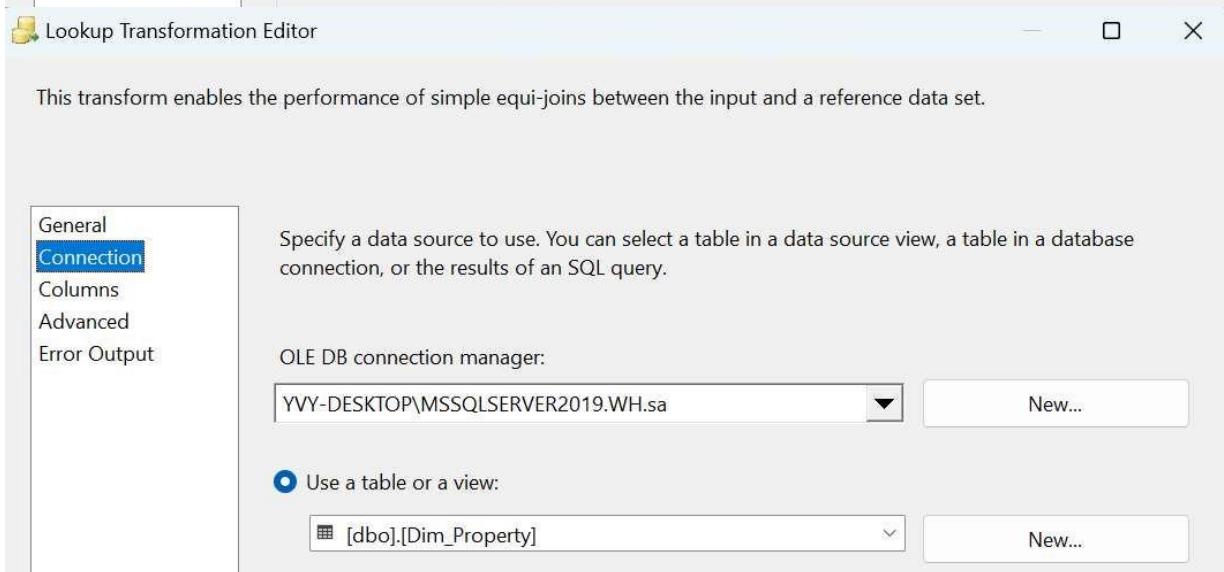
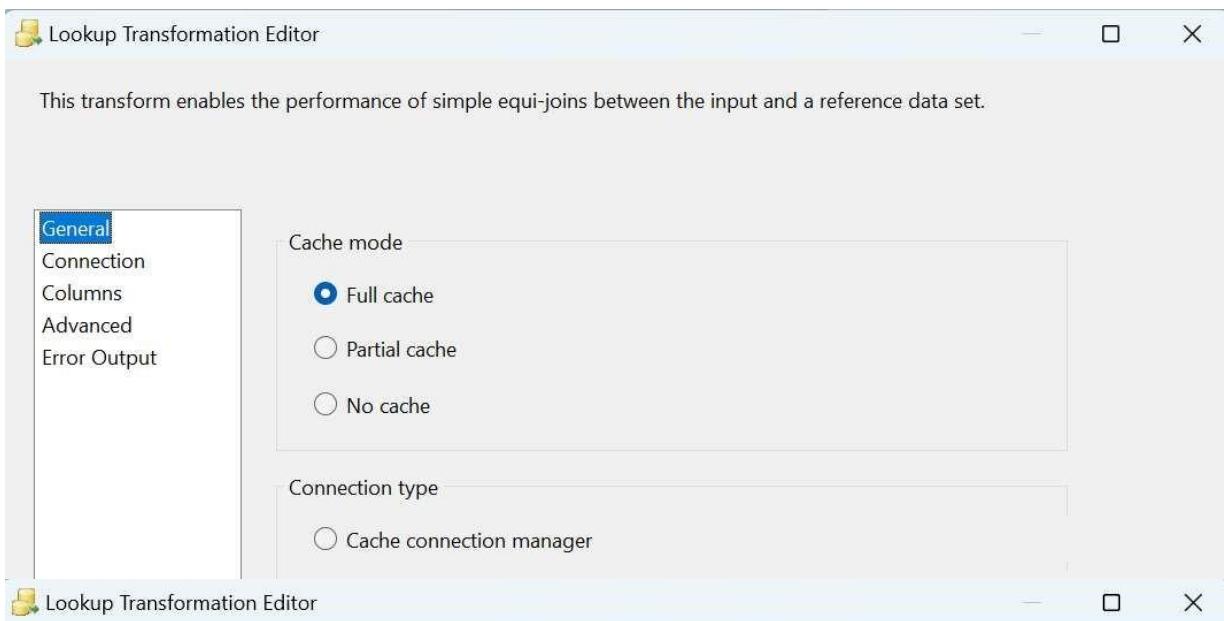
- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.



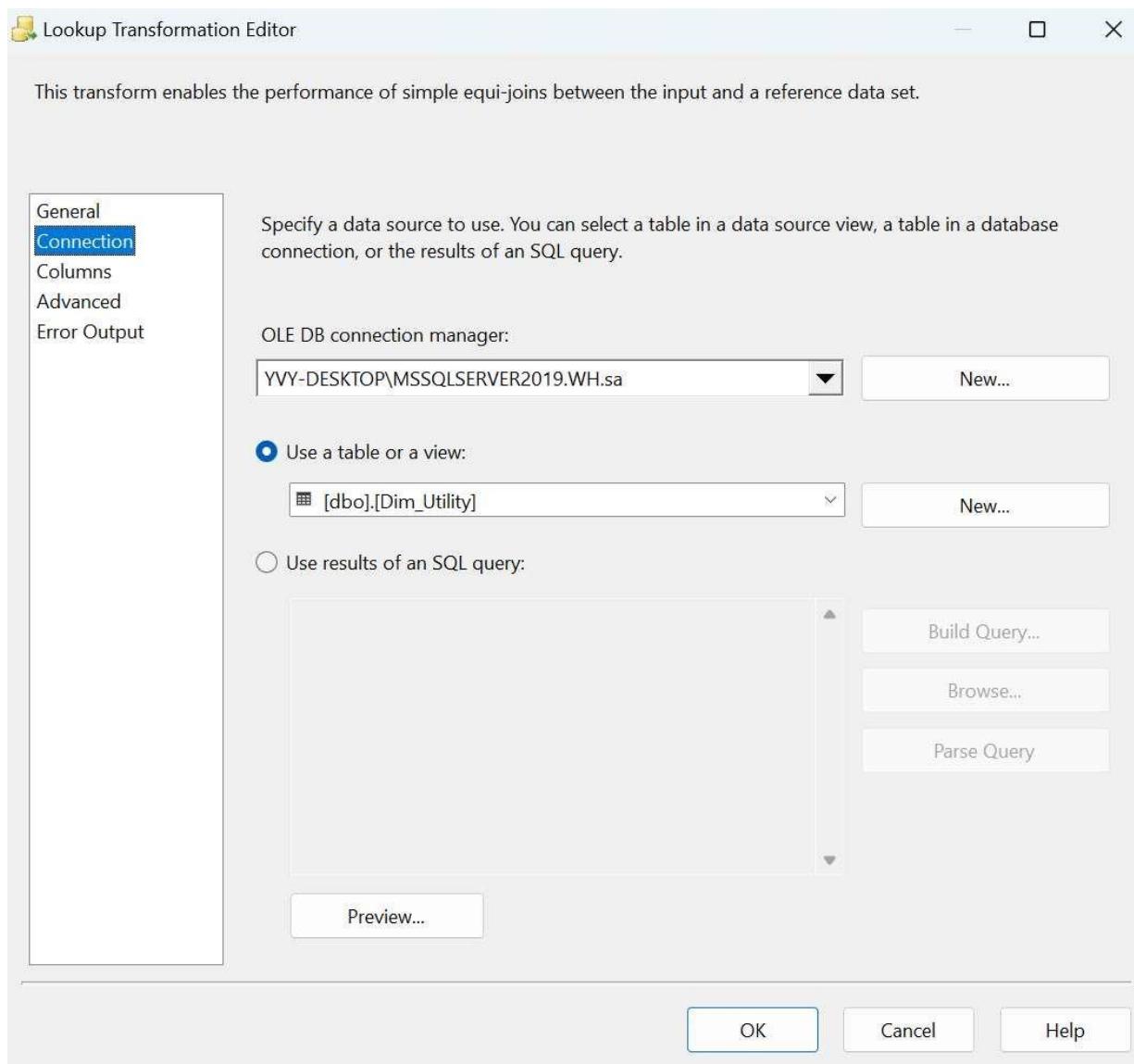
❖ Dim_Utility - Tải

tab General:

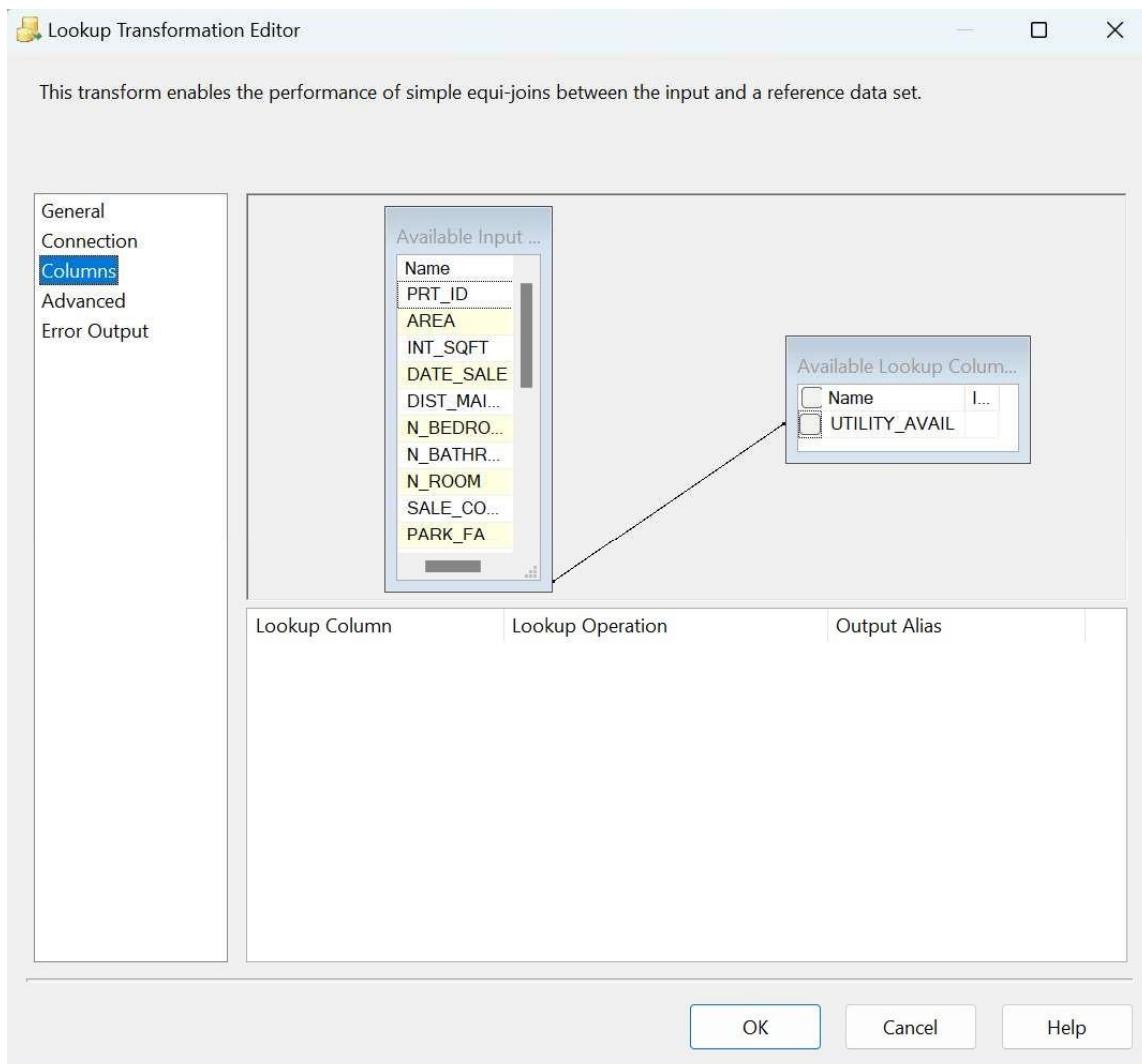
- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.



- Tại tab Connection:
 - OLE DB connection manager chọn WH.
 - Name of the table or the view chọn [dbo].[Dim_Utility]



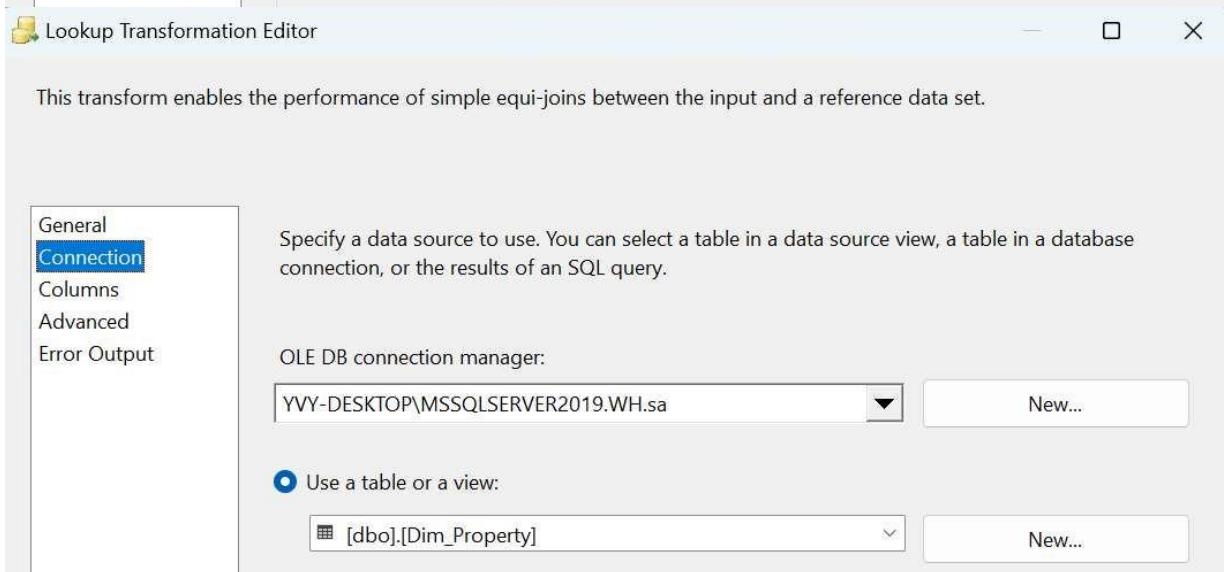
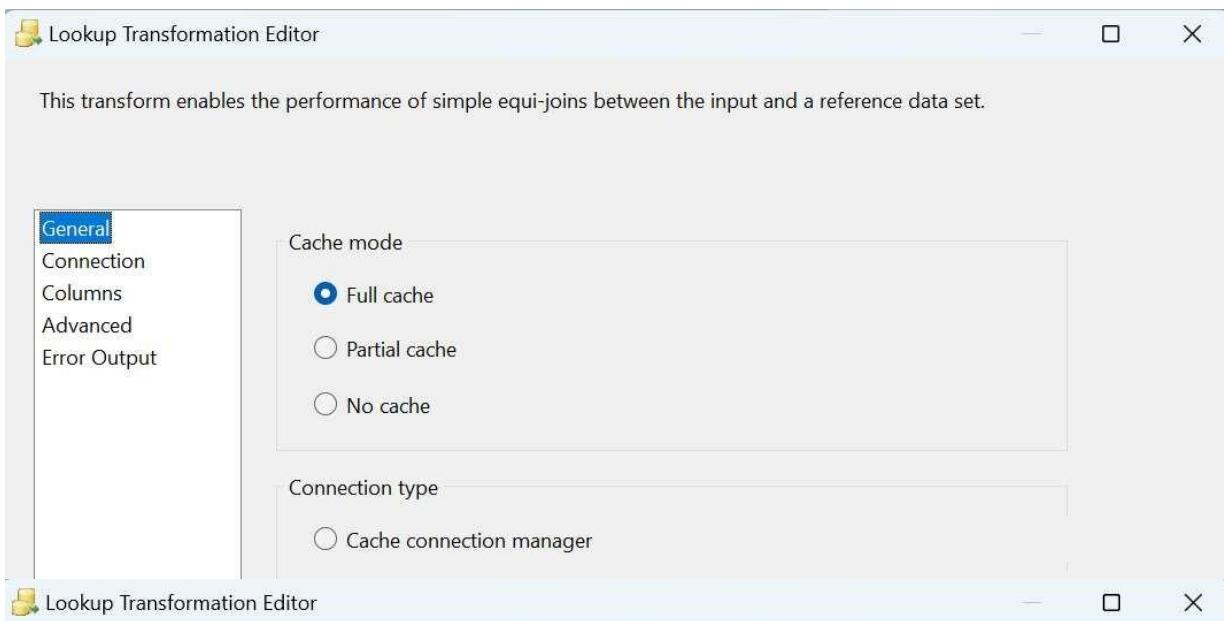
- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.



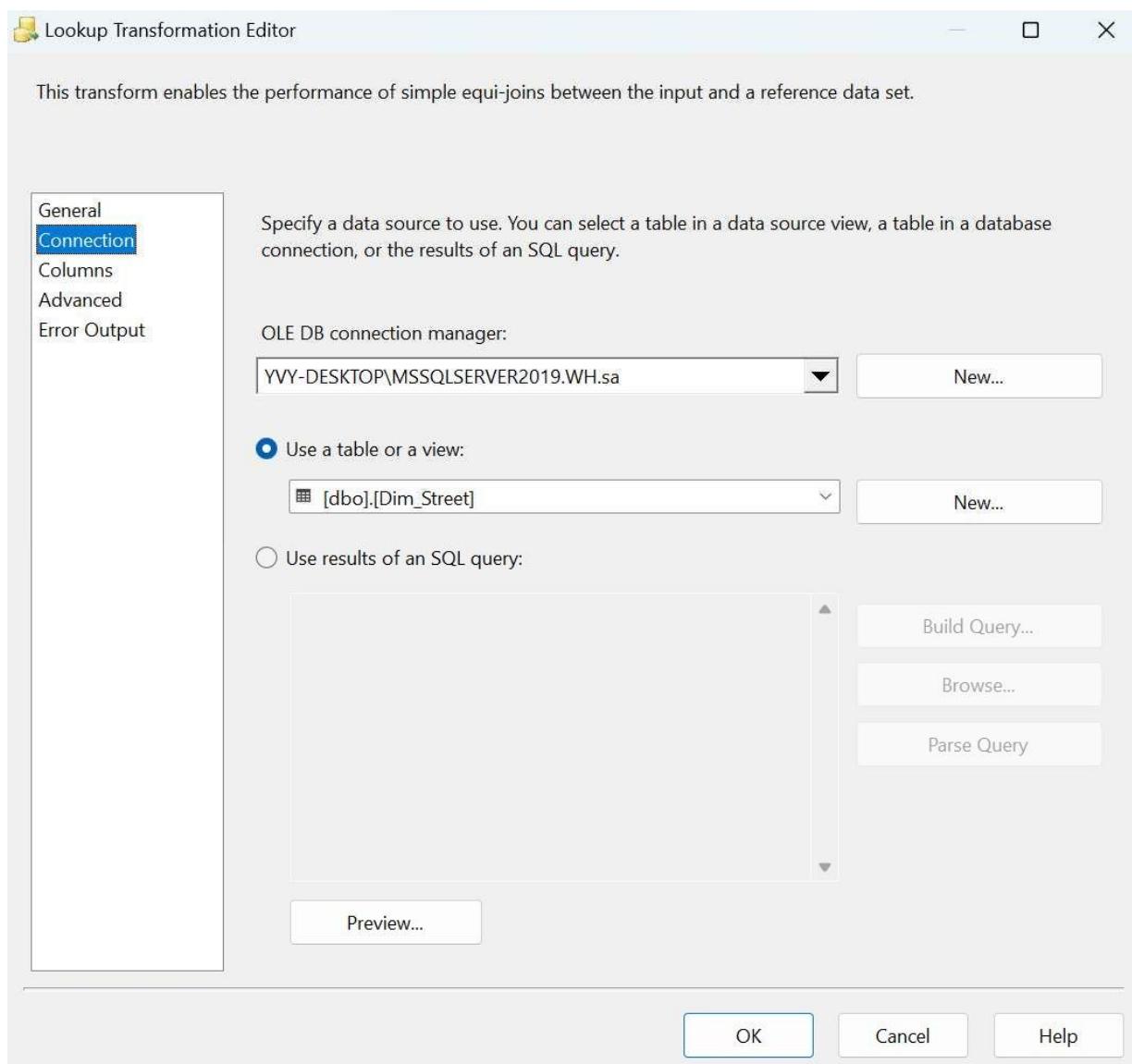
❖ Dim_Street - Tại tab

General:

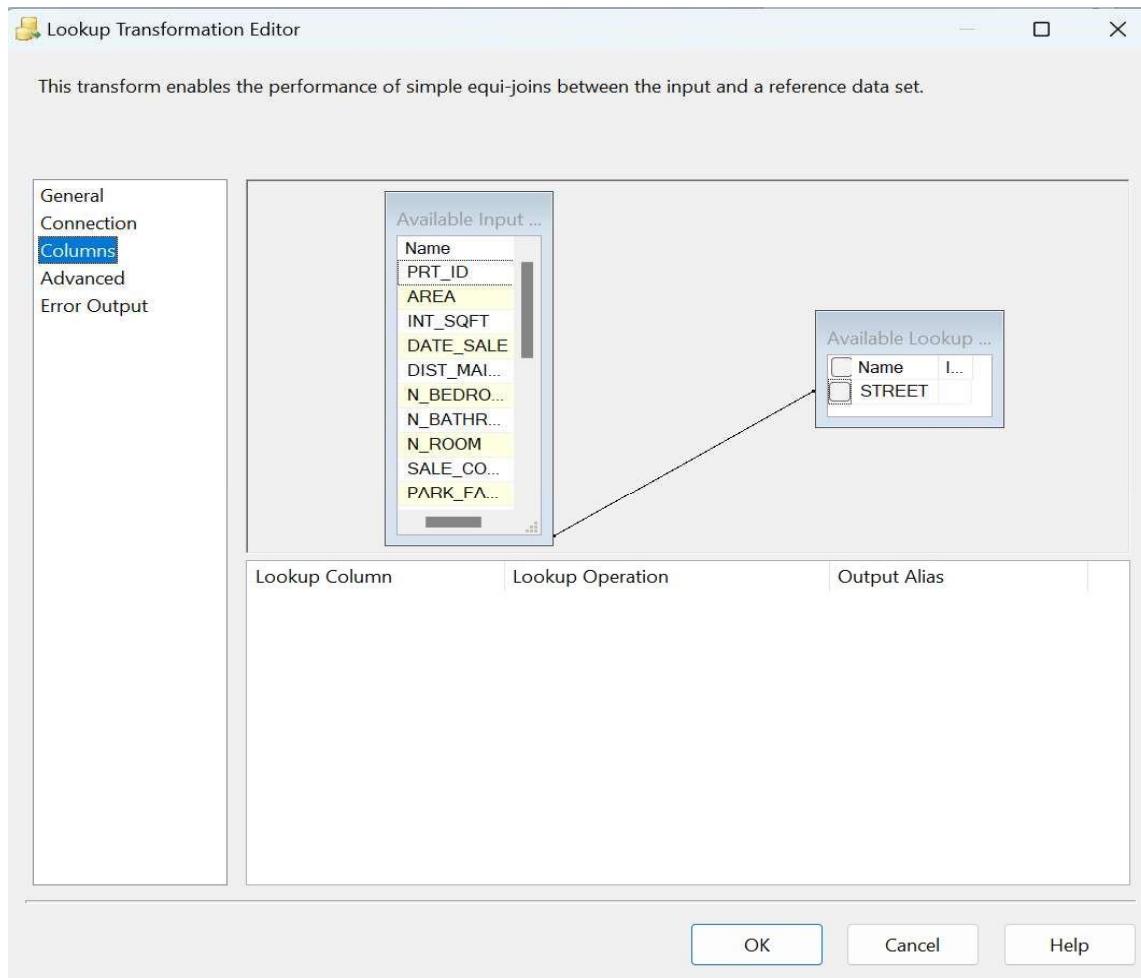
- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.



- Tại tab Connection:
 - OLE DB connection manager chọn WH.
 - Name of the table or the view chọn [dbo].[Dim_Street]



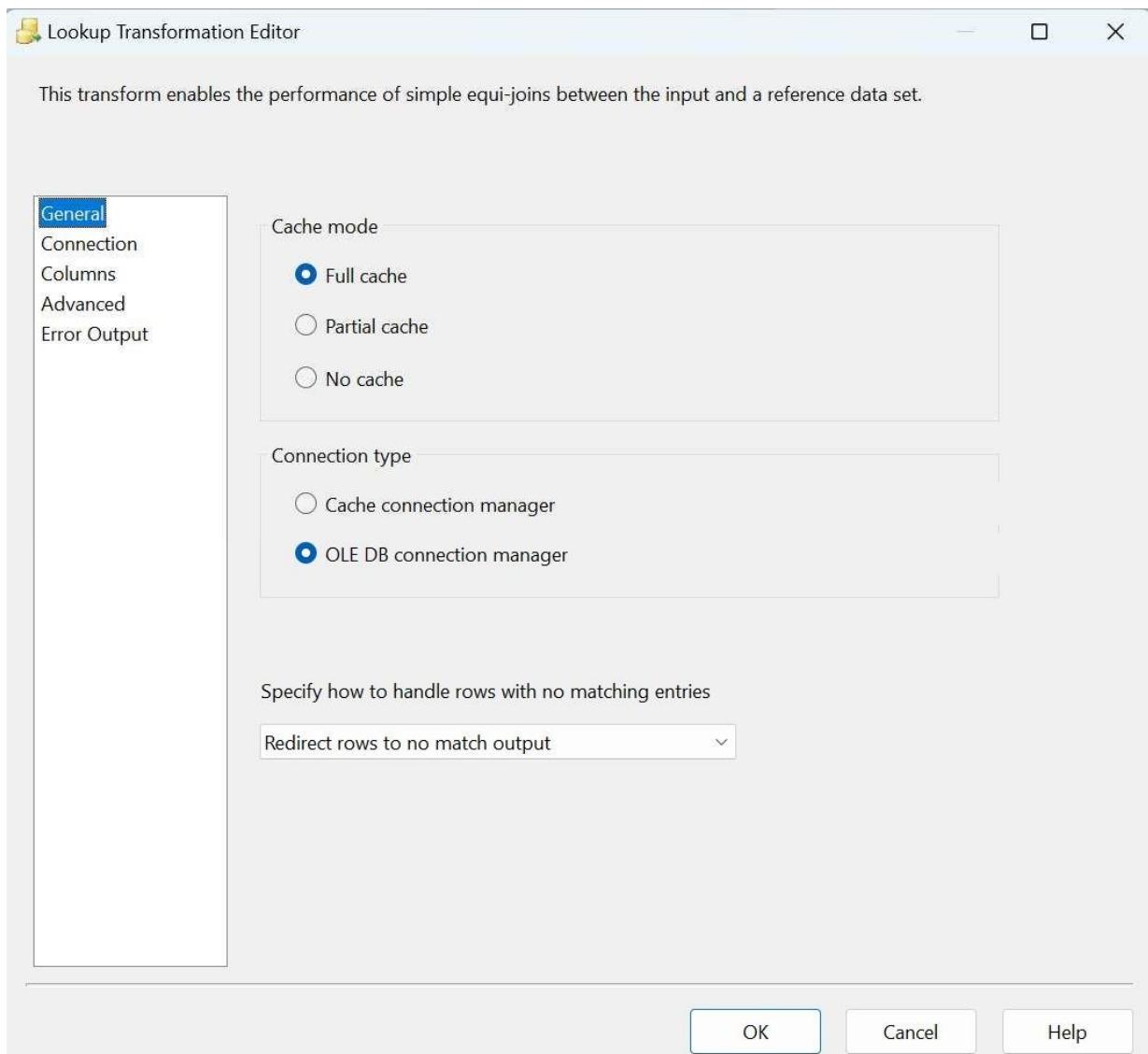
- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.



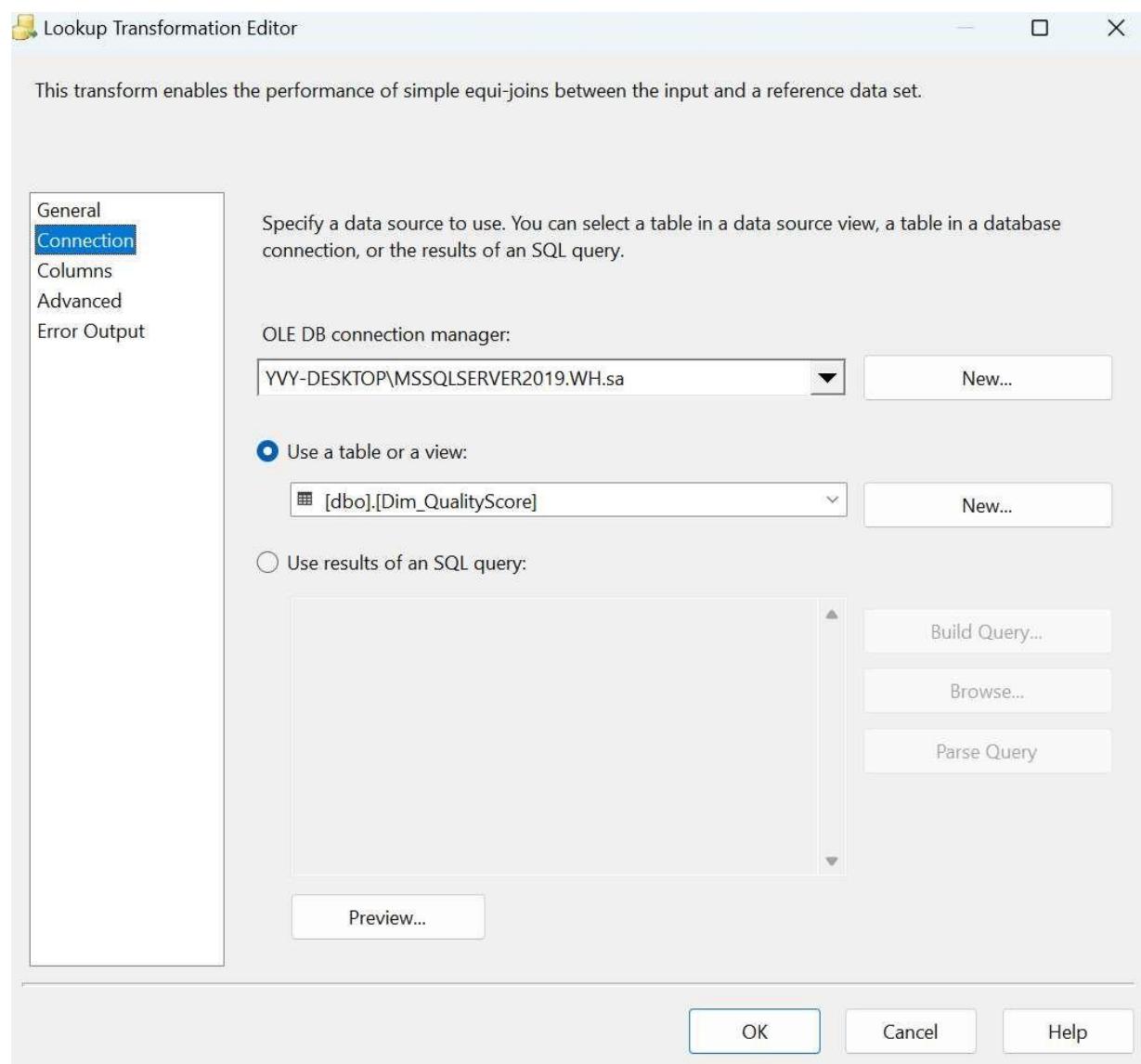
❖ Dim_QualityScore -

Tại tab General:

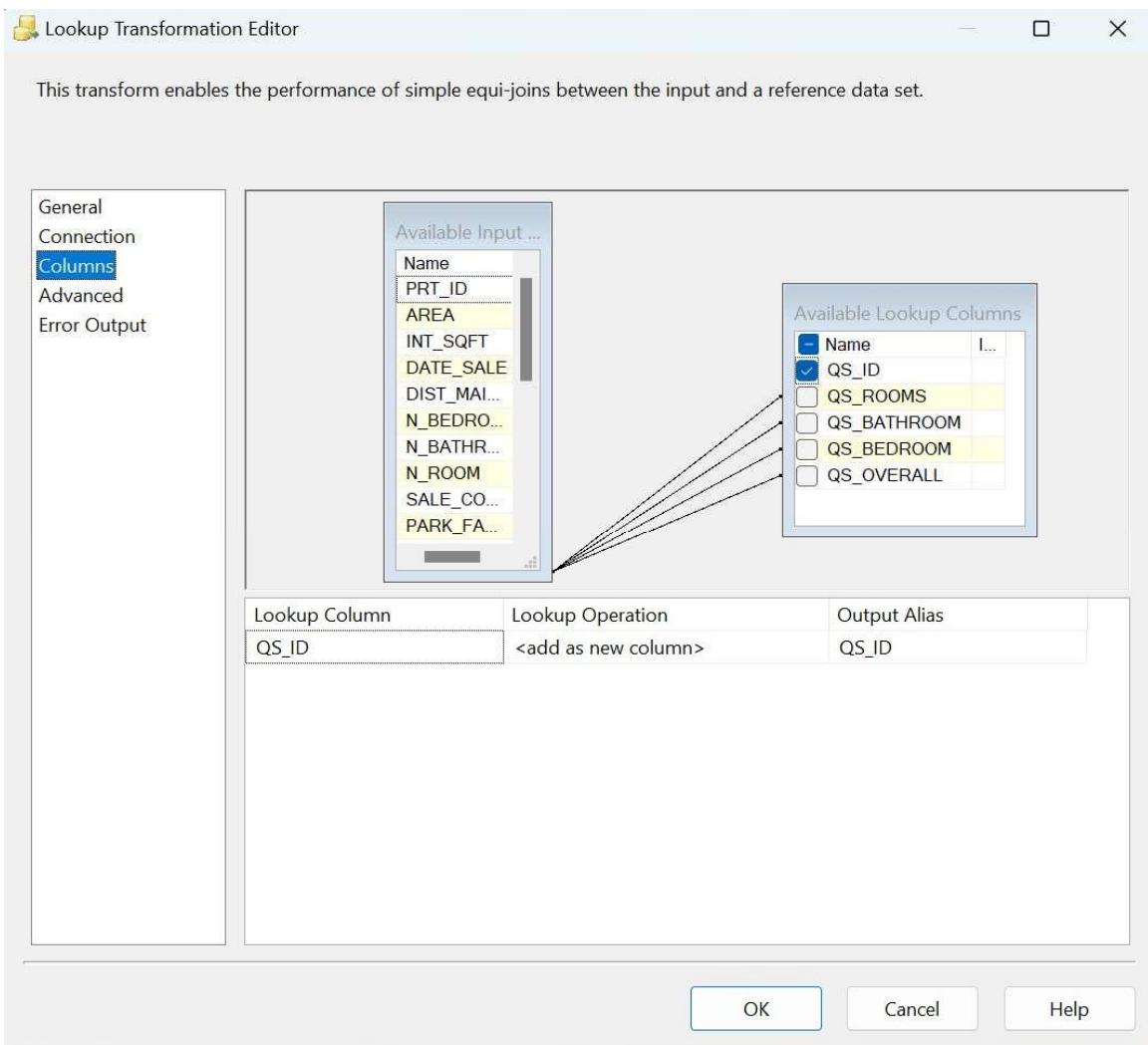
- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.



- Tại tab Connection:
 - OLE DB connection manager chọn WH.
 - Name of the table or the view chọn [dbo].[Dim_QualityScore]



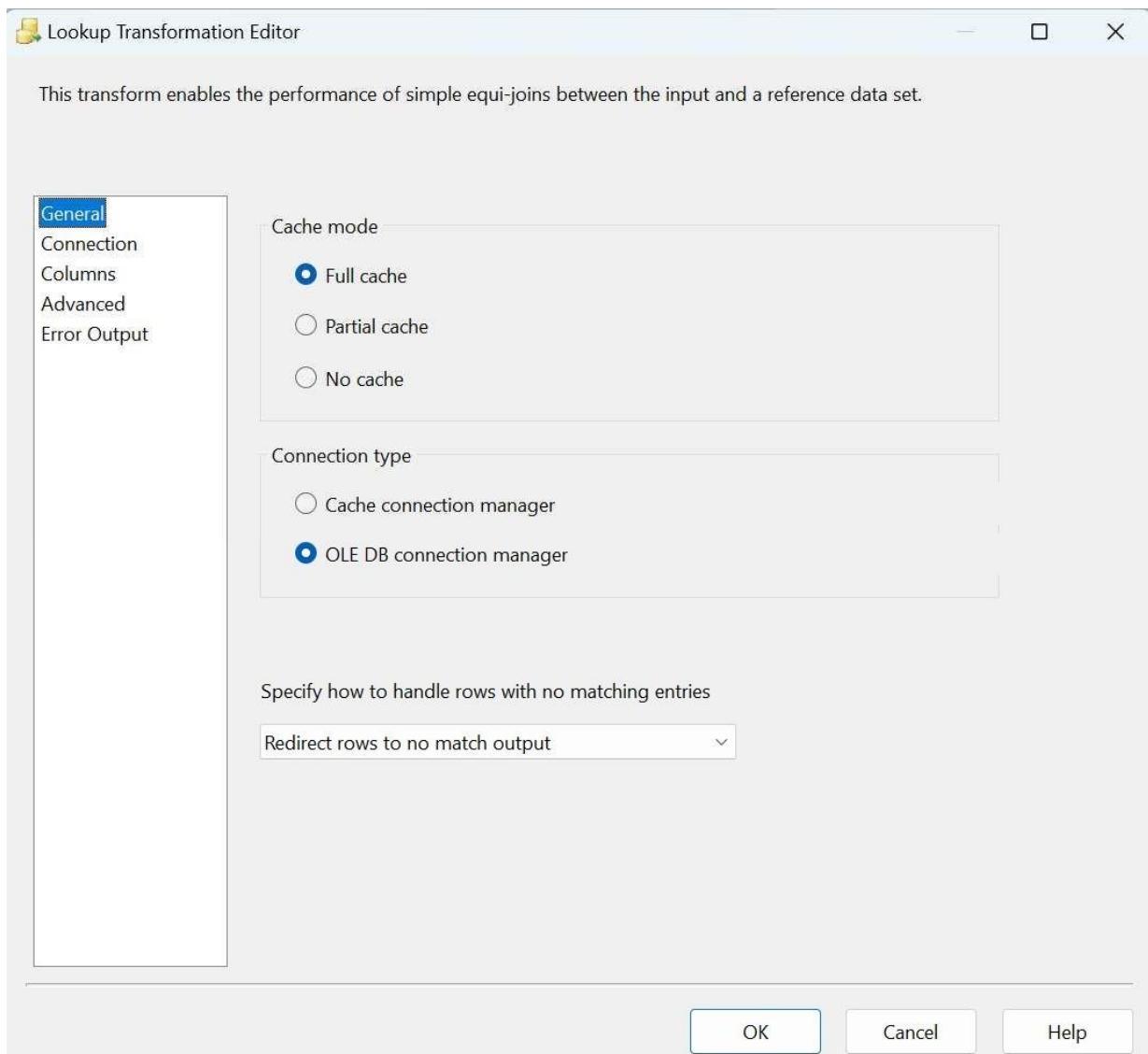
- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.



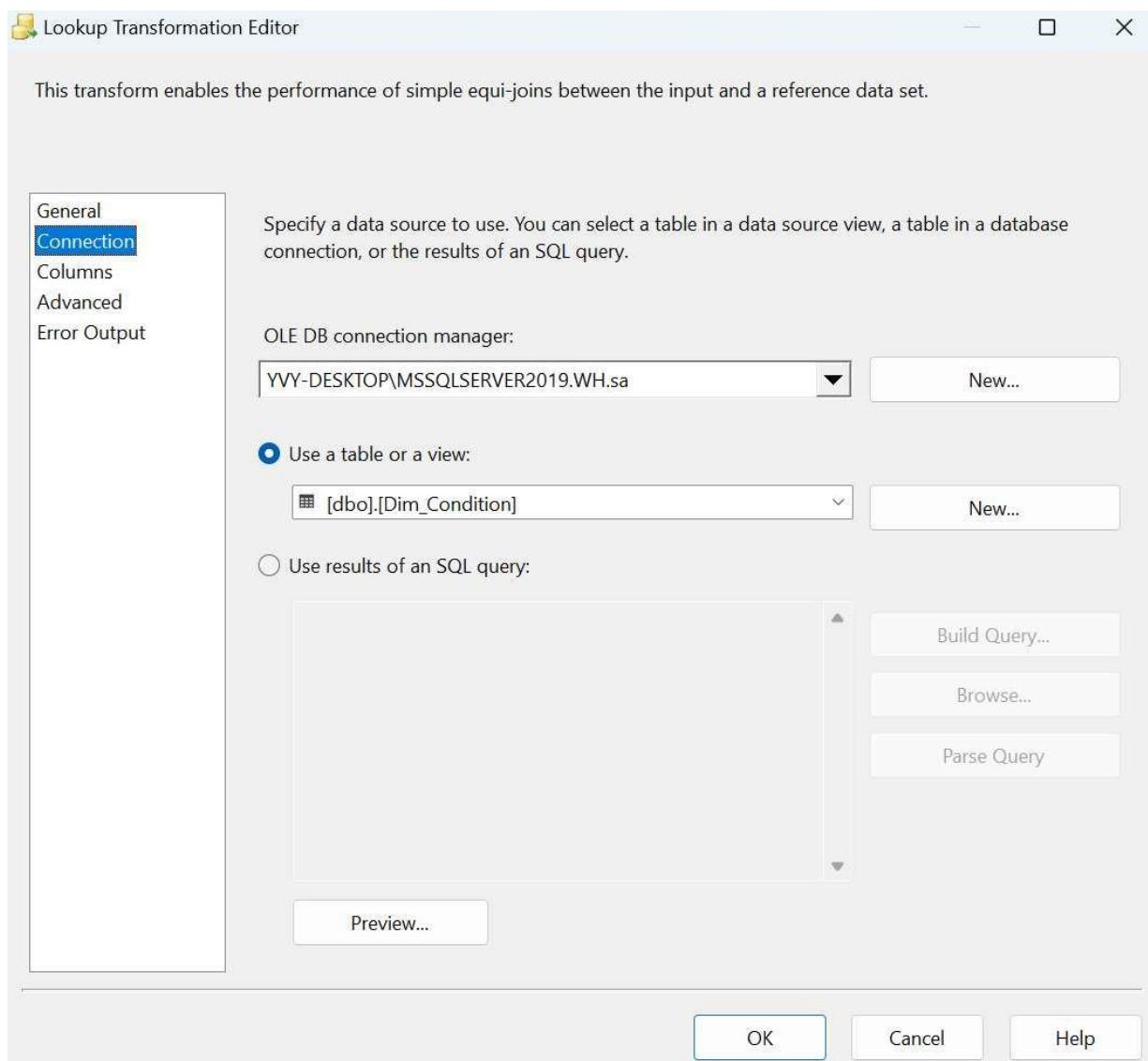
❖ Dim_Condition -

Tại tab General:

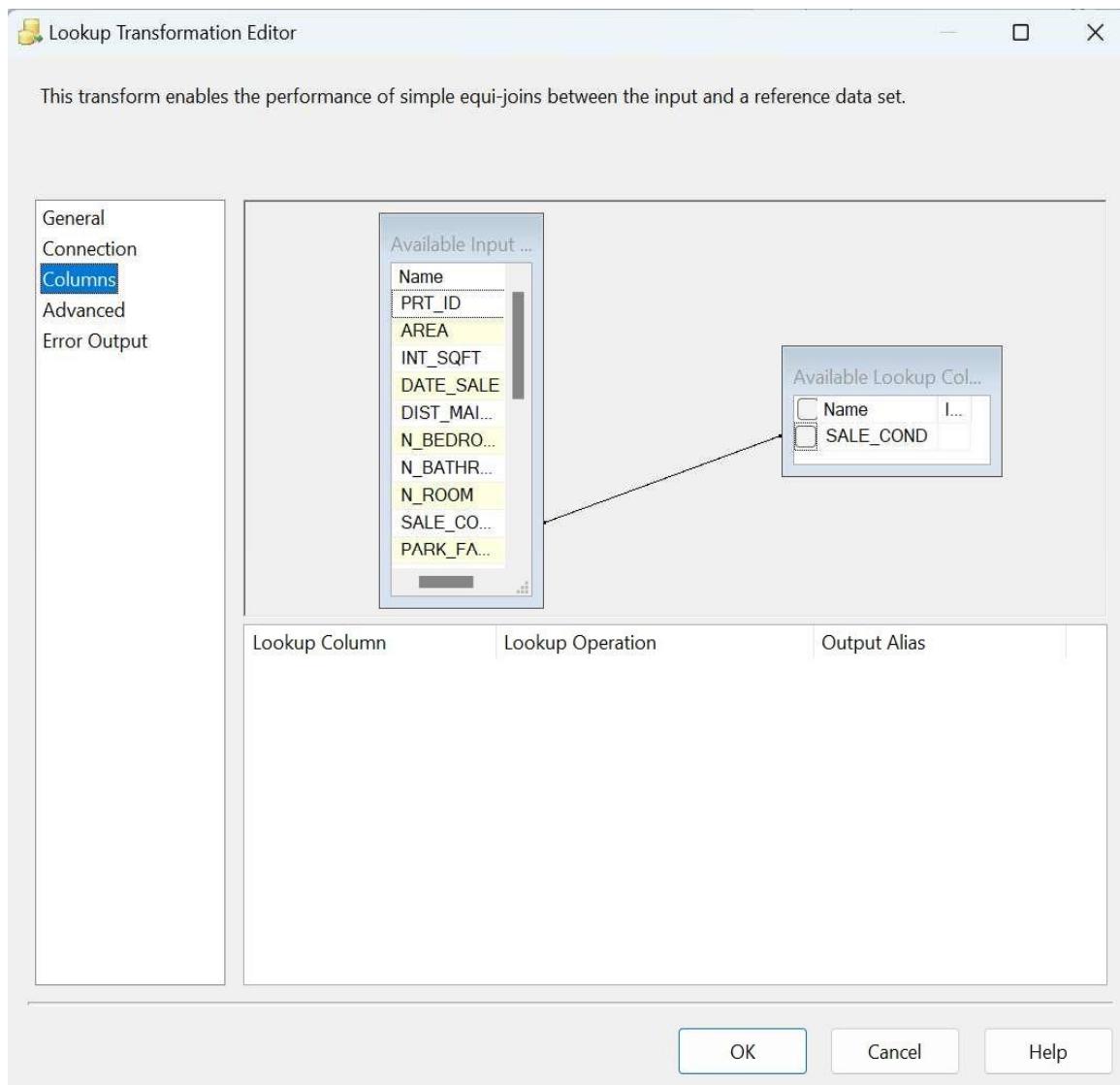
- Specify how to handle rows with no matching entries: Chọn Redirect rows to no match output.



- Tại tab Connection:
 - OLE DB connection manager chọn WH.
 - Name of the table or the view chọn [dbo].[Dim_Condition]

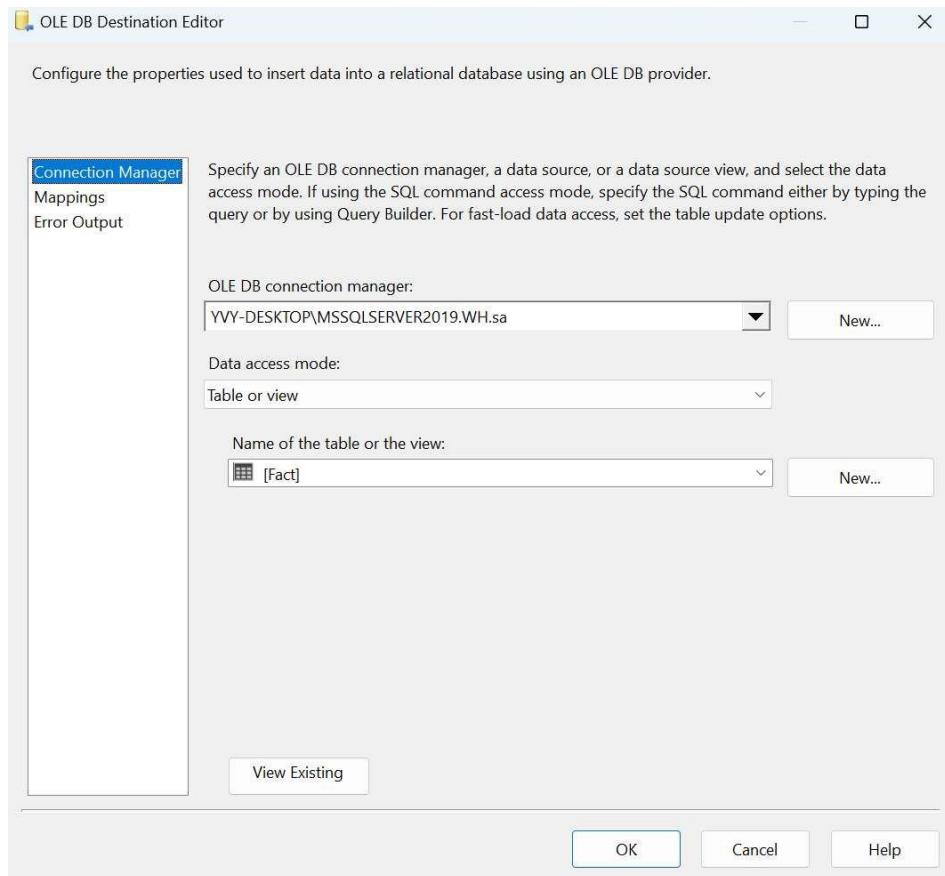


- Kéo thả các cột tương ứng ở Available input columns vào các cột ở Available lookup columns sau đó nhấn OK để hoàn thành.

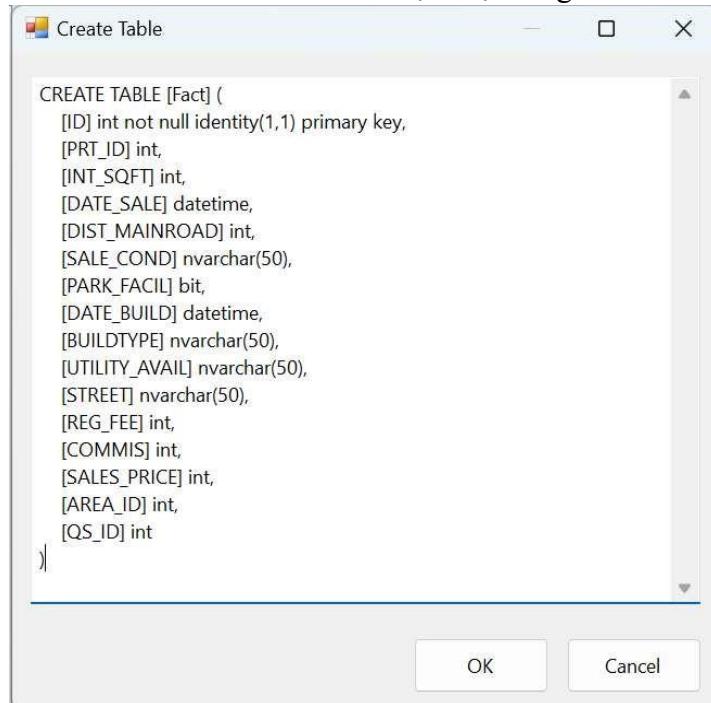


OLE DB Destination:

- OLE DB connection manager chọn WH.

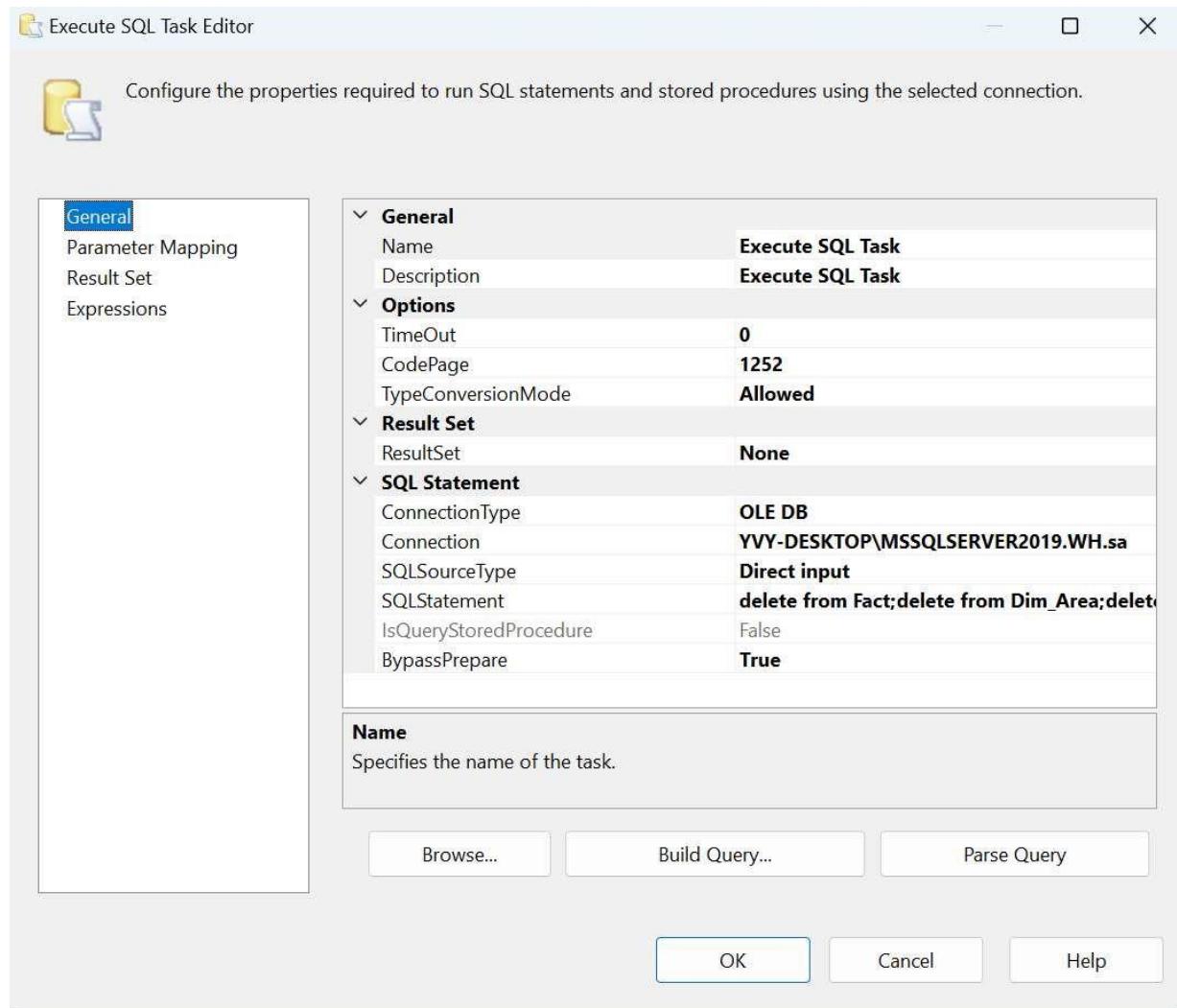


- Name of the table or the view nhấn New để tạo một bảng mới:

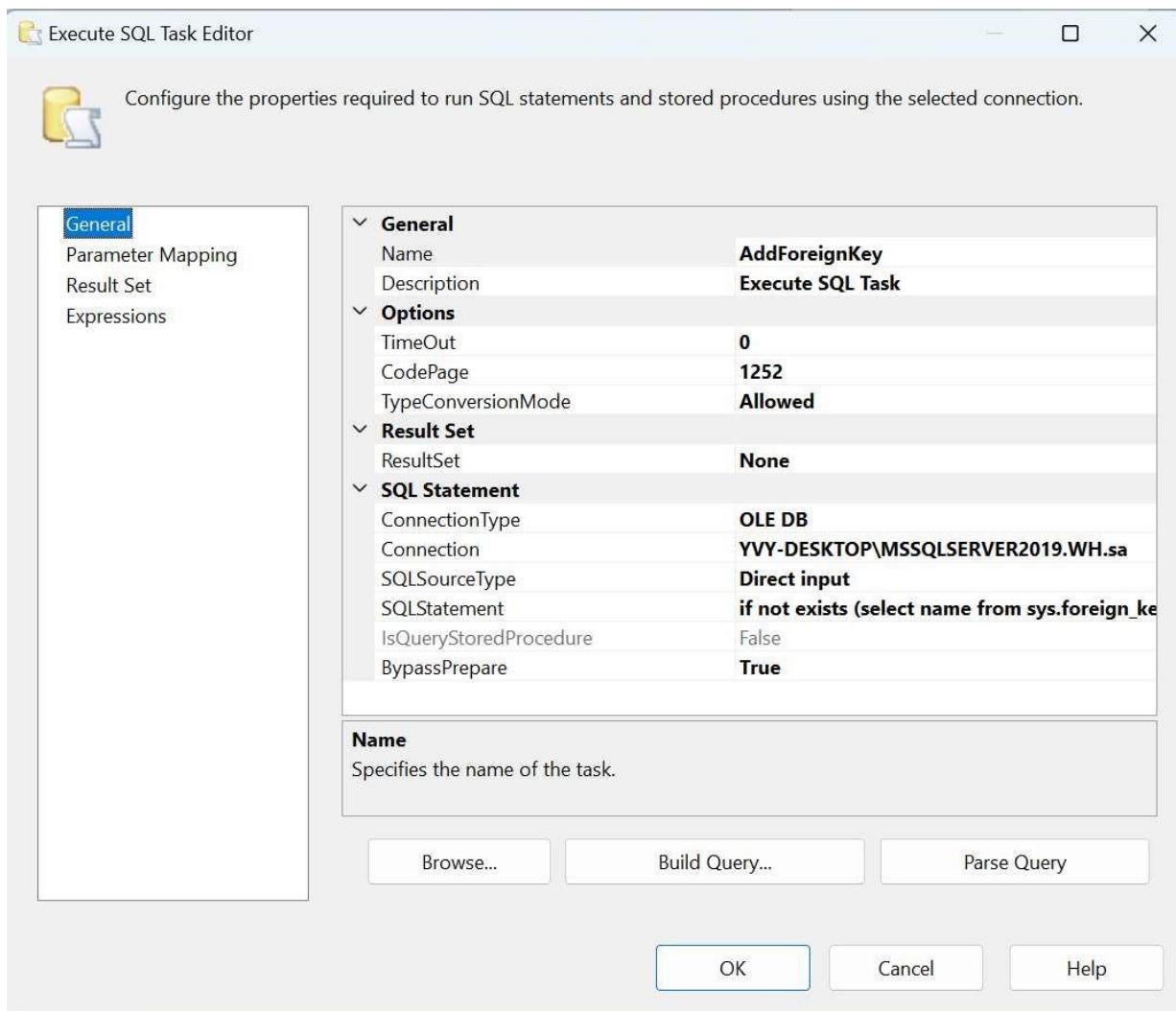


2.6. Viết Execute SQL Task:

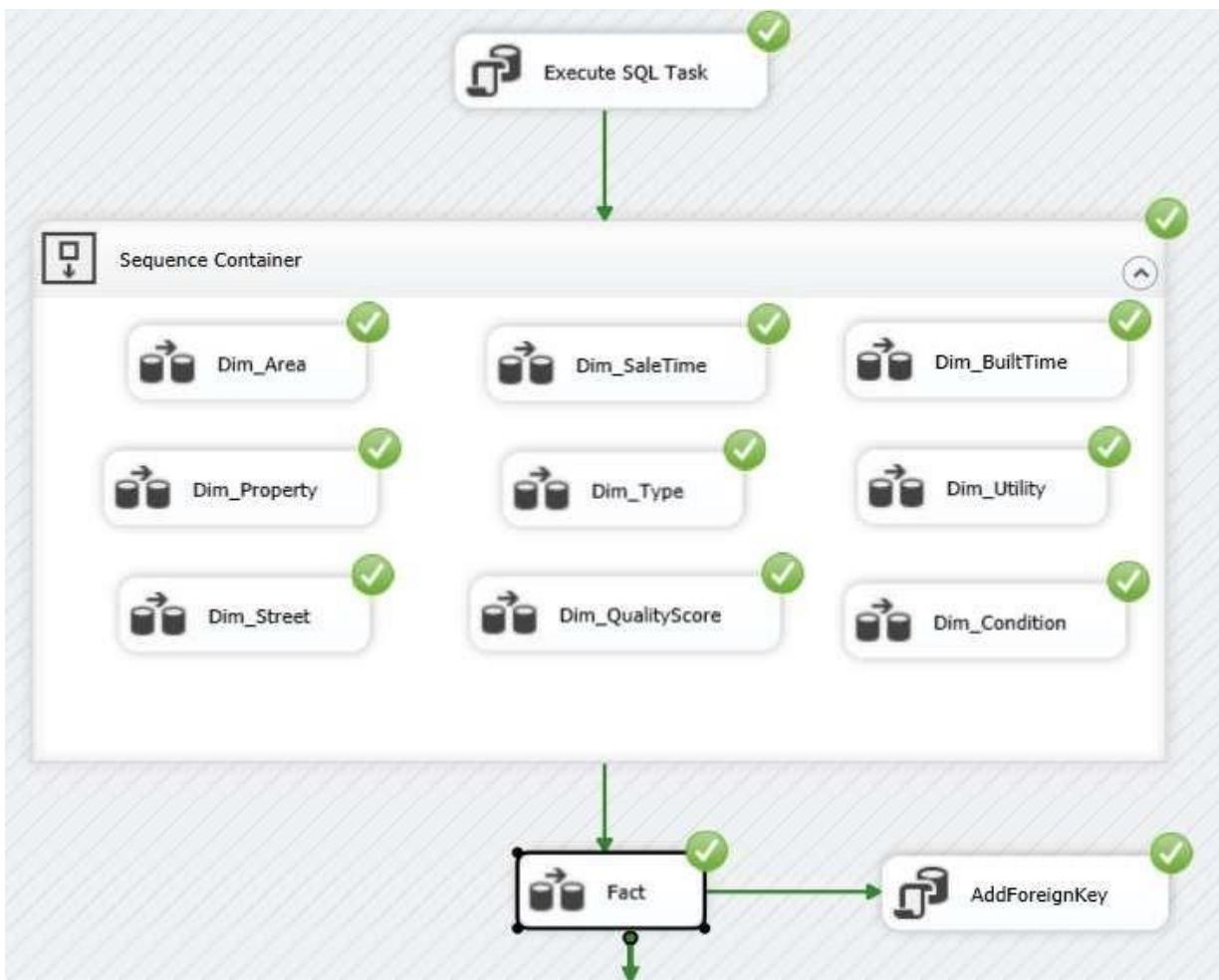
- Sử dụng Execute SQL Task để xóa các bảng mỗi lần Start lại (vì bảng Dimension và Fact sẽ được set khóa chính nên khi chạy lại để tránh việc trùng khóa phải xóa những bảng ở lần Start trước đó) => Click chuột phải vào Execute SQL Task chọn Edit => Chọn Connection là WH => Chính sửa SQLStatement => OK



- Sử dụng Execute SQL Task để tiến hành tạo các khóa ngoại theo đúng thiết kế. Chọn Edit => Chọn Connection là WH => Chính sửa SQLStatement => OK



- Start chương trình



CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN - QUÁ TRÌNH SSAS

3.1. Cấu hình Project

- Đầu tiên, tạo một project Analysis Services với tên là SSAS_ChennaiHousing.

Configure your new project

Analysis Services Multidimensional and Data Mining Project

Project name

Location



Solution name [\(i\)](#)

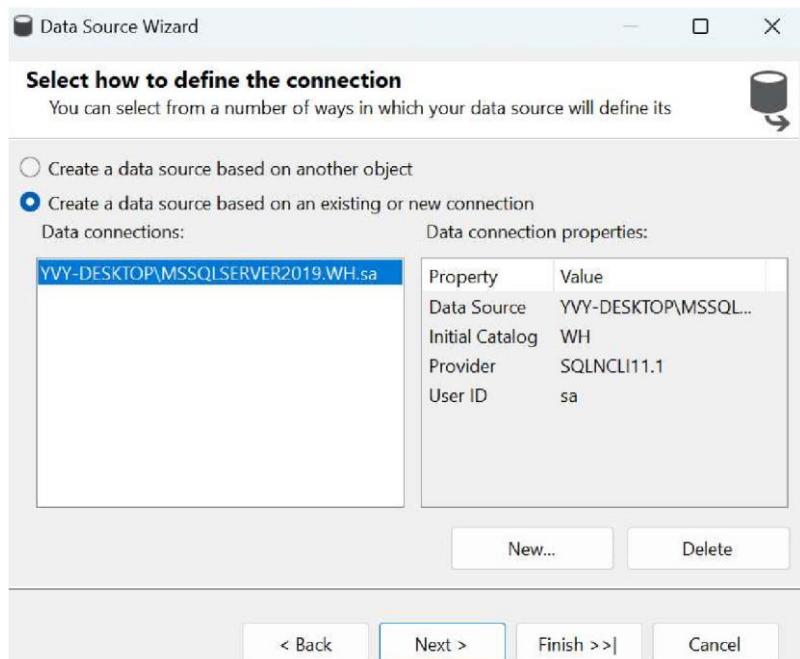
Place solution and project in the same directory

Back

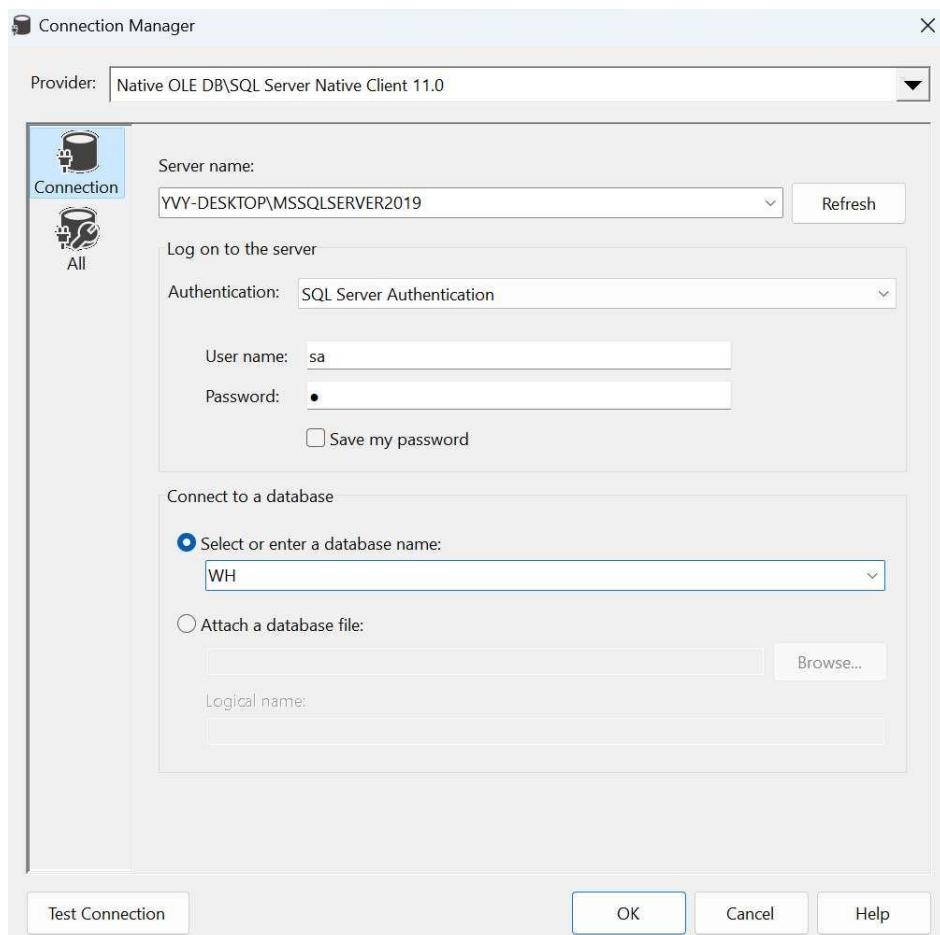
Create

3.1.1. Connect đến Data Sources

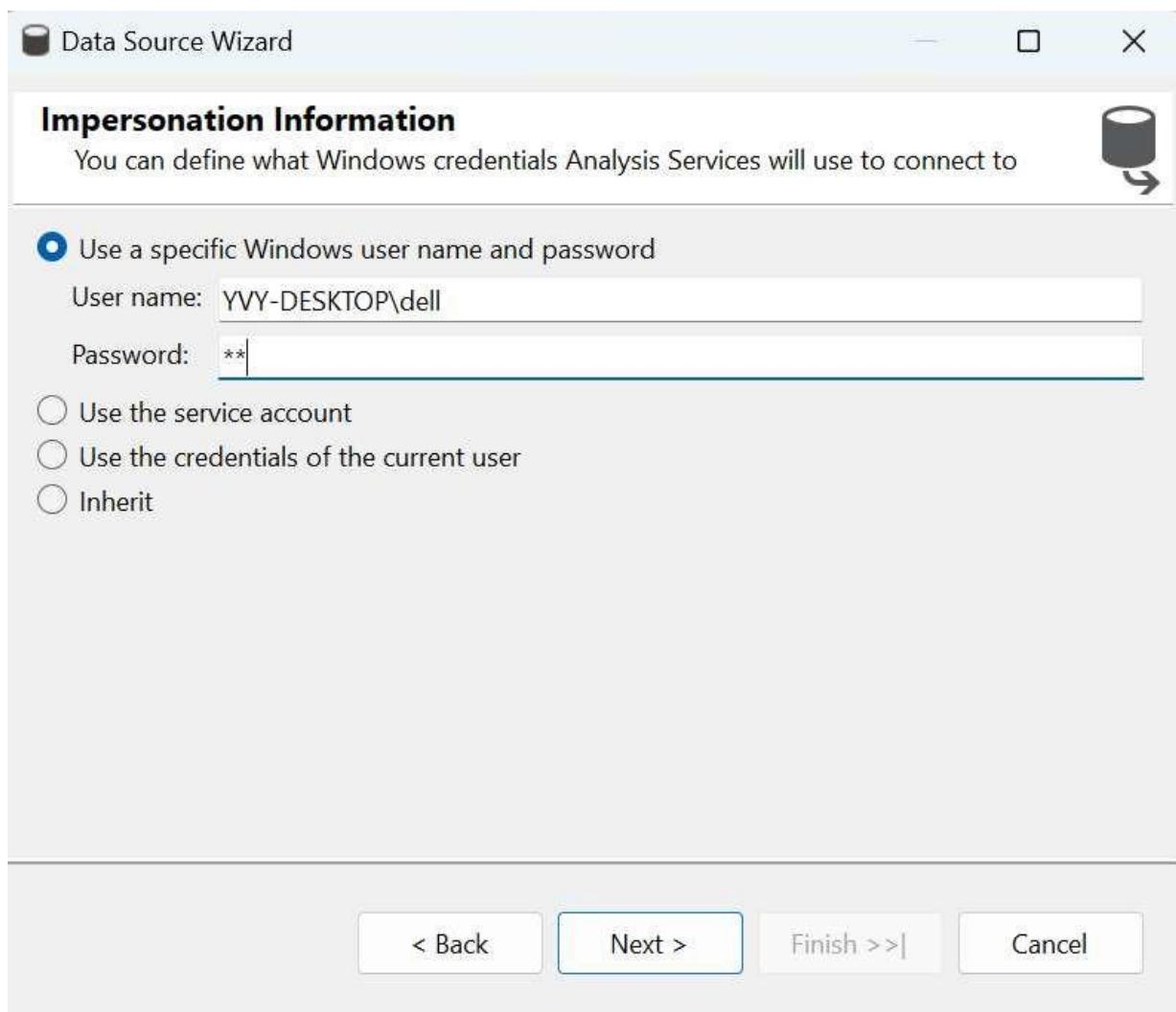
- Define the connection



- Thêm tên server và chọn database WH.



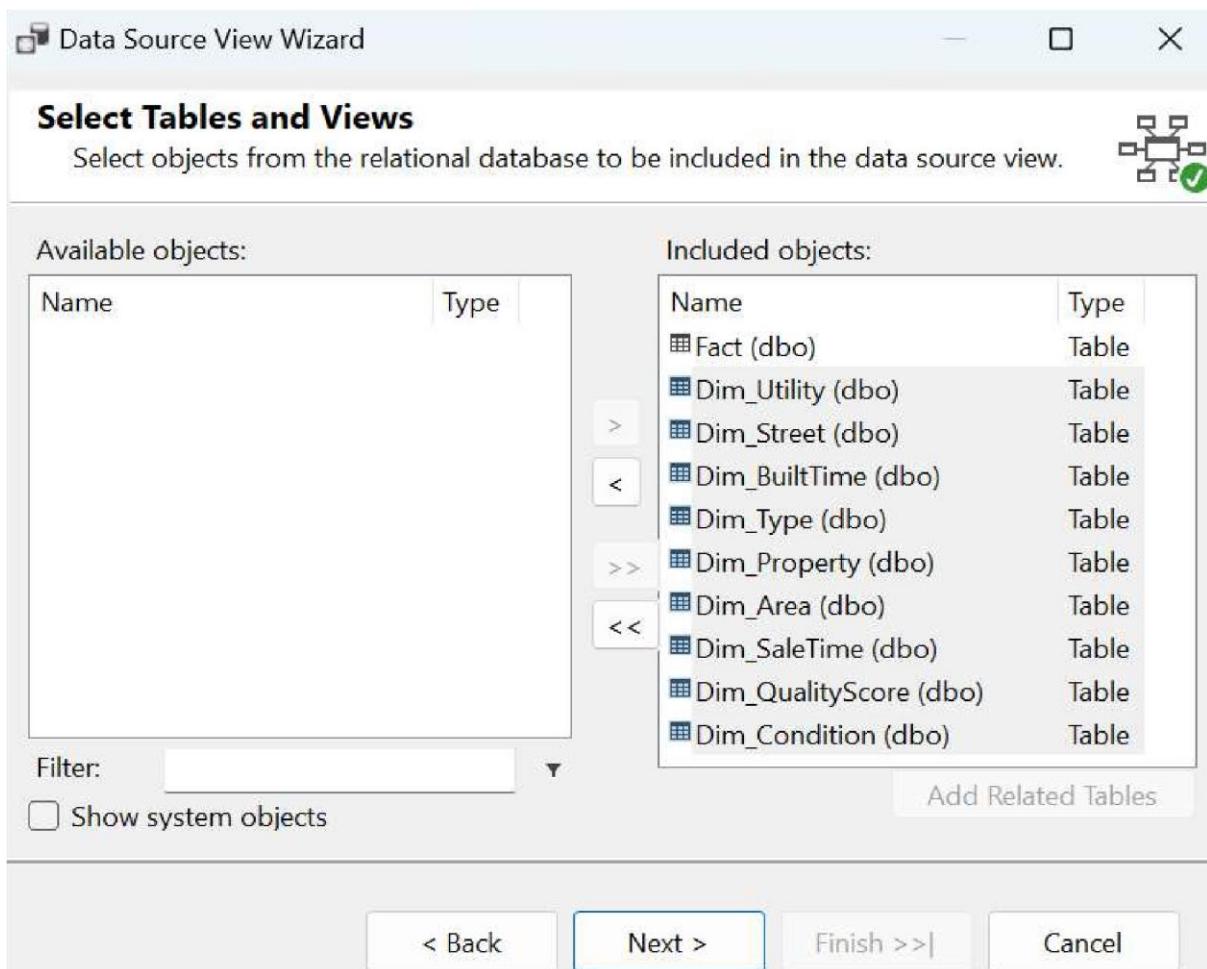
- Chọn Use the specific Windows user name and password.



- Nhấn Next -> Finish để hoàn tất quá trình.

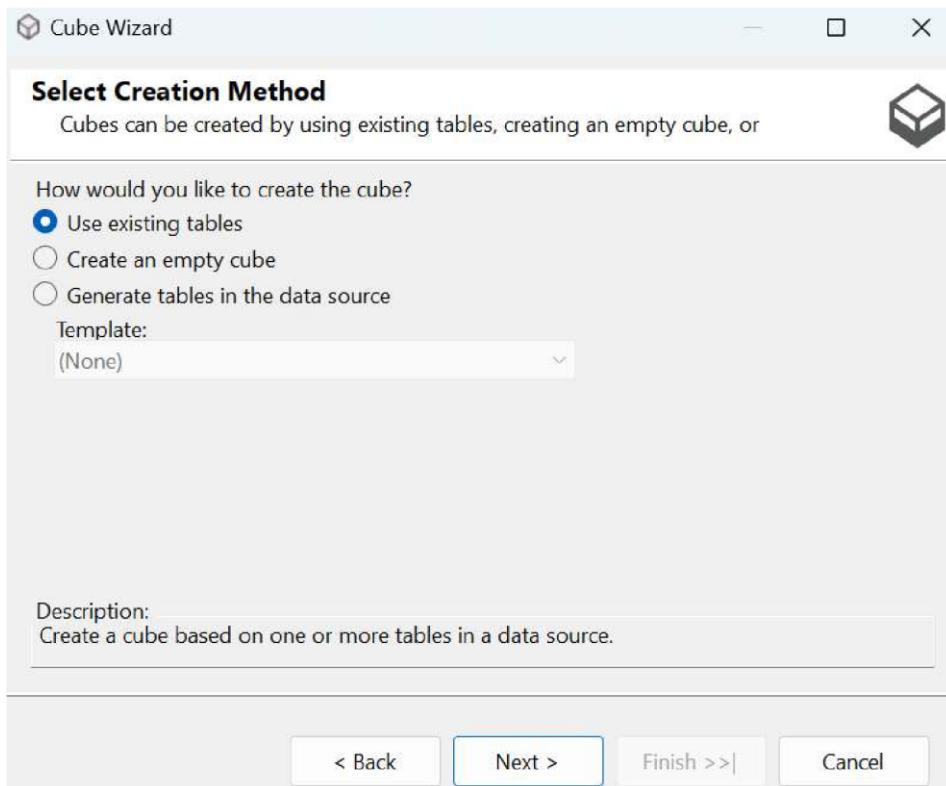
3.1.2. Tạo datasource Views

- Select Tables and Views: Chọn bảng Fact từ Available object chuyển sang Included object. Sau đó click chọn Add Related Tables để thêm các bảng có mối quan hệ với Fact.

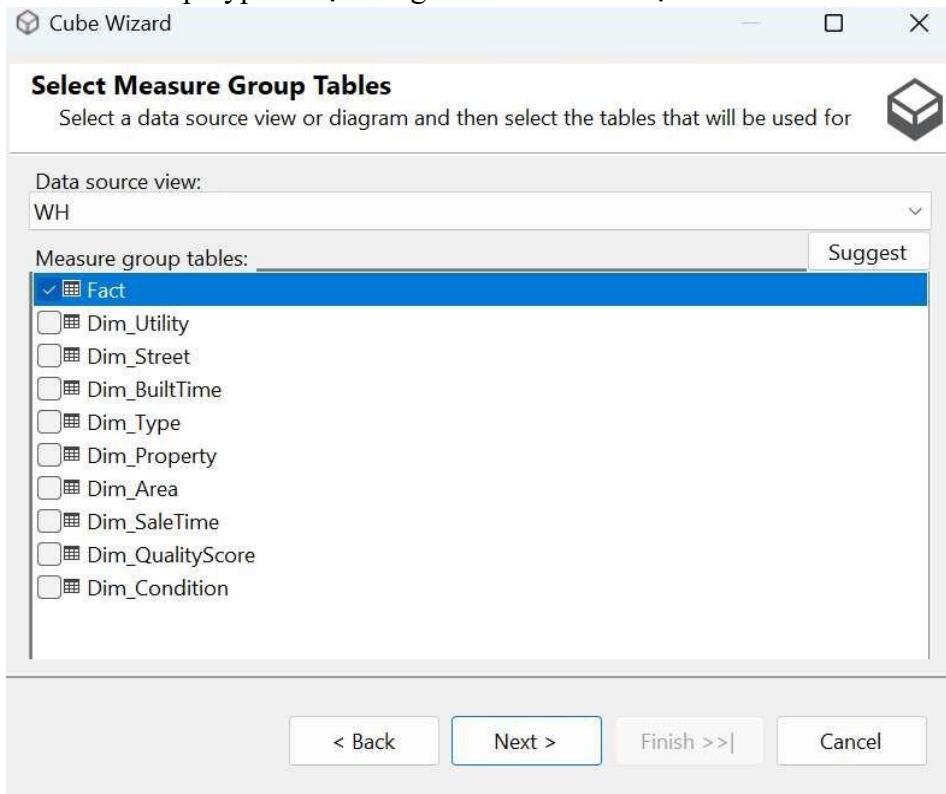


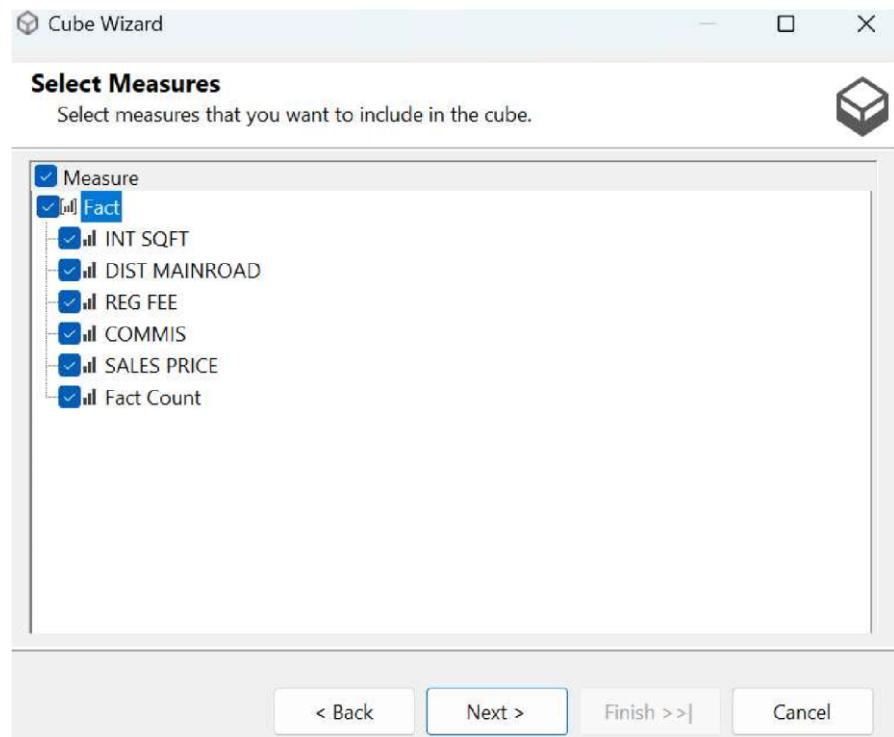
3.1.3. Tạo Cubes

- Select Creation Method: Chọn “Use existing tables” và ấn Next.



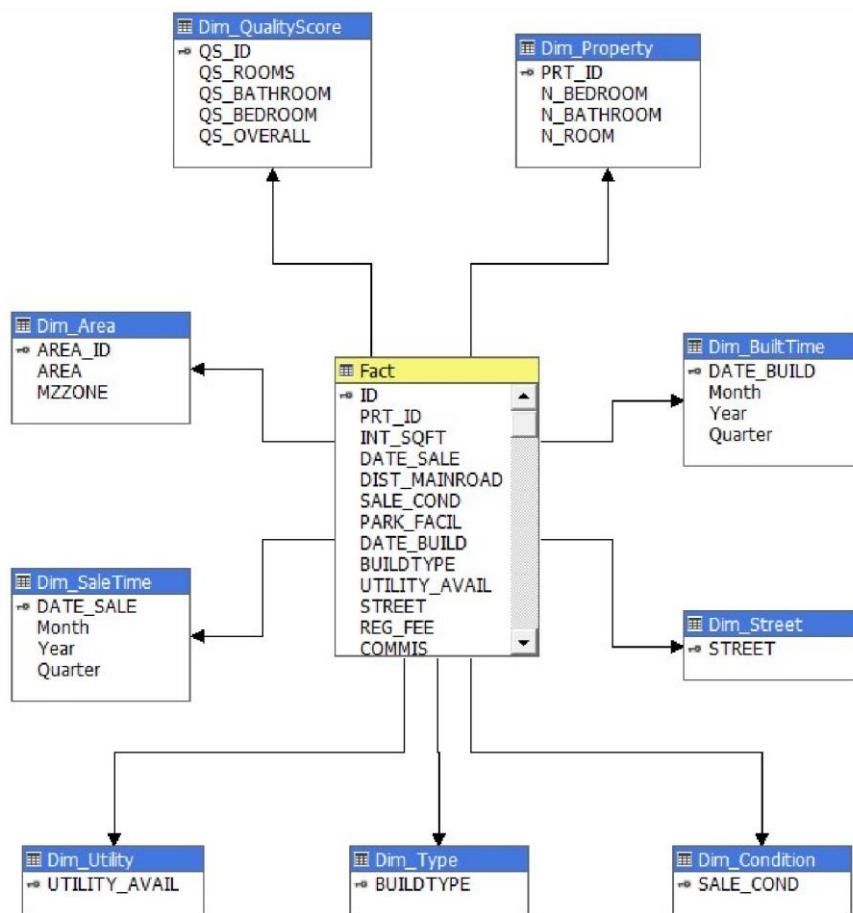
- Select Measure Group Type: Chọn bảng Fact để tìm các độ đo.





- Tiếp tục ấn Next và cuối cùng là Finish để kết thúc quá trình tạo Cube.

Và đây là kết quả sau khi tạo thành công là sơ đồ hình sao.

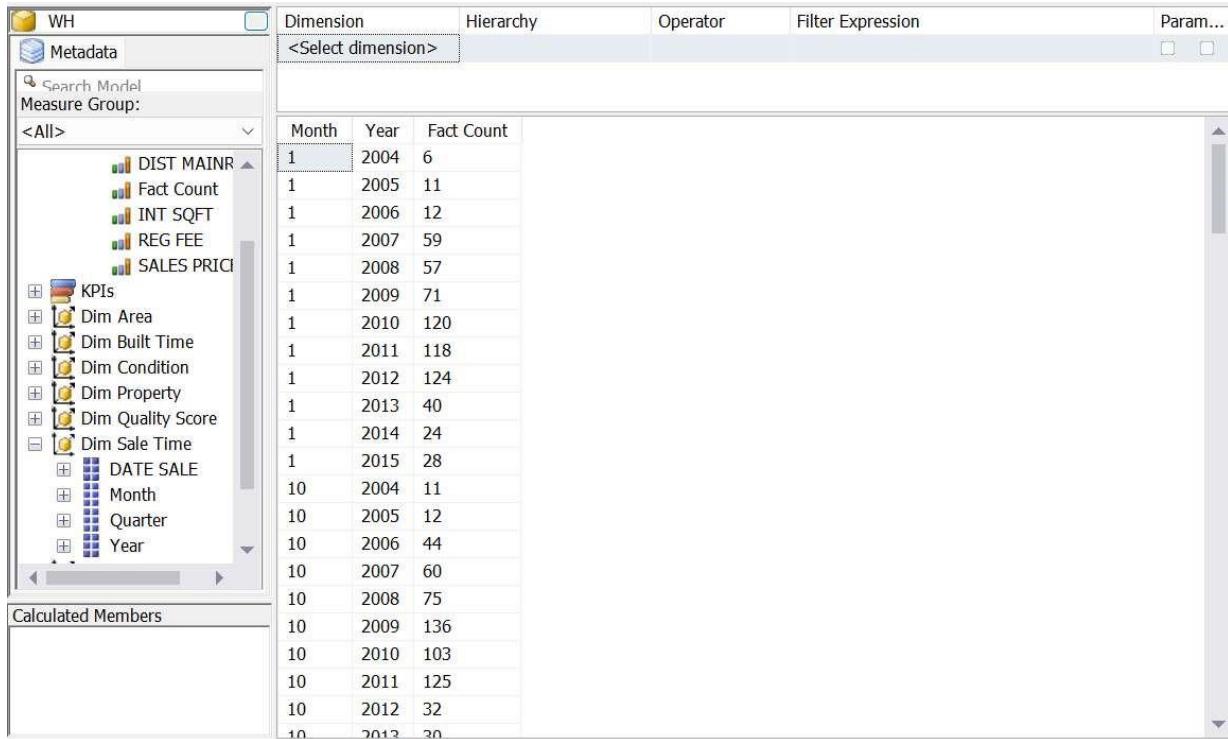


3.2. Thực thi truy vấn trên Visual Studio và Power BI và Excel

- Thao tác và phân tích trên cơ sở dữ liệu nhiều chiều được OLAP cung cấp một số công cụ phân tích từ đơn giản đến phức tạp như: Cuộn lên (Roll up), truy xuống (Drill down), chọn và chiều (Slice and Dice), xoay chiều (Pivot).

3.2.1. Cuộn lên (Roll up)

3.2.1.1. Đếm số bất động sản được bán ra theo từng tháng, năm - Kết quả truy vấn trên Visual Studio:



The screenshot shows the Microsoft Analysis Services Dimension Designer interface. On the left, there's a tree view of the data model with nodes like 'WH', 'Metadata', 'Search Model', 'Measure Group', and various dimensions such as 'KPIs', 'Dim Area', etc. In the center, there's a table titled 'Fact Count' with columns 'Month' and 'Year'. The data is as follows:

Month	Year	Fact Count
1	2004	6
1	2005	11
1	2006	12
1	2007	59
1	2008	57
1	2009	71
1	2010	120
1	2011	118
1	2012	124
1	2013	40
1	2014	24
1	2015	28
10	2004	11
10	2005	12
10	2006	44
10	2007	60
10	2008	75
10	2009	136
10	2010	103
10	2011	125
10	2012	32
10	2013	30

- Kết quả truy vấn trên Power BI (Biểu đồ dạng table):

The screenshot shows the Power BI Desktop interface. On the left, there is a data grid titled "Month Year Fact Count" with the following data:

Month	Year	Fact Count
1	2004	6
1	2005	11
1	2006	12
1	2007	59
1	2008	57
1	2009	71
1	2010	120
1	2011	118
1	2012	124
1	2013	40
1	2014	24
1	2015	28
10	2004	11
Total		7109

The Power BI ribbon is visible at the top, and the right side features the "Filters", "Visualizations", and "Data" panes.

- Kết quả truy vấn trên Excel:

The screenshot shows an Excel spreadsheet with a pivot table named "Q5". The pivot table has "Fact Count" as the Row Labels and "Column Labels" as the Column Labels. The data is summarized by year (2004-2015) and month (1-12). The "Grand Total" row shows the following values:

	10	11	12	2	3	4	5	6	7	8	9	Grand Total
2004	6	11	13	12	6	7	10	9	13	11	4	116
2005	11	12	5	6	6	7	11	9	6	9	12	107
2006	12	44	38	61	11	11	17	10	10	12	9	253
2007	59	60	44	54	57	55	50	63	48	42	57	651
2008	57	75	57	71	57	68	53	77	58	78	72	914
2009	71	136	111	124	57	77	84	86	79	100	94	103
2010	120	103	102	97	121	116	121	117	104	125	123	92
2011	118	125	93	102	86	96	127	116	112	100	109	116
2012	124	32	45	35	123	121	36	31	21	24	34	28
2013	40	30	39	27	23	25	31	36	39	28	30	17
2014	24	25	35	32	24	24	22	34	31	31	28	25
2015	28											51
Grand Total	670	653	582	621	594	607	562	588	521	560	583	568
												7109

3.2.2. Truy xuống (Drill down)

3.2.2.1. Đếm số bất động sản được bán ra theo từng quý, từng tháng - Kết quả truy vấn trên Visual Studio:

The screenshot shows the Power BI Data View interface. On the left, there's a navigation pane with a tree view of the data model, including sections like Metadata, Measure Group, and various dimensions such as Dim Area, Dim Built Time, etc. The main area displays a table with three columns: Quarter, Month, and Fact Count. The data is as follows:

Quarter	Month	Fact Count
1	1	670
1	2	594
1	3	607
2	4	562
2	5	588
2	6	521
3	7	560
3	8	583
3	9	568
4	10	653
4	11	582
4	12	621

- Kết quả truy vấn trên Power BI + Biểu đồ dạng table:

Đếm số bất động sản được bán ra theo từng quý

The screenshot shows the Power BI Desktop interface. A table visualization is displayed on the left, titled "Quarter Fact Count", showing the same data as the previous screenshot. The Data pane on the right is open, showing the data source structure. The "Dim Sale Time" dimension is selected, with its children "DATE SALE", "Month", "Quarter", and "Year" visible. The "Quarter" field is currently selected as a filter.

Đếm số bất động sản được bán ra theo từng tháng

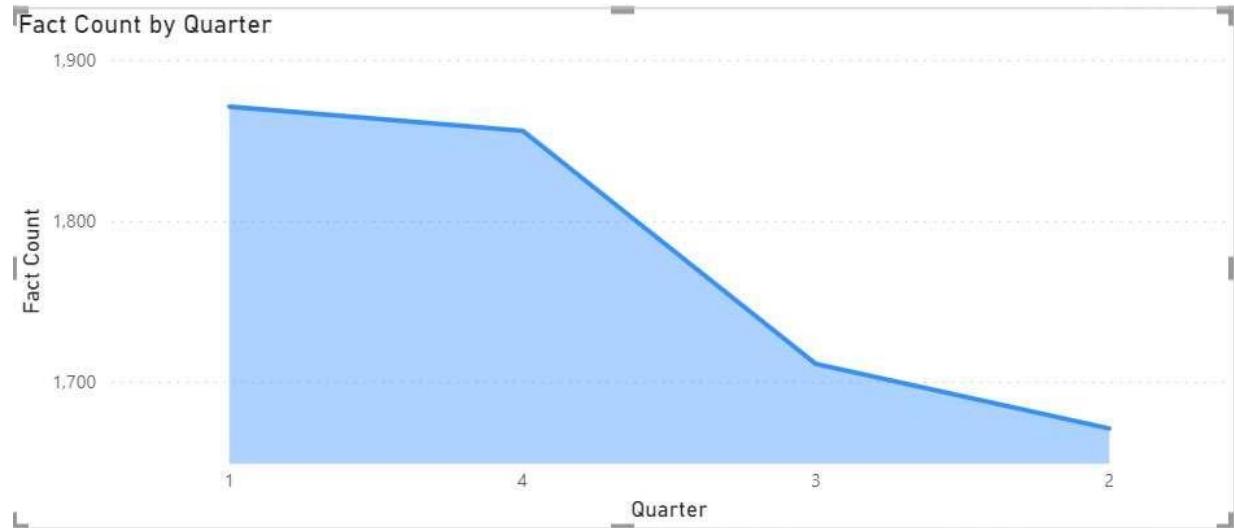
The screenshot shows the Power BI Desktop interface. On the left, there is a table visual titled "Month Fact Count" with the following data:

Month	Fact Count
1	670
10	653
11	582
12	621
2	594
3	607
4	562
5	588
6	521
7	560
8	583
9	568
Total	7109

On the right, the "Filters" pane is open, showing filters applied to the visual. The X-axis is set to "Month" and the Y-axis is set to "Fact Count". The "Visualizations" pane shows various chart types available for selection.

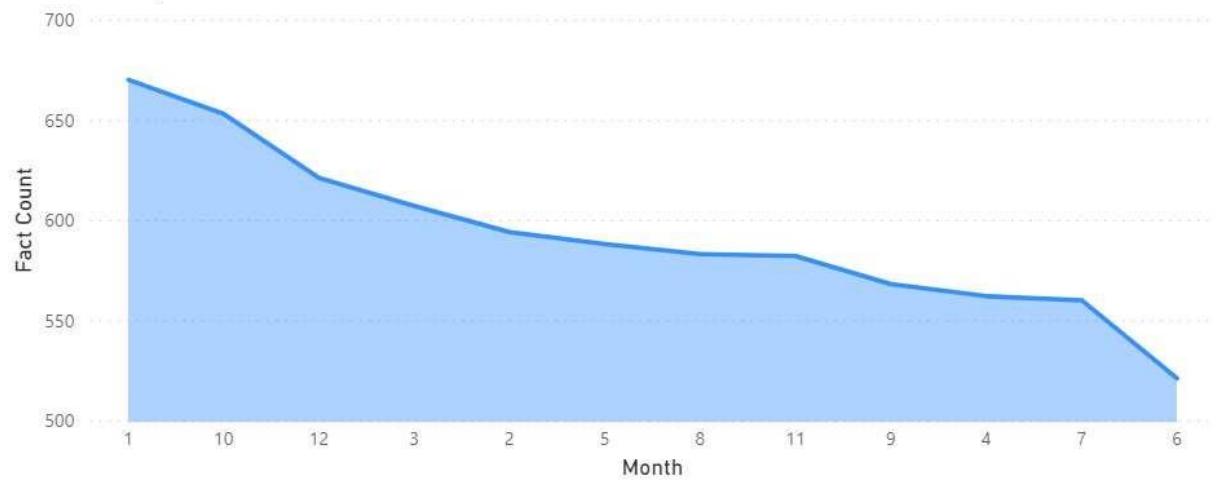
+ Biểu đồ dạng Chart:

Báo cáo dự đoán xu hướng bán bất động sản theo từng quý (Area chart)



Báo cáo dự đoán xu hướng bán bất động sản theo từng tháng (Area chart)

Fact Count by Month



- Kết quả truy vấn trên Excel:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF			
1	Fact Count	Column Labels	1	10	11	12	2	3	4	5	6	7	8	9	Grand Total																				
2	Row Labels	1	670		594	607									1871																				
3	1		670		594	607									1871																				
4	2			562	588	521									1671																				
5	3				560	583	568								1711																				
6	4				653	582	621								1856																				
7	Grand Total		670	653	582	621	594	607	562	588	521	560	583	568	7109																				
8																																			
9																																			
10																																			
11																																			
12																																			
13																																			
14																																			
15																																			
16																																			
17																																			
18																																			
19																																			
20																																			

3.2.3. Chọn và chiếu (Slice and Dice)

3.2.3.1. Trong năm 2009 truy vấn 3 khu vực có nhiều bất động sản được bán ra nhất.

- Kết quả truy vấn trên Visual Studio:

IS217 – Kho dữ liệu và OLAP

The screenshot shows the Analysis Services Management Studio interface. On the left, there's a tree view of the cube structure under 'WH'. The 'Measure Group' section is expanded, showing dimensions like Dim Area, Dim Sale Time, etc. In the center, a table titled 'Fact Count' is displayed with the following data:

Year	AREA	Fact Count
2009	Chromepet	207
2009	Karapakkam	257
2009	KK Nagar	249

- Kết quả truy vấn trên Power BI (Biểu đồ dạng table):

The screenshot shows the Power BI Desktop interface. A table visualization is displayed with the following data:

AREA	Year	Fact Count
Chromepet	2009	207
Karapakkam	2009	257
KK Nagar	2009	249
Total		713

On the right side, the 'Filters' pane is open, showing a filter for 'Year' set to '2009'. The 'Visualizations' pane is also visible, showing various chart and report options.

- Kết quả truy vấn trên Excel:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Year	2009																
2																		
3	Row Labels	Fact Count																
4	Chromepet	207																
5	Karapakkam	257																
6	KK Nagar	249																
7	Grand Total	713																
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		

3.2.3.2. Thống kê những bất động sản có giá bán ra > 23000000. -

Kết quả truy vấn trên Visual Studio:

Dimension	Hierarchy	Operator	Filter Expression
Dim Property	# PRT ID	Custom	
<Select dimension>			
	PRT ID	SALES PRICE	
	67	23314580	
	714	23013500	
	2996	23407860	
	8983	23667340	
	9293	23307000	
	9521	23247590	

- Kết quả truy vấn trên Power BI (Biểu đồ dạng table):

The screenshot shows the Power BI Desktop interface. A table visualization is displayed on the left, titled "PRT ID SALES PRICE". The data includes rows for PRT IDs 67, 714, 2996, 8983, 9293, and 9521, each with a corresponding Sales Price. A total row at the bottom shows "Total 139957870". On the right side, there are three filter panes: "Filters", "Visualizations", and "Data". The "Filters" pane shows filters for "PRT ID" (is All) and "SALES PRICE" (is greater than 23000...). The "Visualizations" pane lists various chart types like bar, line, and map. The "Data" pane shows columns "PRT ID" and "SALES PRICE" and a "Drill through" option.

- Kết quả truy vấn trên Excel:

The screenshot shows a Microsoft Excel spreadsheet. The data is presented in a table with columns labeled A, B, and C. Row 1 contains column headers "Row Labels" and "SALES PRICE". Rows 2 through 10 list individual sales entries with PRT IDs (67, 714, 2996, 8983, 9293, 9521) and their corresponding Sales Prices. Row 10 is a Grand Total row with the value "139957870". The Excel ribbon is visible at the top, showing tabs for Home, Insert, Page Layout, Formulas, Data, Review, View, Automate, Help, and Acrobat.

3.2.3.3. Thống kê tổng tiền phí cho người môi giới theo từng loại bất động sản trong quý 3 năm 2011

- Kết quả truy vấn trên Visual Studio:

The screenshot shows the Analysis Services Management Studio interface. At the top, there are tabs for 'WH.cube [Design]', 'Dim Condition.dim [Design]', 'Dim Type.dim [Design]', and 'Dim Quality Score.dim [Design]'. Below the tabs is a toolbar with various icons for cube management. On the left, a navigation pane displays the database structure under 'WH' and 'Metadata', including 'Dim Condition', 'Dim Property', 'Dim Quality Score', 'Dim Sale Time', and 'Dim Street'. The main area contains an MDX query editor with the following code:

```
SELECT
    {  
        [Dim Sale Time].[Quarter].<All>, [Dim Sale Time].[Year].<All>  
    }  
    ON COLUMNS,  
    {  
        [Dim Sale Time].[Quarter].<All>, [Dim Sale Time].[Year].<All>  
    }  
    ON ROWS  
FROM [WH]
```

Below the query editor is a results grid showing data from the 'Dim Sale Time' dimension:

Quarter	BUILDTYPE	REG FEE
3	Commercial	48097282
3	House	35376231
3	Others	35046876

- Kết quả truy vấn trên Power BI (Biểu đồ dạng table):

The screenshot shows the Power BI Desktop interface. The ribbon is visible with tabs like File, Home, Insert, Modeling, View, Optimize, and Help. The Home tab is selected. In the center, there is a table visualization with the following data:

Year	Quarter	BUILDTYPE	REG FEE
2011	3	Commercial	48097282
2011	3	House	35376231
2011	3	Others	35046876
Total			118520389

On the right side, there are sections for 'Filters', 'Visualizations', and 'Data'. The 'Filters' section shows filters for 'Year', 'Quarter', 'BUILDTYPE', and 'REG FEE'. The 'Visualizations' section shows a list of available visualizations. The 'Data' section shows a hierarchical list of dimensions and their fields.

- Kết quả truy vấn trên Excel:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Year	2011														
2	Quarter	3														
3																
4	Row Labels	REG FEE														
5	Commercial	48097282														
6	House	35376231														
7	Others	35046876														
8	Grand Total	118520389														
9																
10																
11																
12																
13																
14																
15																

3.2.3.4. Thống kê Top 3 tiện ích của bất động sản có số lượng được bán ra cao nhất - Kết quả truy vấn trên Visual Studio:

UTILITY AVAIL	Fact Count
AllPub	1887
No Sewer	1829
NoSeWa	1871

- Kết quả truy vấn trên Power BI:

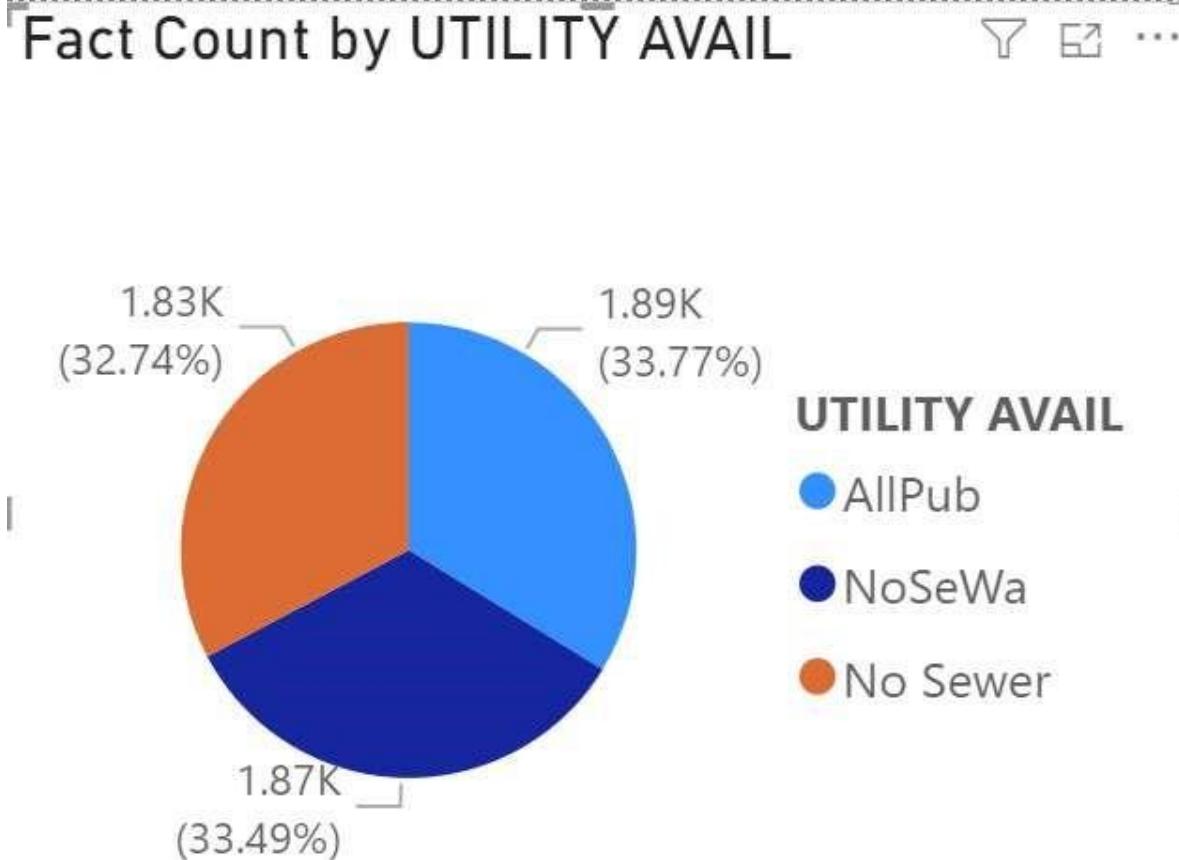
+ Biểu đồ dạng table:

The screenshot shows the Power BI Desktop interface. On the left, there is a table visualization titled "UTILITY AVAIL Fact Count" with the following data:

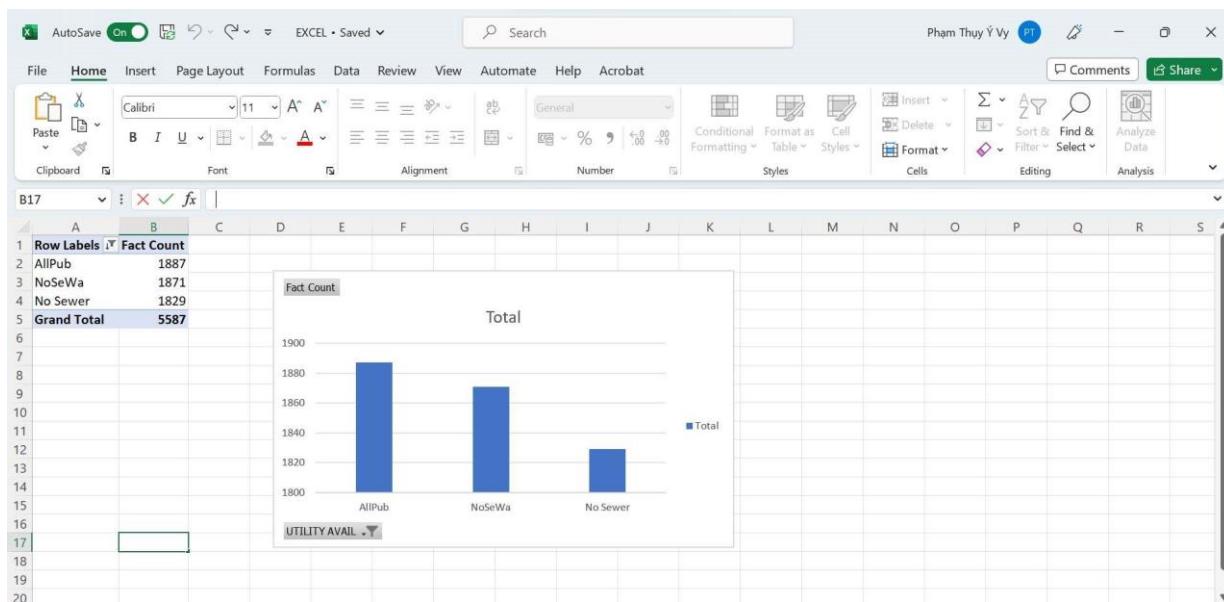
AllPub	1887
No Sewer	1829
NoSeWa	1871
Total	5587

The "Filters" pane on the right shows a filter for "Fact Count" set to "Top 3 by Fact Count". The "Visualizations" pane shows various chart and report options. The "Data" pane on the far right lists dimensions and facts, with "Fact Count" selected.

+ Biểu đồ dạng Chart (Pie chart):



- Kết quả truy vấn trên Excel:



3.2.3.5. Thống kê tổng số bất động sản được bán ra trong năm 2004, 2011, 2014 có chất lượng tất cả các phòng trên 4.0 điểm - Kết quả truy vấn trên Visual Studio:

Year	Dim Quality Score	Fact Count
2004	4.1	2
2004	4.2	4
2004	4.3	2
2004	4.4	1
2004	4.5	5
2011	4.1	49
2011	4.2	34
2011	4.3	35
2011	4.4	21
2011	4.5	12
2011	4.6	10
2011	4.7	5
2011	4.8	3
2011	4.9	1
2014	4.1	8
2014	4.2	15

- Kết quả truy vấn trên Power BI (Biểu đồ dạng table):

The screenshot shows the Power BI Desktop interface. On the left, there is a data grid titled "Year QS OVERALL Fact Count" with the following data:

Year	QS OVERALL	Fact Count
2004	4.1	2
2004	4.2	4
2004	4.3	2
2004	4.4	1
2004	4.5	5
2011	4.1	49
2011	4.2	34
2011	4.3	35
2011	4.4	21
2011	4.5	12
2011	4.6	10
2011	4.7	5
2011	4.8	3
Total		228

The Power BI ribbon is visible at the top, and the right side features the "Filters", "Visualizations", and "Data" panes.

- Kết quả truy vấn trên Excel:

The screenshot shows an Excel spreadsheet with a pivot table. The pivot table has "Fact Count" as the column label and "Row Labels" as the row label. The data is organized by year (2004, 2011, 2014) and includes a Grand Total. The data is as follows:

	Fact Count	Column Labels	2004	2011	2014	Grand Total	
4.1				2	49	8	59
4.2				4	34	15	53
4.3				2	35	9	46
4.4				1	21	4	26
4.5				5	12	4	21
4.6					10	1	11
4.7					5	2	7
4.8					3	1	4
4.9					1		1
Grand Total			14	170	44		228

3.2.3.6. Thống kê tổng số bất động sản đã bán theo từng khu vực với tình trạng của bất động sản khi bán là AbNormal

- Kết quả truy vấn trên Visual Studio:

IS217 – Kho dữ liệu và OLAP

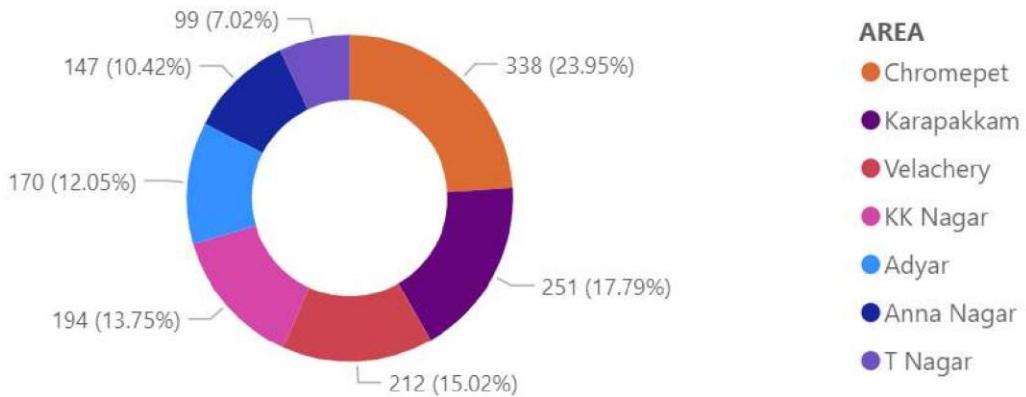
AREA	Fact Count
Adyar	170
Anna Nagar	147
Chromepet	338
Karapakkam	251
KK Nagar	194
T Nagar	99
Velachery	212

- Kết quả truy vấn trên Power BI + Biểu đồ dạng Matrix:

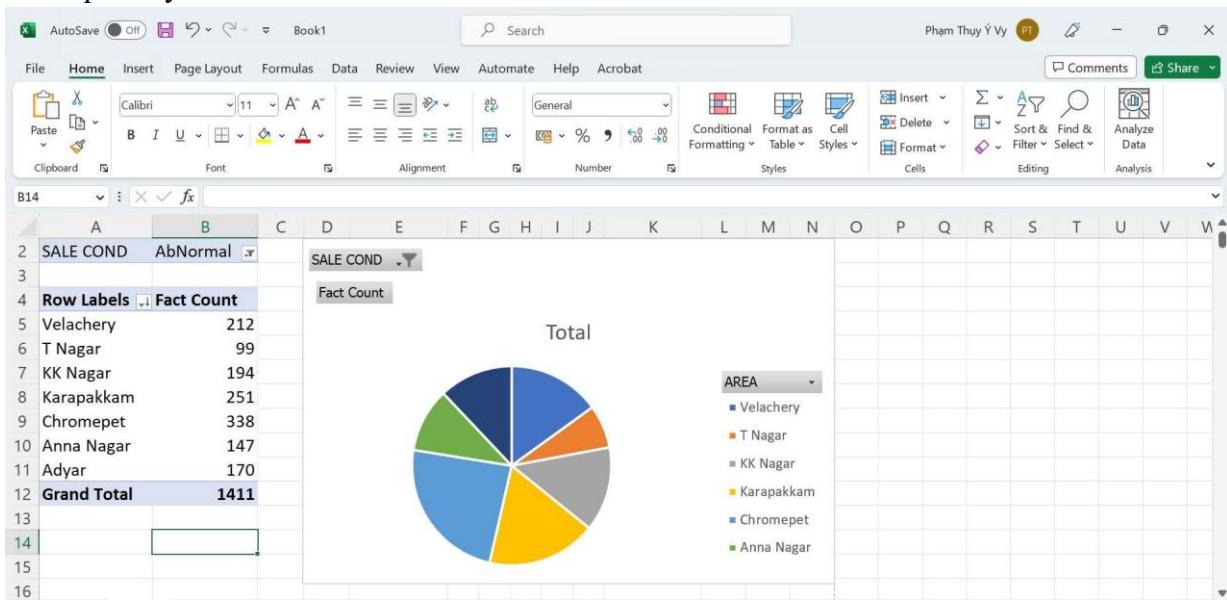
SALE COND	Adyar	Anna Nagar	Chromepet	Karapakkam	KK Nagar	T Nagar	Velachery	Total
AbNormal	170	147	338	251	194	99	212	1411
Total	170	147	338	251	194	99	212	1411

+ Biểu đồ dạng Chart:

Fact Count by AREA



- Kết quả truy vấn trên Excel:



3.2.4. Xoay chiều (Pivot)

3.2.4.1. Thống kê tổng số bất động sản được bán ra theo quý bán và khu vực.

- Kết quả truy vấn trên Visual Studio:

IS217 – Kho dữ liệu và OLAP

Quarter	AREA	Fact Count
1	Adyar	209
1	Anna Nagar	195
1	Chromepet	433
1	Karapakkam	366
1	KK Nagar	282
1	T Nagar	126
1	Velachery	260
2	Adyar	197
2	Anna Nagar	186
2	Chromepet	395
2	Karapakkam	268
2	KK Nagar	250
2	T Nagar	138
2	Velachery	237
3	Adyar	193
3	Anna Nagar	197

 The table has 18 rows in total."/>

- Kết quả truy vấn trên Power BI (Biểu đồ dạng Matrix):

+ Cột là AREA(khu vực) và dòng là Quarter (quý bán)

	1	2	3	4	Total
1	209	197	193	175	774
2	195	186	197	210	788
3	433	395	357	483	1702
4	366	268	229	375	1366
Total	282	138	110	127	997
	126	237	234	250	501
	260	237	234	250	981
	1871	1671	1711	1856	7109

 The matrix has 14 rows in total. The Power BI interface shows various filters and visualizations on the right side."/>

+ Cột là Quarter (quý bán) và dòng là Area (khu vực)

The screenshot shows the Power BI Desktop interface. On the left is a data grid visual titled "AREA" with columns 1, 2, 3, 4, and Total. The data includes rows for Adyar, Anna Nagar, Chromepet, Karapakkam, KK Nagar, T Nagar, and Velachery, with various numerical values. To the right of the visual is the "Filters" pane, which lists "AREA is (All)", "Fact Count is (All)", and "Quarter is (All)". Below it is the "Visualizations" pane showing various chart and report options. The "Data" pane on the far right lists dimensions like Fact Count, Dim Area, and Dim Property, along with their respective fields.

AREA	1	2	3	4	Total
Adyar	209	197	193	175	774
Anna Nagar	195	186	197	210	788
Chromepet	433	395	391	483	1702
Karapakkam	366	268	357	375	1366
KK Nagar	282	250	229	236	997
T Nagar	126	138	110	127	501
Velachery	260	237	234	250	981
Total	1871	1671	1711	1856	7109

- Kết quả truy vấn trên Excel:

+ Cột là AREA(khu vực) và dòng là Quarter (quý bán)

The screenshot shows an Excel spreadsheet with a pivot table selected. The row labels are "Velachery", "T Nagar", "KK Nagar", "Karapakkam", "Chromepet", "Anna Nagar", and "Adyar". The column labels are "Fact Count", "Column Labels", and "Grand Total". The data values correspond to the table in the Power BI screenshot. The Excel ribbon is visible at the top, showing tabs like File, Home, Insert, etc.

	Fact Count	Column Labels							
Row Labels	Velachery	T Nagar	KK Nagar	Karapakkam	Chromepet	Anna Nagar	Adyar	Grand Total	
1		260	126	282	366	433	195	209	1871
2		237	138	250	268	395	186	197	1671
3		234	110	229	357	391	197	193	1711
4		250	127	236	375	483	210	175	1856
10	Grand Total	981	501	997	1366	1702	788	774	7109

+ Cột là Quarter (quý bán) và dòng là Area (khu vực)

	Row Labels	2	3	4	Grand Total
6	Velachery	260	237	234	250
7	T Nagar	126	138	110	127
8	KK Nagar	282	250	229	236
9	Karapakkam	366	268	357	375
10	Chromepet	433	395	391	483
11	Anna Nagar	195	186	197	210
12	Adyar	209	197	193	175
13	Grand Total	1871	1671	1711	1856
14					7109

3.2.4.2. Thống kê tổng số bất động sản được bán ra theo từng khu vực và loại bất động sản.

- Kết quả truy vấn trên Visual Studio:

AREA	BUILDTYPE	Fact Count
Adyar	Commercial	247
Adyar	House	263
Adyar	Others	264
Anna Nagar	Commercial	252
Anna Nagar	House	276
Anna Nagar	Others	260
Chromepet	Commercial	562
Chromepet	House	591
Chromepet	Others	549
Karapakkam	Commercial	433
Karapakkam	House	486
Karapakkam	Others	447
KK Nagar	Commercial	322
KK Nagar	House	348
KK Nagar	Others	327
T Nagar	Commercial	178

- Kết quả truy vấn trên Power BI (Biểu đồ dạng Matrix):

+ Cột là AREA(khu vực) và dòng là BUILDTYPE(loại bất động sản)

BUILDTYPE	Adyar	Anna Nagar	Chromepet	Karapakkam	KK Nagar	T Nagar	Velachery	Total
Commercial	247	252	562	433	322	178	335	2329
House	263	276	591	486	348	156	324	2444
Others	264	260	549	447	327	167	322	2336
Total	774	788	1702	1366	997	501	981	7109

+ Cột là BUILDTYPE(loại bất động sản) và dòng là AREA(khu vực)

AREA	Commercial	House	Others	Total
Adyar	247	263	264	774
Anna Nagar	252	276	260	788
Chromepet	562	591	549	1702
Karapakkam	433	486	447	1366
KK Nagar	322	348	327	997
T Nagar	178	156	167	501
Velachery	335	324	322	981
Total	2329	2444	2336	7109

- Kết quả truy vấn trên Excel:

+ Cột là AREA(khu vực) và dòng là BUILDTYPE(loại bất động sản)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
4	Fact Count	Column Labels													
5	Row Labels	Velachery	T Nagar	KK Nagar	Karapakkam	Chromepet	Anna Nagar	Adyar	Grand Total						
6	Commercial		335	178	322	433	562	252	247	2329					
7	House		324	156	348	486	591	276	263	2444					
8	Others		322	167	327	447	549	260	264	2336					
9	Grand Total		981	501	997	1366	1702	788	774	7109					

+ Cột là BUILDTYPE(loại bất động sản) và dòng là AREA(khu vực)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
4	Fact Count	Column Labels													
5	Row Labels	Commercial	House	Others	Grand Total										
6	Velachery		335	324	322	981									
7	T Nagar		178	156	167	501									
8	KK Nagar		322	348	327	997									
9	Karapakkam		433	486	447	1366									
10	Chromepet		562	591	549	1702									
11	Anna Nagar		252	276	260	788									
12	Adyar		247	263	264	774									
13	Grand Total		2329	2444	2336	7109									

3.3. Ngôn ngữ MDX

3.3.1. Câu 1:

Mô tả: Đếm số bất động sản được bán ra theo từng tháng, năm

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,

NON EMPTY { ([Dim Sale Time].[Month].[Month].ALLMEMBERS * [Dim Sale Time].[Year].[Year].ALLMEMBERS) } ON ROWS

FROM [WH]

		Fact Count
1	2004	6
1	2005	11
1	2006	12
1	2007	59
1	2008	57
1	2009	71
1	2010	120
1	2011	118
1	2012	124
1	2013	40
1	2014	24
1	2015	00

3.3.2. Câu 2:

Mô tả: Thống kê tổng doanh số của bất động sản được bán ra theo loại bất động sản là “Commercial” và tiện ích của bất động sản là “ELO”

```

SELECT NON EMPTY { [Measures].[SALES PRICE] } ON COLUMNS,
NON EMPTY { ([Dim Type].[BUILDTYPE].[BUILDTYPE].ALLMEMBERS * [Dim Utility].[UTILITY AVAIL].[UTILITY AVAIL].ALLMEMBERS ) } ON ROWS
FROM ( SELECT ( { [Dim Utility].[UTILITY AVAIL].&[ELO] } ) ON COLUMNS
FROM ( SELECT ( { [Dim Type].[BUILDTYPE].&[Commercial] } ) ON COLUMNS
FROM [WH]))
```

		SALES PRICE
Commercial	ELO	2125533179

3.3.3. Câu 3:

Mô tả: Trong năm 2009 truy vấn 3 khu vực có nhiều bất động sản được bán ra nhất

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Dim Sale Time].[Year].[Year].ALLMEMBERS *
[Dim Area].[AREA].[AREA].ALLMEMBERS ) } ON ROWS
FROM ( SELECT ( TOPCOUNT( [Dim
Area].[AREA].children,3,[Measures].[Fact Count] ) ) ON COLUMNS
FROM ( SELECT ( { [Dim Sale Time].[Year]&[2009] } ) ON COLUMNS
FROM [WH]))
```

		Fact Count
2009	Chromepet	207
2009	Karapakkam	257
2009	KK Nagar	249

3.3.4. Câu 4:

Mô tả: Thống kê những bất động sản có giá bán ra > 23000000.

```

SELECT NON EMPTY { [Measures].[SALES PRICE] } ON COLUMNS,
NON EMPTY { ([Dim Property].[PRT ID].[PRT ID].ALLMEMBERS ) } ON ROWS
FROM ( SELECT ( FILTER( [Dim Property].[PRT ID].children, [Measures].[SALES PRICE]
> 23000000) ) ON COLUMNS
FROM [WH])

```

SALES PRICE	
67	23314580
714	23013500
2996	23407860
8983	23667340
9293	23307000
9521	23247590

3.3.5. Câu 5:

Mô tả: Thống kê tổng tiền phí cho người môi giới theo từng loại bất động sản trong quý bán 3 năm 2011

```
SELECT NON EMPTY { [Measures].[REG FEE] } ON COLUMNS,  
NON EMPTY { ([Dim Sale Time].[Quarter].[Quarter].ALLMEMBERS * [Dim  
Type].[BUILDTYPE].[BUILDTYPE].ALLMEMBERS ) } ON ROWS  
FROM ( SELECT ( { [Dim Sale Time].[Year].&[2011] } ) ON COLUMNS  
FROM ( SELECT ( { [Dim Sale Time].[Quarter].&[3] } ) ON COLUMNS FROM [WH]))  
WHERE ( [Dim Sale Time].[Year].&[2011] )
```

		REG FEE
3	Commercial	48097282
3	House	35376231
3	Others	35046876

3.3.6. Câu 6:

Mô tả: Thống kê Top 3 tiện ích của bất động sản có số lượng được bán ra cao nhất

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,  
NON EMPTY { ([Dim Utility].[UTILITY AVAIL].[UTILITY AVAIL].ALLMEMBERS ) }  
ON ROWS  
FROM ( SELECT ( TOPCOUNT( [Dim Utility].[UTILITY AVAIL].children,  
3,[Measures].[Fact Count] ) ) ON COLUMNS  
FROM [WH])
```

Messages		Results
	Fact Count	
AllPub	1887	
No Sewer	1829	
NoSeWa	1871	

3.3.7. Câu 7:

Mô tả: Thống kê tổng số bất động sản được bán ra trong năm 2004, 2011, 2014 có chất lượng tất cả các phòng trên 4.0 điểm

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Dim Sale Time].[Year].[Year].ALLMEMBERS * [Dim Quality Score].[QS
OVERALL].[QS OVERALL].ALLMEMBERS) } ON ROWS
FROM ( SELECT ( [Dim Quality Score].[QS OVERALL].&[4.1] : [Dim Quality Score].[QS
OVERALL].&[4.9] ) ON COLUMNS
FROM ( SELECT ( { [Dim Sale Time].[Year].&[2004], [Dim Sale Time].[Year].&[2011],
[Dim Sale Time].[Year].&[2014] } ) ON COLUMNS
FROM [WH]))
```

		Fact Count
2004	4.1	2
2004	4.2	4
2004	4.3	2
2004	4.4	1
2004	4.5	5
2011	4.1	49
2011	4.2	34
2011	4.3	35
2011	4.4	21
2011	4.5	12
2011	4.6	10
2011	4.7	5

3.3.8. Câu 8:

Mô tả: Thống kê tổng số bất động sản đã bán theo từng khu vực với tình trạng của bất động sản khi bán là AbNormal

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Dim Area].[AREA].[AREA].ALLMEMBERS) } ON ROWS
FROM ( SELECT ( { [Dim Condition].[SALE COND].&[AbNormal] } ) ON COLUMNS
FROM [WH])
WHERE ( [Dim Condition].[SALE COND].&[AbNormal] )
```

	Fact Count
Adyar	170
Anna Nagar	147
Chromepet	338
Karapakkam	251
KK Nagar	194
T Nagar	99
Velachery	212

3.3.9. Câu 9:

Mô tả: Thống kê tổng số bất động sản được bán ra theo quý bán và khu vực.

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,  
NON EMPTY { ([Dim Sale Time].[Quarter].[Quarter].ALLMEMBERS * [Dim  
Area].[AREA].[AREA].ALLMEMBERS ) } ON ROWS  
FROM [WH]
```

		Fact Count
1	Adyar	209
1	Anna Nagar	195
1	Chromepet	433
1	Karapakkam	366
1	KK Nagar	282
1	T Nagar	126
1	Velachery	260
2	Adyar	197
2	Anna Nagar	186
2	Chromepet	395
2	Karapakkam	268
		250

3.3.10. Câu 10:

Mô tả: Thống kê tổng số bất động sản được bán ra theo từng khu vực và loại bất động sản.

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,  
NON EMPTY { ([Dim Area].[AREA].[AREA].ALLMEMBERS * [Dim  
Type].[BUILDTYPE].[BUILDTYPE].ALLMEMBERS ) } ON ROWS  
FROM [WH]
```

		Fact Count
Adyar	Commercial	247
Adyar	House	263
Adyar	Others	264
Anna Nagar	Commercial	252
Anna Nagar	House	276
Anna Nagar	Others	260
Chromepet	Commercial	562
Chromepet	House	591
Chromepet	Others	549
Karapakkam	Commercial	433
Karapakkam	House	486
		447

CHƯƠNG 4: KHAI PHÁ DỮ LIỆU - DATA MINING

Mục tiêu: Dự đoán giá nhà dựa trên

4.1. Nhập thư viện

```
↳ Import libraries

↳ pip install graphviz
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: graphviz in /usr/local/lib/python3.10/dist-packages (0.20.1)

↳ import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import tree
import graphviz
from sklearn.tree import export_graphviz
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV

from sklearn import preprocessing
from sklearn.metrics import mean_absolute_error
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
```

4.2. Tiền xử lý dữ liệu

4.2.1. Đọc bộ dữ liệu:

```
↳ Read data

↳ data = pd.read_csv("/content/drive/MyDrive/IS217 OLAP/20522183_20521938_BaoCaoOLAP/DATA/Chennai_Housing_Sale.csv")
data
```

PRT_ID	AREA	INT_SFQT	DATE_SALE	DIST_MAINROAD	N_BEDROOM	N_BATHROOM	N_ROOM	SALE_COND	PARK_FACIL	...	UTILITY_AVAIL	STREET	MZONE	QS_ROOMS	QS_BATHROOM	QS_BEDROOM	
0	P03210	Karapakkam	1004	04/05/2011	131	1	1	3	AbNormal	Yes	...	AllPub	Paved	A	4.0	3.9	4.9
1	P09411	Anna Nagar	1986	19/12/2006	26	2	1	5	AbNormal	No	...	AllPub	Gravel	RH	4.9	4.2	2.5
2	P01812	Adyar	909	04/02/2012	70	1	1	3	AbNormal	Yes	...	ELO	Gravel	RL	4.1	3.8	2.2
3	P05346	Velachery	1855	13/03/2010	14	3	2	5	Family	No	...	No Sewer	Paved	I	4.7	3.9	3.6
4	P06210	Karapakkam	1226	05/10/2009	84	1	1	3	AbNormal	Yes	...	AllPub	Gravel	C	3.0	2.5	4.1
...	
7104	P03834	Karapakkam	598	03/01/2011	51	1	1	2	AdjLand	No	...	ELO	No Access	RM	3.0	2.2	2.4
7105	P10000	Velachery	1897	08/04/2004	52	3	2	5	Family	Yes	...	NoSeWa	No Access	RH	3.6	4.5	3.3
7106	P09594	Velachery	1614	25/08/2006	152	2	1	4	Normal Sale	No	...	NoSeWa	Gravel	I	4.3	4.2	2.9
7107	P06508	Karapakkam	787	03/08/2009	40	1	1	2	Partial	Yes	...	ELO	Paved	RL	4.6	3.8	4.1
7108	P09794	Velachery	1896	13/07/2005	156	3	2	5	Partial	Yes	...	ELO	Paved	I	3.1	3.5	4.3

7109 rows × 22 columns

4.2.2. Chọn các đặc trưng đầu vào và đặc trưng cần dự đoán:

X, y

```
x = data[['AREA', 'INT_SQFT', 'PARK_FACIL', 'BUILDTYPE', 'UTILITY_AVAIL']]  
x
```

	AREA	INT_SQFT	PARK_FACIL	BUILDTYPE	UTILITY_AVAIL
0	Karapakkam	1004	Yes	Commercial	AllPub
1	Anna Nagar	1986	No	Commercial	AllPub
2	Adyar	909	Yes	Commercial	ELO
3	Velachery	1855	No	Others	No Sewer
4	Karapakkam	1226	Yes	Others	AllPub
...
7104	Karapakkam	598	No	Others	ELO
7105	Velachery	1897	Yes	Others	NoSeWa
7106	Velachery	1614	No	House	NoSeWa
7107	Karapakkam	787	Yes	Commercial	ELO
7108	Velachery	1896	Yes	Others	ELO

7109 rows × 5 columns

```
y = pd.DataFrame(data['SALES_PRICE'])
y
```

	SALES_PRICE
0	7600000
1	21717770
2	13159200
3	9630290
4	7406250
...	...
7104	5353000
7105	10818480
7106	8351410
7107	8507000
7108	9976480

7109 rows × 1 columns

Do cả 2 mô hình Decision Tree và Random Forest chỉ nhận đặc trưng đầu vào ở dạng số nên ta cần xử lý dữ liệu ở X bằng cách sử dụng LabelEncoder() như sau:

▼ Note

- Park Facil: 1-Yes, 2-No
- BUILD TYPE: 0-Commercial, 1-House, 2-Others
- AREA: 0-Adyar, 1-Anna Nagar, 2-Chromepet, 3-KK Nagar, 4-Karapakkam, 5-T Nagar, 6-Velachery
- UTILITY AVAIL: 0-AlIPub, 1-ELO, 2-No Sewer, 3-NoSeWa

```
▶ label_encoder = preprocessing.LabelEncoder()

X[ 'AREA' ]= label_encoder.fit_transform(X[ 'AREA' ])
X[ 'BUILDTYPE' ]= label_encoder.fit_transform(X[ 'BUILDTYPE' ])
X[ 'UTILITY_AVAIL' ]= label_encoder.fit_transform(X[ 'UTILITY_AVAIL' ])
X[ 'PARK_FACIL' ]= label_encoder.fit_transform(X[ 'PARK_FACIL' ])

X
```

	AREA	INT_SQFT	PARK_FACIL	BUILDTYPE	UTILITY_AVAIL
0	4	1004	1	0	0
1	1	1986	0	0	0
2	0	909	1	0	1
3	6	1855	0	2	2
4	4	1226	1	2	0
...
7104	4	598	0	2	1
7105	6	1897	1	2	3
7106	6	1614	0	1	3
7107	4	787	1	0	1
7108	6	1896	1	2	1
7109 rows × 5 columns					

Và do là đối với 2 mô hình đó thì khoảng giá trị giữa các giá trị của thuộc tính cần dự đoán nên gần nhau để mô hình đạt kết quả tốt bằng cách dùng MinMaxScaler() như sau:

```
[ ] scaler = MinMaxScaler(feature_range=(0, 15))
y = scaler.fit_transform(y)
y

[ ] array([[ 3.79568154],
       [13.64049661],
       [ 7.67230625],
       ...,
       [ 4.31966603],
       [ 4.42816438],
       [ 5.45288421]])
```

4.2.3. Chia tập huấn luyện và tập kiểm thử:

▼ Train, Test - 7:3

```
[ ] X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.3)
```

4.3. Huấn luyện mô hình

Ta đều huấn luyện cả 2 mô hình Decision Tree và Random Forest theo 2 parameters sau
max_depth = 3 và random_state = 0

- Decision Tree

▼ Train model

```
[ ] regressor = DecisionTreeRegressor(max_depth=3, random_state = 0)
```

```
[ ] regressor.fit(X_train_dt,y_train_dt)
```

```
▼          DecisionTreeRegressor
DecisionTreeRegressor(max_depth=3, random_state=0)
```

- Random Forest

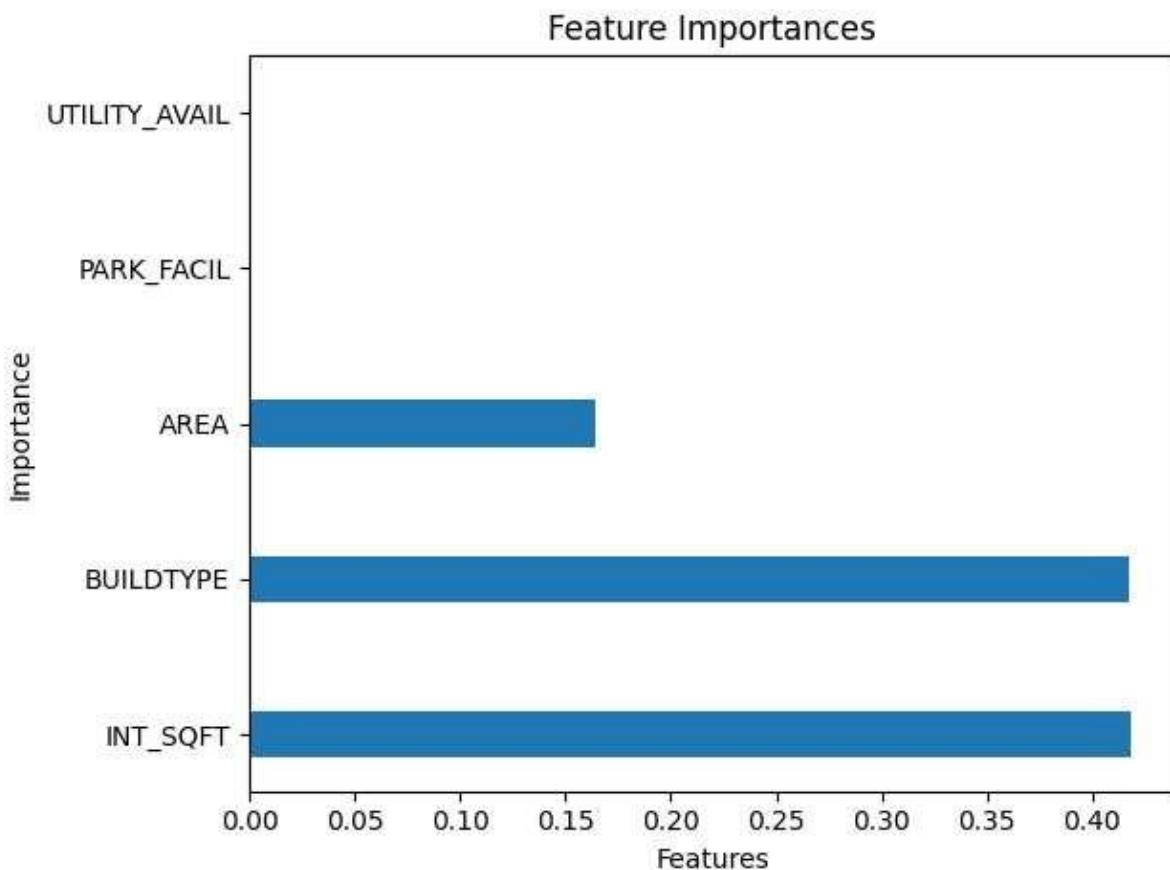
▼ Train model

```
[ ] rfr = RandomForestRegressor(max_depth=3, random_state = 0)
rfr.fit(x_train_rf, y_train_rf)

<ipython-input-51-78ad37cb8930>:2: DataConversionWarning: A
    rfr.fit(x_train_rf, y_train_rf)
    ^
    RandomForestRegressor
    RandomForestRegressor(max_depth=3, random_state=0)
```

Sau khi huấn luyện, ta tiến hành lọc ra những đặc trưng không ảnh hưởng với mô hình thông qua `feature_importances_`.

Từ kết quả nhận thấy cả 2 mô hình đều có 2 đặc trưng `PARK_FACIL` và `UTILITY_AVAIL` không ảnh hưởng:



Nên ta tiến hành loại bỏ 2 đặc trưng đó cả trên tập huấn luyện và tập kiểm thử:

▼ Drop low-ranking features

```
▶ x_train_dt = x_train_dt.drop(columns=['PARK_FACIL','UTILITY_AVAIL'])
x_train_dt
```

	AREA	INT_SQFT	BUILDTYPE
3585	2	1250	0
2627	2	1263	1
1068	5	1854	2
5334	6	1738	0
1473	2	790	2
...
4500	0	1230	2
1143	4	1647	1
1626	2	1081	0
2987	2	1237	0
5741	3	2042	2

4976 rows × 3 columns

```
[ ] x_test_dt = x_test_dt.drop(columns=['PARK_FACIL','UTILITY_AVAIL'])
x_test_dt
```

	AREA	INT_SQFT	BUILDTYPE
5353	6	1910	0
6535	1	1864	2
405	1	1882	2
538	0	1089	2
4565	3	1820	0
...
4463	5	1584	1
6913	6	1533	2
4056	2	810	0
6671	2	813	0
4201	4	1639	1

2133 rows × 3 columns

Tiếp theo ta tiến hành tuning mô hình bằng cách dùng GridSearchCV cho mô hình Decision Tree và RandomizedSearchCV cho mô hình Random Forest để đạt được kết quả tốt nhất có thể như sau:

- Decision Tree

- Tuning model

```
[17] parameters = {"splitter":["best","random"],
                   "max_depth" : [1,3,5,7,9,11,12],
                   "min_samples_leaf": [1,2,3,4,5,6,7,8,9,10],
                   "min_weight_fraction_leaf": [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9],
                   "max_features": ["auto","log2","sqrt",None],
                   "max_leaf_nodes": [None,10,20,30,40,50,60,70,80,90] }
```

```
[18] tun_regr = GridSearchCV(regressor,param_grid=parameters,cv=3,verbose=3)
```

```
[19] tun_regr.fit(x_train_dt,y_train_dt)
```

- Random Forest

- Tuning model

```
[ ] n_estimators = [int(x) for x in np.linspace(start = 5 , stop = 15, num = 10)]

r_grid = {'n_estimators': n_estimators,
          "max_depth" : [1,3,5,7,9,11,12],
          "max_features": ["auto", "log2"],
          'bootstrap': [True, False]}

[ ] rfr_random = RandomizedSearchCV(estimator=rfr, param_distributions=r_grid, n_iter = 20, cv = 3, verbose=2, random_state=42, n_jobs=-1, return_train_score=True)

[ ] rfr_random.fit(X_train_rf,y_train_rf)
```

Thông qua best_params_ ta được các parameters tốt nhất sau khi tuning của 2 mô hình sau:

- Decision Tree

```
[20] tun_regr.best_params_
{'max_depth': 5,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_samples_leaf': 1,
 'min_weight_fraction_leaf': 0.1,
 'splitter': 'best'}
```

- Random Forest

```
[ ] rfr_random.best_params_
{'n_estimators': 11, 'max_features': 'auto', 'max_depth': 7, 'bootstrap': True}
```

Cuối cùng ta tiến hành huấn luyện lại 2 mô hình với các parameters trên:

- Decision Tree

```
[21] regr_tun = DecisionTreeRegressor(max_depth=5, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_weight_fraction_leaf=0.1, splitter='best')
regr_tun.fit(X_train_dt,y_train_dt)
/usr/local/lib/python3.10/dist-packages/sklearn/tree/_classes.py:277: FutureWarning: `max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To
warnings.warn(
    DecisionTreeRegressor(max_depth=5, max_features='auto',
        min_weight_fraction_leaf=0.1)
```

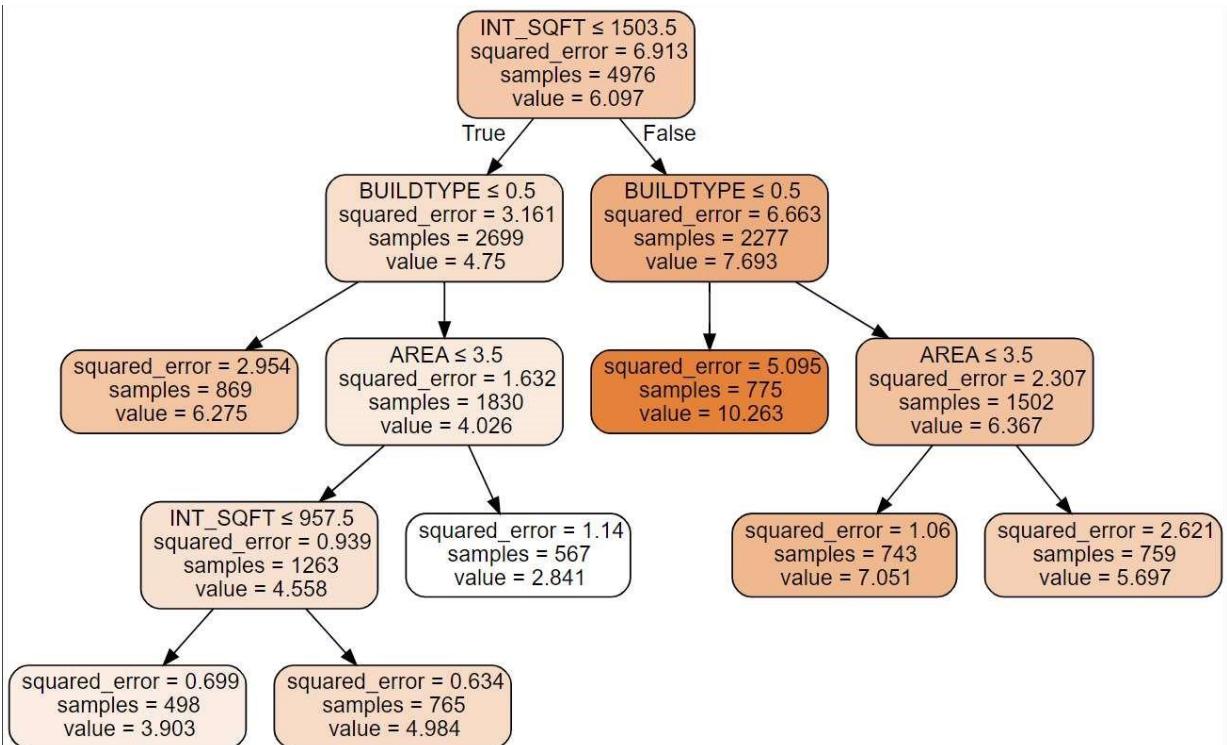
- Random Forest

```
[ ] rfr_tun = RandomForestRegressor(n_estimators=11,max_features='auto',max_depth=7,bootstrap=True)
rfr_tun.fit(X_train_rf,y_train_rf)
<ipython-input-35-e44ea4f6f585>:2: DataConversionWarning: A column-vector y was passed when a 1d array was
    rfr_tun.fit(X_train_rf,y_train_rf)
/usr/local/lib/python3.10/dist-packages/sklearn/ensemble/_forest.py:413: FutureWarning: `max_features='aut
warn(
    RandomForestRegressor
RandomForestRegressor(max_depth=7, max_features='auto', n_estimators=11)
```

4.4. Xuất ra tập luật

Do đây là 2 mô hình dựa trên ý tưởng là ‘cây’ nên ta tiến hành plot() ra 2 cây của 2 mô hình:

- Decision Tree



Do trước đó ta đã tiến hành Scaler() nên ta kết hợp dựa vào cây trên và inverse() lại giá trị thì

ta có thể rút ra được tập luật như sau

1. Diện tích ≤ 1503.5 và thuộc loại bất động sản Commercial → giá nhà 11155419.525
2. Diện tích 1503.5 và thuộc loại bất động sản Commercial → giá nhà 16874335.153
3. Diện tích 957.5, thuộc loại bất động sản House, Others và nằm trong khu vực Anna
4. Diện tích ≥ 1503.5 , thuộc loại bất động sản House, Others và nằm trong khu vực Anna Nagar, Chromepet, KK Nagar → giá nhà 7753897.993
5. Diện tích ≤ 1503.5 , thuộc loại bất động sản House, Others và nằm trong khu vực Nagar, Chromepet, KK Nagar → giá nhà 12268227.581
6. Diện tích ≥ 1503.5 , thuộc loại bất động sản House, Others và nằm trong khu vực Karapakkam, T Nagar, Velachery → giá nhà 6230957.071
7. $957.5 \leq$ Diện tích ≤ 1503.5 , thuộc loại bất động sản House, Others và nằm trong khu vực Karapakkam, T Nagar, Velachery → giá nhà 10326549.607
vực Anna Nagar, Chromepet, KK Nagar → giá nhà 9304085.504

- Random Forest

Do sau khi fine_tune() nên tìm thấy được max_depth = 7 nên cây khá sâu. Nên ta xuất ảnh ở trong file code với đường link sau: [DataMining](#)

4.5. Đánh giá mô hình

Đây là bài toán hồi quy nên ta chọn 1 trong các độ đo đánh giá cho bài toán này là MAE (Mean Absolute Error) thể hiện được chênh lệch của giá trị thực và giá trị dự đoán mà ít bị ảnh hưởng bởi các outlier.

- Decision Tree

```

✓ [24] y_pred_dt = regr_tun.predict(x_test_dt)
          mean_absolute_error(y_test_dt, y_pred_dt)

1.1525599976404457

```

- Random Forest

```
[ ] y_pred_rf = rfr_tun.predict(x_test_rf)
mean_absolute_error(y_test_rf, y_pred_rf)
```

```
0.6329363359303528
```

=> Từ 2 kết quả trên ta thấy được bài toán này phù hợp với mô hình Random Forest hơn Decision Tree.