

ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN



BUILDING AND MINING DATA
WAREHOUSE
BÁO CÁO ĐỒ ÁN THỰC HÀNH

Môn: Hệ thống thông tin phục vụ trí tuệ kinh doanh

Giảng viên hướng dẫn:

ThS. Hồ Thị Hoàng Vy
ThS. Nguyễn Ngọc Minh Châu
ThS. Tiết Gia Hồng

TP Hồ Chí Minh, 03 tháng 01, 2024

MỤC LỤC

A. GIỚI THIỆU.....	6
1. Thông tin nhóm	6
2. Phân công công việc	6
B. QUÁ TRÌNH ETL SOURCE → STAGE → NDS.....	7
1. Mô tả dữ liệu (Air Quality Data)	7
1.1. (2B)uscounties.csv	7
1.2. 10_state_aqi_[2021, 2022, 2023].csv	7
2. ETL từ Source → Stage.....	8
2.1. ETL là gì?	8
2.2. Các bước thực hiện đổ từ bảng nguồn vào stage	9
2.2.1. Chuẩn bị cơ sở dữ liệu	9
2.2.1.1. Metadata.....	9
2.2.1.2. uscounties_Stage	9
2.2.1.3. aqi2021_Stage	9
2.2.2. Đổ dữ liệu vào stage	10
2.2.2.1. (2B)uncounties.csv	10
2.2.2.2. 10_state_aqi_[2021, 2022, 2023].csv	11
3. ETL từ Stage → NDS.....	12
3.1. Mô hình dữ liệu trong NDS.....	12
3.2. Phân tích phụ thuộc hàm – dạng chuẩn	12
3.2.1. Bảng SOURCE	12
3.2.2. Bảng STATE.....	13
3.2.3. Bảng COUNTY	13
3.2.4. Bảng CATEGORY	13
3.2.5. Bảng PARAM.....	13
3.2.6. Bảng SITE.....	13
3.2.7. Bảng AQI	13
3.3. Mô tả cấu trúc bảng trong NDS.....	13
3.3.1. Cấu trúc bảng SOURCE	13
3.3.2. Cấu trúc bảng STATE.....	14
3.3.3. Cấu trúc bảng COUNTY	14

3.3.4.	Cấu trúc bảng CATEGORY	14
3.3.5.	Cấu trúc bảng PARAM.....	15
3.3.6.	Cấu trúc bảng SITE.....	15
3.3.7.	Cấu trúc bảng AQI.....	15
4.	Chuyển đổi dữ liệu từ Stage → NDS.....	16
4.1.	Bảng STATE trong NDS:.....	16
4.2.	Bảng COUNTY trong NDS:.....	21
4.3.	Bảng SITE trong NDS:.....	27
4.4.	Bảng CATEGORY trong NDS:.....	33
4.5.	Bảng PARAM trong NDS:	37
4.6.	Bảng AQI trong NDS:	38
5.	ETL từ NDS → DDS.....	47
5.1.	Cấu trúc dữ liệu nguồn NDS.....	48
5.2.	Quy trình chuyển đổi dữ liệu từ NDS sang DDS.....	49
5.2.1.	Mô hình bông tuyết (Snowflake Schema)	49
5.2.2.	Phân cấp chiều	49
5.2.3.	Facts	49
5.2.4.	Flow cơ bản.....	50
5.2.5.	Tạo bảng chiều (Dimensions)	53
5.2.5.1.	<i>Dim_Date</i>	53
5.2.5.2.	<i>Dim_State</i>	54
5.2.5.3.	<i>Dim_County</i>	55
5.2.5.4.	<i>Dim_Category</i>	57
5.2.5.5.	<i>Dim_Site</i>	57
5.2.5.6.	<i>Dim_Parameter</i>	58
5.2.6.	Tạo bảng sự kiện (Fact_AQI)	58
6.	OLAP, MDX, and Reporting	62
6.1.	OLAP	62
6.1.1.	Hướng dẫn cài đặt	62
a)	Các công cụ cần thiết	62
b)	Các bước cài đặt.....	62
6.1.2.	Thực hiện OLAP	62

6.2. MDX	78
6.2.1. Report the min and max of AQI value for each State during each quarter of years. <i>Analysis hints:</i> How do the AQI values fluctuate during the year? Pay attention to the values (max, min). Are any unusually large or small?	78
6.2.2. Report the mean and the standard deviation of AQI value for each State during each quarter of years. <i>Analysis hints:</i> How do the AQI values fluctuate during the year? Pay attention to the values (mean, std, max, min). Are any unusually large or small?	81
6.2.3. Report the number of days, and the mean AQI value where the air quality is rated as "very unhealthy" or worse for each State and County. <i>Analysis hint:</i> What is the AQI limit above which air quality is "very unhealthy" or worse?	83
6.2.4. For the four following states: Hawaii, Alaska, Illinois and Delaware, count the number of days in each air quality Category (Good, Moderate,etc.) by County. <i>Analysis hints:</i> Comparing the data of the states and counties, focus on the distribution of the harmful air condition. What could you conclude about the differences?)	85
6.2.5. For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the mean AQI value by quarters. <i>Analysis hints:</i> Comparing the data of the states over the year. What could you conclude about the fluctuations?	89
6.2.6. Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California. <i>Analysis hint:</i> Give your opinion about the fluctuations of AQI value.....	91
6.2.7. Build graphs/charts for the above reports	92
6.2.8. Use a regional map to visually represent (by color) the mean AQI value in regions during a year. <i>Example:</i>	92
6.2.9. Report the mean, the standard deviation, min and max of AQI value group by State and County during each quarter of the year. <i>Analysis hints:</i> Pay attention to the values (mean, std, max, min). Are any unusually large or small? Compare the standard deviation values between question 1 and 2, explain.....	93
6.2.10. Create a new attribute, DayLightSaving, in a suitable table. DayLightSaving may have two values: True: Between March 12, 2023, and November 5, 2023; False: Otherwise. Report the mean AQI value by State, Category, DayLightSaving over years. <i>Analysis hint:</i> Is there any notable difference on the air quality during the DaylightSaving period compared to the other?	
95	
6.2.11. Count the number of days by State, Category in each month.....	96
6.2.12. Report the number of days by Category and Defining Parameter. <i>Analysis hints:</i> What is your opinion on the pollution situation in the United States as a whole? Additionally, please identify the primary factors that the country should consider in order to enhance air quality..	98
7. Data Mining.....	99
7.1. Tổng quan.....	99

7.2. Mô hình và Thuật toán sử dụng	99
7.2.1. Thuật toán đề xuất.....	99
7.2.2. Lý do chọn thuật toán	99
7.3. Các bước thực hiện	99
7.4. Kết quả và Đánh giá	101
7.4.1. Kết quả dự đoán	101
7.4.1.1. <i>Mô hình Naïve Bayes</i>	101
7.4.1.2. <i>Mô hình SARIMA</i>	102
7.4.1.3. <i>Mô hình Random Forest</i>	102
7.4.2. Đánh giá mô hình.....	102
7.4.2.1. <i>Mô hình Naïve Bayes</i>	102
7.4.2.2. <i>Mô hình SARIMA</i>	103
7.4.2.3. <i>Mô hình Random Forest</i>	103
8. Tài liệu tham khảo	104

A. GIỚI THIỆU

1. Thông tin nhóm

NHÓM 09			
STT	MSSV	Họ và tên	Email
1	21127072	Nguyễn Hữu Khánh	nhkhanh21@clc.fitus.edu.vn
2	21127234	Nguyễn Lê Anh Chi	nlachi21@clc.fitus.edu.vn
3	21127235	Nguyễn Xuân Quỳnh Chi	nxqchi21@clc.fitus.edu.vn
4	21127495	Lê Ngô Song Cát	lncat21@clc.fitus.edu.vn
5	21127659	Bùi Ngọc Kiều Nhi	bnknnhi21@clc.fitus.edu.vn

2. Phân công công việc

STT	Thành viên	Nội dung công việc	% Đóng góp
1	Nguyễn Hữu Khánh	Thiết kế NDS, DDS Đỗ dữ liệu vào NDS Data Mining Viết báo cáo	20%
2	Nguyễn Lê Anh Chi	Thiết kế NDS, DDS Viết Script tạo NDS, DDS OLAP, MDX Viết báo cáo	20%
3	Nguyễn Xuân Quỳnh Chi	Thiết kế NDS, DDS Đỗ dữ liệu từ Source vào Stage Đỗ dữ liệu vào NDS, DDS Viết báo cáo	20%
4	Lê Ngô Song Cát	Thiết kế NDS, DDS Tạo metadata OLAP, MDX Viết báo cáo	20%
5	Bùi Ngọc Kiều Nhi	Thiết kế NDS, DDS Đỗ dữ liệu vào NDS, DDS OLAP, MDX Viết báo cáo	20%

- Demo video: [\[BI - 21HTTT1 - Nhóm 9\] Demo đồ án thực hành](#)

B. QUÁ TRÌNH ETL SOURCE → STAGE → NDS

1. Mô tả dữ liệu (Air Quality Data)

1.1. (2B)uscounties.csv

ID	Thuộc tính	Kiểu dữ liệu	Mô tả
1	county	VARCHAR	Tên của quận
2	county_ascii	VARCHAR	Tên quận ở dạng ký tự ASCII
3	county_full	VARCHAR	Tên quận đầy đủ
4	county_fips	VARCHAR	Mã FIPS của quận
5	state_id	VARCHAR	ID của tiểu bang
6	state_name	VARCHAR	Tên của tiểu bang
7	lat	FLOAT	Vĩ độ (latitude)
8	lng	FLOAT	Kinh độ (longitude)
9	population	INT	Mật độ dân số

1.2. 10_state_aqi_[2021, 2022, 2023].csv

ID	Thuộc tính	Kiểu dữ liệu	Mô tả
1	State Name	VARCHAR	Tên của tiểu bang nơi đo chỉ số AQI.
2	County Name	VARCHAR	Tên của quận trong tiểu bang nơi đo chỉ số AQI.
3	State Code	VARCHAR	Mã FIPS dùng để xác định duy nhất mỗi tiểu bang ở Mỹ.
4	County Code	VARCHAR	FIPS dùng để xác định mỗi quận trong một tiểu bang.
5	Date	DATE	Ngày mà chỉ số AQI được ghi nhận
6	AQI	INT	Giá trị số đại diện cho mức độ chất lượng không khí, được tính toán dựa trên nồng độ các chất ô nhiễm. Thang đo AQI thường dao động từ 0 đến 500, với các giá trị cao hơn chỉ ra chất lượng không khí xấu hơn.
7	Category	VARCHAR	Danh mục liên quan đến sức khỏe tương ứng với giá trị AQI. Các danh mục phổ biến bao gồm "Good", "Moderate", "Unhealthy for Sensitive Groups", "Unhealthy", "Very Unhealthy", và "Hazardous". Xem bảng AQI basics ở dưới để có thông tin chi tiết.
8	Defining Parameter	VARCHAR	Chất ô nhiễm cụ thể (ví dụ: ozone, PM2.5, PM10) có ảnh hưởng lớn nhất đến AQI tại vị trí và ngày đó.

ID	Thuộc tính	Kiểu dữ liệu	Mô tả
9	Defining Site	VARCHAR	Mã định danh của trạm quan trắc cụ thể chịu trách nhiệm báo cáo chỉ số AQI (Air Quality Index) cho bản ghi đó
10	Number of Sites Reporting	INT	Tổng số trạm đo trong quận báo cáo dữ liệu vào ngày hôm đó. Nhiều trạm đo có thể báo cáo AQI, và trường này chỉ ra số lượng trạm đóng góp vào dữ liệu của ngày hôm đó.
11	Created	DATETIME	Ngày và giờ khi bản ghi dữ liệu AQI này được tạo lần đầu trong hệ thống (dữ liệu tổng hợp tạo ra cho bài tập).
12	Last Updated	DATETIME	Ngày và giờ khi bản ghi dữ liệu AQI này được cập nhật lần cuối trong hệ thống (dữ liệu tổng hợp tạo ra cho bài tập).

AQI Basics for Ozone and Particle Pollution			
Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Hình 1. AQI basics. Nguồn: [AQI Basics / AirNow.gov](#)

2. ETL từ Source → Stage

2.1. ETL là gì?

- **ETL** (Extract, Transform, Load) là quá trình lấy dữ liệu từ các nguồn (Extract), chuyển đổi dữ liệu thành dạng phù hợp (Transform), và cuối cùng tải dữ liệu vào một kho dữ liệu hoặc hệ thống lưu trữ (Load).
- **Extract (Lấy dữ liệu):** Lấy dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, API hoặc các nguồn dữ liệu khác.

- Transform (Chuyển đổi dữ liệu):** Dữ liệu được xử lý, chuyển đổi, và làm sạch sao cho phù hợp với yêu cầu của kho dữ liệu hoặc mục đích phân tích.
- Load (Tải dữ liệu):** Dữ liệu sau khi chuyển đổi sẽ được tải vào hệ thống kho dữ liệu, như cơ sở dữ liệu, data warehouse (DW), hoặc hệ thống báo cáo.

Quá trình **ETL** đóng vai trò quan trọng trong việc thu thập, làm sạch và chuẩn bị dữ liệu từ các nguồn đầu vào khác nhau để sử dụng trong các phân tích và báo cáo.

2.2. Các bước thực hiện đổ từ bảng nguồn vào stage

2.2.1. Chuẩn bị cơ sở dữ liệu

2.2.1.1. Metadata

ID	Thuộc tính	Kiểu dữ liệu	Mô tả
1	id	INT	Khóa tự tăng của bảng Metadata
2	table_name	VARCHAR	Tên bảng dữ liệu
3	LSET	DATETIME	Thời điểm cuối cùng dữ liệu được trích xuất thành công từ nguồn
4	CET	DATETIME	Thời điểm dữ liệu được tạo hoặc ghi nhận ban đầu từ nguồn

2.2.1.2. uscounties_Stage

ID	Thuộc tính	Kiểu dữ liệu	Mô tả
1	county	VARCHAR	Tên của quận
2	county_ascii	VARCHAR	Tên quận ở dạng ký tự ASCII
3	county_full	VARCHAR	Tên quận đầy đủ
4	county_fips	VARCHAR	Mã FIPS của quận
5	state_id	VARCHAR	ID của tiểu bang
6	state_name	VARCHAR	Tên của tiểu bang
7	lat	FLOAT	Vĩ độ (latitude)
8	lng	FLOAT	Kinh độ (longitude)
9	population	INT	Mật độ dân số

2.2.1.3. aqi2021_Stage

ID	Thuộc tính	Kiểu dữ liệu	Mô tả
1	State Name	VARCHAR	Tên của tiểu bang nơi đo chỉ số AQI.
2	County Name	VARCHAR	Tên của quận trong tiểu bang nơi đo chỉ số AQI.
3	State Code	VARCHAR	Mã FIPS dùng để xác định duy nhất mỗi tiểu bang ở Mỹ.
4	County Code	VARCHAR	FIPS dùng để xác định mỗi quận trong một tiểu bang.
5	Date	DATE	Ngày mà chỉ số AQI được ghi nhận
6	AQI	INT	Giá trị số đại diện cho mức độ chất lượng không khí, được tính toán dựa trên nồng độ các chất ô nhiễm.

ID	Thuộc tính	Kiểu dữ liệu	Mô tả
7	Category	VARCHAR	Danh mục liên quan đến sức khỏe tương ứng với giá trị AQI.
8	Defining Parameter	VARCHAR	Chất ô nhiễm cụ thể (ví dụ: ozone, PM2.5, PM10) có ảnh hưởng lớn nhất đến AQI tại vị trí và ngày đó.
9	Defining Site	VARCHAR	Mã định danh của trạm quan trắc cụ thể chịu trách nhiệm báo cáo chỉ số AQI (Air Quality Index) cho bản ghi đó
10	Number of Sites Reporting	INT	Tổng số trạm đo trong quận báo cáo dữ liệu vào ngày hôm đó. Nhiều trạm đo có thể báo cáo AQI, và trường này chỉ ra số lượng trạm đóng góp vào dữ liệu của ngày hôm đó.
11	Created	DATETIME	Ngày và giờ khi bản ghi dữ liệu AQI này được tạo lần đầu trong hệ thống.
12	Last Updated	DATETIME	Ngày và giờ khi bản ghi dữ liệu AQI này được cập nhật lần cuối trong hệ thống.

Ghi chú: Bảng aqi2022_Stage và aqi2023_Stage có cấu trúc và được xử lý tương tự như aqi2021_Stage.

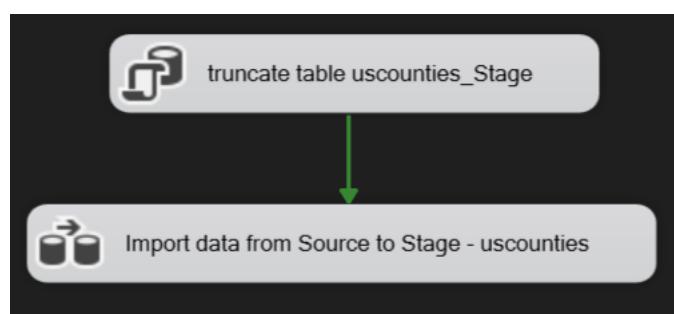
2.2.2. Đỗ dữ liệu vào stage

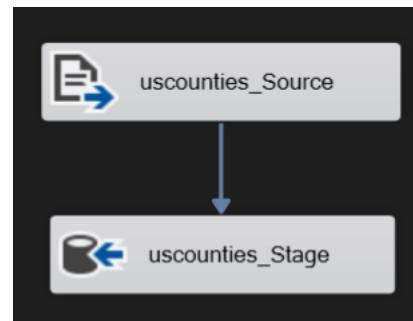
2.2.2.1. (2B)uncounties.csv

Các bước thực hiện:

- Truncate bảng Stage hiện tại: TRUNCATE TABLE <BangStage>
- Rút trích dữ liệu từ flat file csv sang bảng Stage.

Ghi chú: Do dữ liệu trong file này không lưu trữ ngày tháng tạo cũng như cập nhật, chỉ lưu trữ dữ liệu thuần về các quận trong tiểu bang, không đòi hỏi phải cập nhật thường xuyên hoặc real-time nên Whole Table Extract sẽ là lựa chọn tốt hơn.



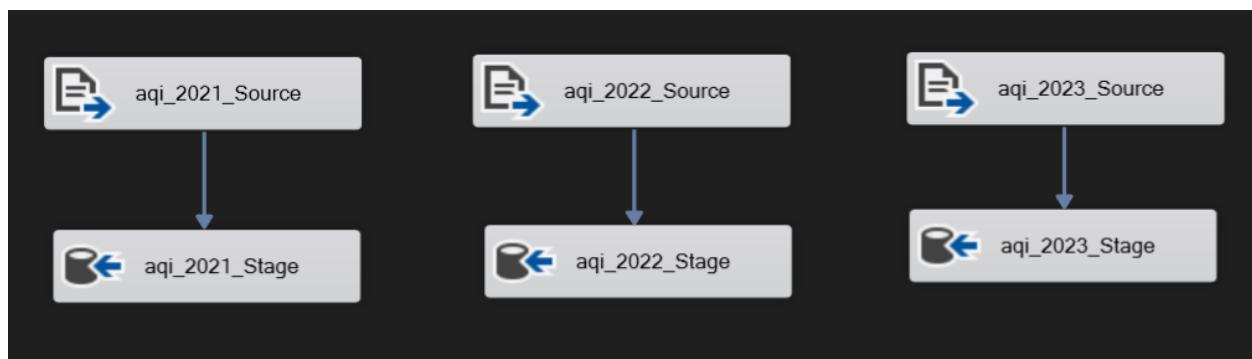
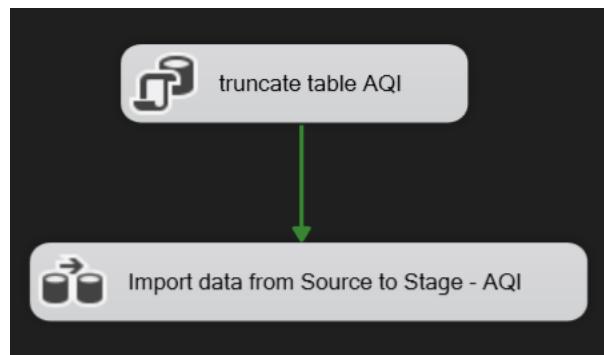


2.2.2.2. 10_state_aqi_[2021, 2022, 2023].csv

Các bước thực hiện:

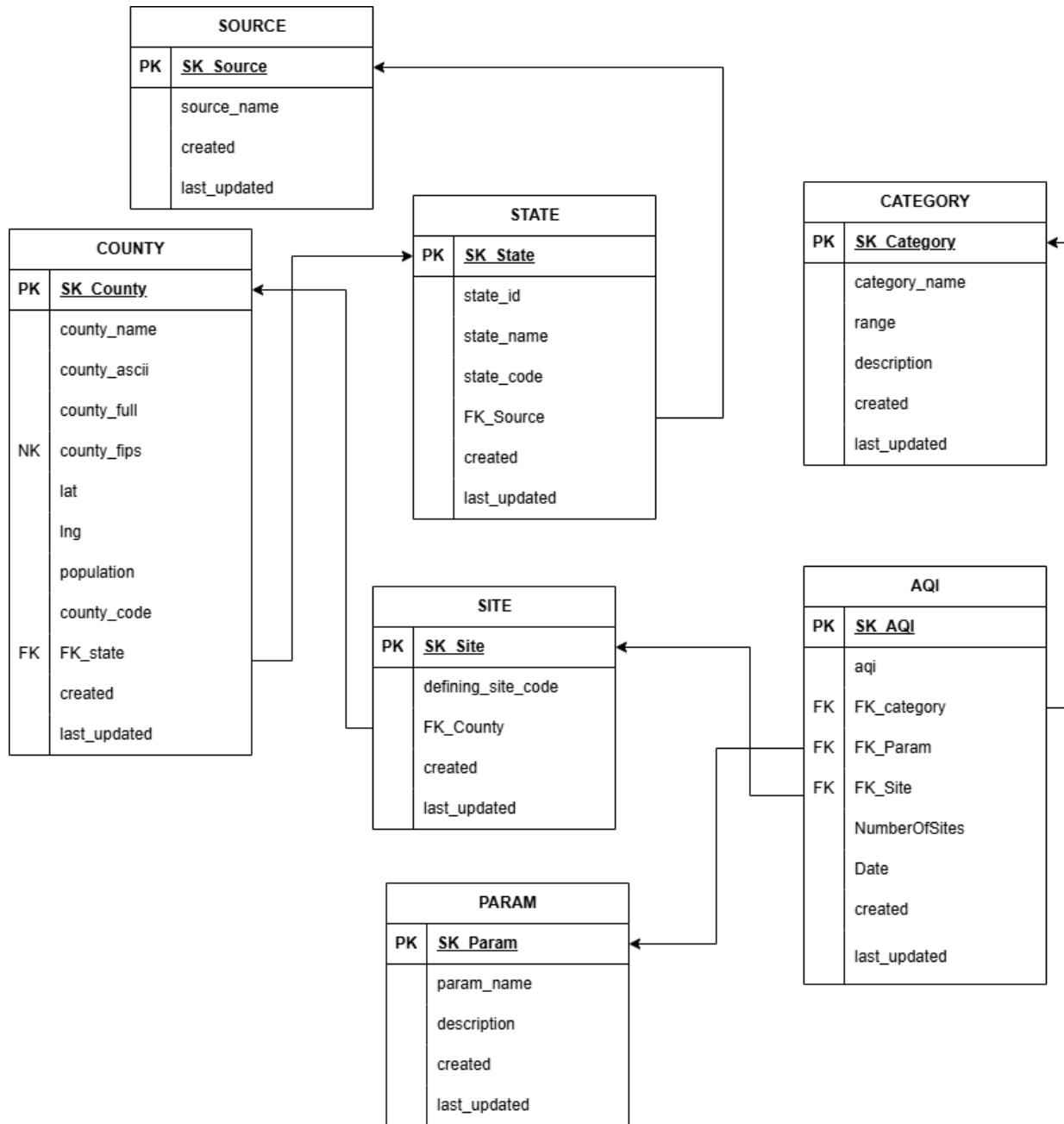
1. Truncate các bảng AQI 2021, 2022 và 2023 hiện tại: TRUNCATE TABLE <BangStage>
2. Rút trích dữ liệu từ flat file csv sang bảng Stage.

Ghi chú: Do dữ liệu trong file ở Source không lưu trữ ngày tháng tạo cũng như cập nhật khi import data vào Stage mà chỉ lưu thời gian tạo và cập nhật liên quan đến source, nên không có thông tin về LSET và CET cần thiết. Do đó, dữ liệu trong 3 bảng này sẽ được whole table extract.



3. ETL từ Stage → NDS

3.1. Mô hình dữ liệu trong NDS



3.2. Phân tích phụ thuộc hàm – dạng chuẩn

3.2.1. Bảng SOURCE

- **SOURCE (SK_Source, source_name, created, last_updated)**
- Phụ thuộc hàm F1 = {SK_Source → source_name, created, last_updated}
- ➔ Dạng chuẩn: 3NF (không có phụ thuộc bắc cầu).

3.2.2. Bảng STATE

- **STATE (SK State, state_id, state_name, state_code, FK_Source, created, last_updated)**
- Phụ thuộc hàm F2 = {SK_State → state_id, state_name, state_code, FK_Source, created, last_updated}
- ➔ Dạng chuẩn: **3NF** (không có phụ thuộc bắc cầu).

3.2.3. Bảng COUNTY

- **COUNTY (SK County, county_name, county_ascii, county_full, county_fips, lat, lng, population, county_code, FK_State, created, last_updated)**
- Phụ thuộc hàm F3 = {SK_County → county_name, county_ascii, county_full, county_fips, lat, lng, population, county_code, FK_State, created, last_updated}
- ➔ Dạng chuẩn: **3NF** (không có phụ thuộc bắc cầu).

3.2.4. Bảng CATEGORY

- **CATEGORY (SK Category, category_name, range, description, created, last_updated)**
- Phụ thuộc hàm F4 = {SK_Category → category_name, range, description, created, last_updated}
- ➔ Dạng chuẩn: **3NF** (không có phụ thuộc bắc cầu).

3.2.5. Bảng PARAM

- **PARAM (SK Param, param_name, description, created, last_updated)**
- Phụ thuộc hàm F5 = {SK_Param → param_name, description, created, last_updated}
- ➔ Dạng chuẩn: **3NF** (không có phụ thuộc bắc cầu).

3.2.6. Bảng SITE

- **SITE (SK Site, defining_site_code, FK_County, created, last_updated)**
- Phụ thuộc hàm F6 = {SK_Site → defining_site_code, FK_County, created, last_updated}
- ➔ Dạng chuẩn: **3NF** (không có phụ thuộc bắc cầu).

3.2.7. Bảng AQI

- **AQI (SK AQI, AQI, FK_Category, FK_Param, FK_Site, NumberOfSites, created, last_updated)**
- Phụ thuộc hàm F7 = {SK_AQI → AQI, FK_Category, FK_Param, FK_Site, NumberOfSites, created, last_updated}
- ➔ Dạng chuẩn: **3NF** (không có phụ thuộc bắc cầu).

3.3. Mô tả cấu trúc bảng trong NDS

3.3.1. Cấu trúc bảng SOURCE

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
1	SK_Source	Khóa chính của nguồn dữ liệu.	INT	

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
2	source_name	Tên của nguồn dữ liệu.	VARCHAR	
3	created	Ngày và giờ tạo	DATETIME	
4	last_updated	Ngày và giờ cập nhật lần cuối của bản ghi.	DATETIME	

3.3.2. Cấu trúc bảng STATE

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
1	SK_State	Khóa chính tự động tăng của bảng Stage	INT	
2	state_id	Tên tiêu bang viết tắt	VARCHAR	County_Stage.state_id
3	state_name	Tên của tiêu bang	VARCHAR	County_Stage.state_name
4	state_code	Mã FIPS xác định duy nhất mỗi tiêu bang	VARCHAR	AQI_Stage.state_code
5	fk_source	ID của source	INT	Source.SK_Source
6	created	Ngày và giờ tạo	DATETIME	
7	last_updated	Ngày và giờ cập nhật lần cuối của bản ghi.	DATETIME	

3.3.3. Cấu trúc bảng COUNTY

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
1	SK_County	Khóa chính tự động tăng của bảng County	INT	
2	county_name	Tên của quận	VARCHAR	County_Stage.county_name
3	county_ascii	Tên quận ở dạng ký tự ASCII	VARCHAR	County_Stage.county_ascii
4	county_full	Tên quận đầy đủ	VARCHAR	County_Stage.county_full
5	county_fips	Mã FIPS của quận	VARCHAR	County_Stage.county_fips
6	lat	Vĩ độ	FLOAT	County_Stage.lat
7	lng	Kinh độ	FLOAT	County_Stage.lng
8	population	Mật độ dân số	INT	County_Stage.population
9	county_code	Mã quận	VARCHAR	County_Stage.county
10	fk_state	Khóa ngoại tham chiếu đến bảng STATE	INT	State.SK_State
11	created	Ngày và giờ tạo	DATETIME	
12	last_updated	Ngày và giờ cập nhật lần cuối của bản ghi.	DATETIME	

3.3.4. Cấu trúc bảng CATEGORY

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
1	SK_Category	Khóa chính tự động tăng của bảng Category	INT	
2	category_name	Tên phân loại	VARCHAR	AQI_Stage.Category

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
3	range	Vùng giá trị	VARCHAR	
4	description	Mô tả về phân loại	VARCHAR	
5	created	Ngày và giờ tạo	DATETIME	
6	last_updated	Ngày và giờ cập nhật lần cuối của bản ghi.	DATETIME	

3.3.5. Cấu trúc bảng PARAM

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
1	SK_Param	Khóa chính tự động tăng của bảng Param	VARCHAR	
2	param_name	Tên chất ô nhiễm	VARCHAR	AQI_Stage.defining_param
3	description	Mô tả chất ô nhiễm	VARCHAR	
4	created	Ngày và giờ tạo	DATETIME	
5	last_updated	Ngày và giờ cập nhật lần cuối của bản ghi.	DATETIME	

3.3.6. Cấu trúc bảng SITE

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
1	SK_Site	Khóa chính tự động tăng của bảng Site	INT	
2	defining_site_code	Mã trạm	VARCHAR	AQI_Stage.defining_site
3	FK_County	Khóa ngoại tham chiếu đến bảng COUNTY	INT	County.SK_County
4	created	Ngày và giờ tạo	DATETIME	
5	last_updated	Ngày và giờ cập nhật lần cuối của bản ghi.	DATETIME	

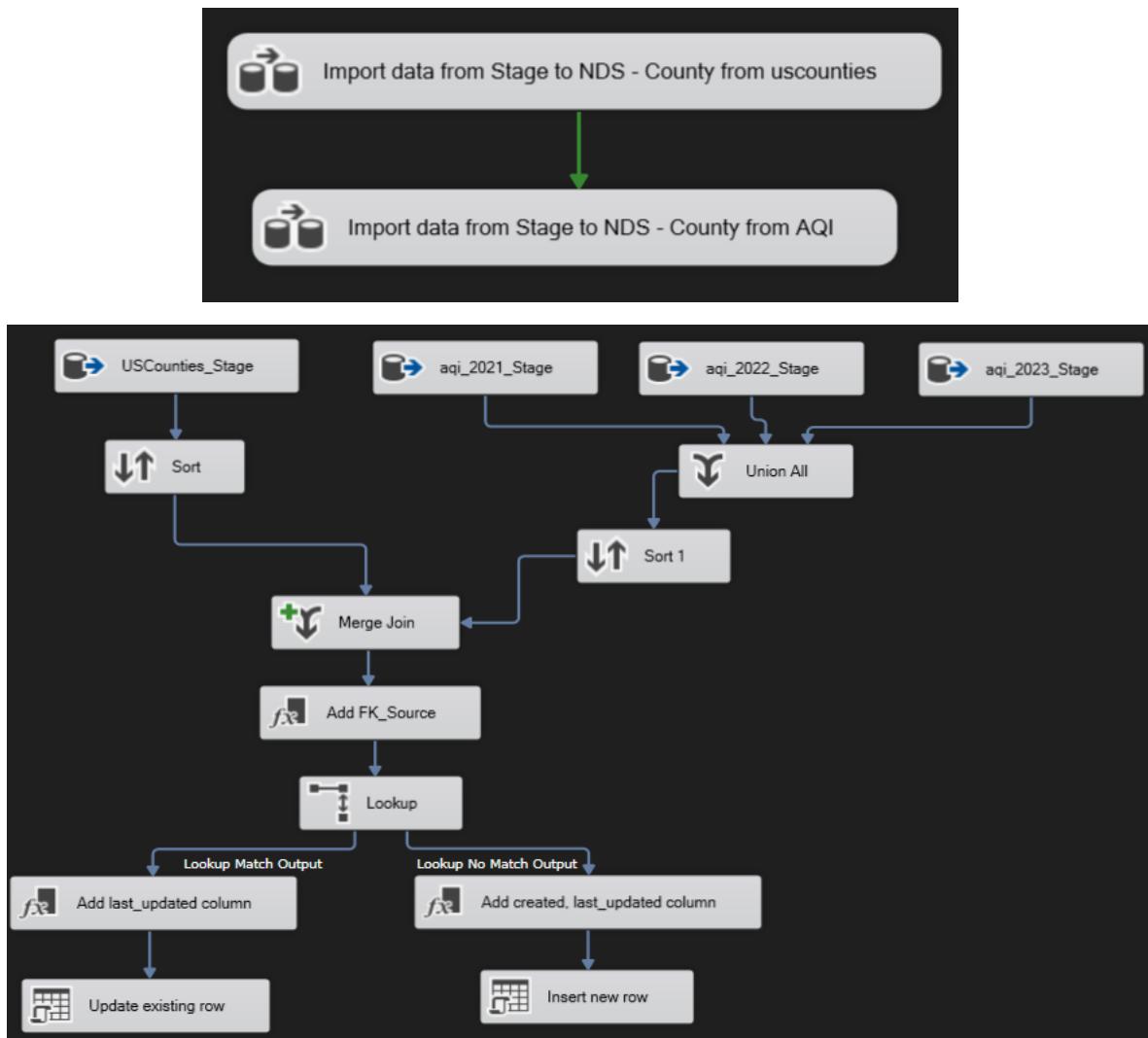
3.3.7. Cấu trúc bảng AQI

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
1	SK_AQI	Khóa chính tự động tăng của bảng AQI	INT	
2	AQI	Chỉ số mức độ chất lượng không khí	INT	AQI_Stage.AQI
3	FK_Category	Khóa ngoại tham chiếu đến bảng CATEGORY	INT	Category.SK_Category
4	FK_Param	Khóa ngoại tham chiếu đến bảng PARAMETER	INT	Param.SK_Param
5	FK_Site	Khóa ngoại tham chiếu đến bảng SITE	INT	Site.SK_Site

ID	Thuộc tính	Mô tả	Kiểu dữ liệu	Nguồn
6	NumberOfSites	Tổng số trạm đo trong quận báo cáo dữ liệu	INT	AQI_Stage.NumberOfSites
7	Date	Ngày ghi nhận số liệu tại trạm	DATE	
8	created	Ngày và giờ tạo	DATETIME	
9	last_updated	Ngày và giờ cập nhật lần cuối của bản ghi.	DATETIME	

4. Chuyển đổi dữ liệu từ Stage → NDS

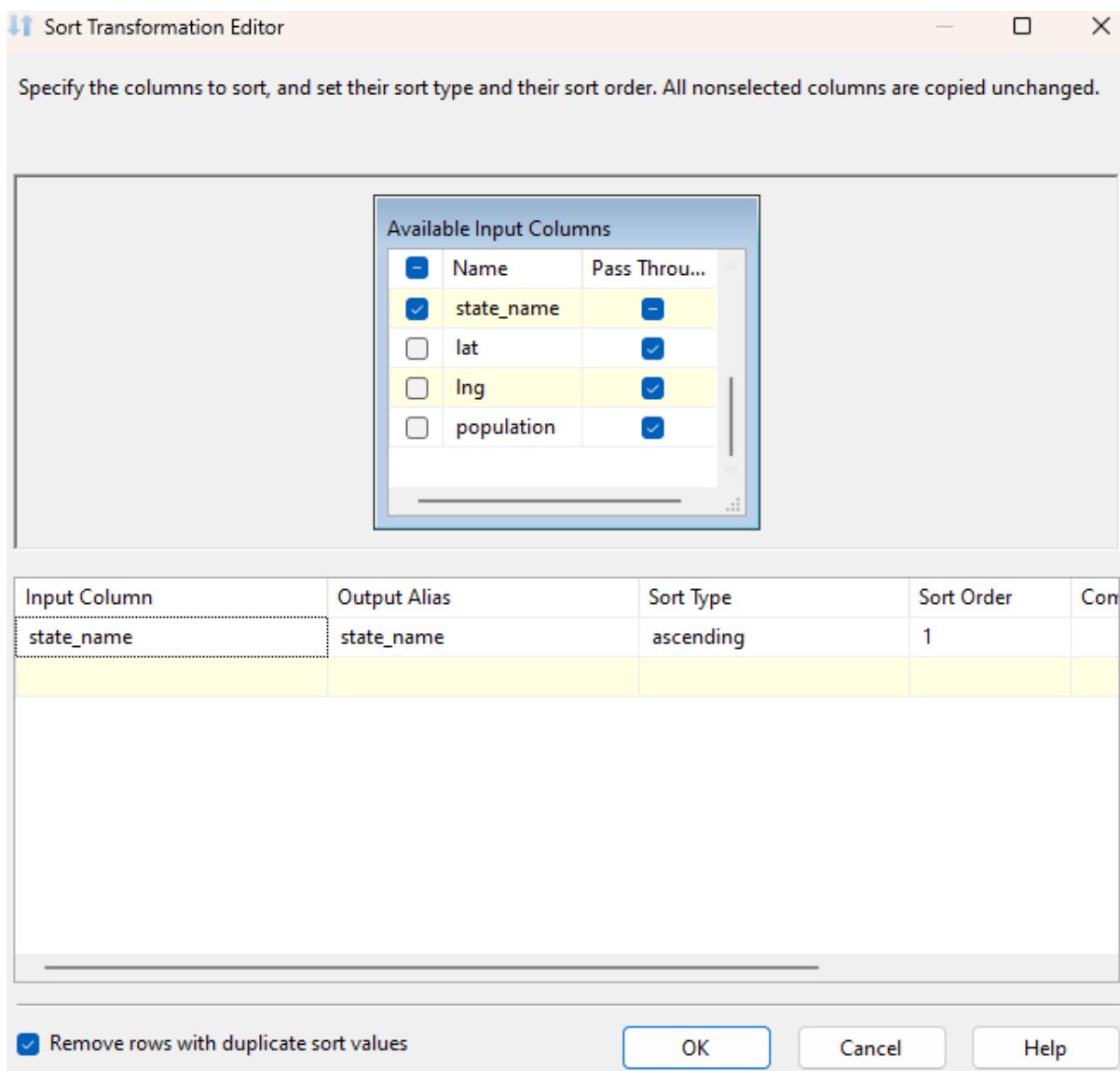
4.1. Bảng STATE trong NDS:

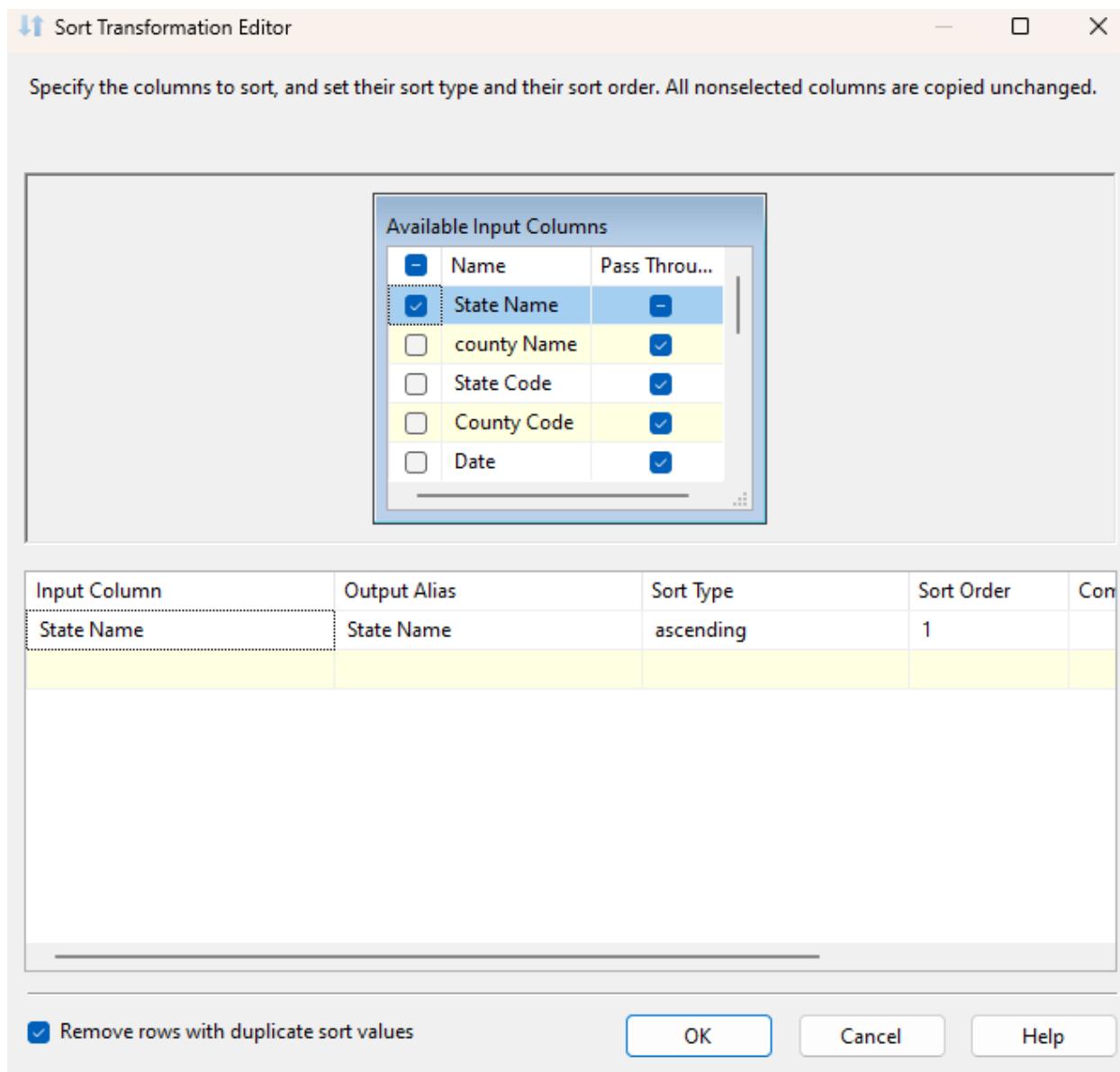


Các bước thực hiện:

- Đỗ dữ liệu từ Stage vào, bao gồm dữ liệu từ 4 bảng Stage là **uscounties_Stage**, **aqi2021_Stage**, **aqi2022_Stage** và **aqi2023_Stage**.

2. Ở nhánh bên phải, sau khi đổ dữ liệu từ 3 bảng aqi của cả 3 năm vào, thực hiện UNION ALL vì thông tin cơ bản của cả 3 bảng là giống nhau và tương đồng về số lượng cột dữ liệu.
3. Với dữ liệu đổ từ các bảng Stage, theo như thiết kế NDS dự tính của nhóm, cần phải MERGE dữ liệu về State ở 3 bảng và cho ra bảng mới là SITE trong NDS và trước khi merge cần có bước sort, với thuộc tính được chọn là **state_name** vì tồn tại trong cả 2 bảng là **uscounties** và **aqi** của cả 3 năm đã được UNION.





4. Sau khi thực hiện SORT ở các bảng cần thiết, thiết lập kiểu MERGE JOIN là Left Outer Join vì nhóm muốn giữ tất cả các dữ liệu có trong bảng uscounties_Stage, với điều kiện join theo thứ tự sort ưu tiên là state_id, state_name và State Code.

Merge Join Transformation Editor

Configure the properties used to join two sources of sorted data. Select the join type and then specify the columns to be used as the join key. Join keys must be used in the order specified by the sort-key position of the column.

Join type: Left outer join Swap Inputs

Sort

Name	Order	Join K...
county	0	<input type="checkbox"/>
county_ascii	0	<input type="checkbox"/>
county_full	0	<input type="checkbox"/>
county_fips	0	<input type="checkbox"/>
state_id	0	<input checked="" type="checkbox"/>

Sort 1

Name	Order	Join K...
State Name	1	<input checked="" type="checkbox"/>
county Name	0	<input type="checkbox"/>
State Code	0	<input type="checkbox"/>
County Code	0	<input type="checkbox"/>
Date	0	<input type="checkbox"/>

Input

Input	Input Column	Output Alias
Sort	state_id	state_id
Sort	state_name	state_name
Sort 1	State Code	State Code

5. Sau khi MERGE JOIN, thêm khóa FK_Source tham chiếu tới bảng SOURCE như là một Derived Columnn.

Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

Variables and Parameters

Mathematical Functions

String Functions

Date/Time Functions

NULL Functions

Type Casts

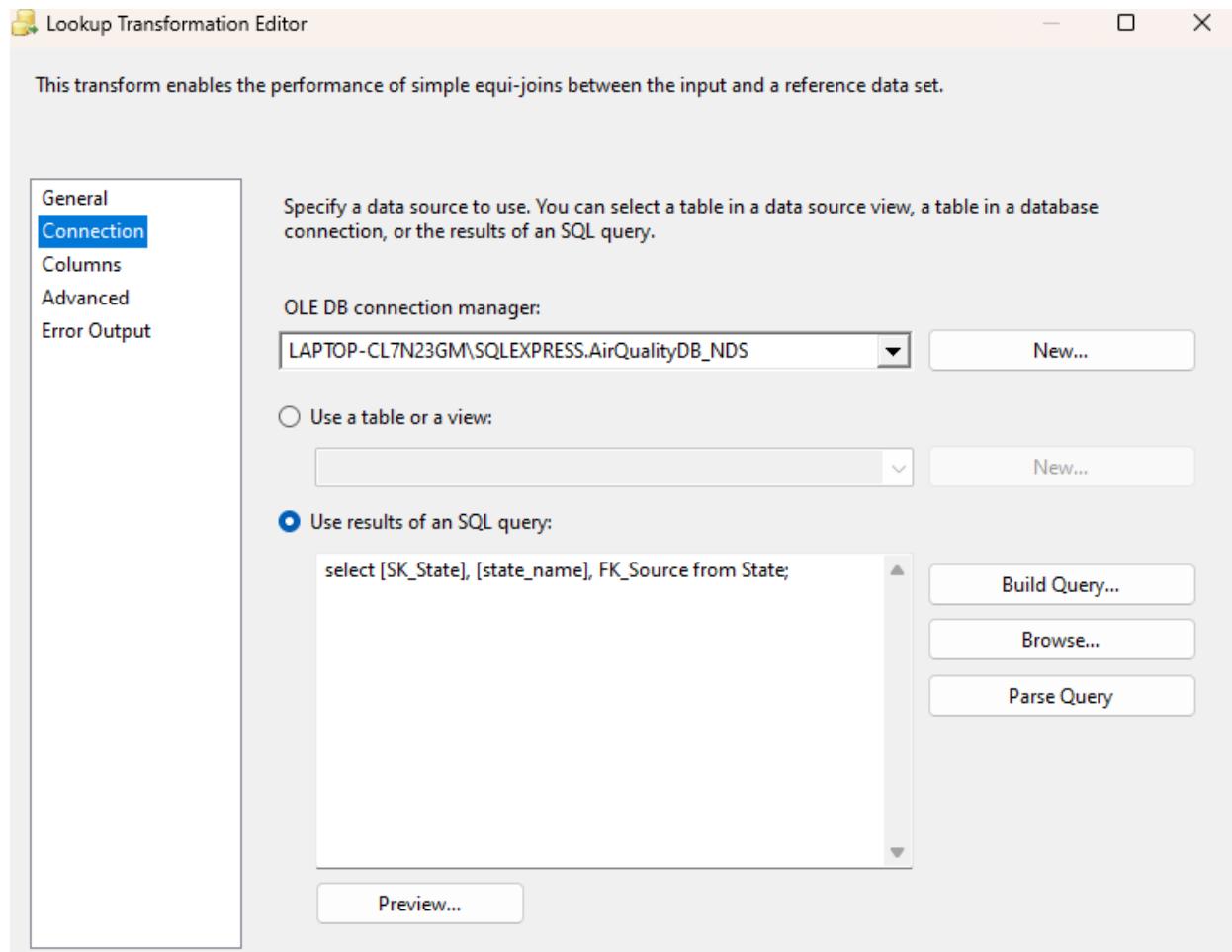
Operators

Description:

Derived Column Name | **Derived Column** | **Expression** | **Data Type**

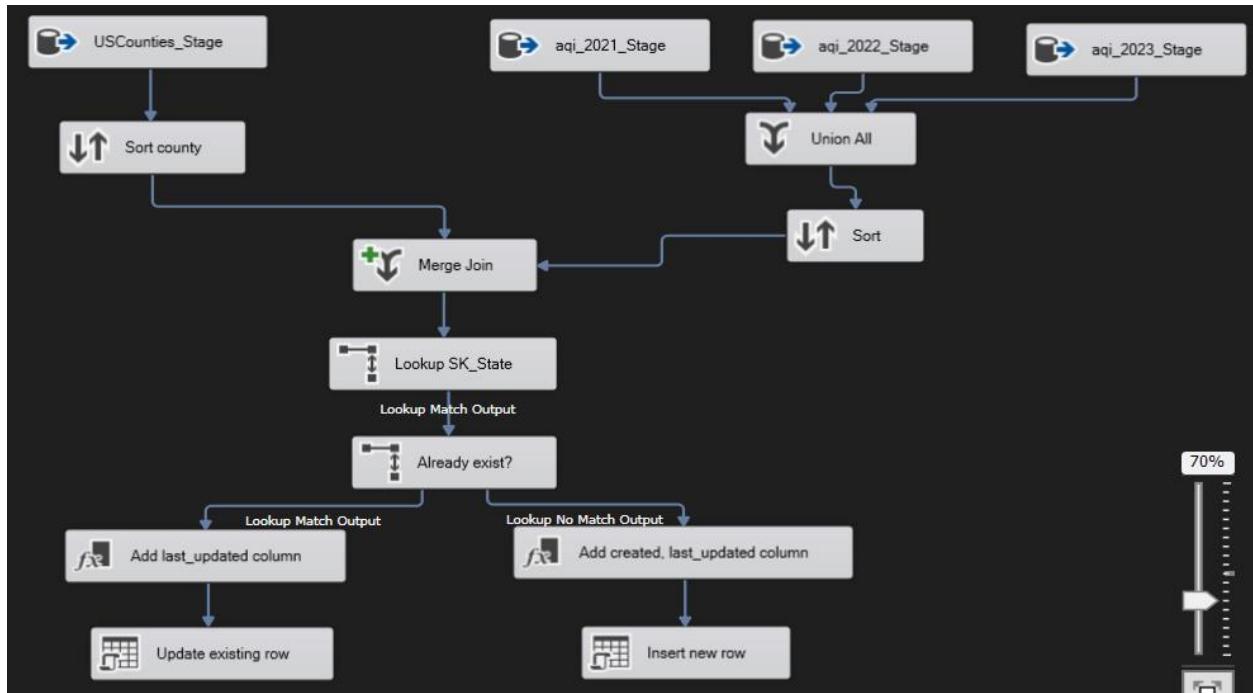
FK_Source	<add as new column>	1	four-byte signed integer
-----------	---------------------	---	--------------------------

6. Tìm trong bảng bảng LOOKUP xem có tồn tại thông tin tương ứng với câu query trong hình để xác định đây là bản ghi mới hay đã tồn tại chỉ cần cập nhật mới.



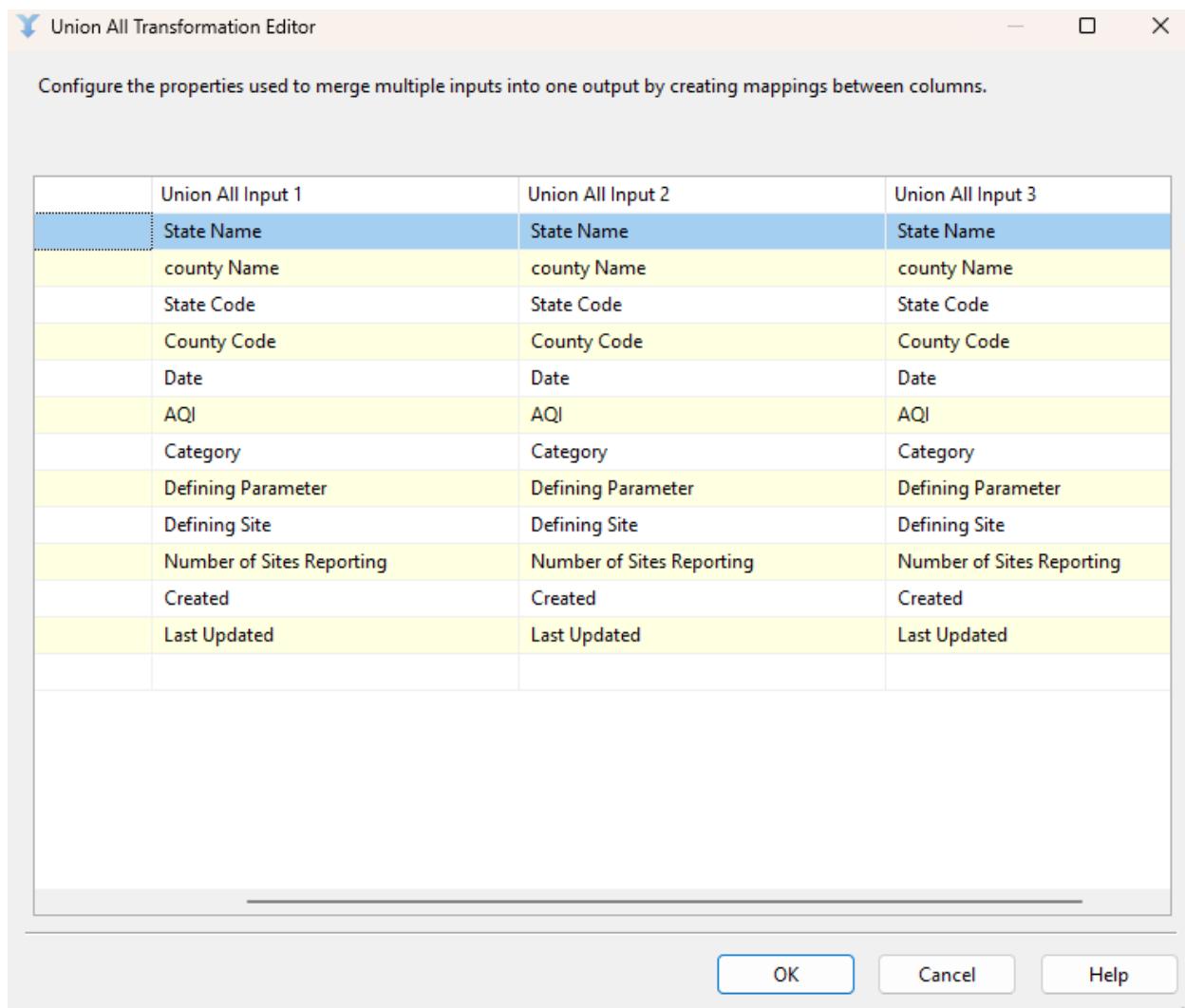
7. Nếu như dữ liệu không tồn tại, tạo thêm dòng mới để chứa dữ liệu mới đồng thời cập nhật ngày giờ tạo bản ghi và lần cập nhật gần nhất. Nếu như dữ liệu đã tồn tại thì chỉnh sửa bản ghi đã có và cập nhật lại thời gian cập nhật cuối cùng.

4.2. Bảng COUNTY trong NDS:

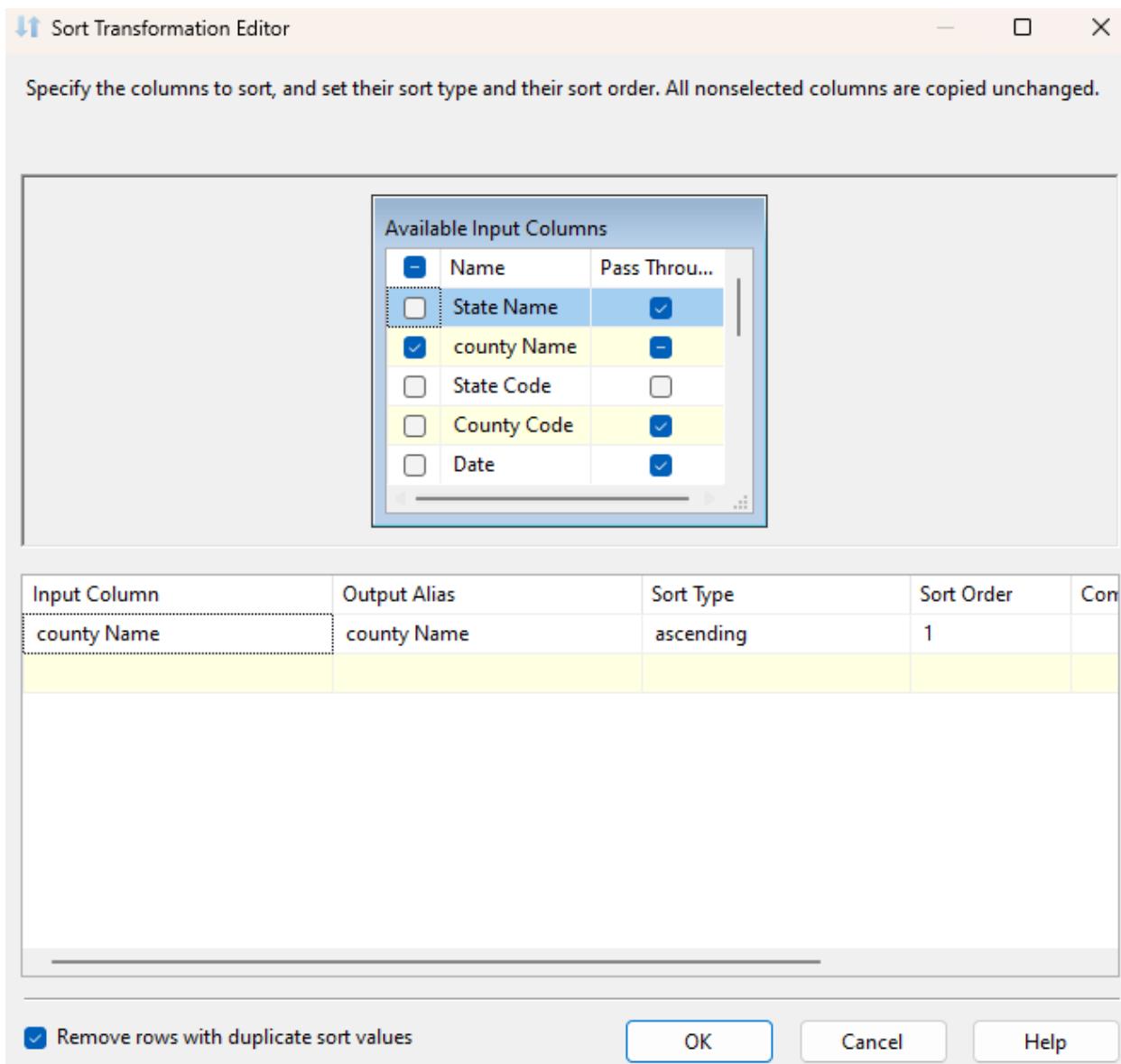


Các bước thực hiện:

- Đỗ dữ liệu từ Stage vào, bao gồm dữ liệu từ 4 bảng Stage là **uscounties_Stage**, **aqi2021_Stage**, **aqi2022_Stage** và **aqi2023_Stage**.
- Ở nhánh bên phải, sau khi đỗ dữ liệu từ 3 bảng aqi của cả 3 năm vào, thực hiện UNION ALL vì thông tin cơ bản của cả 3 bảng là giống nhau và tương đồng về số lượng cột dữ liệu.



3. Với dữ liệu đó từ các bảng Stage, theo nhu thiết kế NDS dự tính của nhóm, cần phải MERGE dữ liệu về State ở 3 bảng và cho ra bảng mới là COUNTY trong NDS và trước khi merge cần có bước sort, với thuộc tính được chọn là **county_name** vì tồn tại trong cả 2 bảng là **uscounties** và **aqi** của cả 3 năm đã được UNION.



4. Thiết lập kiểu MERGE JOIN là Full Outer Join vì nhóm muốn giữ tất cả các dữ liệu có trong bảng uscounties_Stage và aqi của cả 3 năm, với điều kiện join theo thứ tự sort ưu tiên như hình dưới.

Merge Join Transformation Editor

Configure the properties used to join two sources of sorted data. Select the join type and then specify the columns to be used as the join key. Join keys must be used in the order specified by the sort-key position of the column.

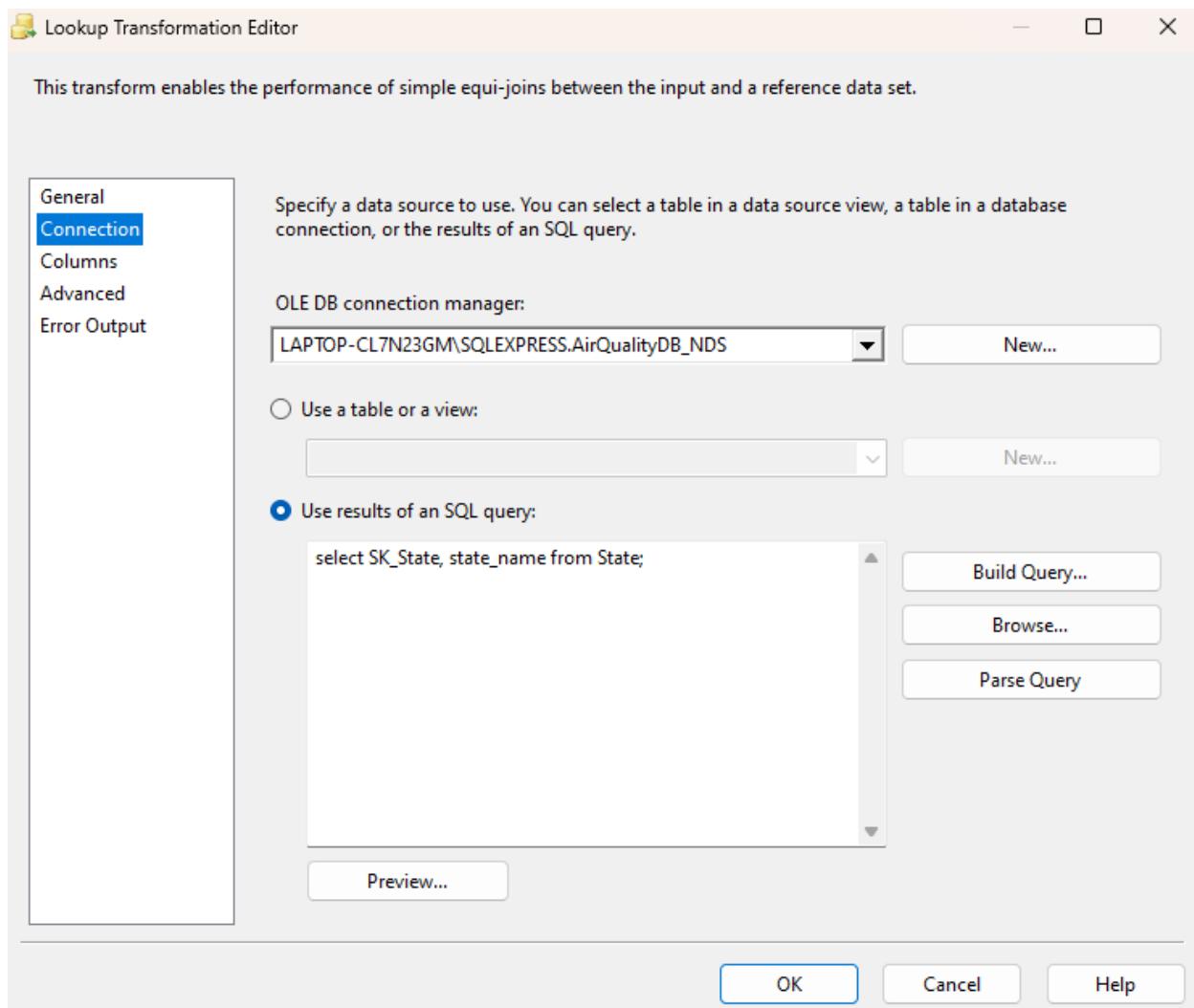
Join type: Full outer join Swap Inputs

The screenshot shows the configuration of a Merge Join Transformation. At the top, the 'Join type' is set to 'Full outer join'. Below this are two 'Sort' components. The left 'Sort county' component has five columns: 'Name' (checkbox checked), 'Order' (value 1), and 'Join K...' (checkbox checked). The right 'Sort' component has four columns: 'Name' (checkbox checked), 'Order' (value 1), and 'Join K...' (checkbox checked). A connection line links the 'Join K...' column of the first sort to the 'Join K...' column of the second sort. Below these components is a table mapping input columns to output aliases:

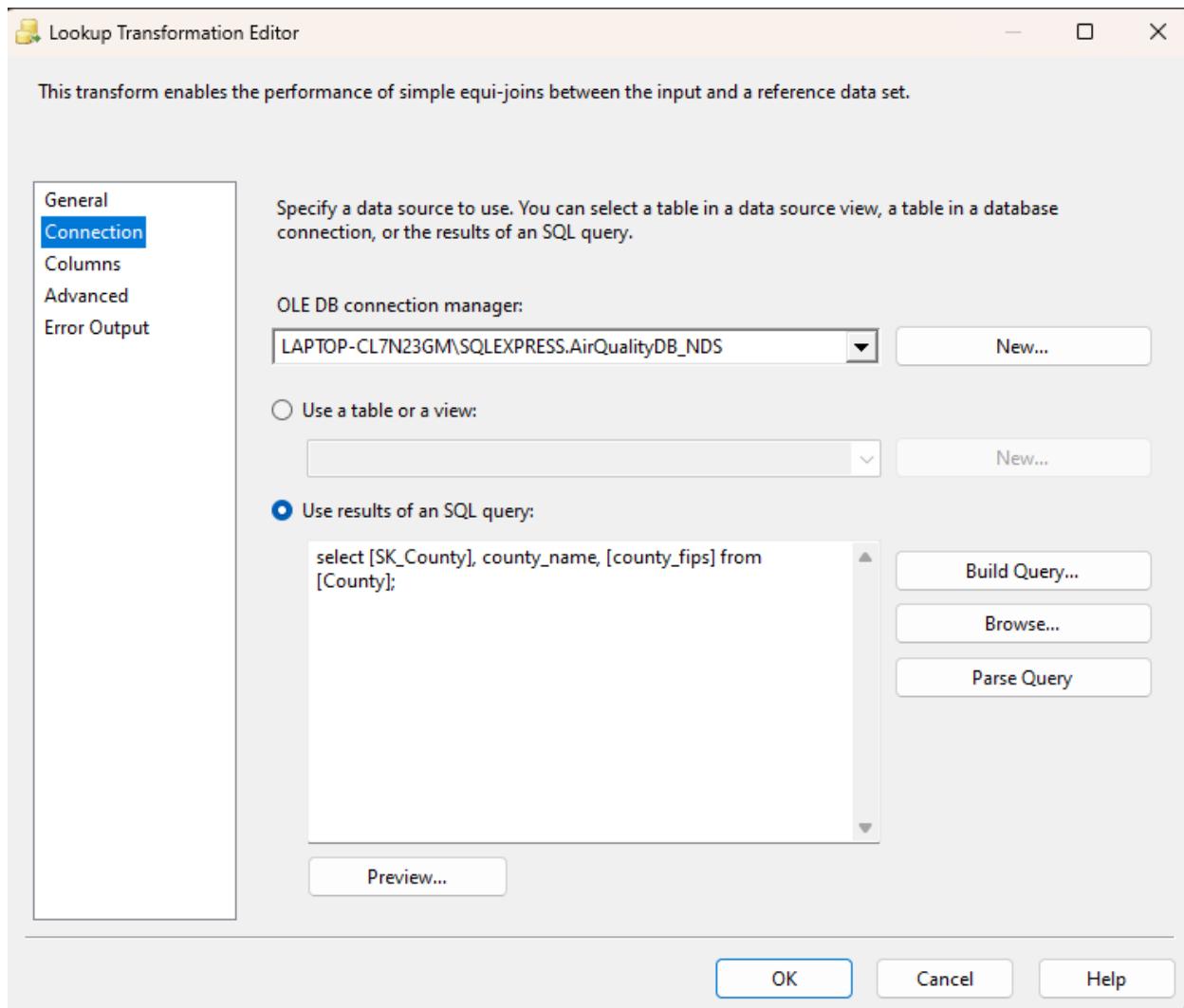
Input	Input Column	Output Alias
Sort county	county	county
Sort county	county_ascii	county_ascii
Sort county	county_full	county_full
Sort county	county_fips	county_fips
Sort county	lat	lat
Sort county	lng	lng
Sort county	population	population
Sort county	state_name	state_name
Sort	County Code	County Code
Sort	county Name	county Name
Sort	State Name	State Name

At the bottom are buttons for 'OK', 'Cancel', and 'Help'.

5. Tìm trong bảng bảng LOOKUP xem có tồn tại thông tin tương ứng với câu query trong hình, ở đây là tìm xem có tồn tại thông tin về tiểu bang được truy vấn hay không, nếu có dữ liệu tương thích sẽ tiến đến bước tiếp theo.

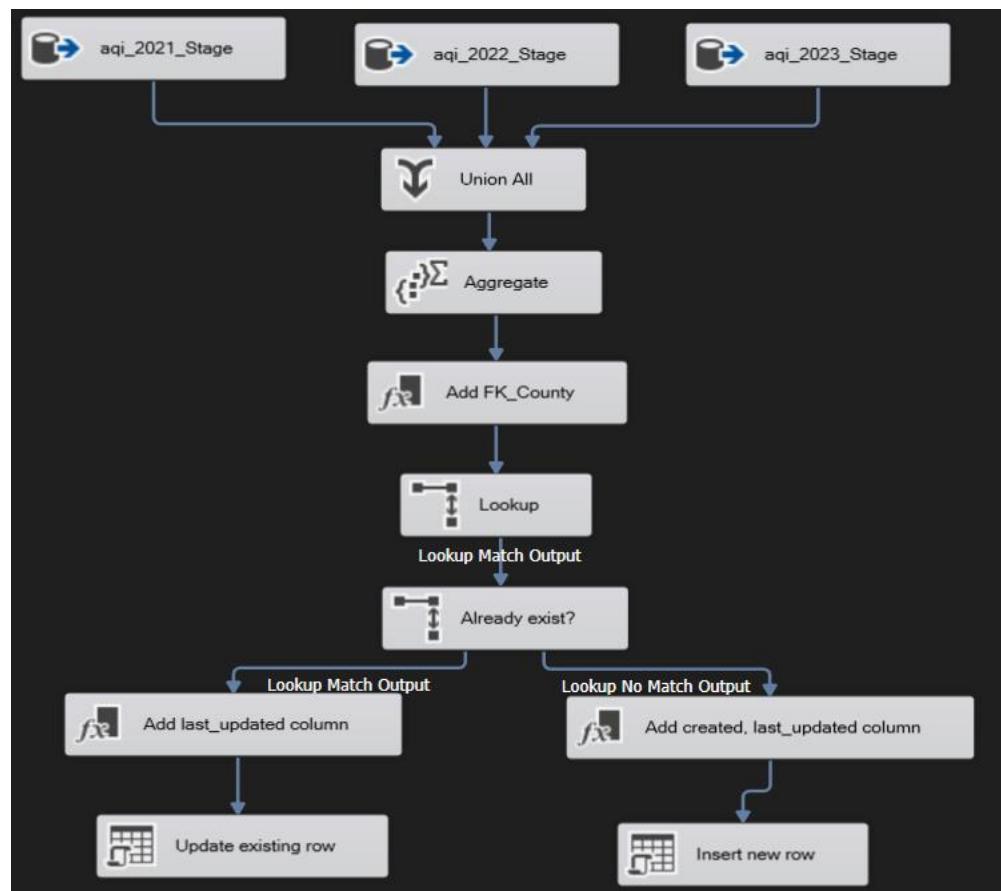


6. Tìm trong bảng bảng LOOKUP xem có tồn tại thông tin tương ứng với câu query trong hình, ở đây là tìm xem có tồn tại thông tin về tiểu bang đã được truy vấn ở trên hay không.



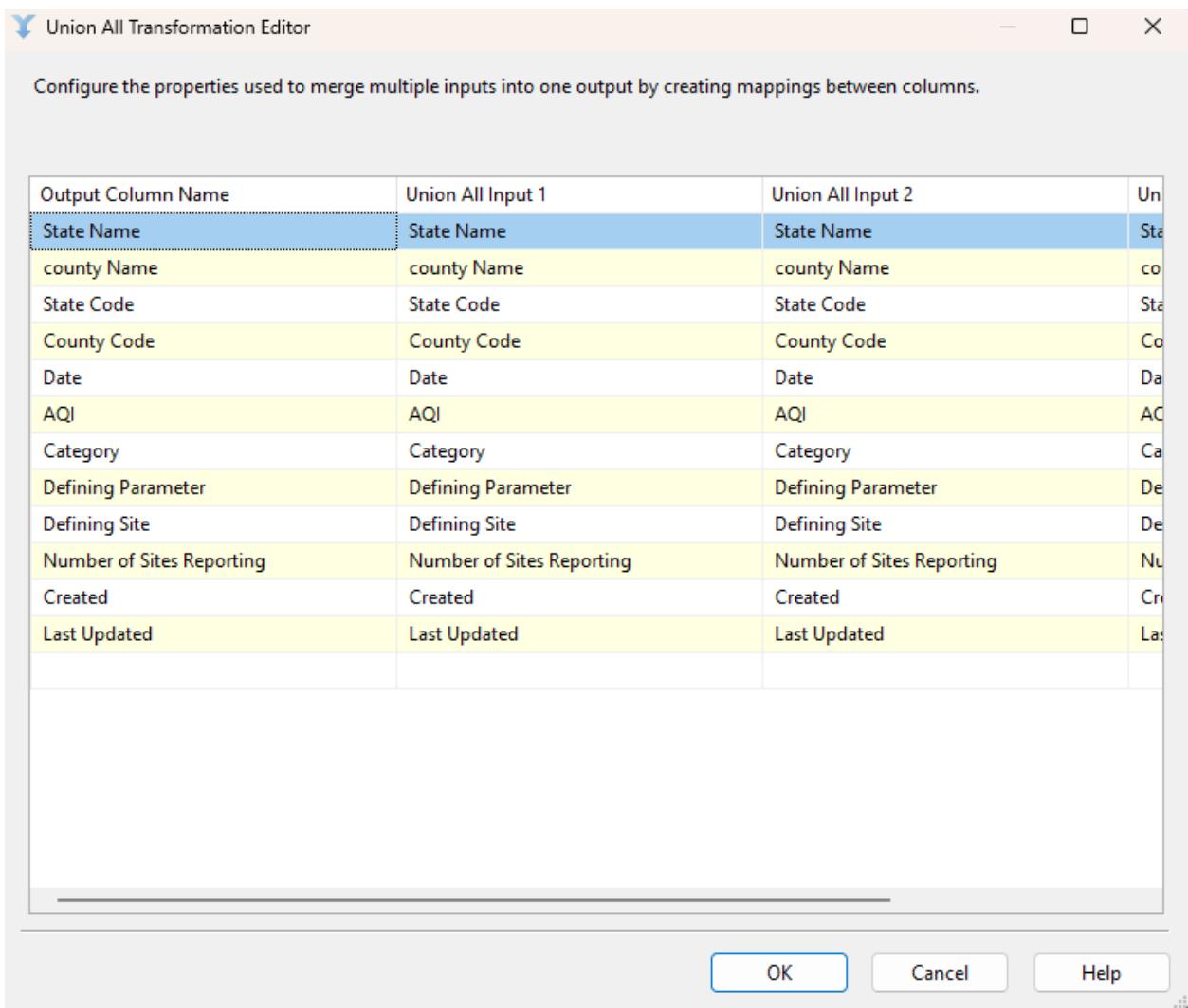
7. Nếu như dữ liệu không tồn tại, tạo thêm dòng mới để chứa dữ liệu mới đồng thời cập nhật ngày giờ tạo bản ghi và lần cập nhật gần nhất. Nếu như dữ liệu đã tồn tại thì chỉnh sửa bản ghi đã có và cập nhật lại thời gian cập nhật cuối cùng

4.3. Bảng SITE trong NDS:

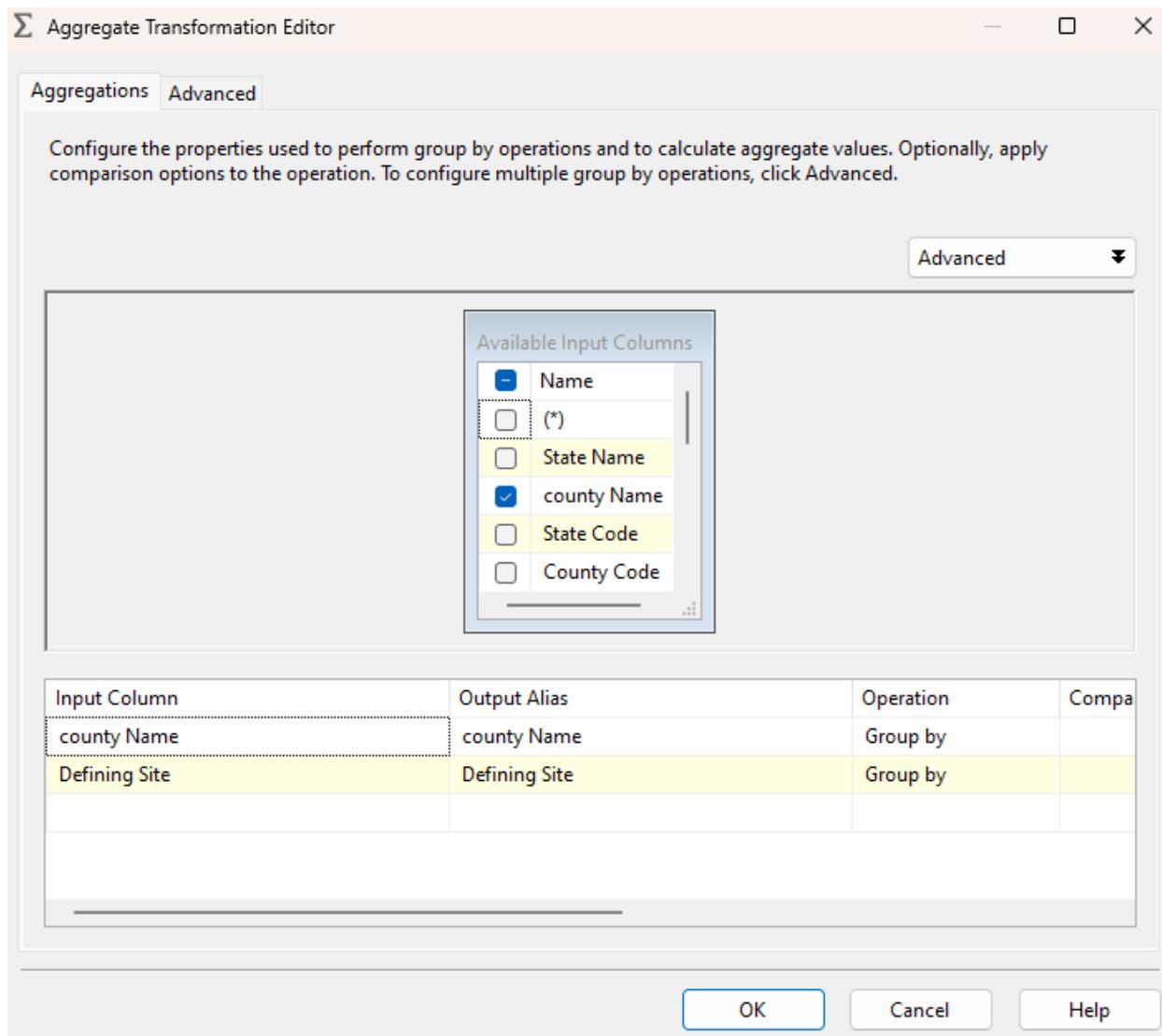


Các bước thực hiện:

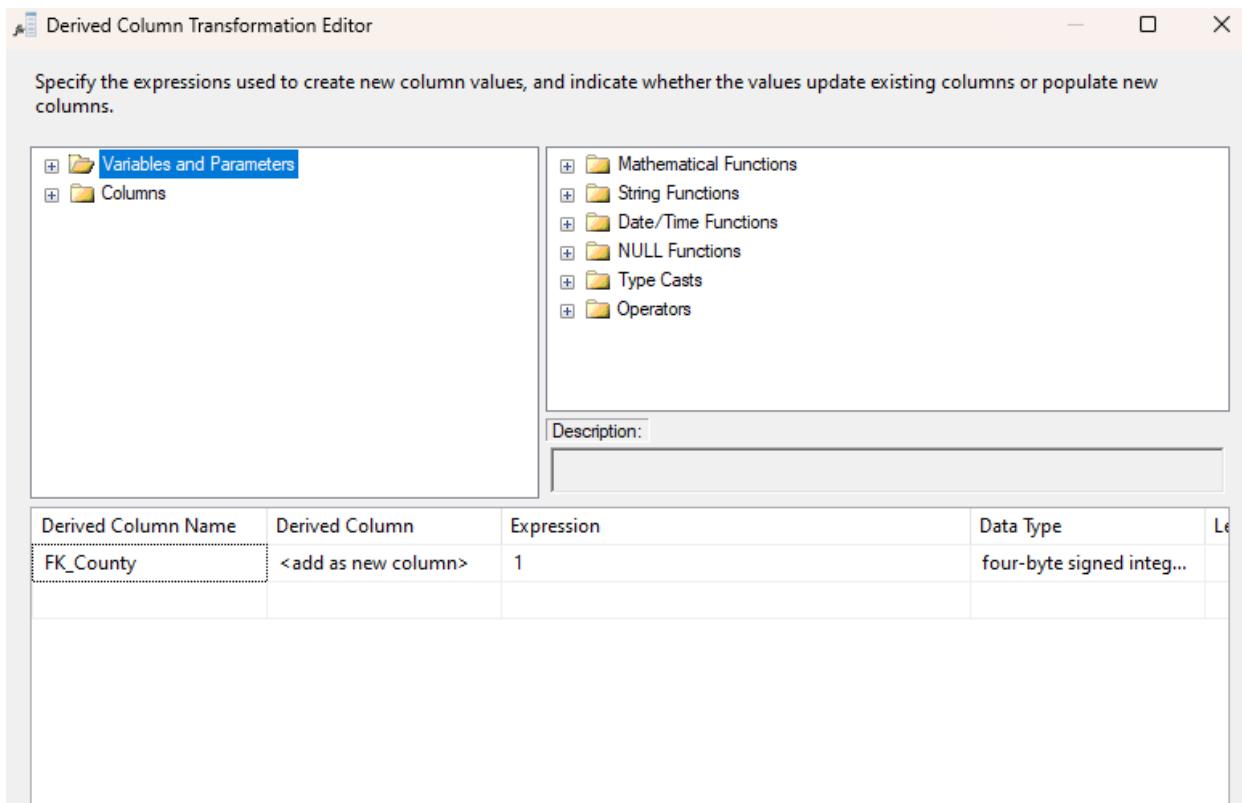
- Đỗ dữ liệu từ Stage vào, bao gồm dữ liệu từ 3 bảng Stage là **aqi2021_Sstage**, **aqi2022_Sstage** và **aqi2023_Sstage** (không có bảng uscounties_Sstage vì dữ liệu liên quan đến Site chỉ có trong 3 bảng còn lại).
- Sau khi đỗ dữ liệu từ 3 bảng aqi của cả 3 năm vào, thực hiện UNION ALL vì thông tin cơ bản của cả 3 bảng là giống nhau và tương đồng về số lượng cột dữ liệu.



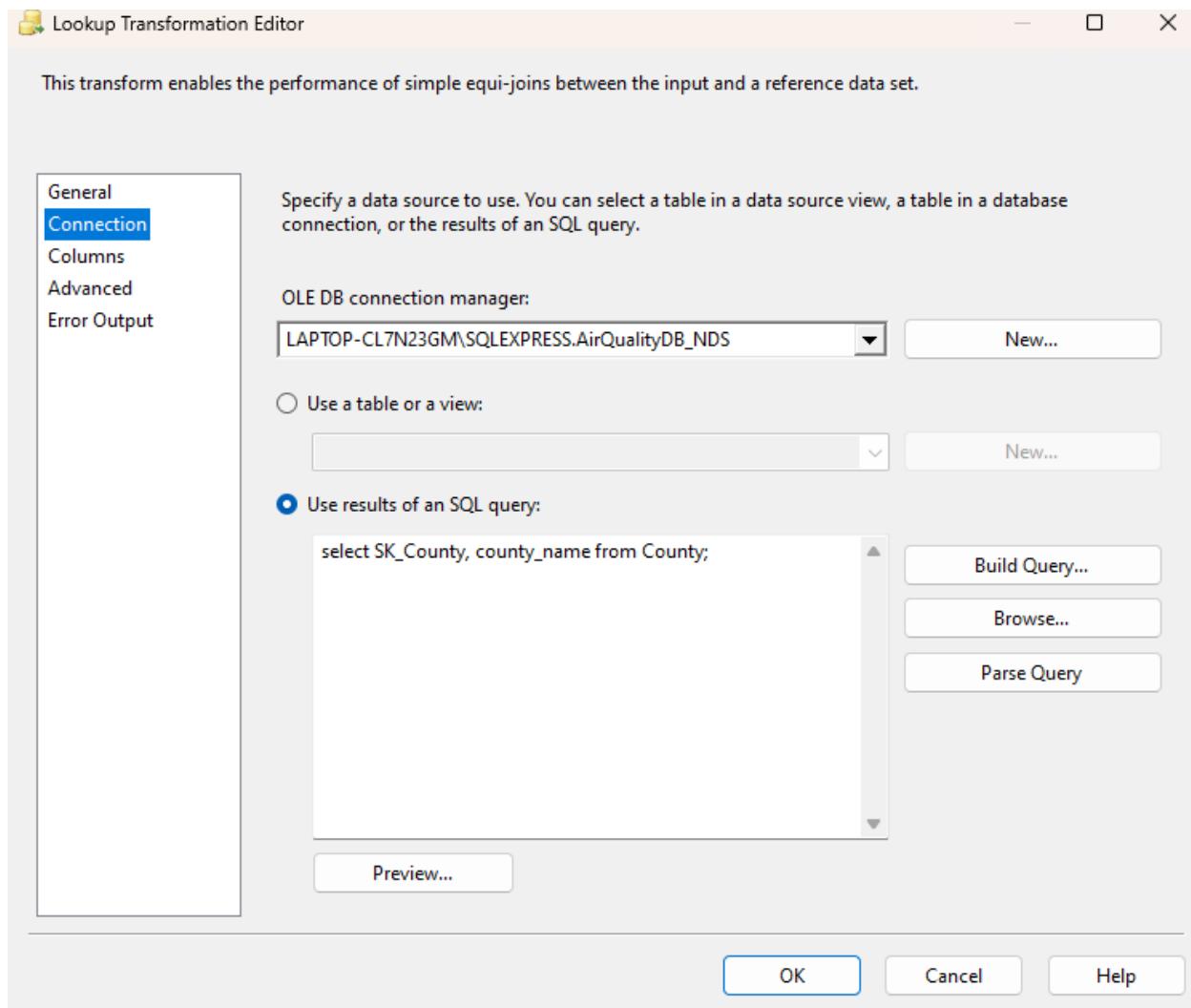
3. Sau khi gộp bảng, dùng lệnh AGGREGATE để thực hiện nhóm các dòng dữ liệu liên quan về trạm quan trắc theo từng quận bằng GROUP BY.



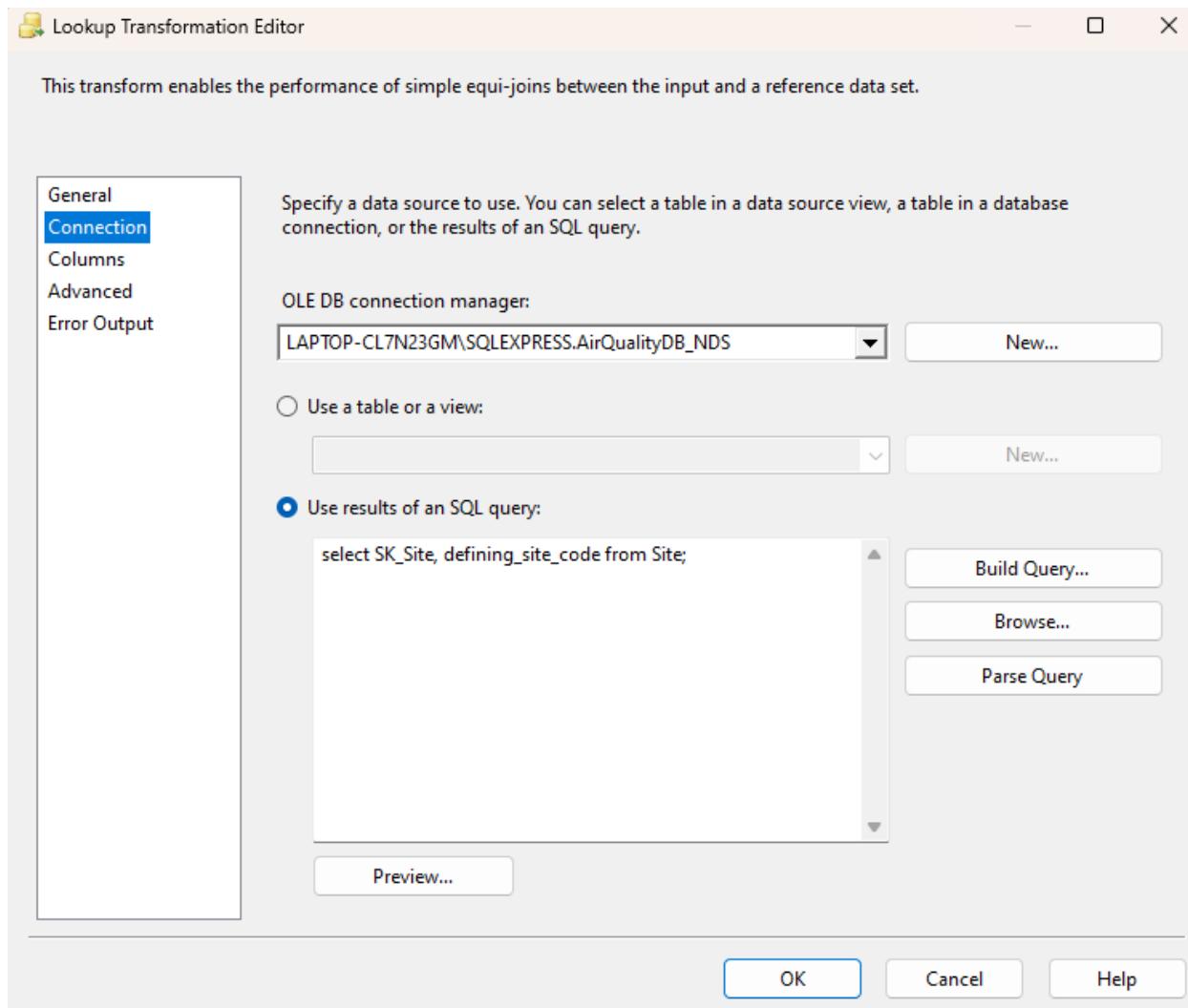
- Sau khi MERGE JOIN, thêm khóa **FK_County** tham chiếu tới bảng COUNTY như là một Derived Column.



5. Tìm trong bảng bảng LOOKUP xem có tồn tại thông tin tương ứng với câu query trong hình, ở đây là tìm xem có tồn tại thông tin về quận được truy vấn hay không, nếu có dữ liệu tương thích sẽ tiến đến bước tiếp theo.

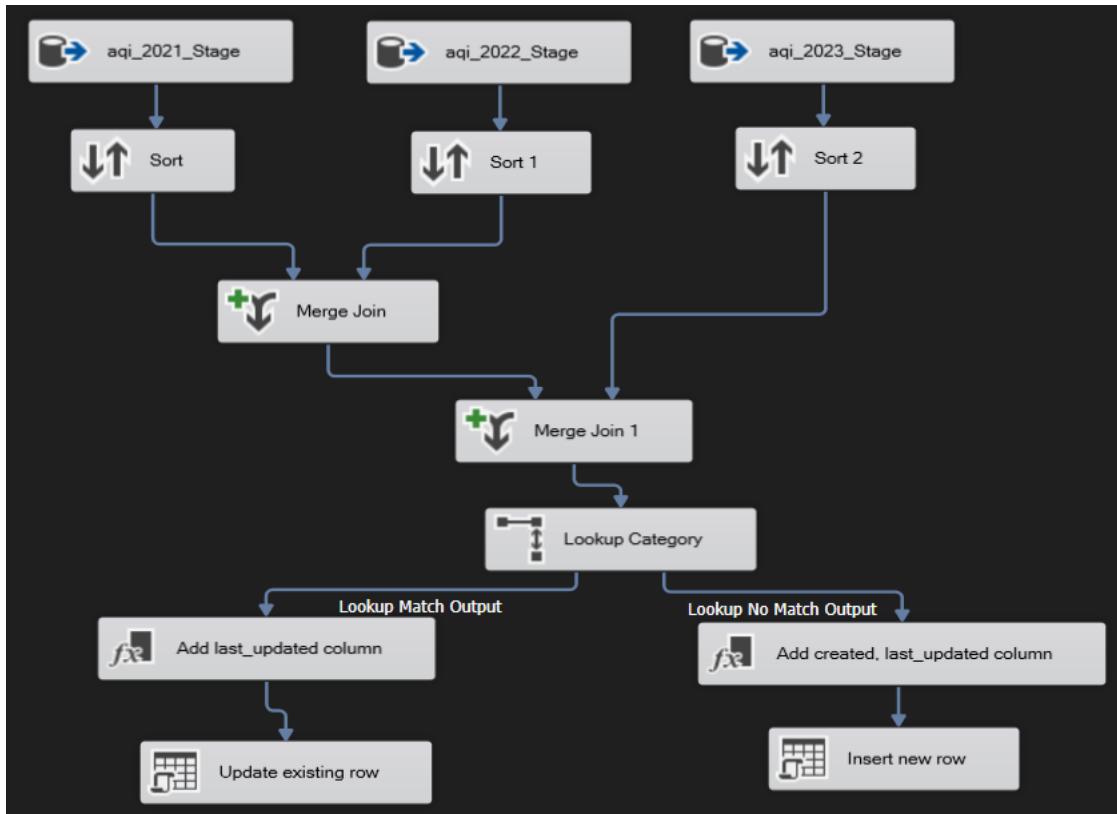


6. Tìm trong bảng bảng LOOKUP xem có tồn tại thông tin tương ứng với câu query trong hình, ở đây là tìm xem có tồn tại thông tin về trạm quan trắc thuộc quận đã được truy vấn ở trên hay không.



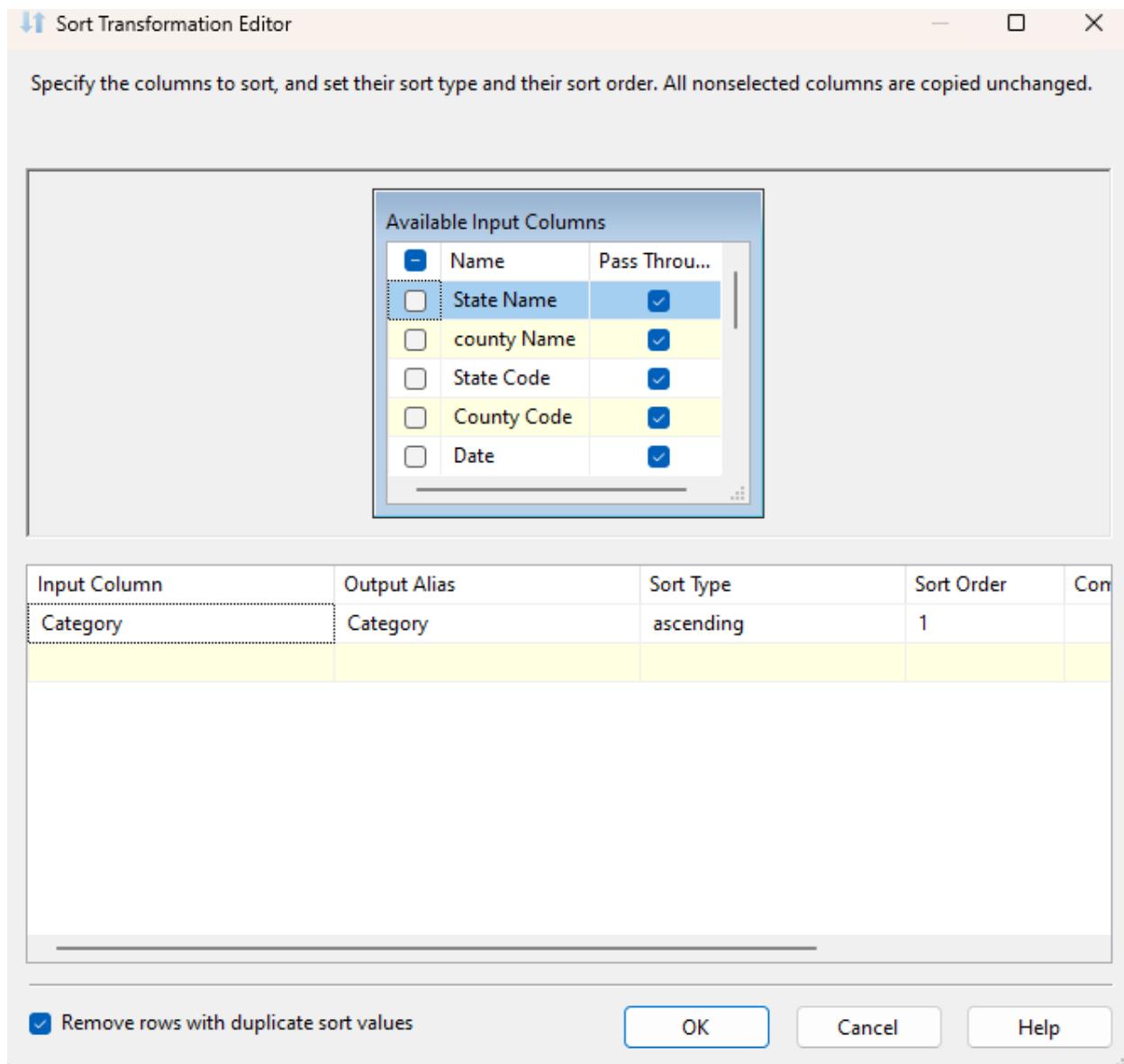
7. Nếu như dữ liệu không tồn tại, tạo thêm dòng mới để chứa dữ liệu mới đồng thời cập nhật ngày giờ tạo bản ghi và lần cập nhật gần nhất. Nếu như dữ liệu đã tồn tại thì chỉnh sửa bản ghi đã có và cập nhật lại thời gian cập nhật cuối cùng.

4.4. Bảng CATEGORY trong NDS:

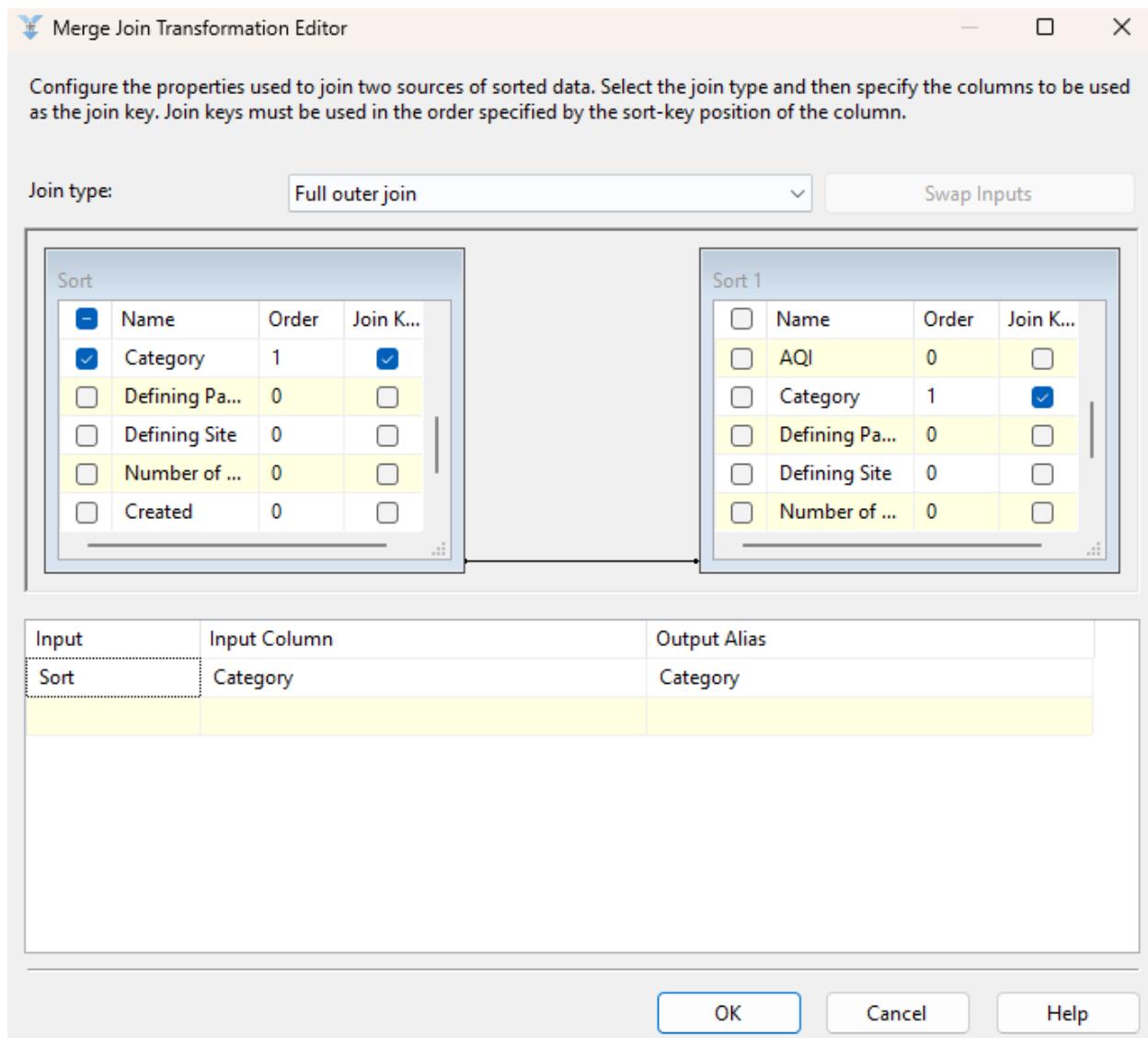


Các bước thực hiện:

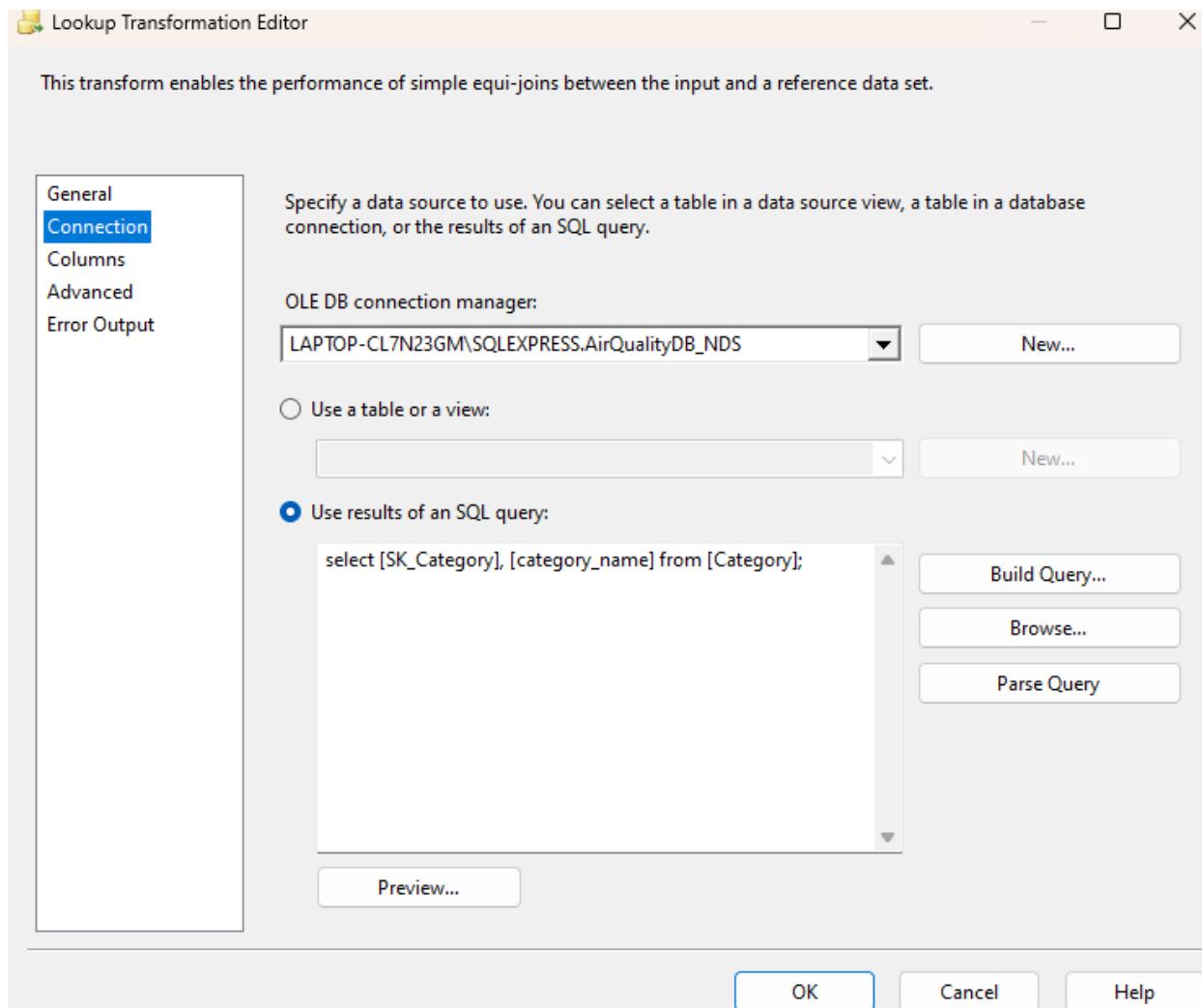
- Đỗ dữ liệu từ Stage vào, bao gồm dữ liệu từ 3 bảng Stage là **aqi2021_Sstage**, **aqi2022_Sstage** và **aqi2023_Sstage** (không có bảng uscounties_Sstage vì dữ liệu liên quan đến Site chỉ có trong 3 bảng còn lại).
- Cần MERGE JOIN dữ liệu ở cả 3 bảng này, và trước khi MERGE dữ liệu cần được SORT theo thuộc tính key để MERGE, ở đây là thuộc tính **category** cho cả 3 bảng.



3. Vì MERGE JOIN chỉ cho phép 2 bảng một lúc, nên cần MERGE 2 bảng của năm 2021 và 2022 trước, sau đó mới tiếp tục MERGE JOIN với bảng của năm 2023, sử dụng kiểu JOIN là Full Outer Join.

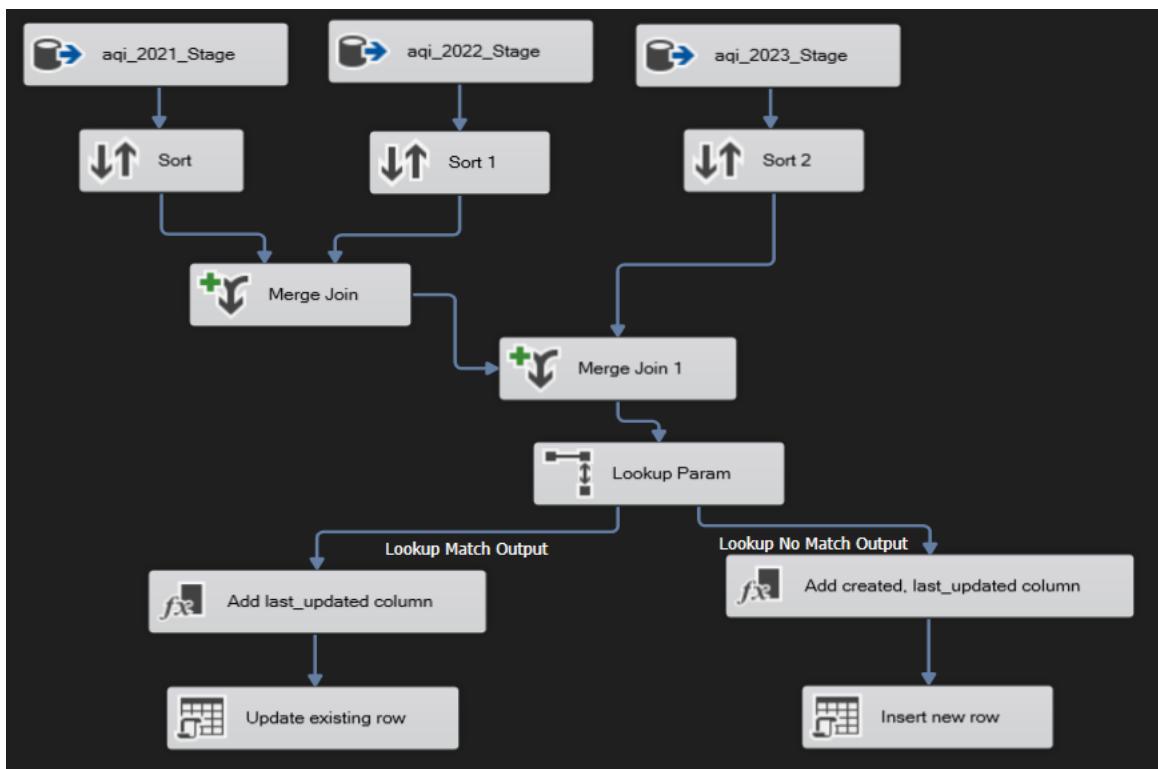


4. Sau khi MERGE JOIN thành 1 bảng dữ liệu duy nhất, tiến hành LOOKUP xem có tồn tại thông tin tương ứng với câu query trong hình.



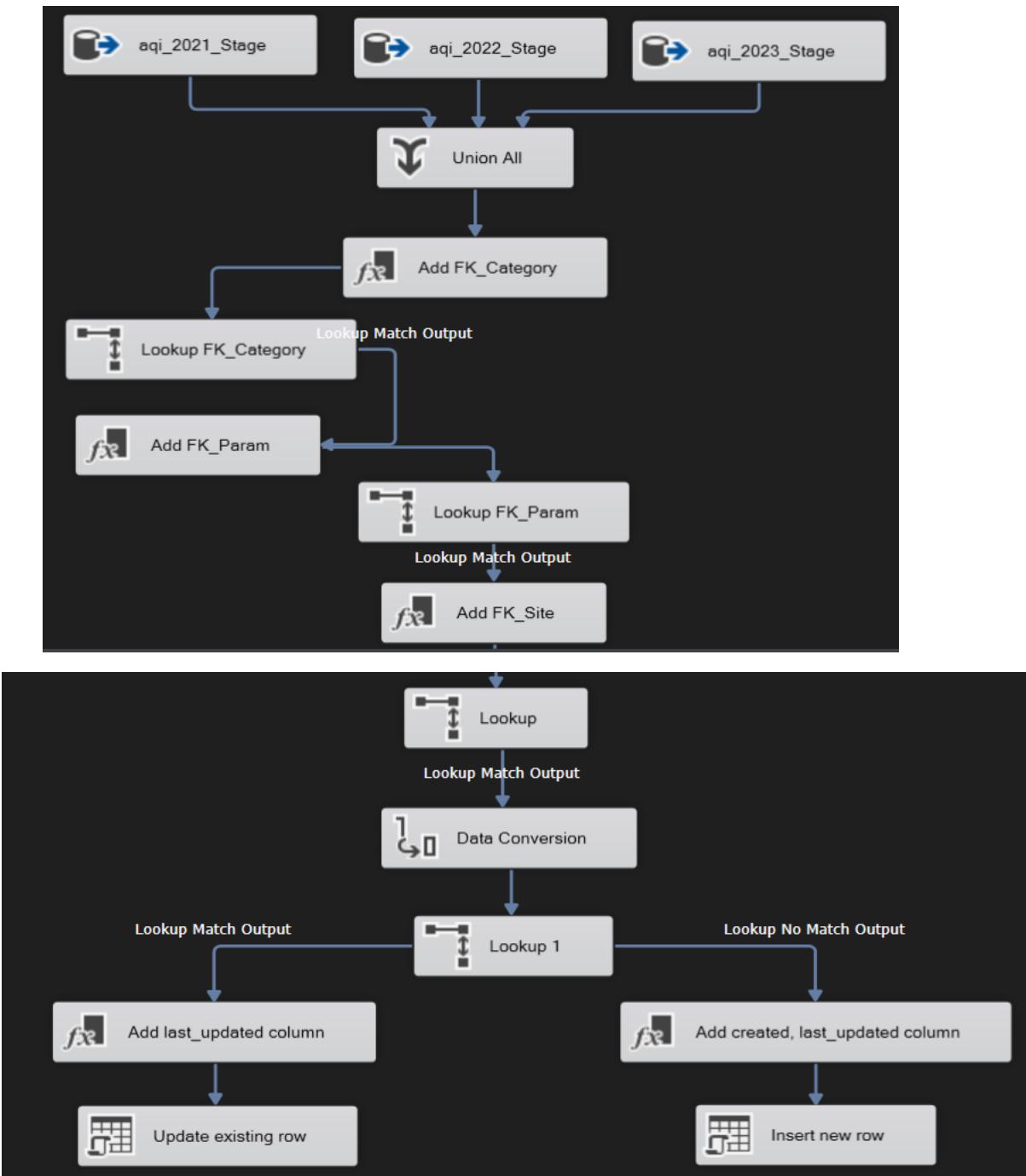
5. Nếu như dữ liệu không tồn tại, tạo thêm dòng mới để chứa dữ liệu mới đồng thời cập nhật ngày giờ tạo bản ghi và lần cập nhật gần nhất. Nếu như dữ liệu đã tồn tại thì chỉnh sửa bản ghi đã có và cập nhật lại thời gian cập nhật cuối cùng.

4.5. Bảng PARAM trong NDS:



Các bước thực hiện của bảng này tương tự như CATEGORY, chỉ thay các thuộc tính chọn từ CATEGORY thành DEFINE PARAMETERS.

4.6. Bảng AQI trong NDS:



Các bước thực hiện:

- Đỗ dữ liệu từ Stage vào, bao gồm dữ liệu từ 3 bảng Stage là **aqi2021_Sstage**, **aqi2022_Sstage** và **aqi2023_Sstage** (không có bảng uscounties_Sstage vì dữ liệu liên quan đến Site chỉ có trong 3 bảng còn lại).
- Sau khi đỗ dữ liệu từ 3 bảng aqi của cả 3 năm vào, thực hiện UNION ALL vì thông tin cơ bản của cả 3 bảng là giống nhau và tương đồng về số lượng cột dữ liệu.

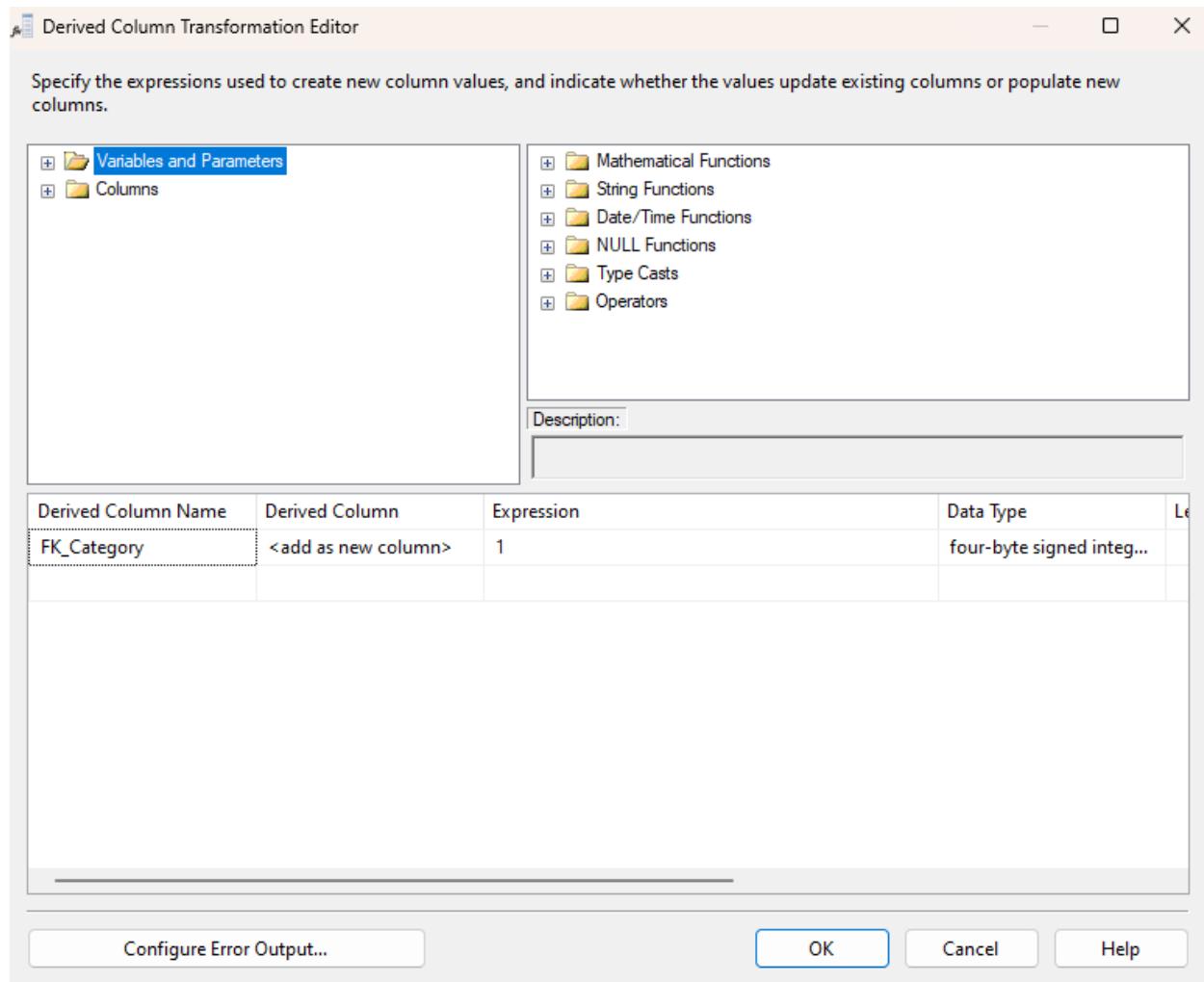
Union All Transformation Editor

Configure the properties used to merge multiple inputs into one output by creating mappings between columns.

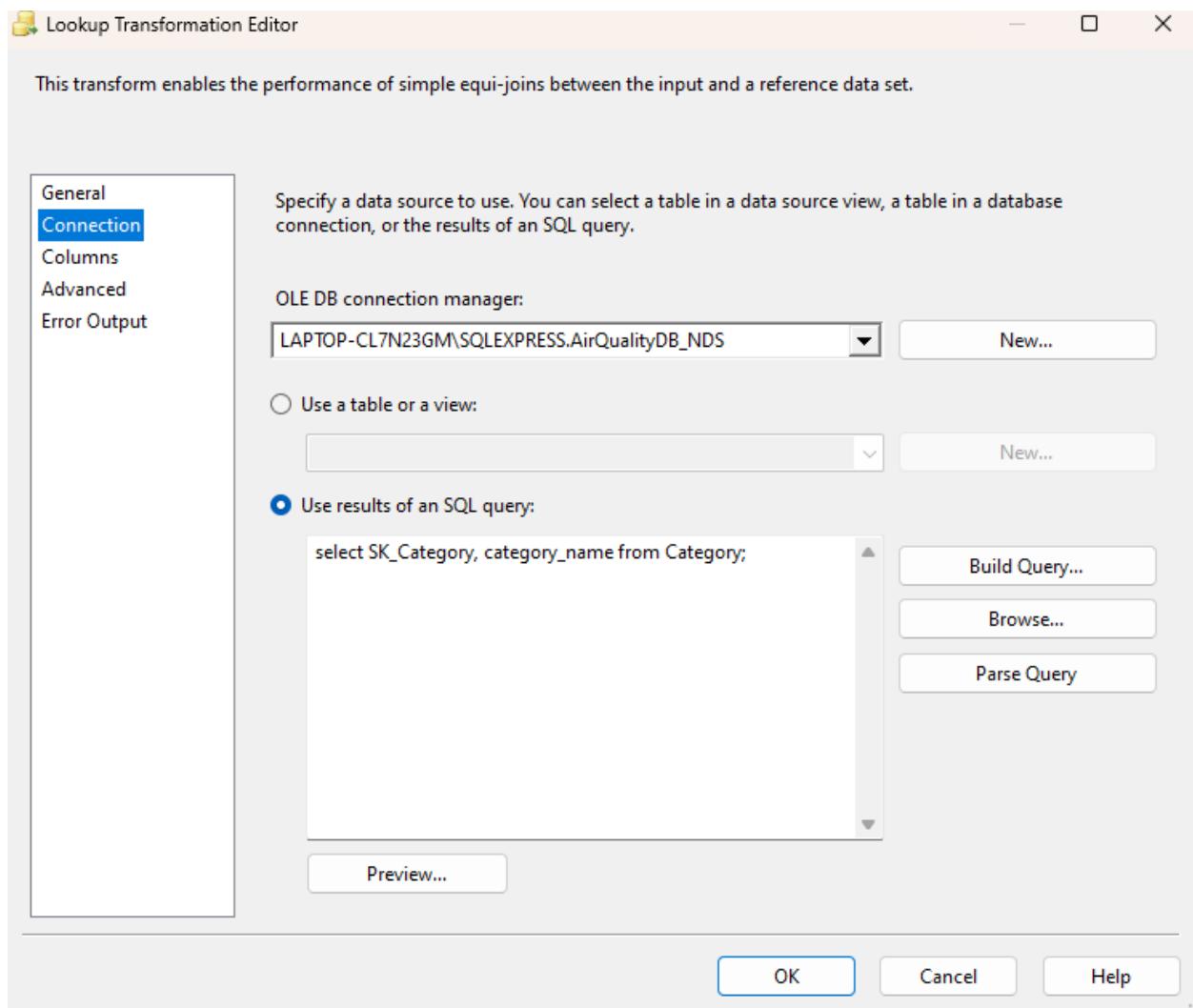
	Union All Input 1	Union All Input 2	Union All Input 3
	State Name	State Name	State Name
	county Name	county Name	county Name
	State Code	State Code	State Code
	County Code	County Code	County Code
	Date	Date	Date
	AQI	AQI	AQI
	Category	Category	Category
	Defining Parameter	Defining Parameter	Defining Parameter
	Defining Site	Defining Site	Defining Site
	Number of Sites Reporting	Number of Sites Reporting	Number of Sites Reporting
	Created	Created	Created
	Last Updated	Last Updated	Last Updated

OK Cancel Help

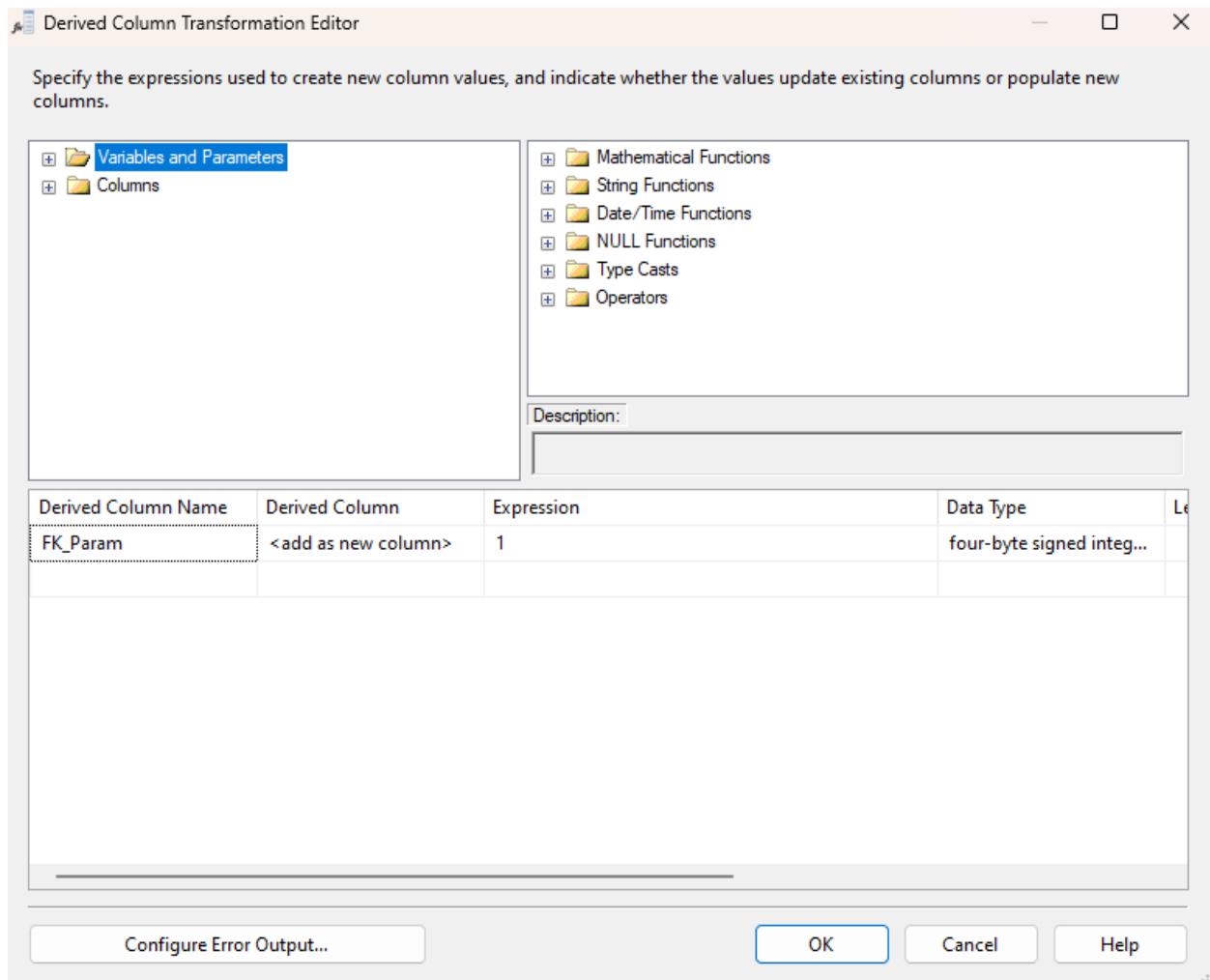
3. Thêm khóa **FK_Category** tham chiếu tới bảng CATEGORY như là một Derived Column.

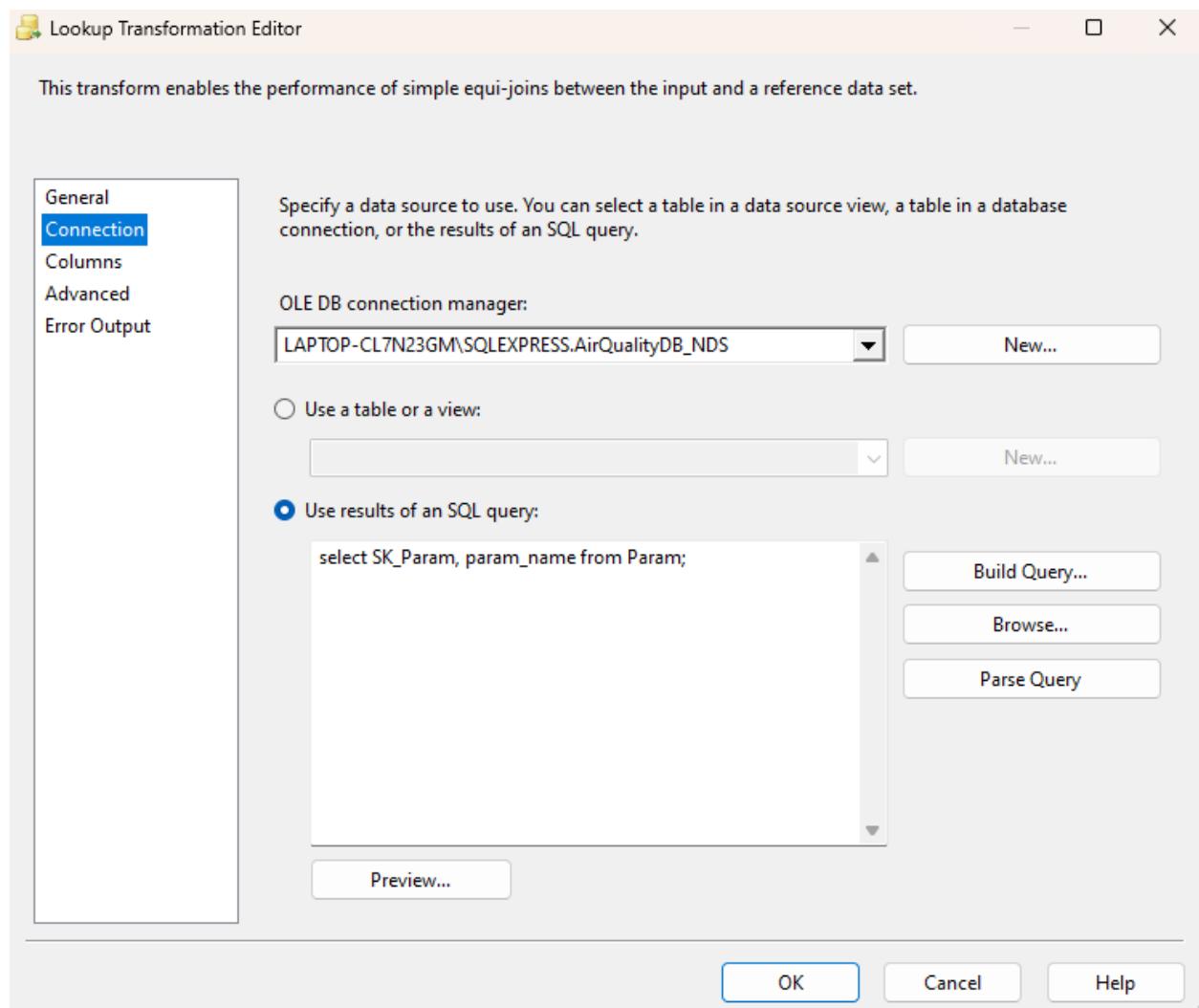


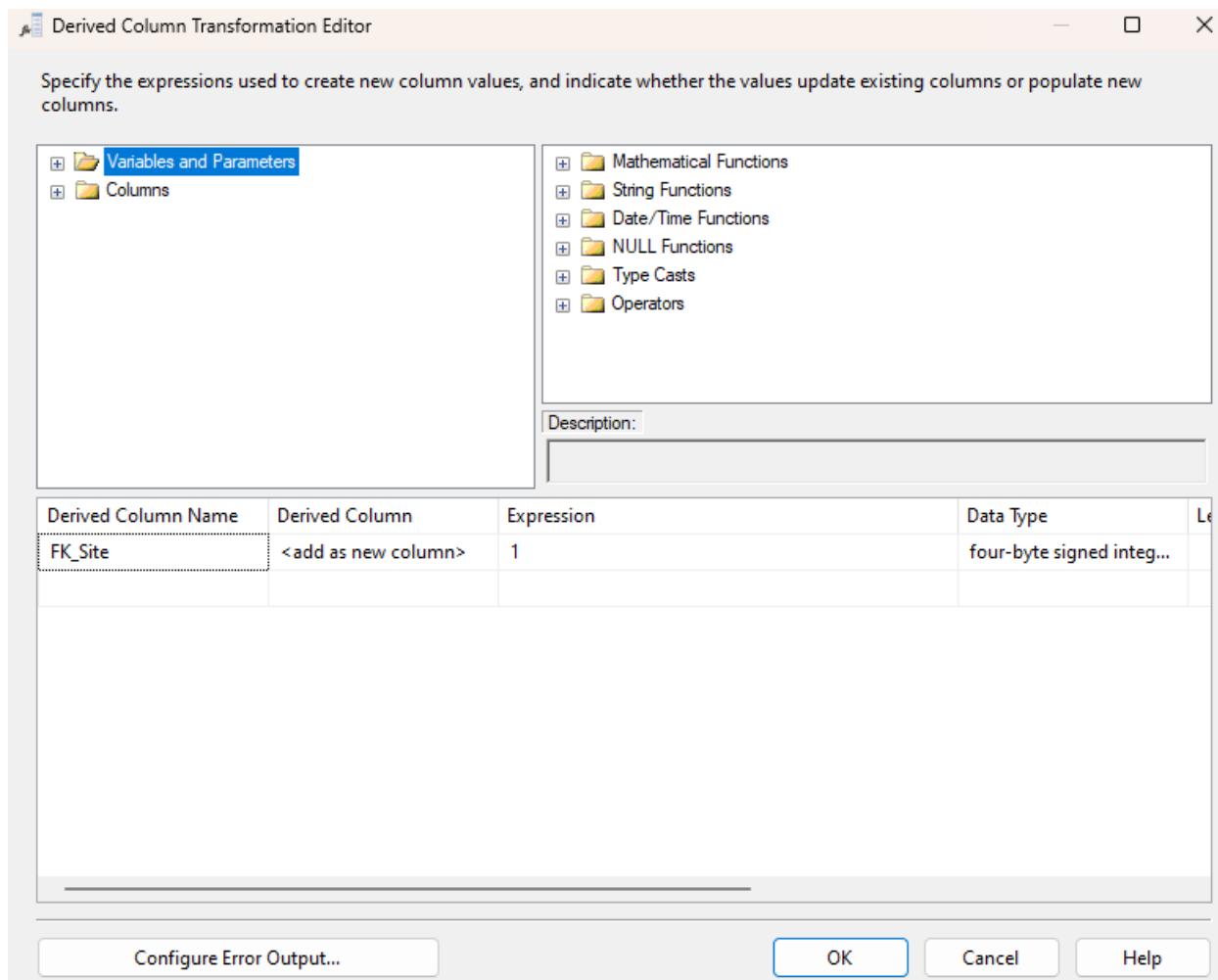
4. Tìm trong bảng bằng LOOKUP xem có tồn tại thông tin tương ứng với câu query trong hình, ở đây là tìm xem có tồn tại thông tin về phân loại được truy vấn hay không, nếu có dữ liệu tương thích sẽ tiến đến bước tiếp theo.

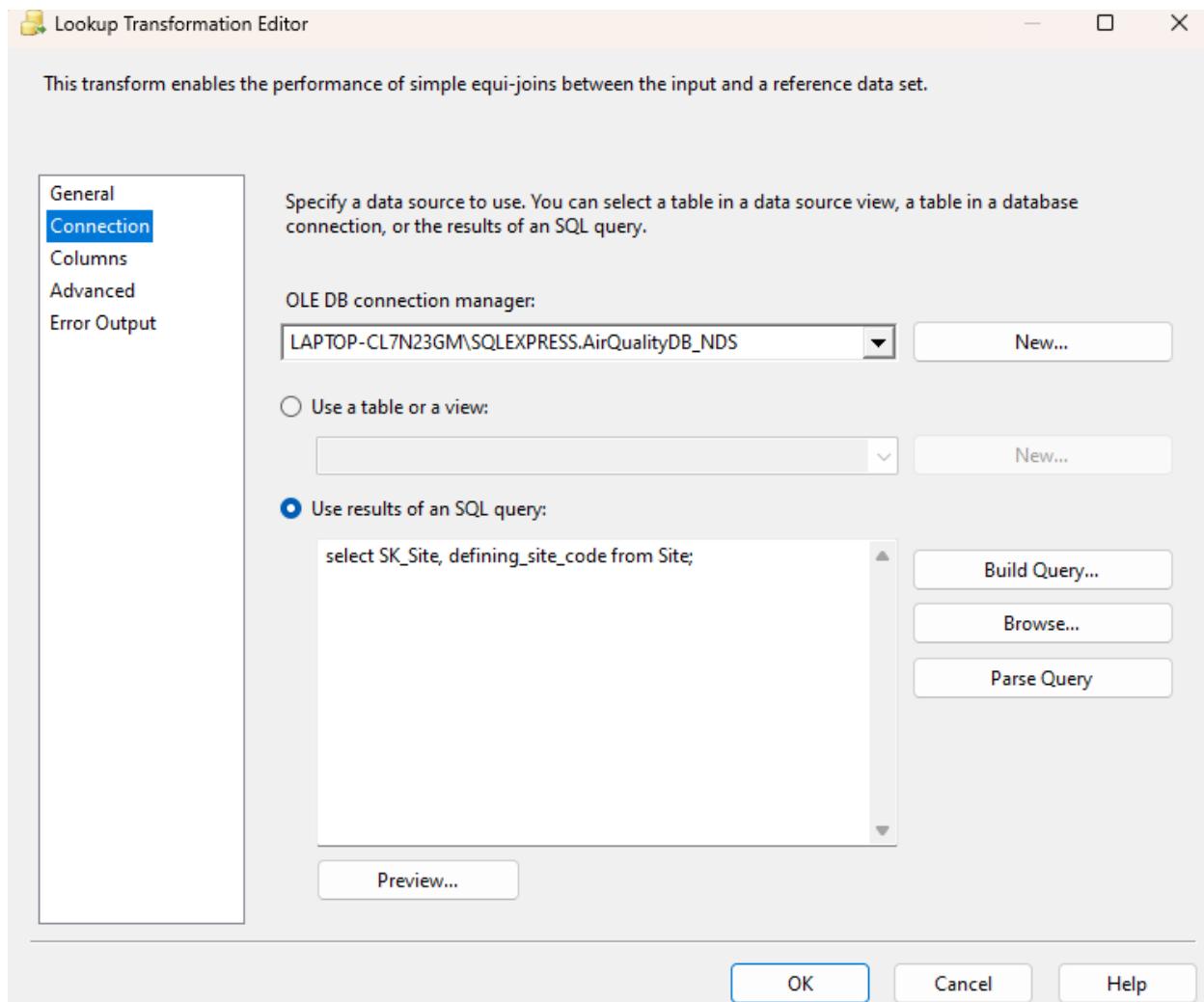


5. Lặp lại tương tự cho FK_Param tương ứng khóa ngoại đến bảng PARAMETER và FK_Site tương ứng khóa ngoại đến bảng SITE.

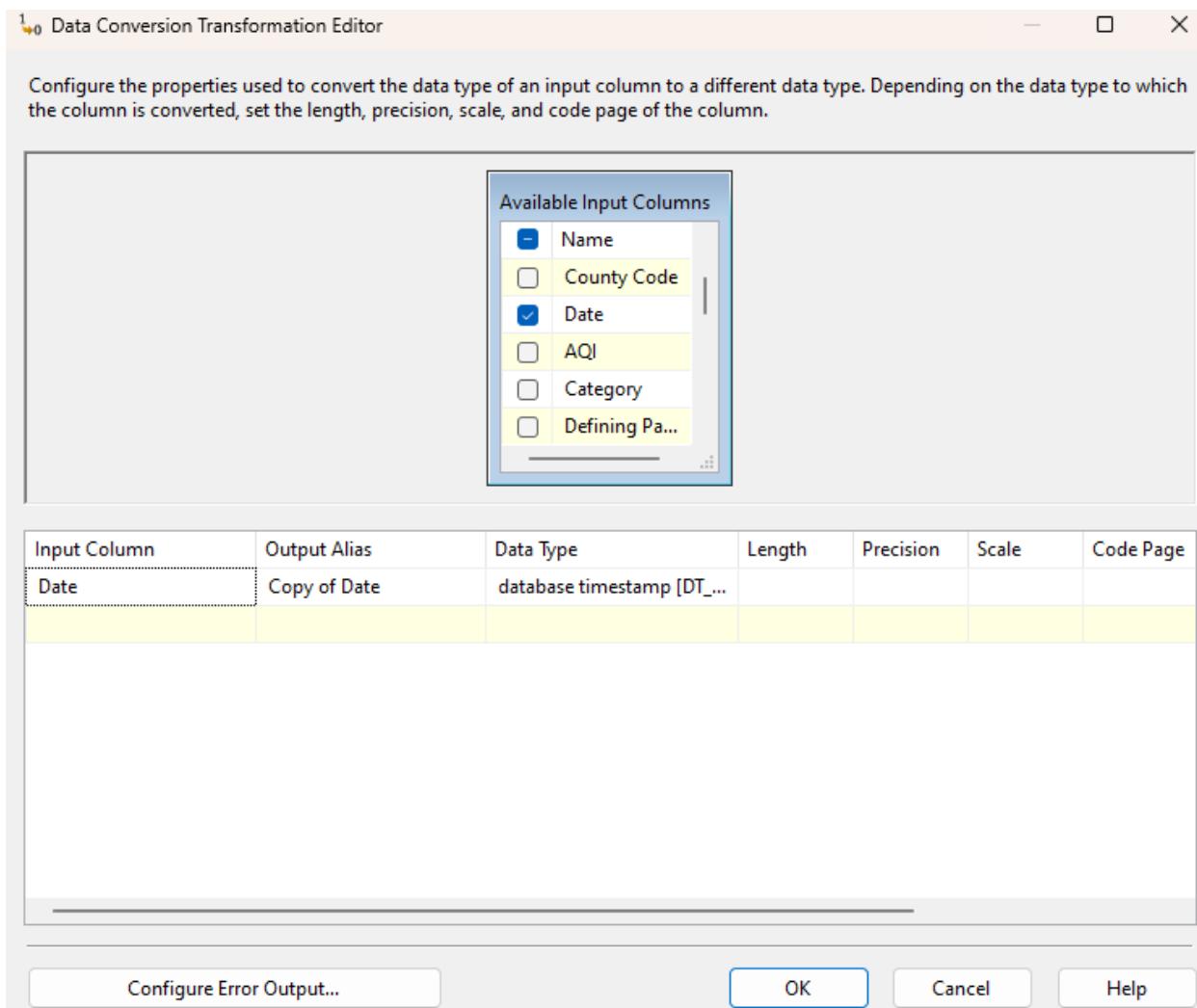




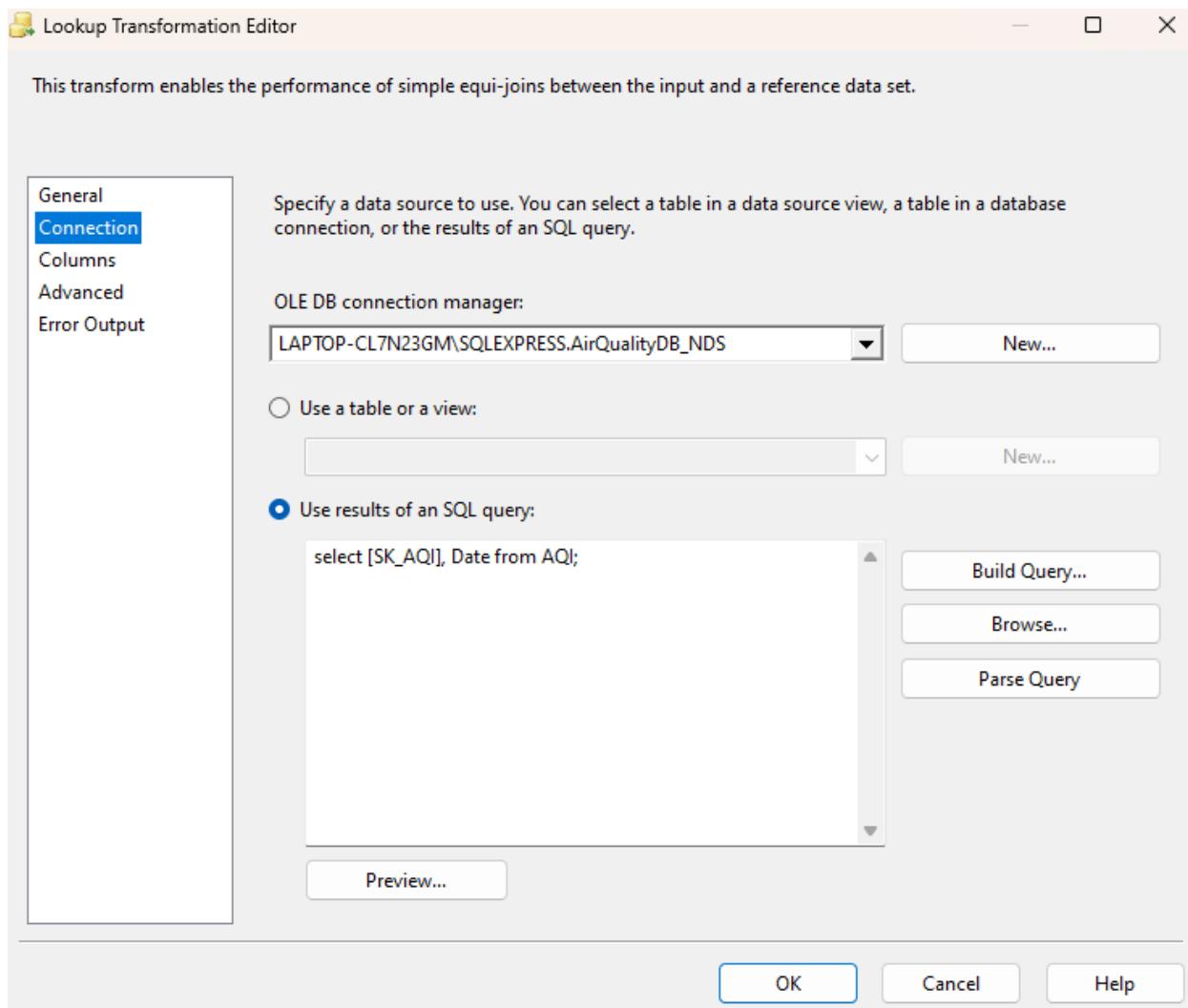




6. Thực hiện Data Conversion để lấy dữ liệu về Date và lưu đúng định dạng.



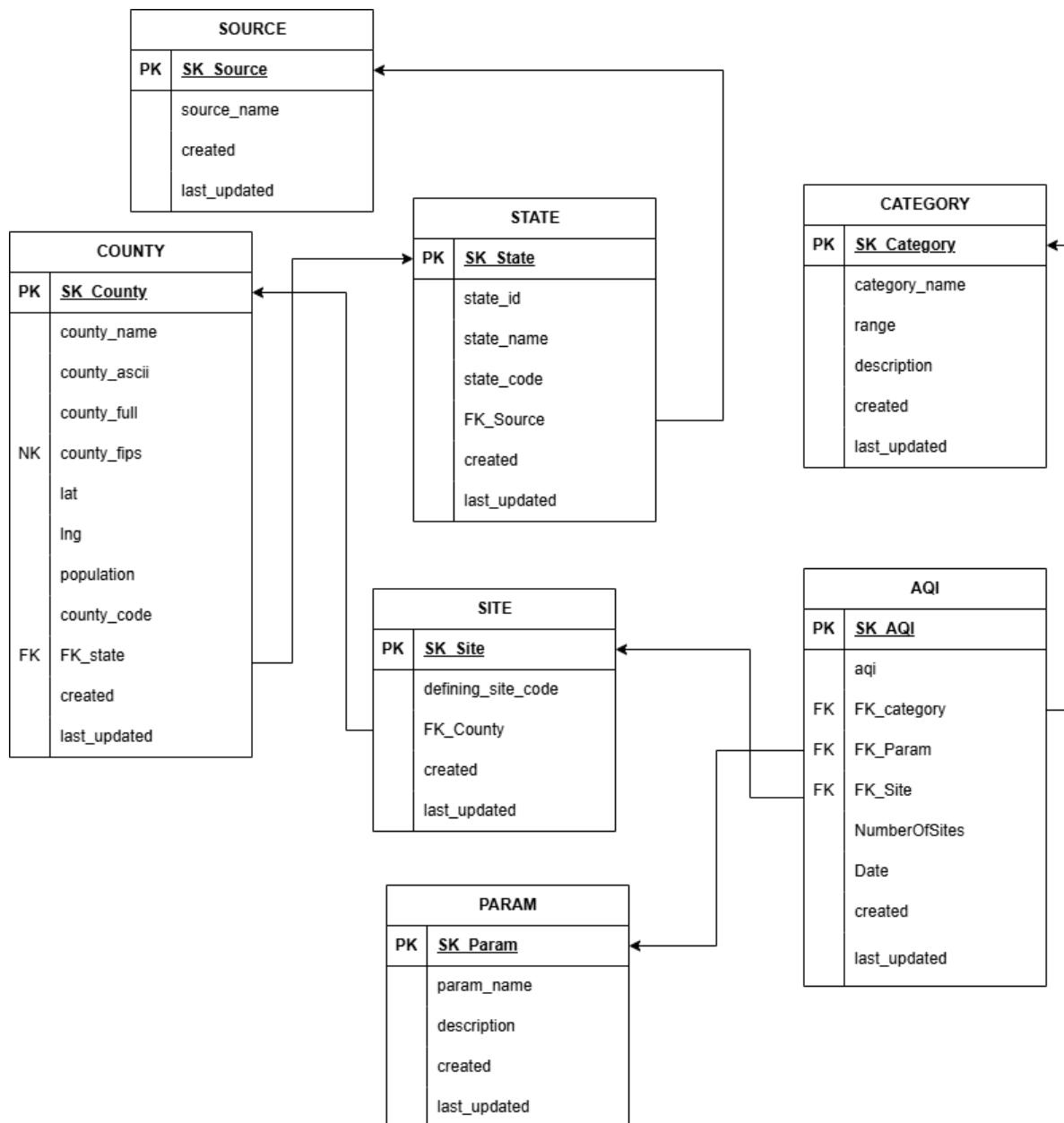
7. Sau khi thêm đầy đủ các cột dữ liệu có tham chiếu đến các bảng khác, tiến hành LOOKUP xem có tồn tại thông tin tương ứng với câu query trong hình, ở đây là tìm xem có tồn tại thông tin chi tiết về AQI được truy vấn đến hay không.



8. Nếu như dữ liệu không tồn tại, tạo thêm dòng mới để chứa dữ liệu mới đồng thời cập nhật ngày giờ tạo bản ghi và lần cập nhật gần nhất. Nếu như dữ liệu đã tồn tại thì chỉnh sửa bản ghi đã có và cập nhật lại thời gian cập nhật cuối cùng.

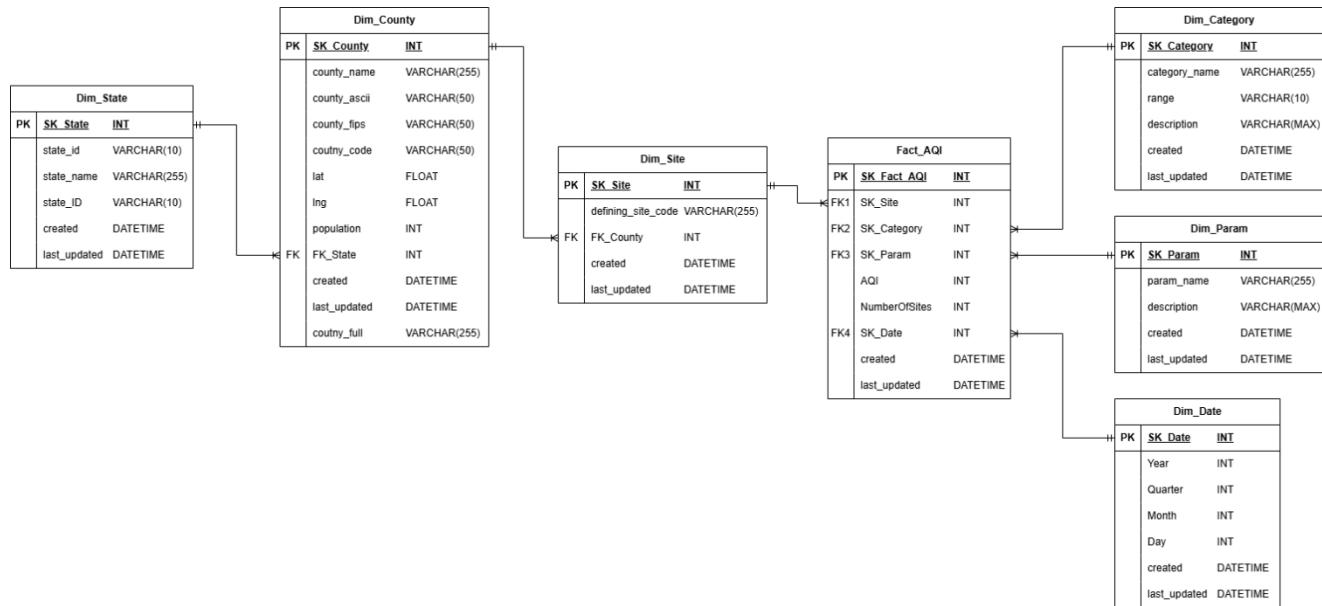
5. ETL từ NDS → DDS

5.1. Cấu trúc dữ liệu nguồn NDS



5.2. Quy trình chuyển đổi dữ liệu từ NDS sang DDS

5.2.1. Mô hình bông tuyết (Snowflake Schema)



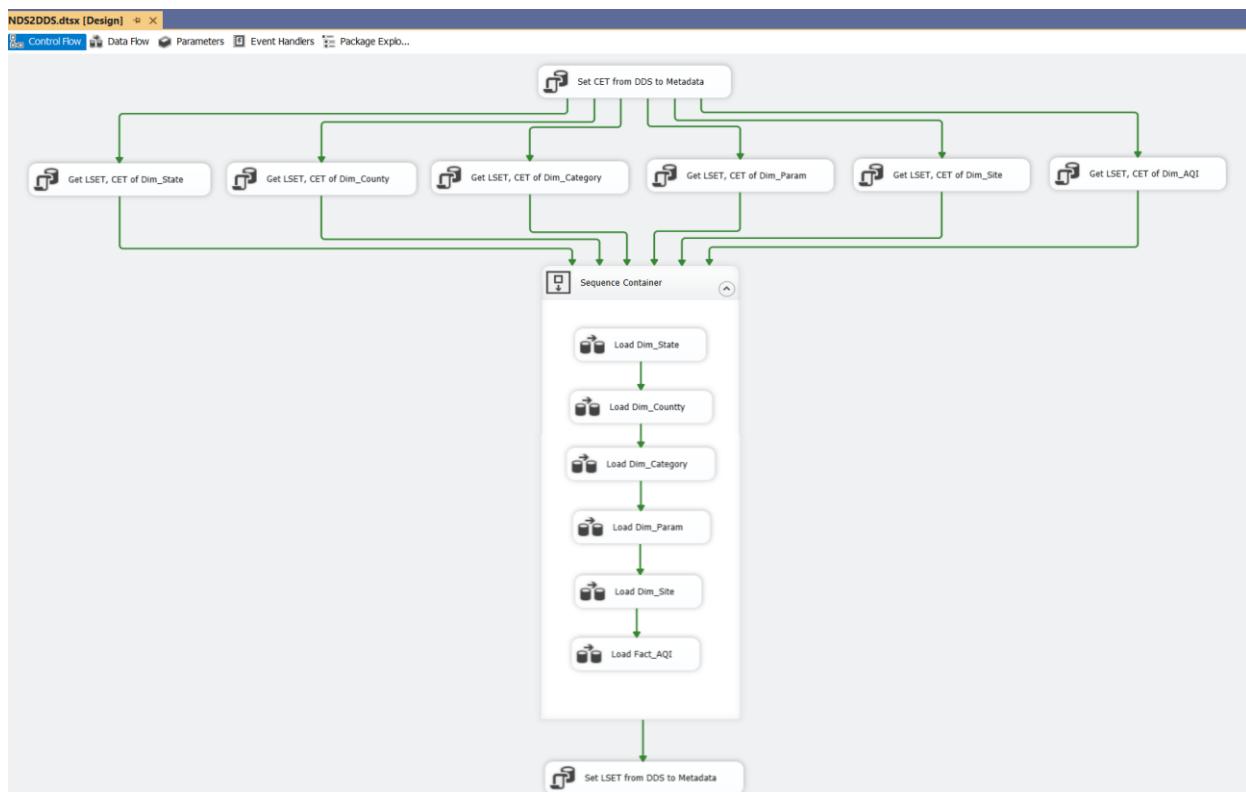
5.2.2. Phân cấp chiều

- **Dim_Date**: Day → Month → Quarter → Year

5.2.3. Facts

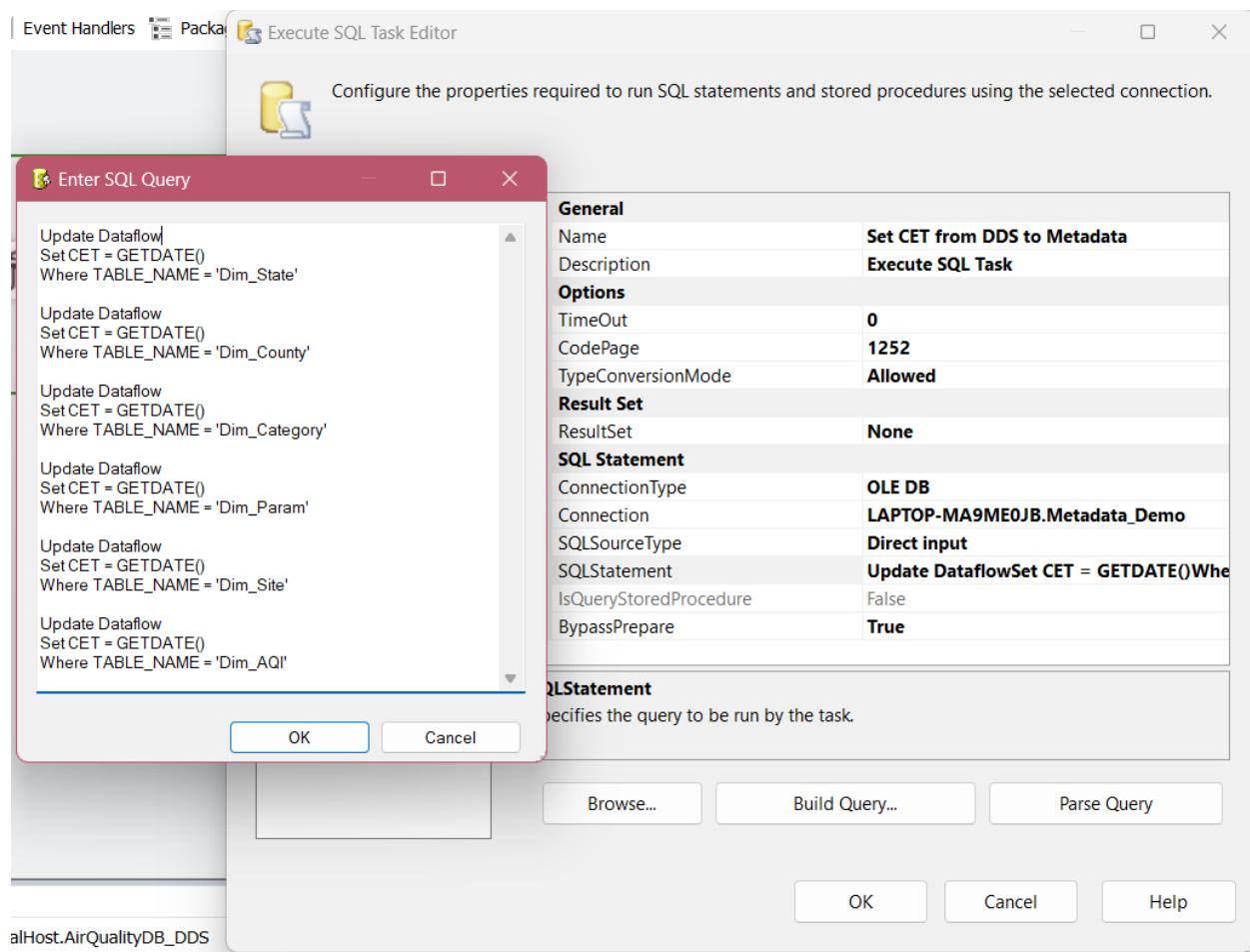
- **Keys:**
 - SK_Site
 - SK_Category
 - SK_Param
 - SK_Date
- **Measures:**
 - **AQI (Air Quality Index)**
 - Có sẵn từ nguồn
 - Semi-additive: có thể cộng dồn theo SK_Site hoặc SK_Category nhưng không nên cộng dồn theo SK_Date
 - **NumberOfSites**
 - Có sẵn từ nguồn
 - Non-additive: Số lượng site không thể cộng dồn theo thời gian hoặc các chiều khác

5.2.4. Flow cơ bản

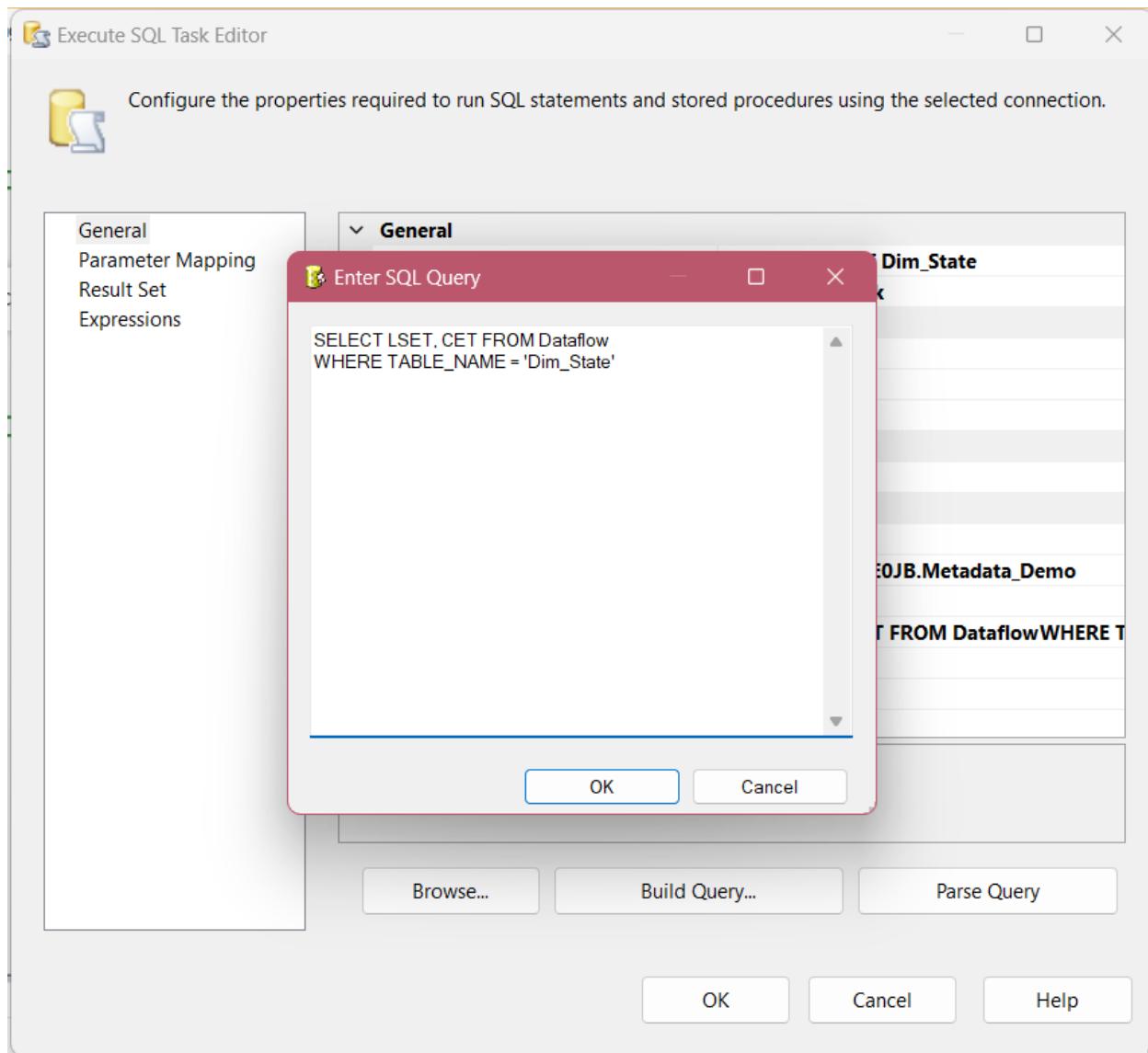


Các bước thực hiện:

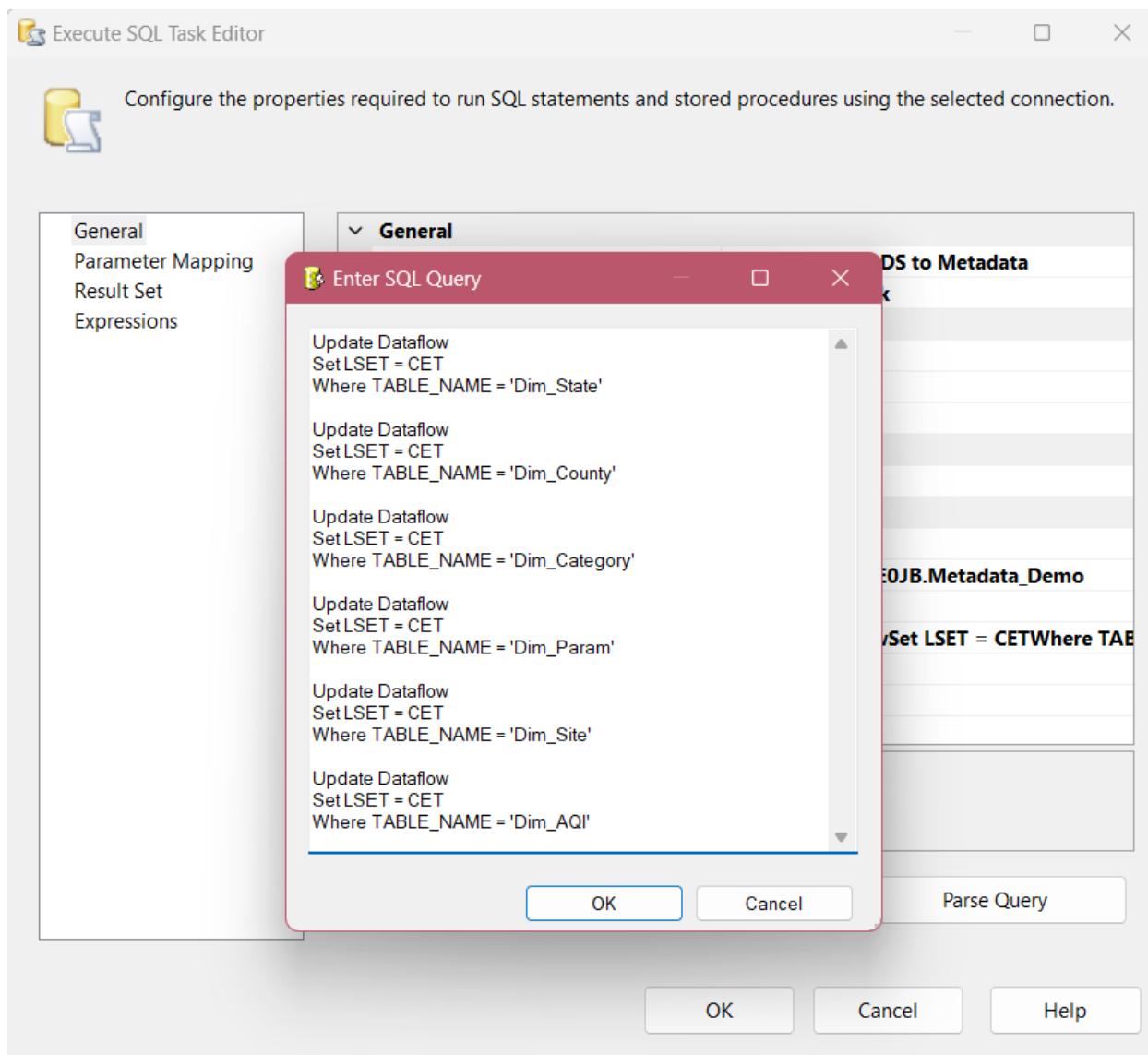
3. Cập nhật LSET = GETDATE() của từng bảng được lưu trong metadata.



4. Lấy LSET, CET của từng bảng lưu vào các biến.



5. Đổ dữ liệu vào các bảng chiều (Dimensions), bảng Fact.
6. Cập nhật lại LSET = CET.



5.2.5. Tạo bảng chiều (Dimensions)

5.2.5.1. Dim_Date

Các bước thực hiện:

3. Viết Stored Procedure để thêm dữ liệu tự động vào bảng Dim_Date.

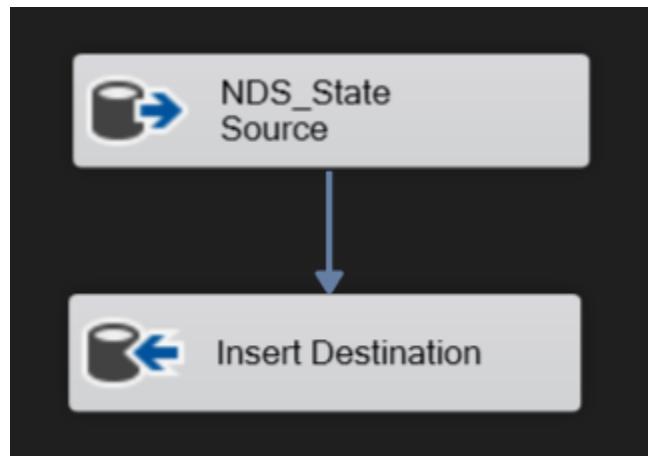
```

CREATE OR ALTER PROCEDURE PopulateDateDimension
AS
BEGIN
    INSERT INTO Dim_Date (Day, Month, Quarter, Year)
    SELECT
        DAY(Date) AS Day,
        MONTH(Date) AS Month,
        CASE
            WHEN MONTH(Date) IN (1, 2, 3) THEN 1
            WHEN MONTH(Date) IN (4, 5, 6) THEN 2
            WHEN MONTH(Date) IN (7, 8, 9) THEN 3
            ELSE 4
        END AS Quarter,
        YEAR(Date) AS Year
        /*CASE
            -- Generalized logic for Daylight Saving Time (example for U.S.)
            WHEN Date >= DATEFROMPARTS(YEAR(Date), 3, 8) AND Date <= DATEFROMPARTS(YEAR(Date), 11, 1) THEN 1
            ELSE 0
        END AS DayLightSaving*/
    FROM
        (SELECT DISTINCT Date FROM AirQualityDB_NDS.dbo.AQI) AS AQI
    ORDER BY
        Date ASC;
END;
EXECUTE PopulateDateDimension

```

4. Nguyên lí: Lấy dữ liệu phân biệt từ trường Date thuộc bảng AQI trong NDS, với giá trị của Day, Month, Quarter và Year được xử lý từ hàm cùng tên là dữ liệu được rút trích ra từ trường Date.

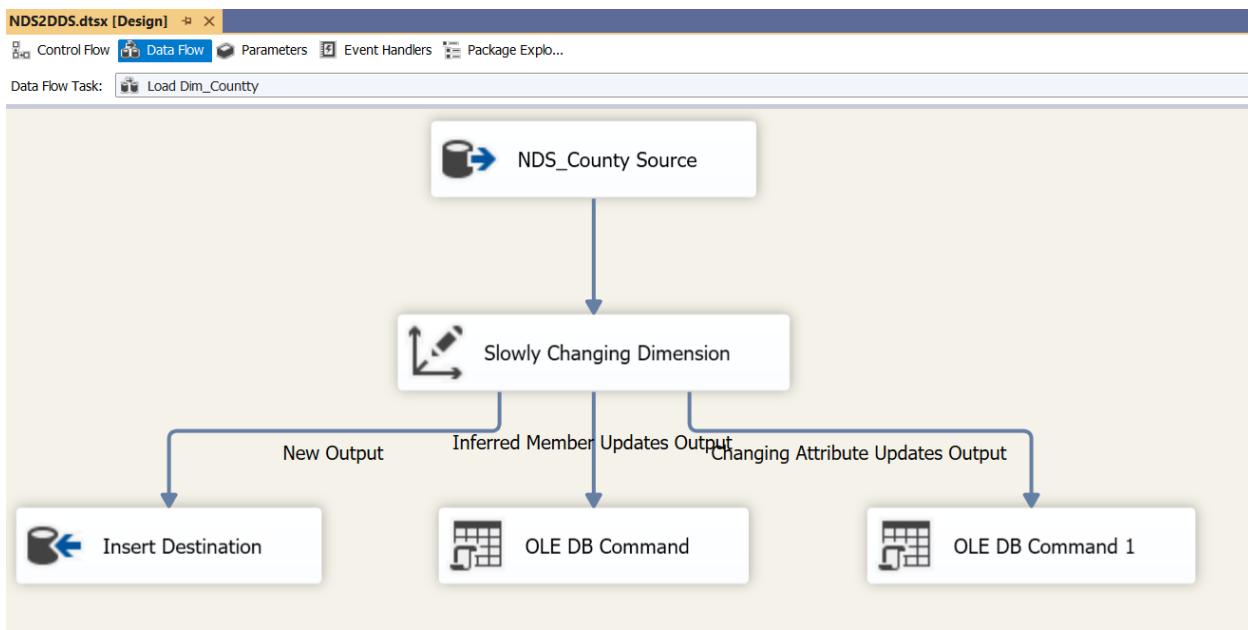
5.2.5.2. Dim_State



Các bước thực hiện:

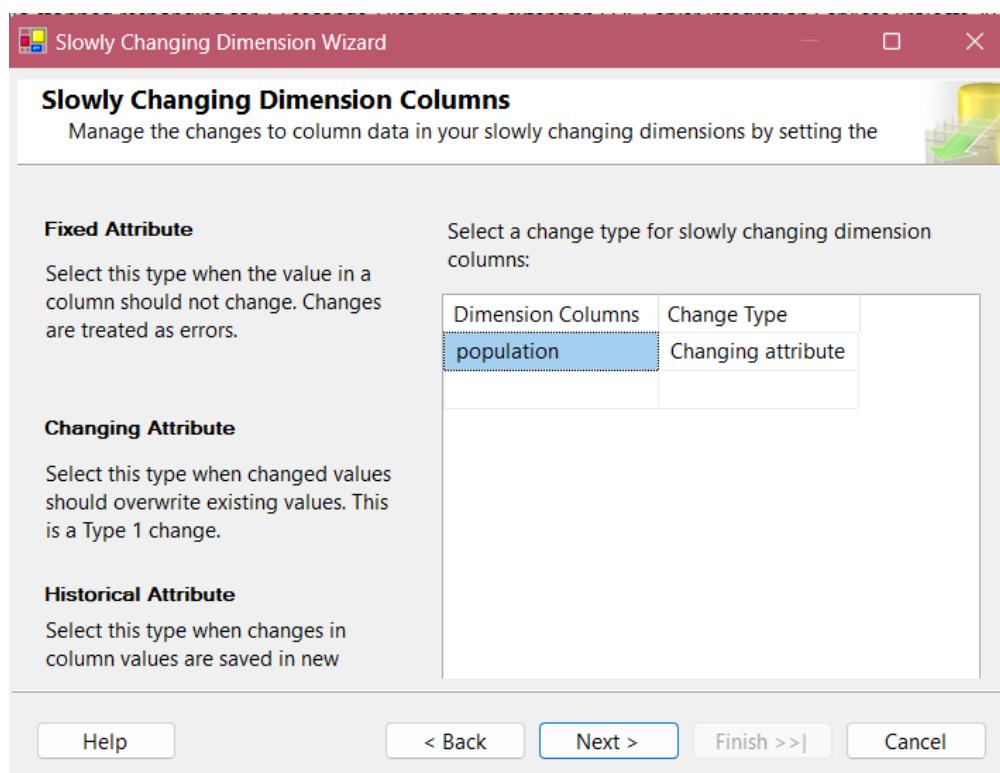
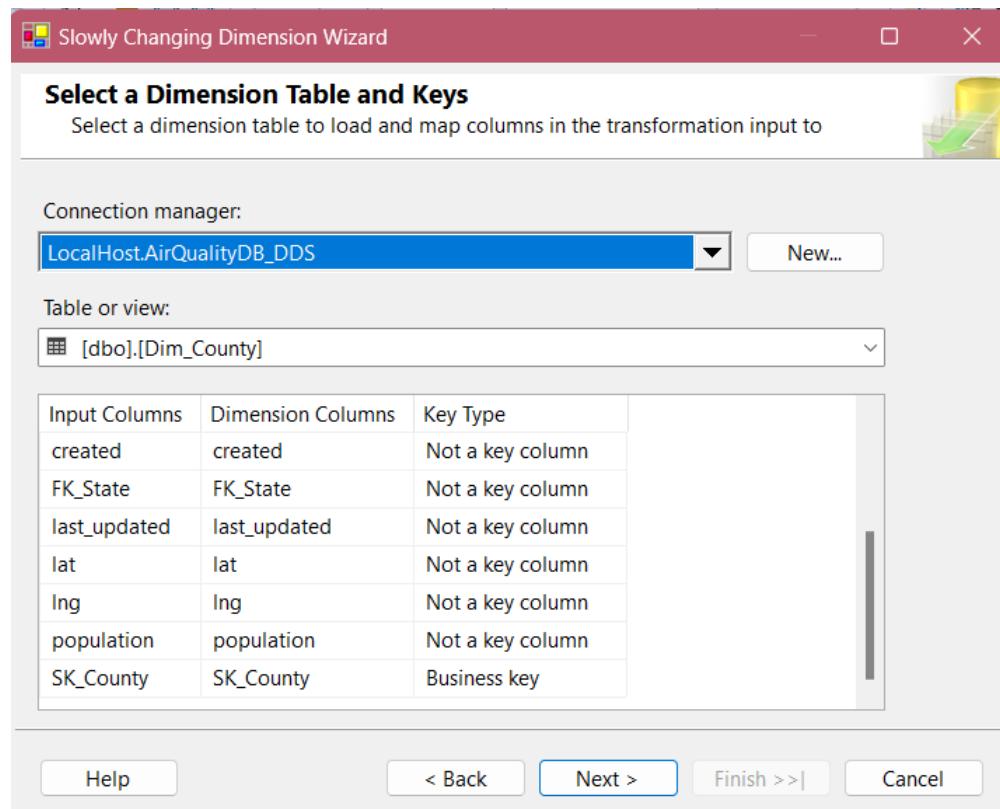
1. Rút trích dữ liệu từ NDS sang DDS:
Select * From State
Where (created > LSET And created <= CET)
Or (last_updated > LSET And last_updated <= CET)
2. Kiểm tra ánh xạ vào bảng DDS.

5.2.5.3. Dim_County



Các bước thực hiện:

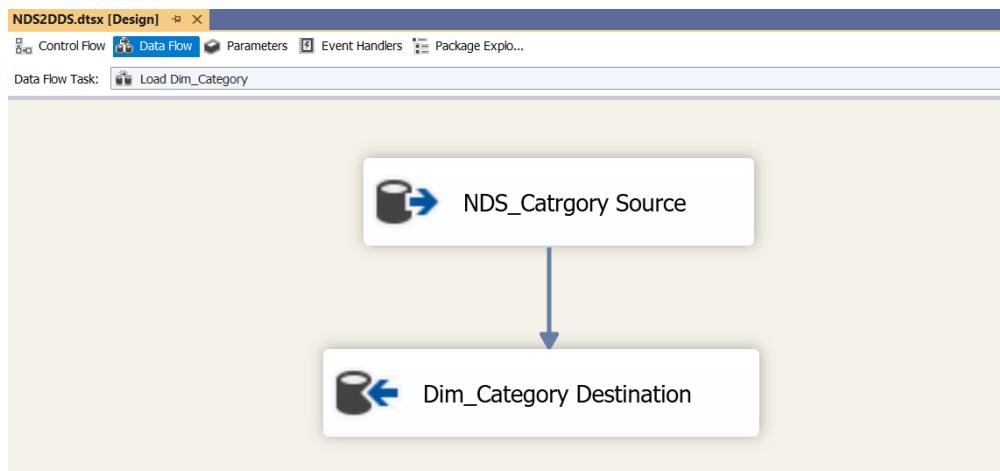
1. Rút trích dữ liệu từ NDS sang DDS:
Select * From County
Where (created > LSET And created <= CET)
Or (last_updated > LSET And last_updated <= CET)
2. Thiết lập SCD (Slow Changing Dimension), chọn khóa chính SK_County làm Business key, chọn population là thuộc tính kiểu Changing attribute (Ghi đè lại giá trị cũ nếu có sự thay đổi dữ liệu).



3. Sau khi setup SCD, hệ thống tự động tạo thêm 3 nhánh nhỏ, bên phải là update thuộc tính ghi đè, ở giữa là update tất cả các thuộc tính, bên trái là thêm dòng dữ liệu mới.

4. Kiểm tra ánh xạ vào bảng DDS.

5.2.5.4. Dim_Category



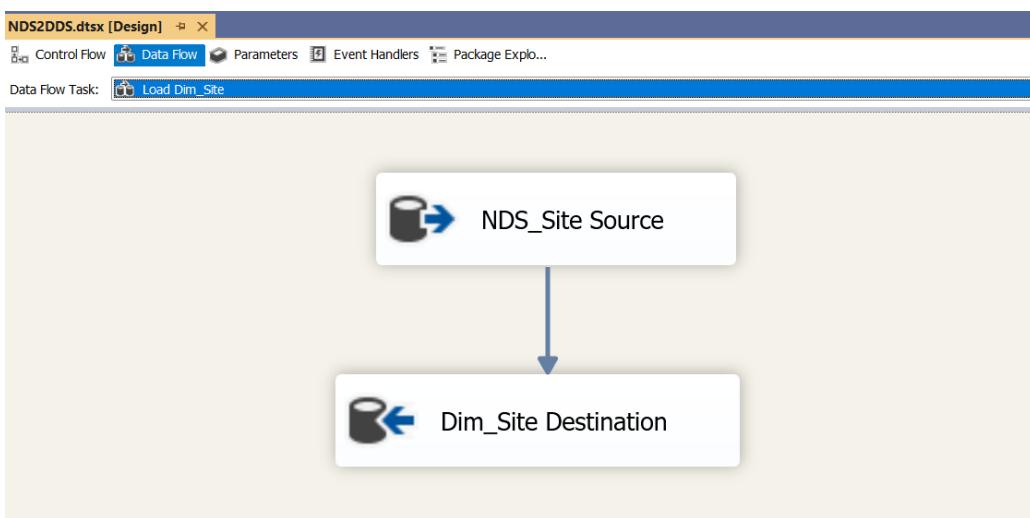
Các bước thực hiện:

3. Rút trích dữ liệu từ NDS sang DDS:

Select * From Category
Where (created > LSET And created <= CET)
Or (last_updated > LSET And last_updated <= CET)

4. Kiểm tra ánh xạ vào bảng DDS.

5.2.5.5. Dim_Site



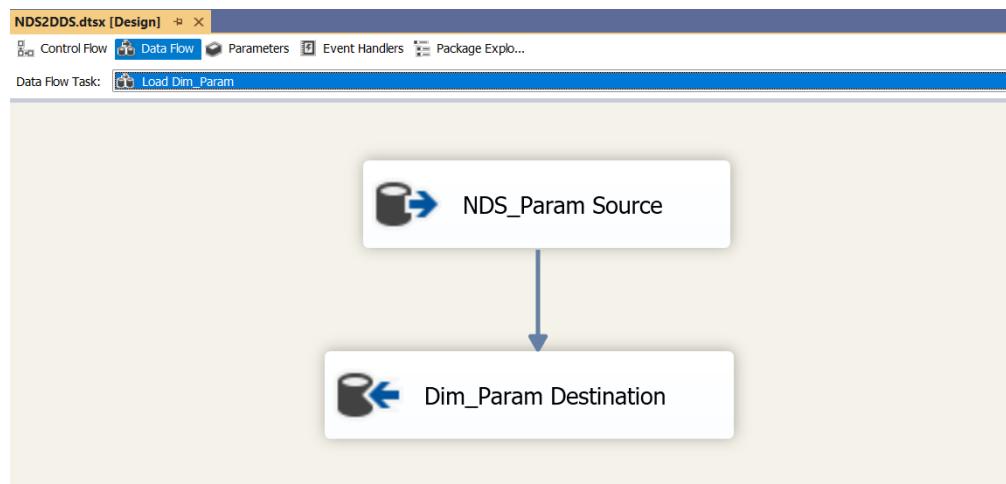
Các bước thực hiện:

1. Rút trích dữ liệu từ NDS sang DDS:

Select * From Site
Where (created > LSET And created <= CET)
Or (last_updated > LSET And last_updated <= CET)

2. Kiểm tra ánh xạ vào bảng DDS.

5.2.5.6. Dim_Parameter



Các bước thực hiện:

- Rút trích dữ liệu từ NDS sang DDS:

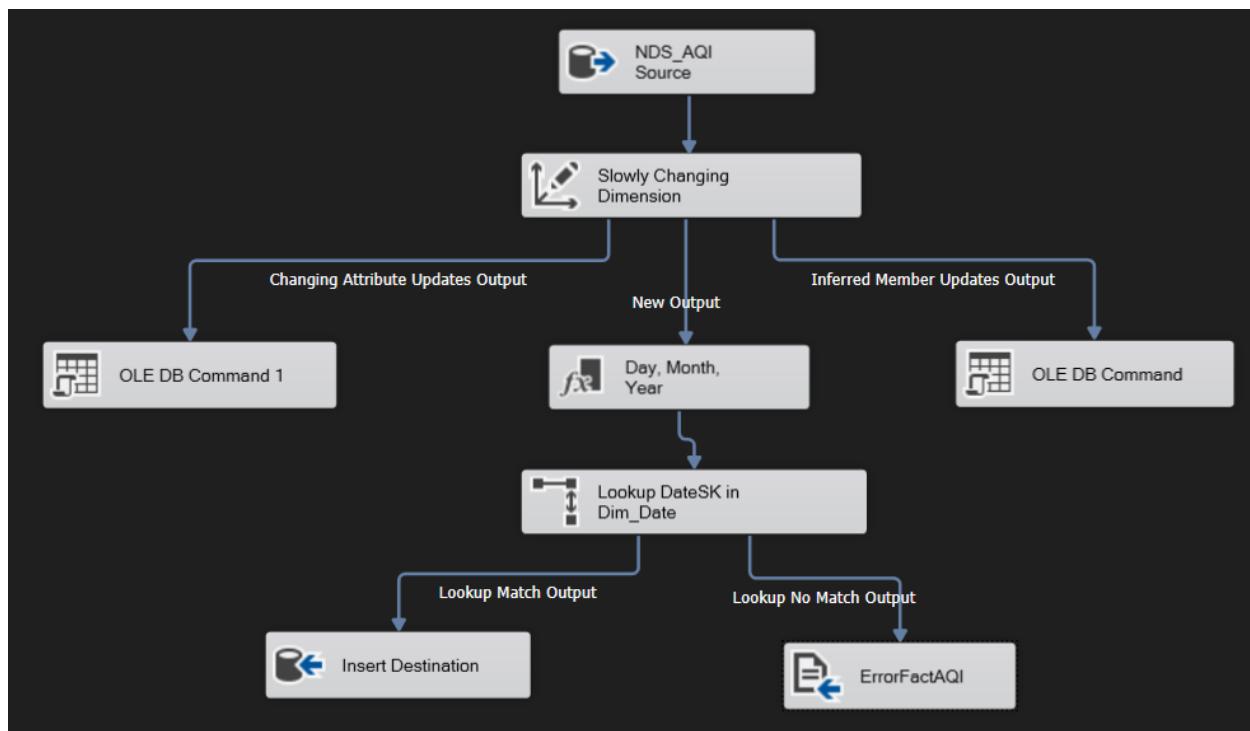
Select * From Param

Where (created > LSET And created <= CET)

Or (last_updated > LSET And last_updated <= CET)

- Kiểm tra ánh xạ vào bảng DDS.

5.2.6. Tạo bảng sự kiện (Fact_AQI)

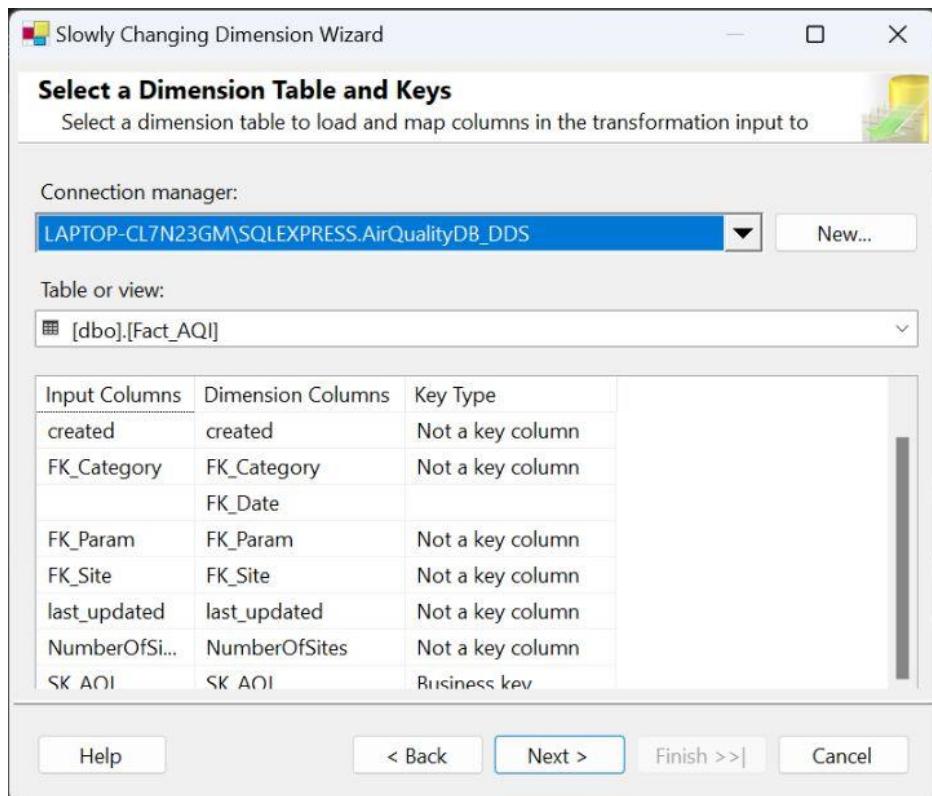


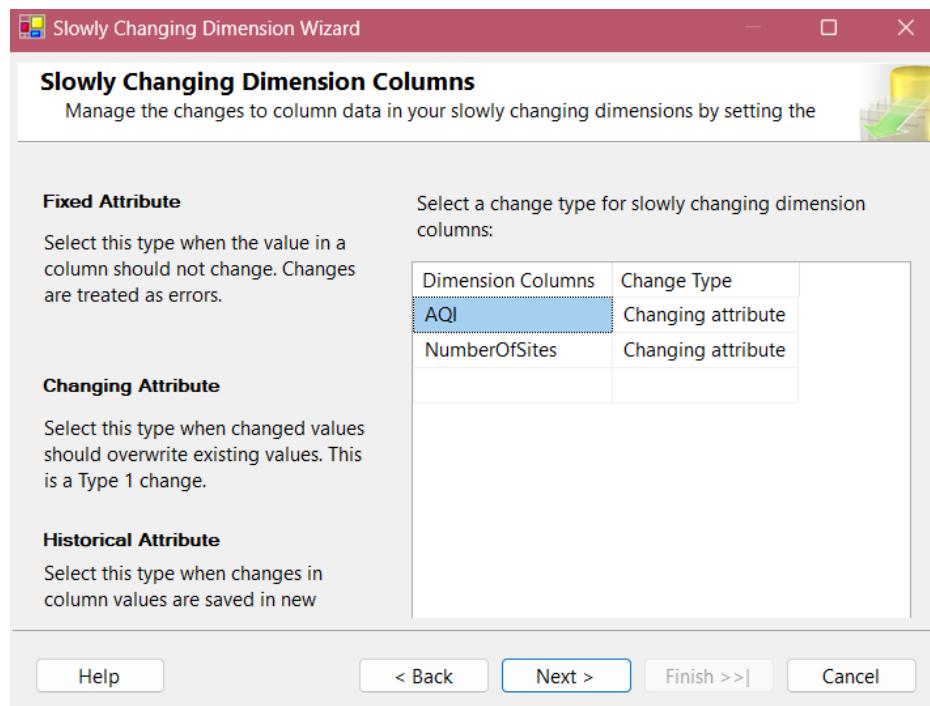
Các bước thực hiện:

- Rút trích dữ liệu từ NDS sang DDS:

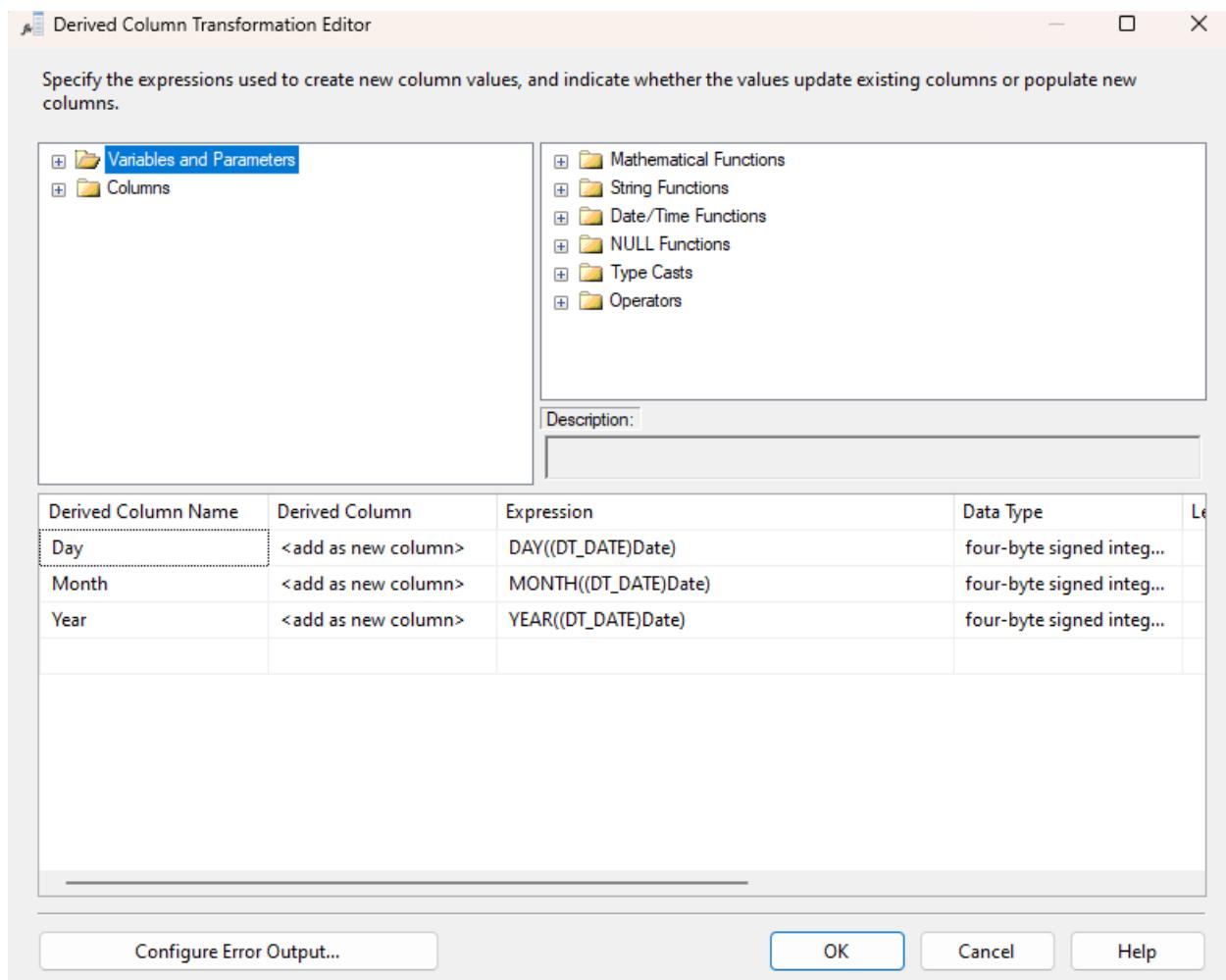
Select * From AQI
Where (created > LSET And created <= CET)
Or (last_updated > LSET And last_updated <= CET)

2. Thiết lập SCD (Slow Changing Dimension), chọn khóa chính SK_AQI làm Business key, với AQI và NumberOfSites là Changing Attributes.

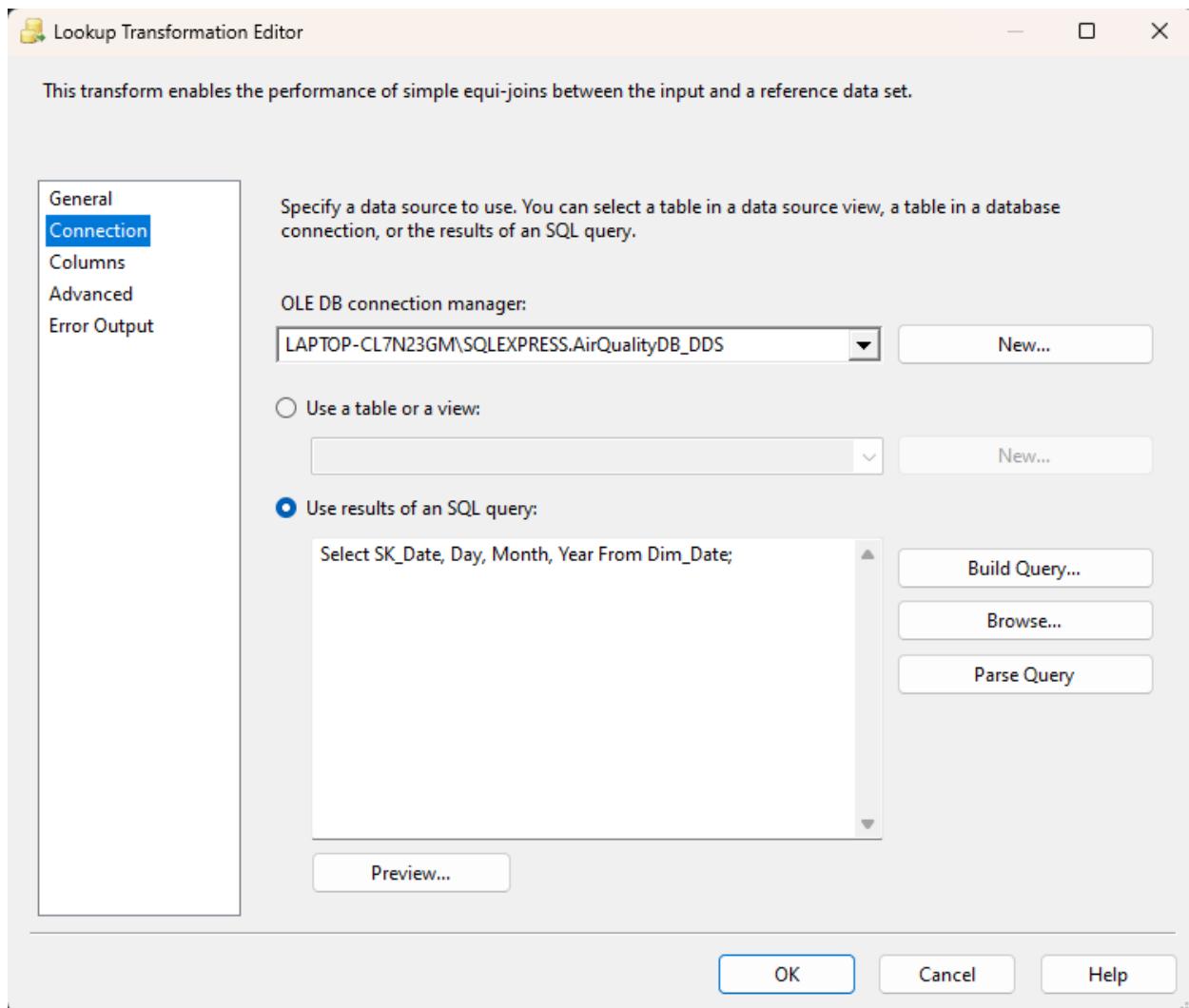




3. Sau khi setup SCD, hệ thống tự động tạo thêm 3 nhánh nhỏ, bên trái là update thuộc tính ghi đè, bên phải là update tất cả các thuộc tính, nhánh giữa là thêm dòng dữ liệu mới nếu đáp ứng được các điều kiện.
4. Ở nhánh giữa, tạo Derived Column với dữ liệu nguồn là từ cột Date để tạo dữ liệu cho các trường Day, Month, Year.



5. Thực hiện Lookup trong bảng Dim_Date để tìm SK_Date với giá trị Day, Month, Year tương ứng, nếu có giá trị SK_Date tương ứng sẽ thực hiện thêm dữ liệu bảng, ngược lại sẽ ghi lại các dữ liệu bị lỗi ở vị trí khác.



6. Kiểm tra ánh xạ vào bảng DDS.

6. OLAP, MDX, and Reporting

6.1. OLAP

6.1.1. Hướng dẫn cài đặt

a) Các công cụ cần thiết

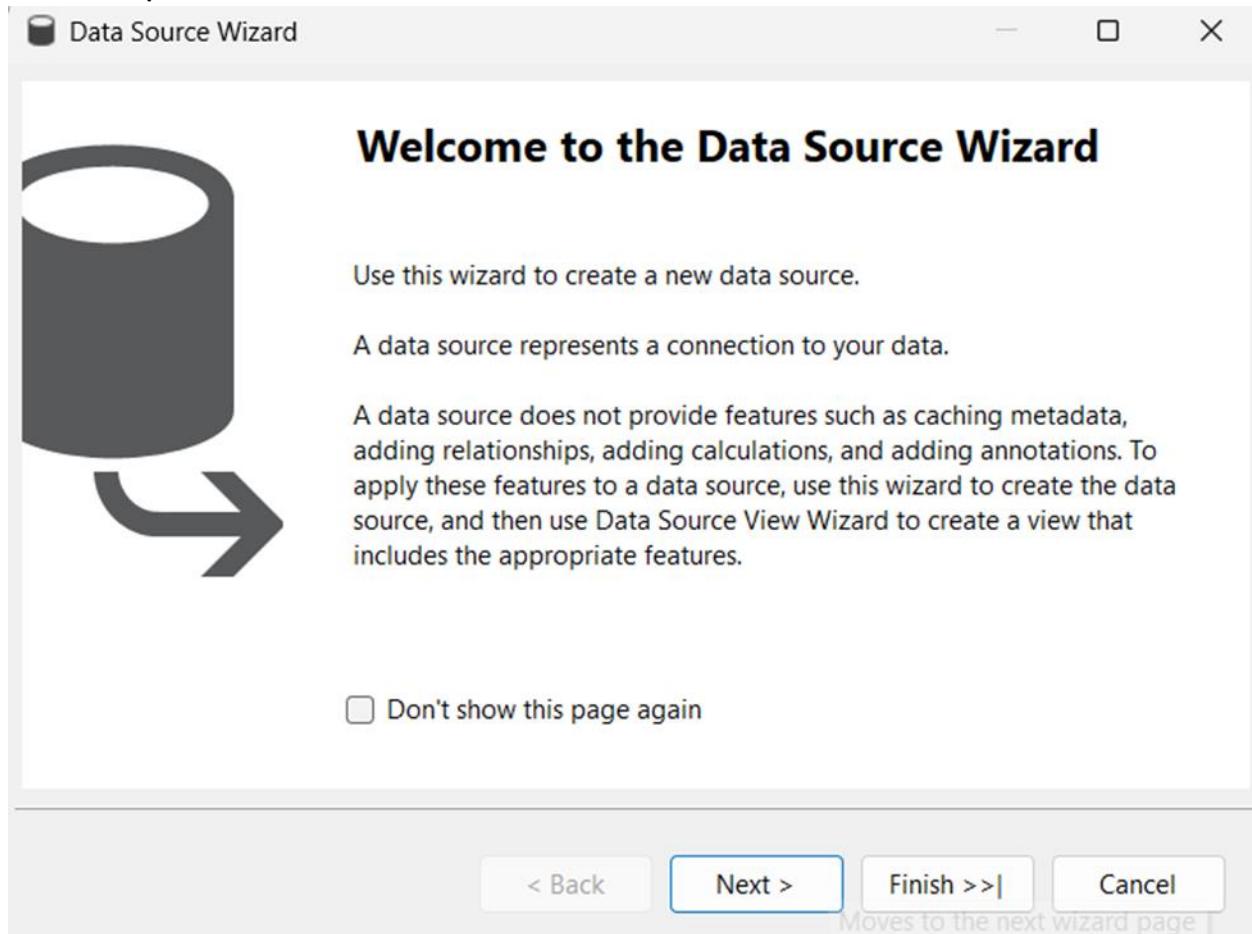
- Microsoft Visual Studio phiên bản 2019.
- SQL Server Data Tools (SSDT): Đây là phần mở rộng dành cho Visual Studio để làm việc với các dự án BI.

b) Các bước cài đặt

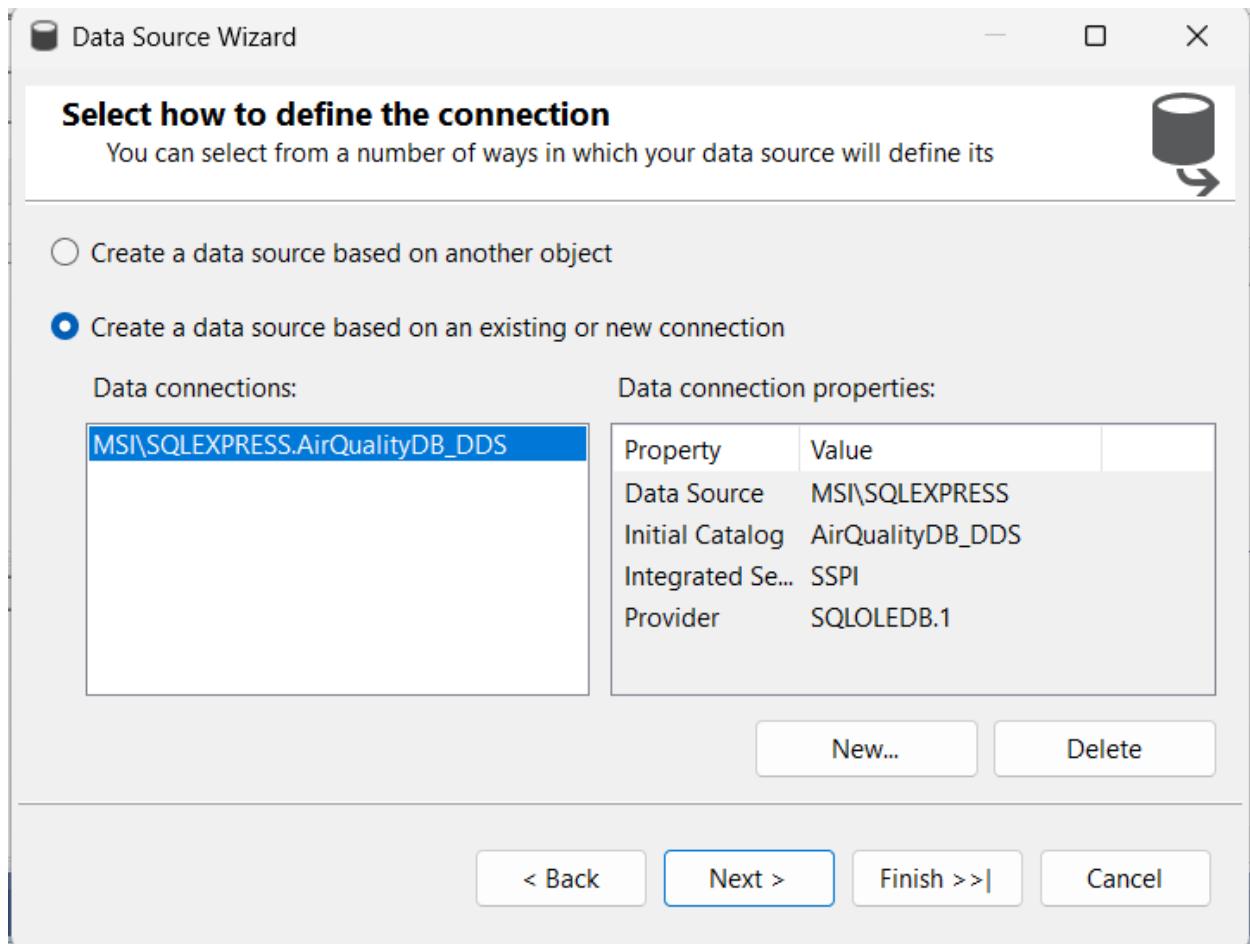
- **Bước 1:** Mở Visual Studio Installer.
- **Bước 2:** Chọn Modify.
- **Bước 3:** Trong Workloads, chọn thêm **Data storage and processing**.
- **Bước 4:** Chọn Modify để cài đặt.

6.1.2. Thực hiện OLAP

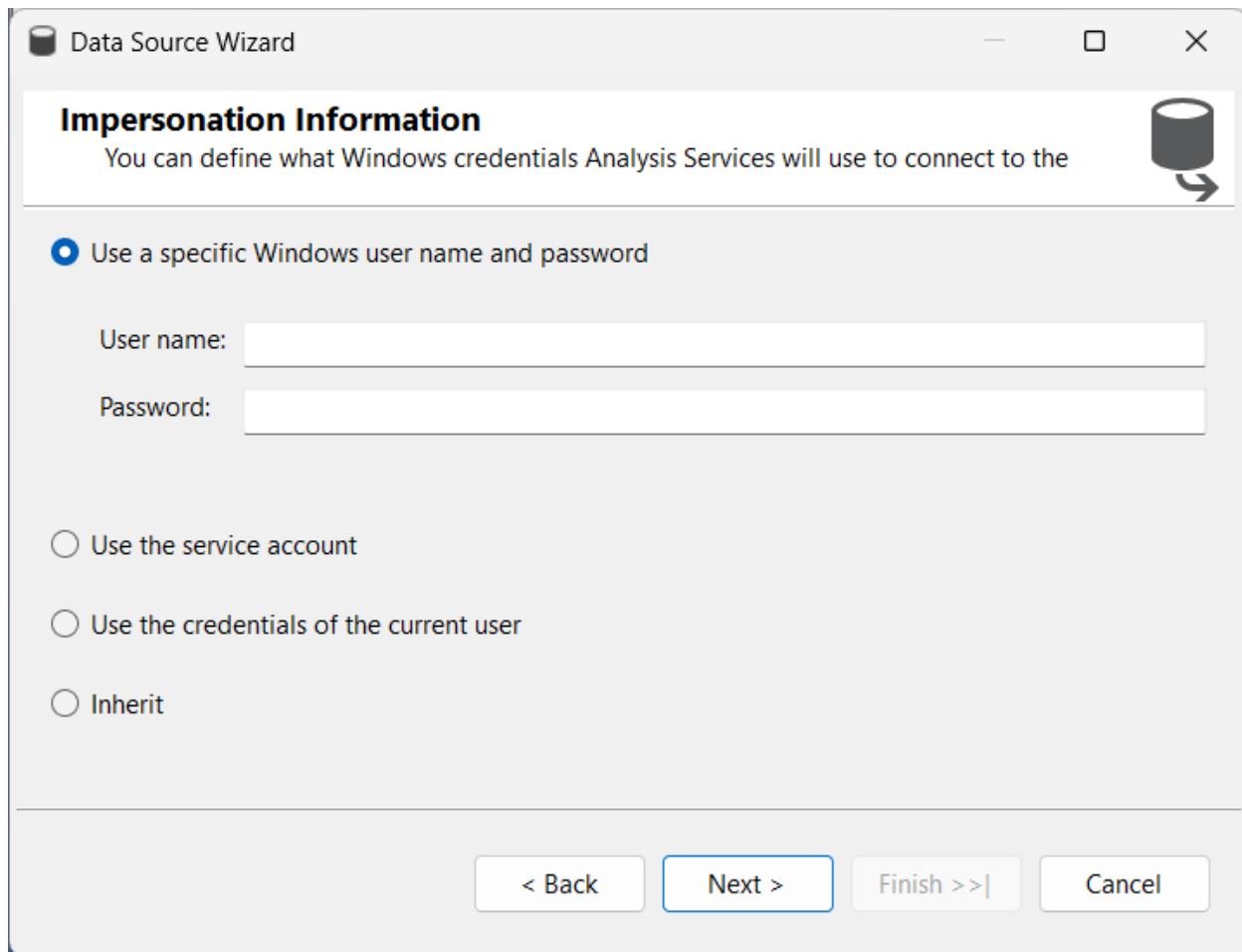
- **Bước 1:** Tạo dự án *Analysis Services Multidimensional and Data Mining Project* mới trong Visual Studio.
- **Bước 2:** Nhấn chuột phải vào **Project Name** chọn **Properties > Deployment**.
- **Bước 3:** Thực hiện cấu hình ở mục target Server và Database tương ứng với Server và Database trong SQL Server Analysis Services.
- **Bước 4:** Kết nối dữ liệu nguồn, tạo Data Source mới
 - Chọn **New Data Source > Next**.



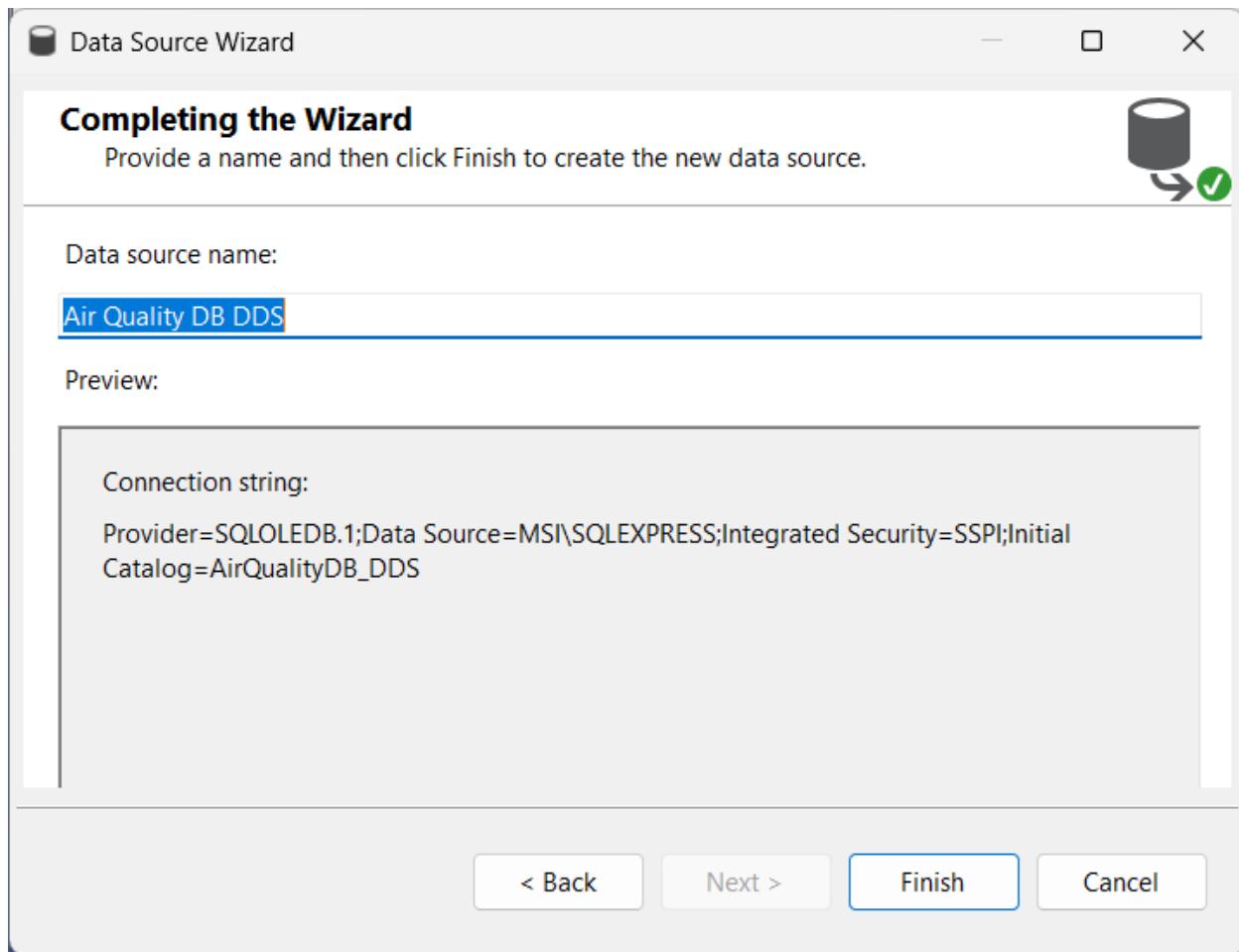
- Chọn ‘Create a data source based on an existing or new connection’ và thực hiện kết nối đến Server trong SQL Server Data Engine tương ứng.



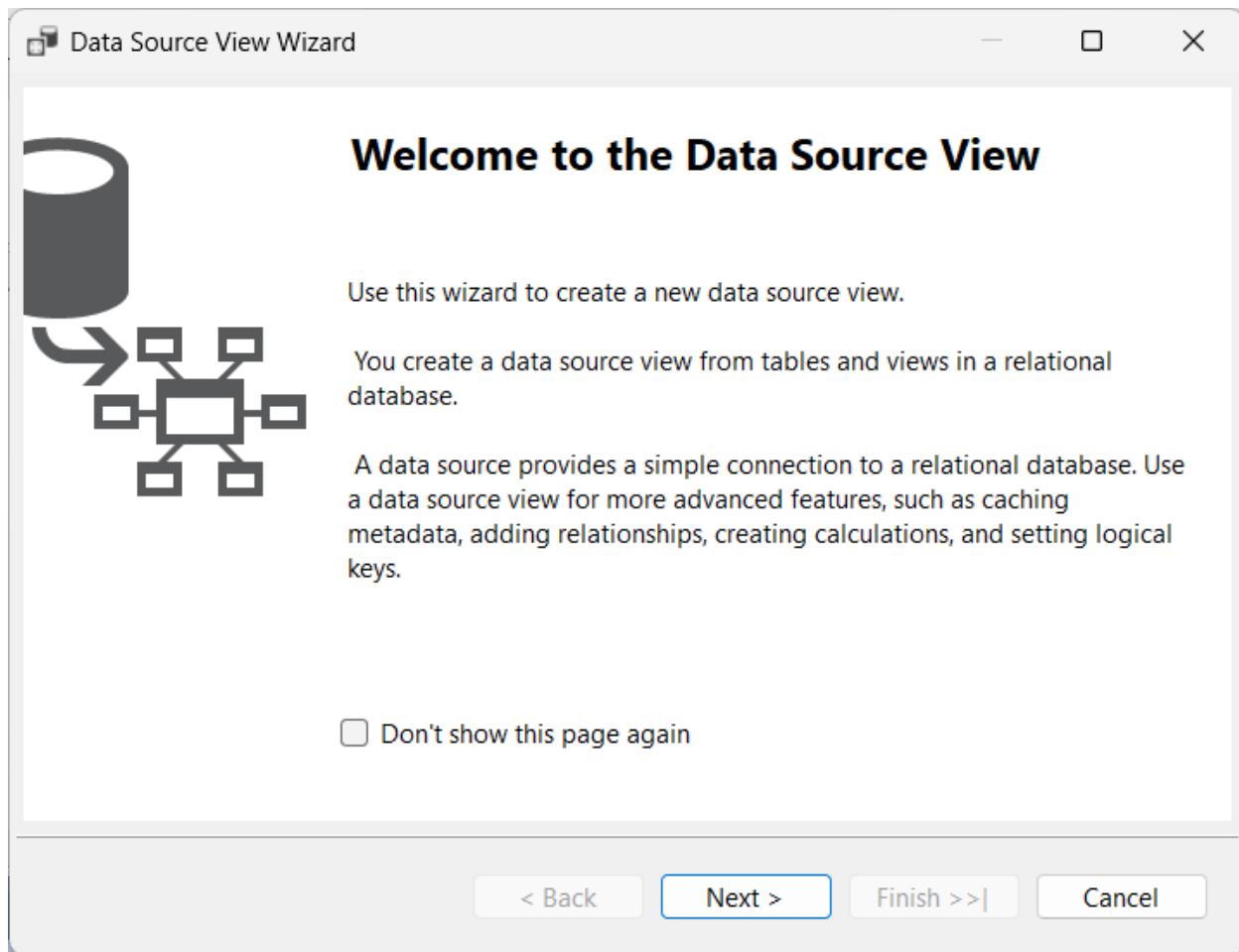
- Nhập thông tin kết nối Windows *username* và *password*.



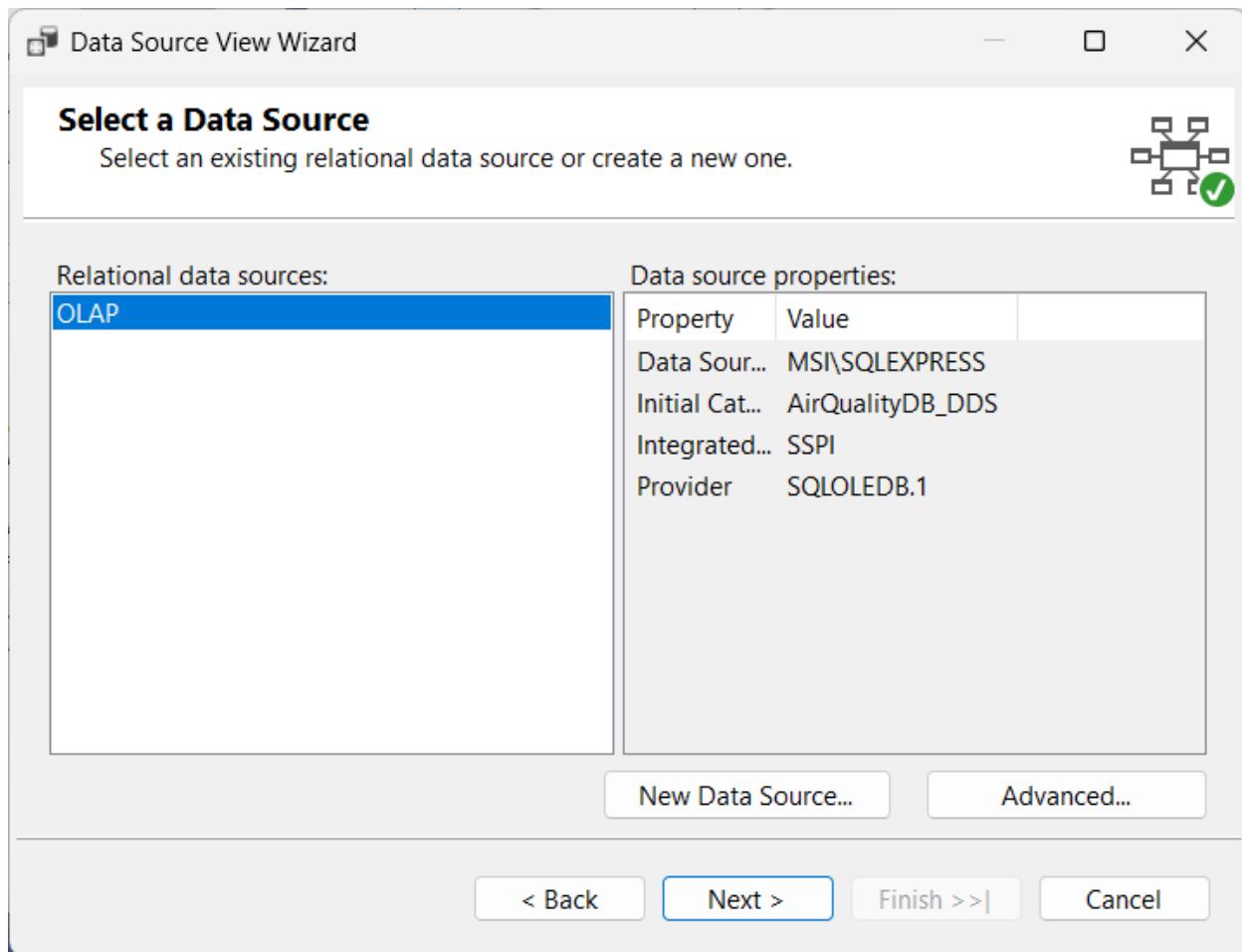
- Đặt tên Data Source và chọn **Finish**.



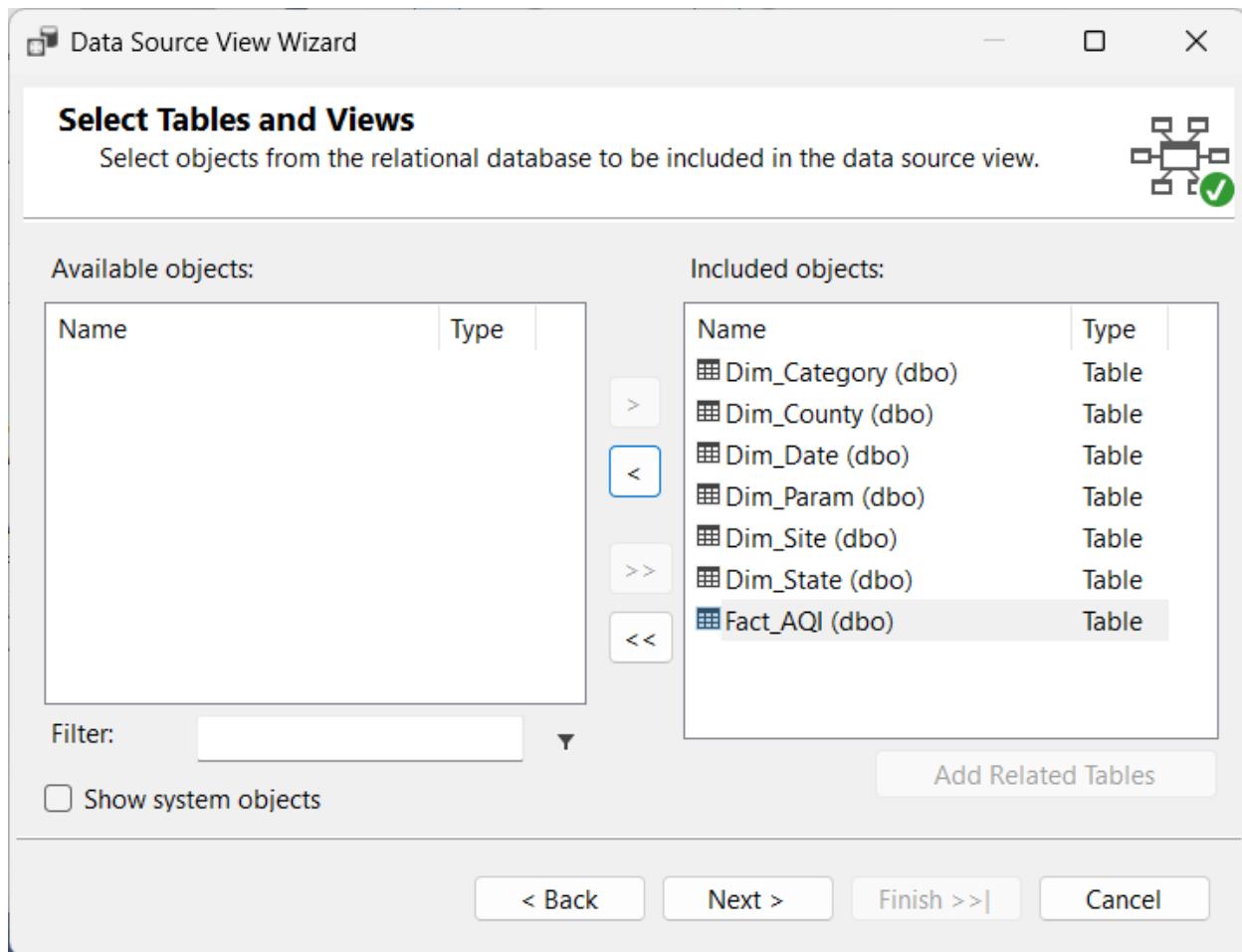
- **Bước 5:** Tạo Data Source View mới.
- Chọn New Data Source View > Next.



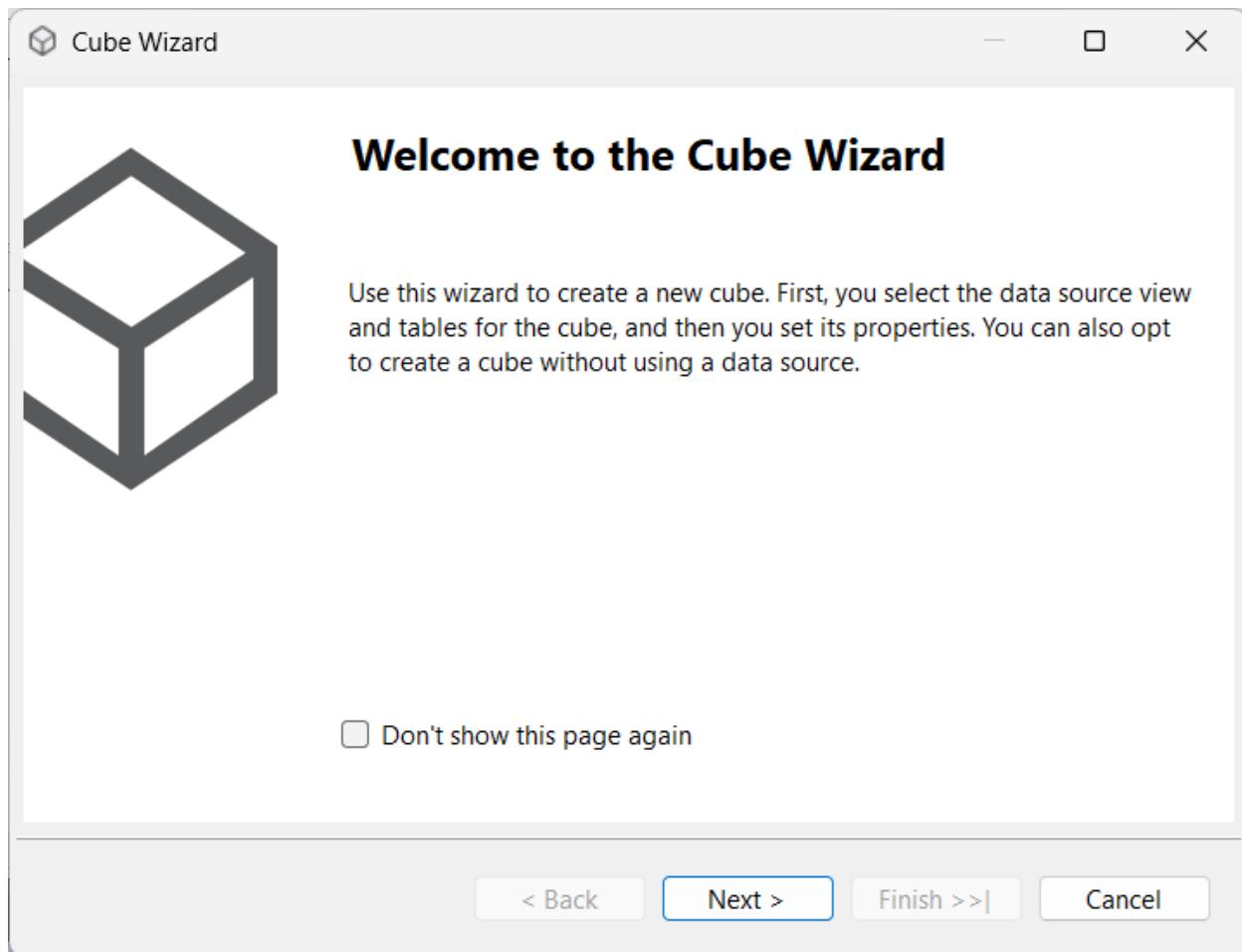
- Chọn Data Source vừa tạo ở **bước 4** và nhấn **Next**.



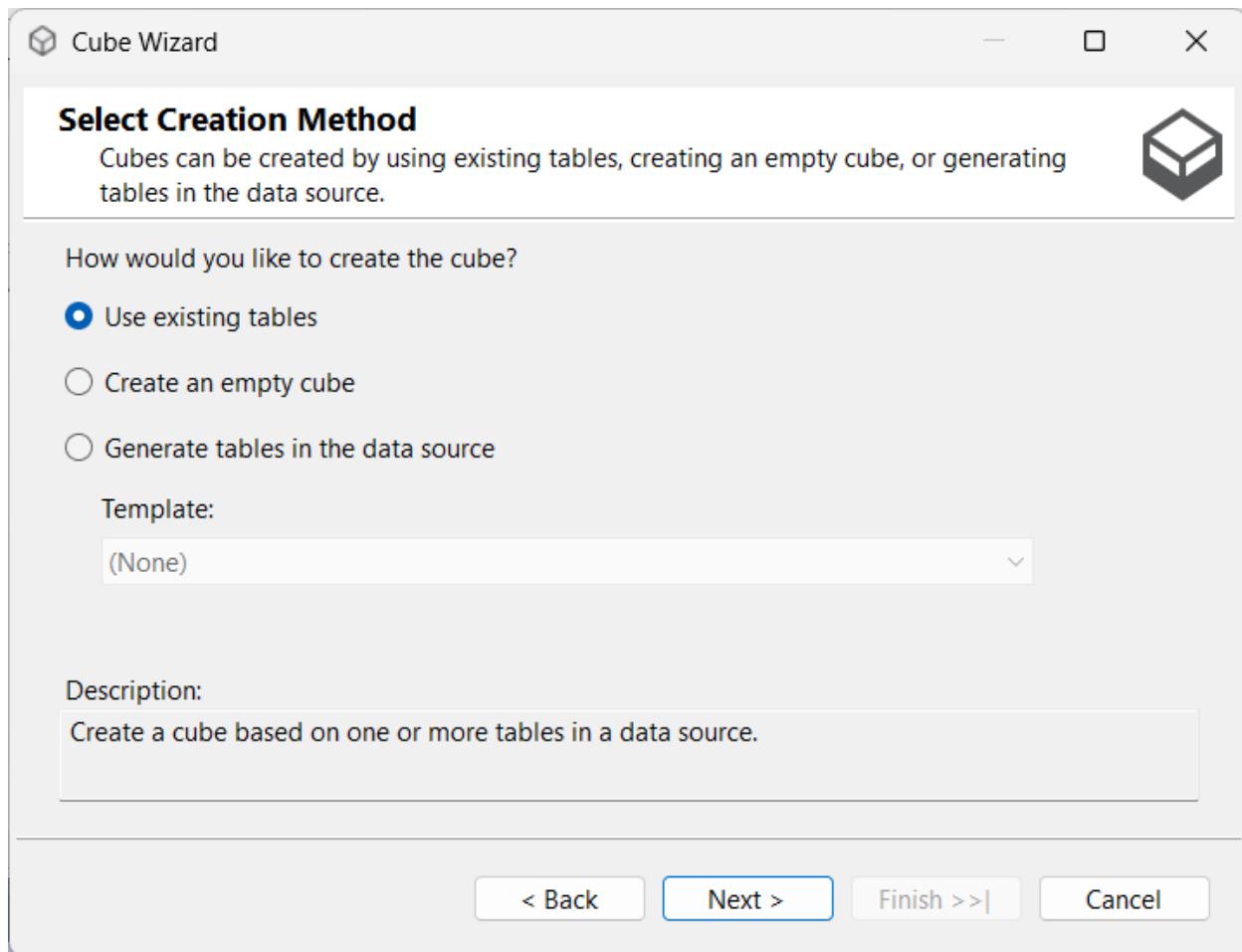
- Thực hiện chọn các bảng để thực hiện OLAP và bấm **Next**.



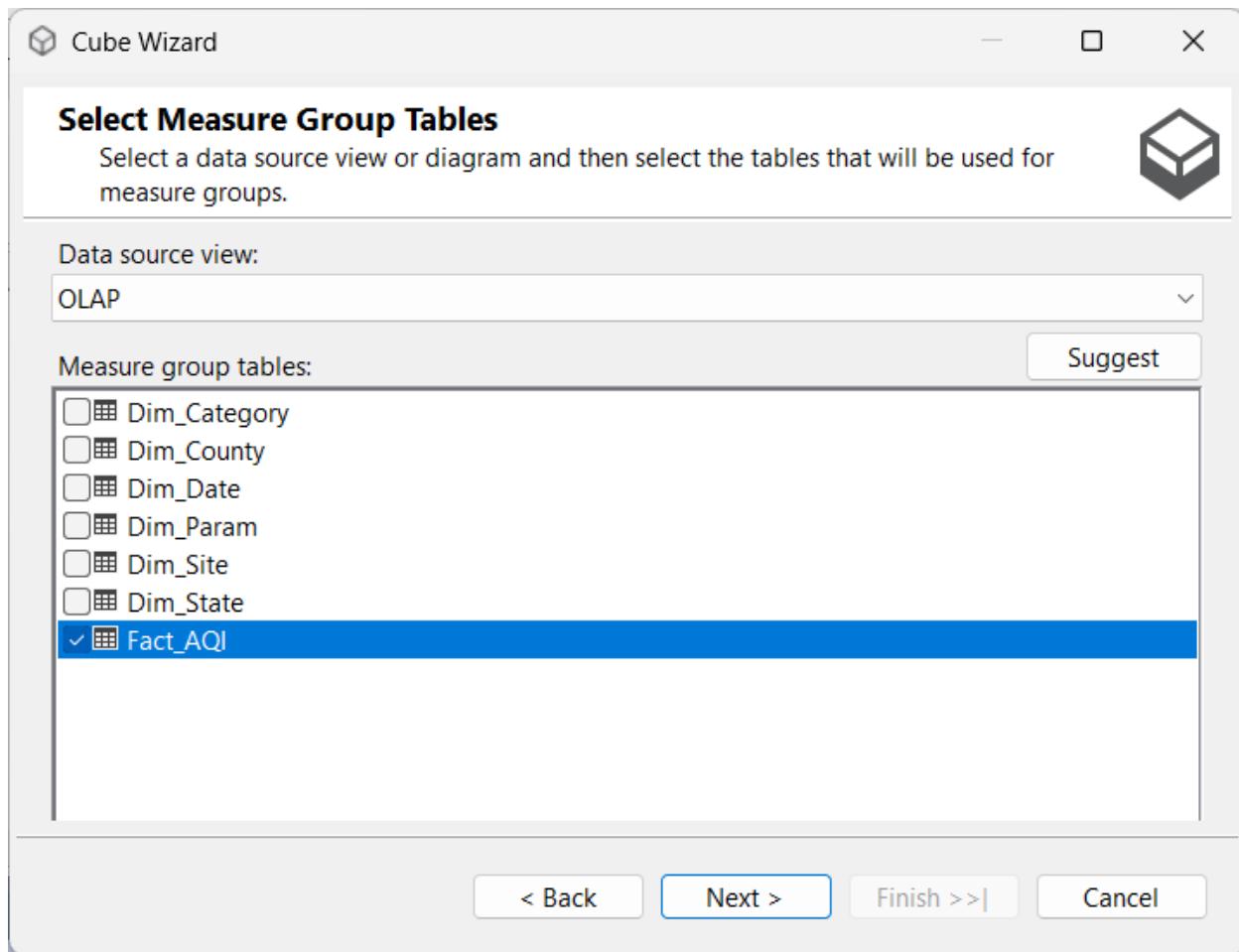
- Đặt tên cho Data Source View và bấm **Finish**.
- **Bước 6:** Tạo Cube mới.
- Chọn **New Cube > Next**.



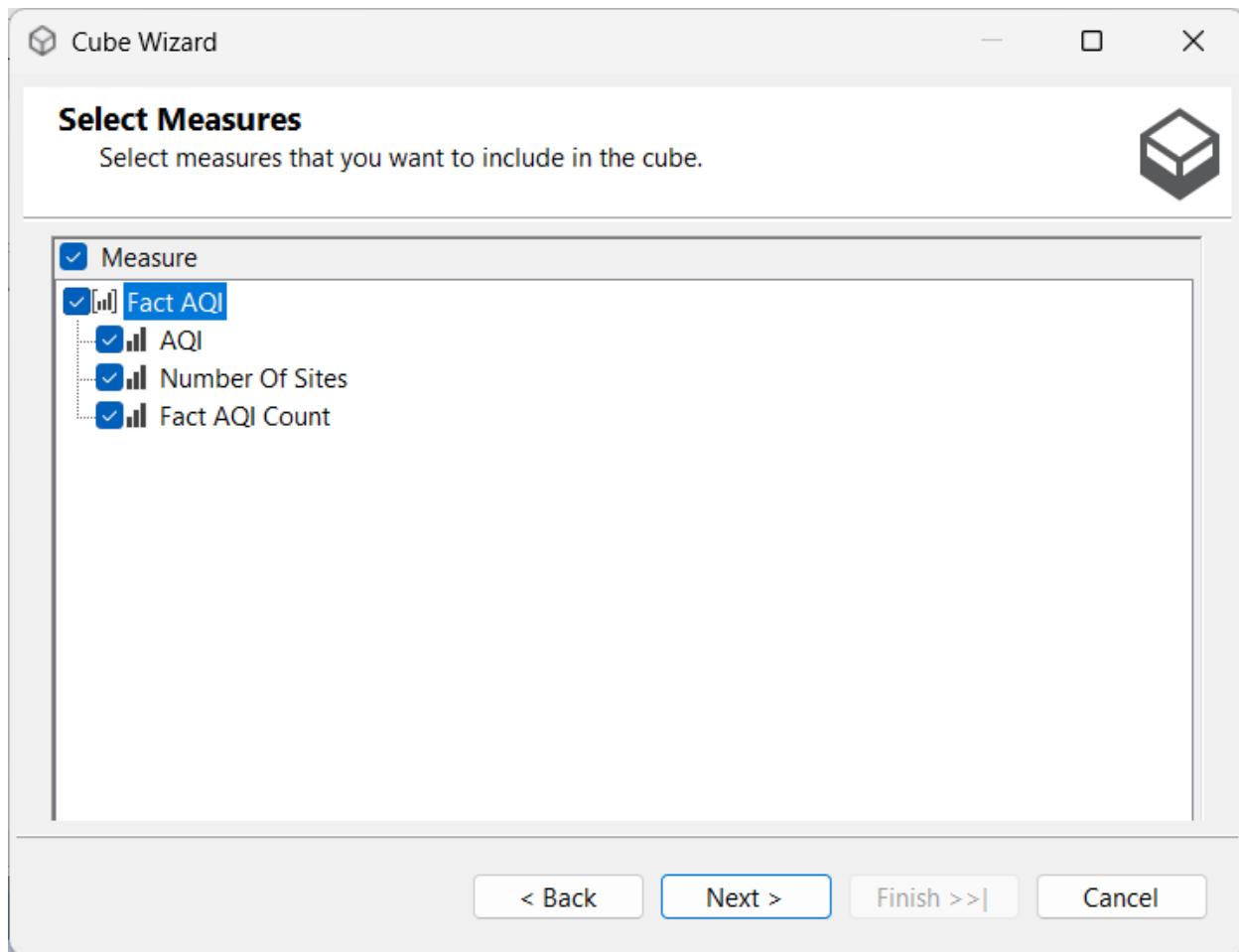
- Chọn 'Use existing tables' và bấm Next.



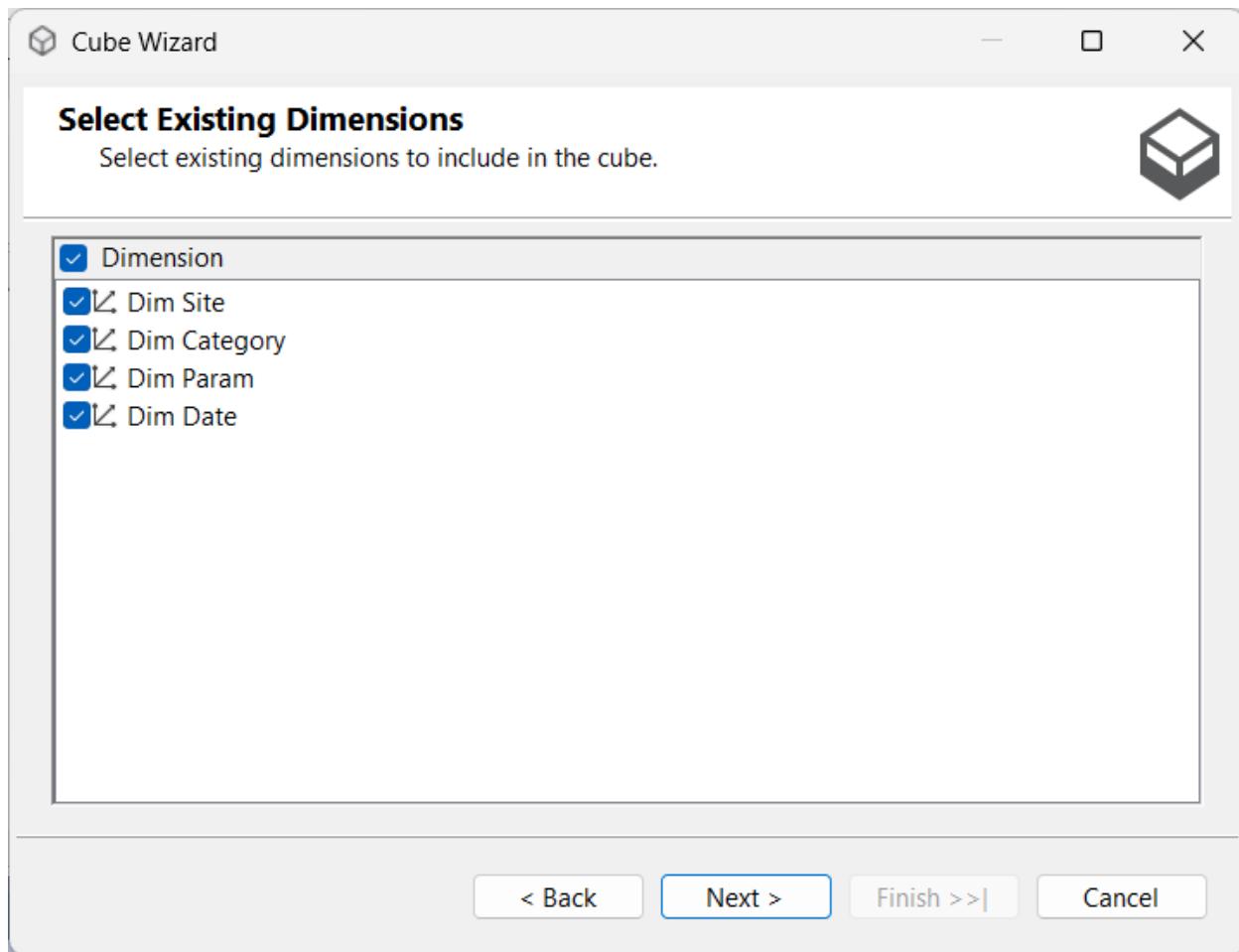
- Chọn bảng chứa các measures là bảng **Fact_AQI** và bấm **Next**.



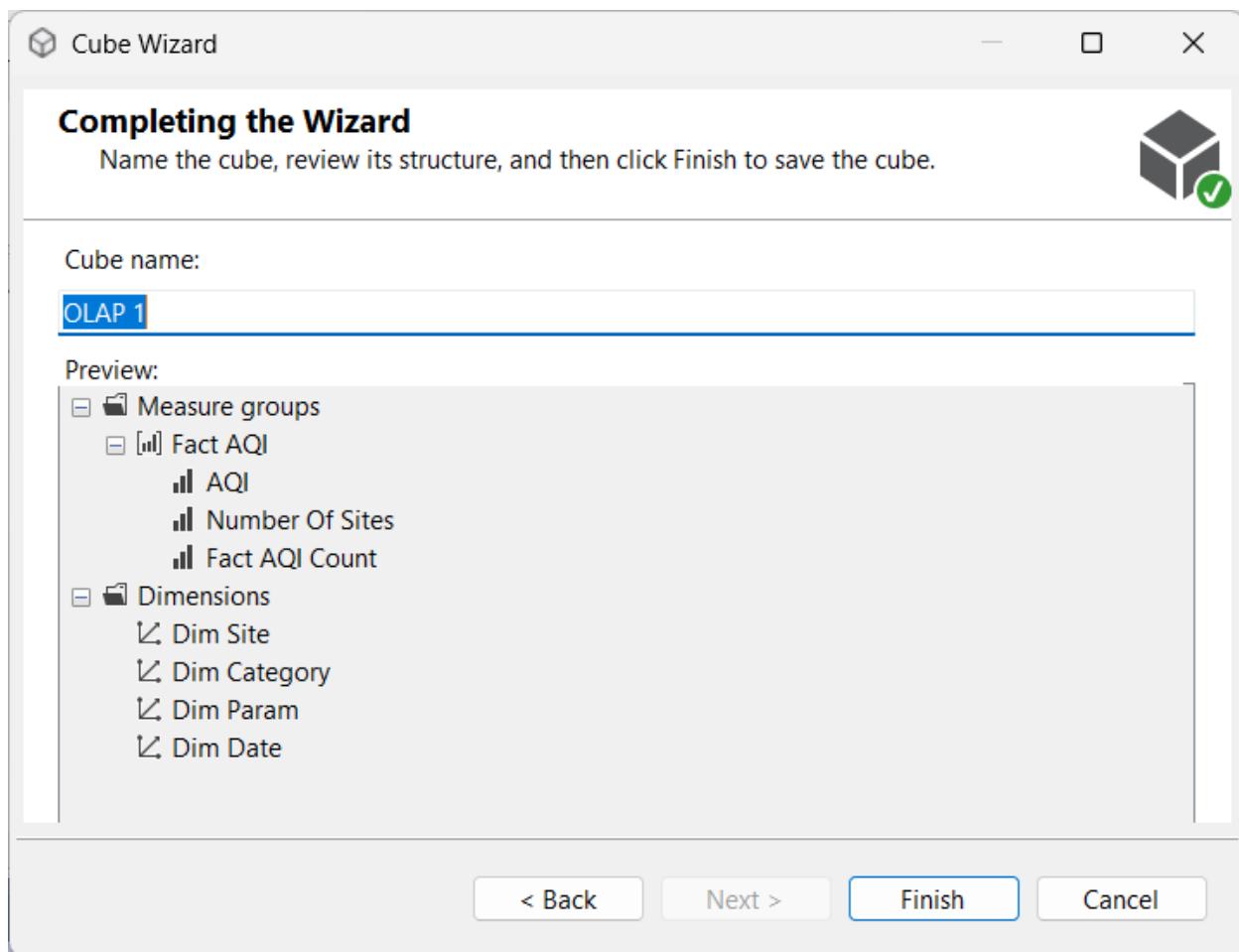
- Thực hiện chọn các **measures** cần sử dụng trong bảng **Fact** và bấm **Next**.



- Thực hiện chọn các **Dimension** và bấm **Next**.



- Đặt tên cho Cube và bấm **Finish**.



- Bước 7:** Thực hiện phân cấp chiều tương ứng cho các Dimension.

SSIS... File Edit View Git Project Build Debug Format Test Analyze Tools Extensions Window Help Search OLAP

Solution Explorer

Dim Date.dim [Design]

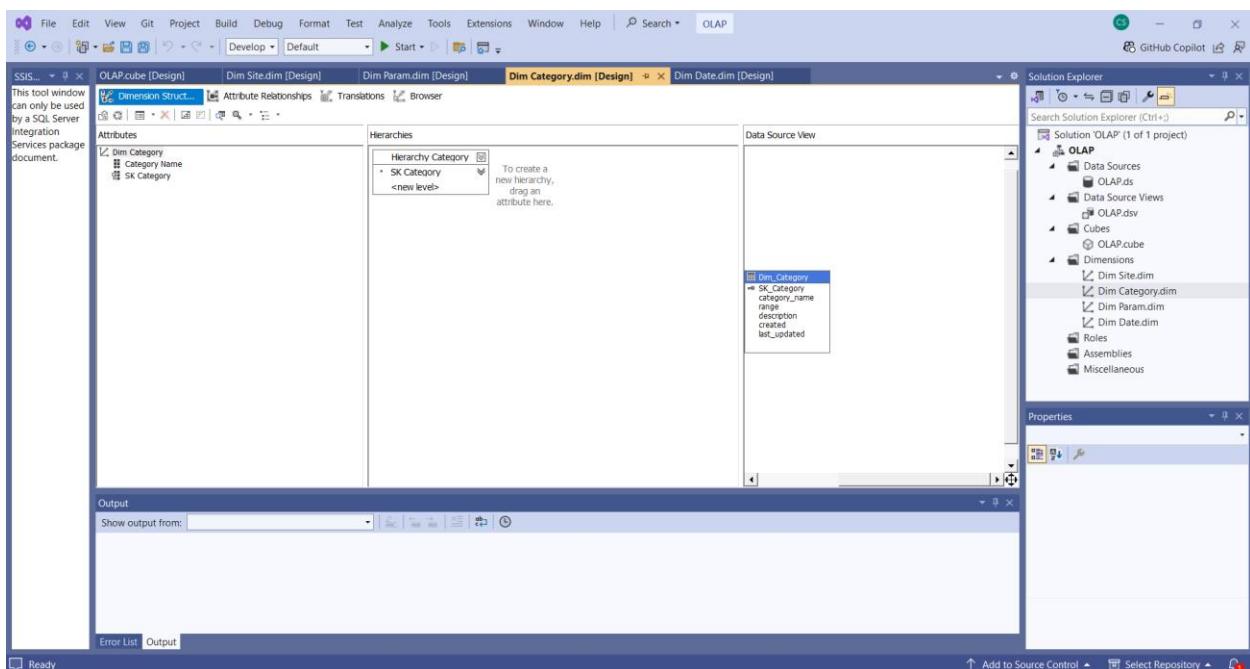
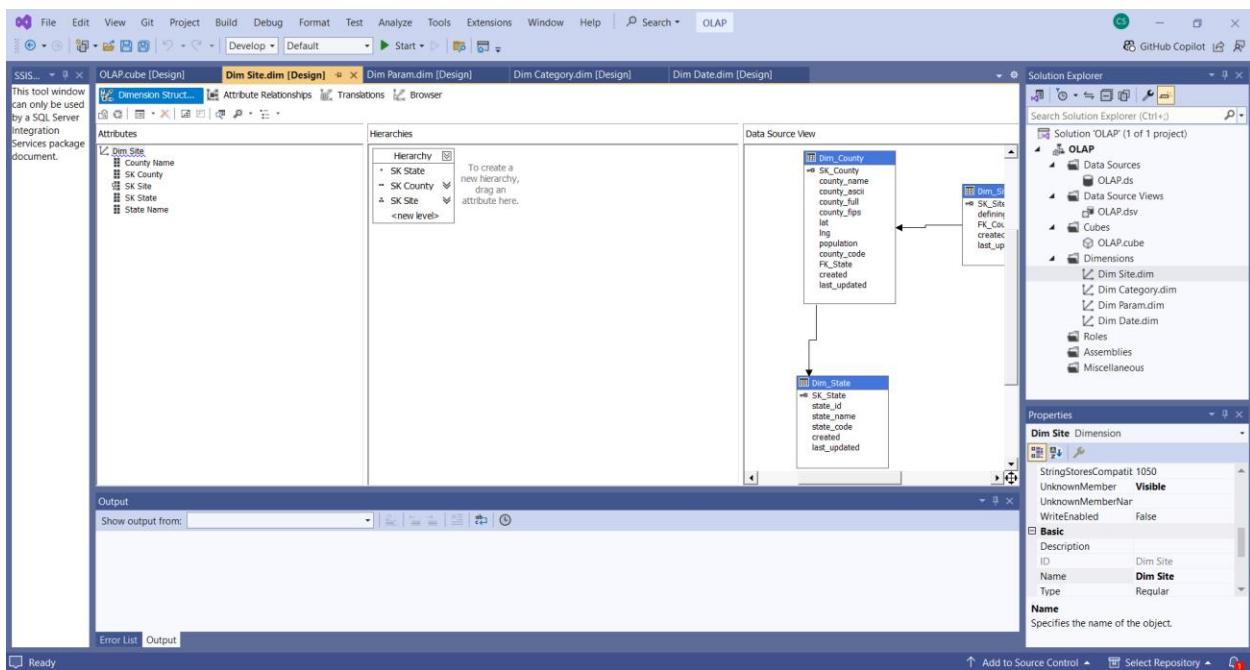
Properties

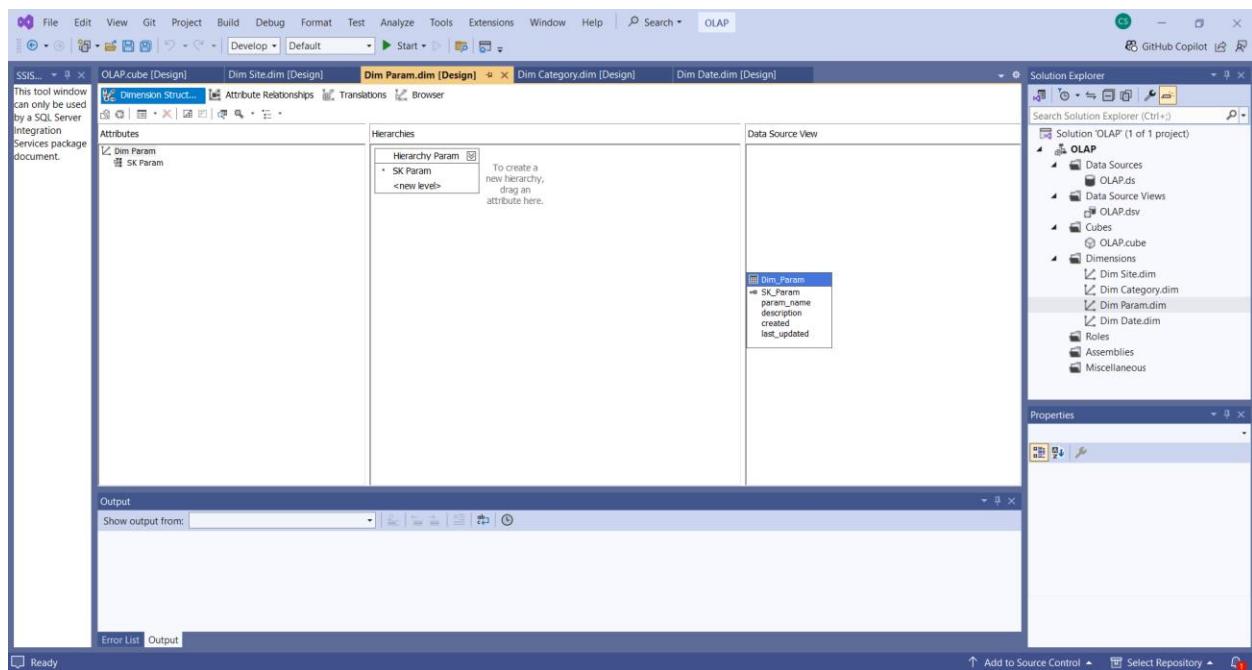
Dim Date Dimension

StringStoreCompatibility 1050
UnknownMemberVisible
UnknownMemberName
WriteEnabled False

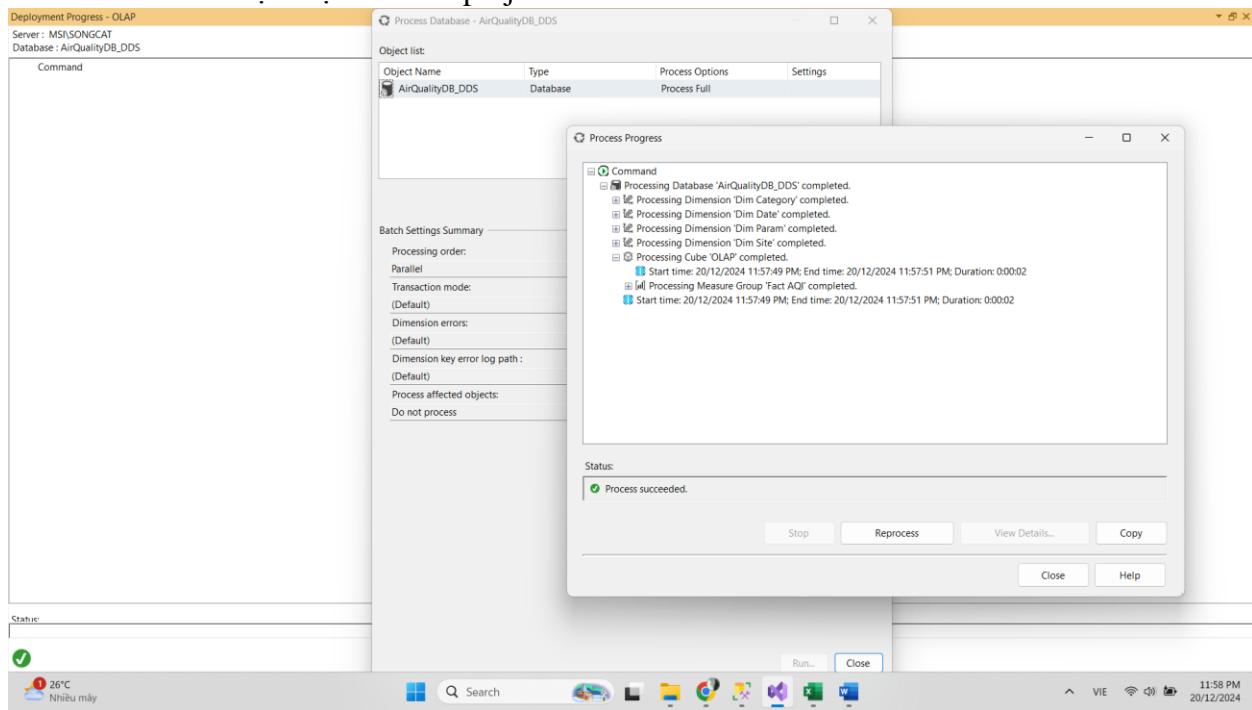
Basic

Description
ID
Name Dim Date
Type Regular
Name





- **Bước 8:** Thực hiện Process project.



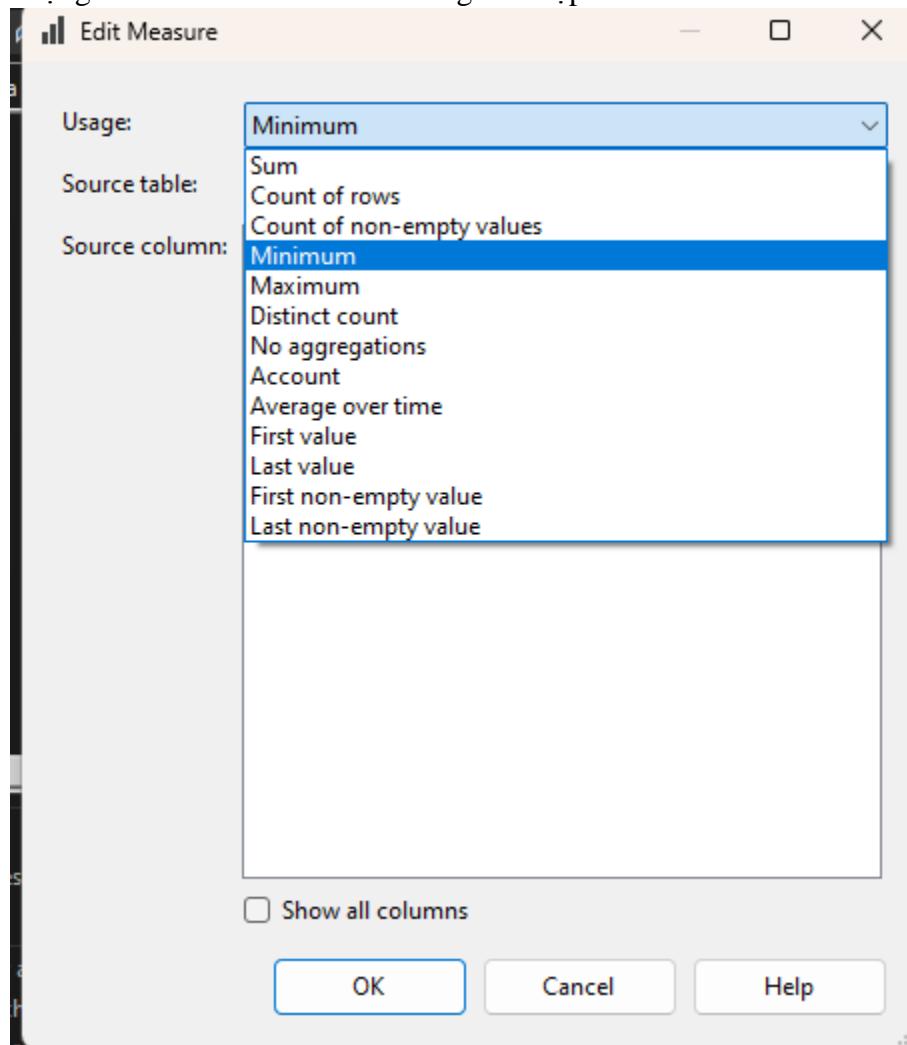
6.2. MDX

6.2.1. Report the min and max of AQI value for each State during each quarter of years.

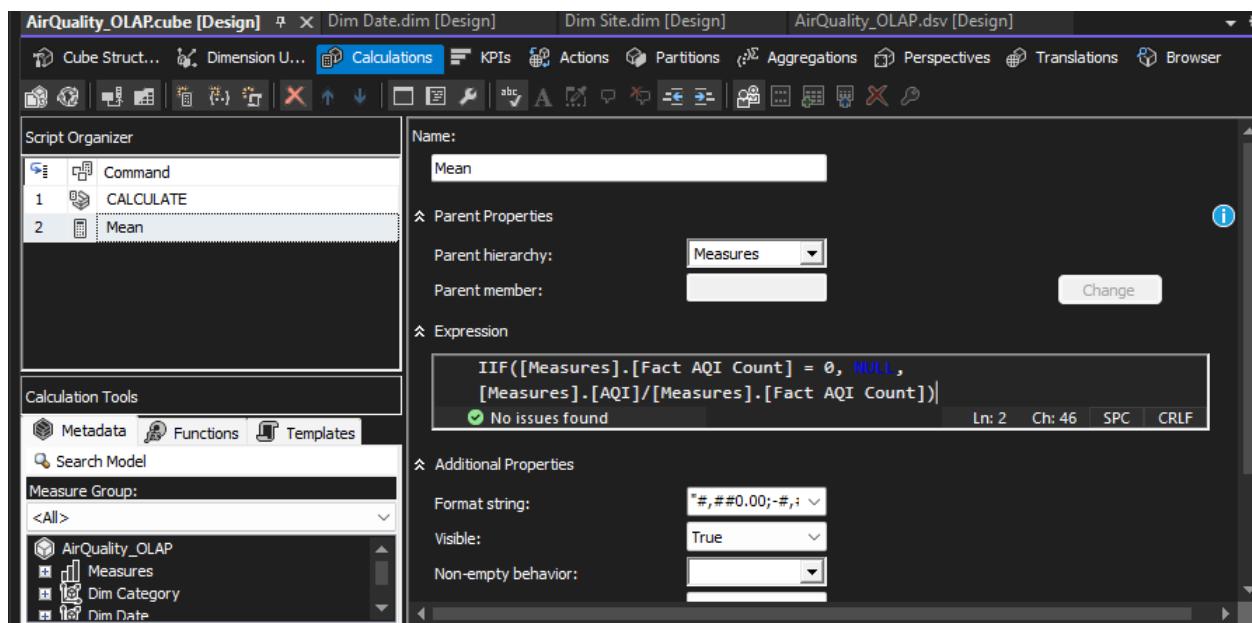
Analysis hints: How do the AQI values fluctuate during the year? Pay attention to the values (max, min). Are any unusually large or small?

- **Các bước thực hiện**

1. Tạo measure mới là Maximum AQI, Minimum AQI trực tiếp vào bảng Fact_AQI bằng cách sử dụng built-in function có sẵn trong thiết lập của SSDT.



2. Giá trị Mean không có thiết lập sẵn nên sẽ cần tạo thủ công ở mục Calculations, nhập công thức để tính Mean.



3. Thực hiện query MDX

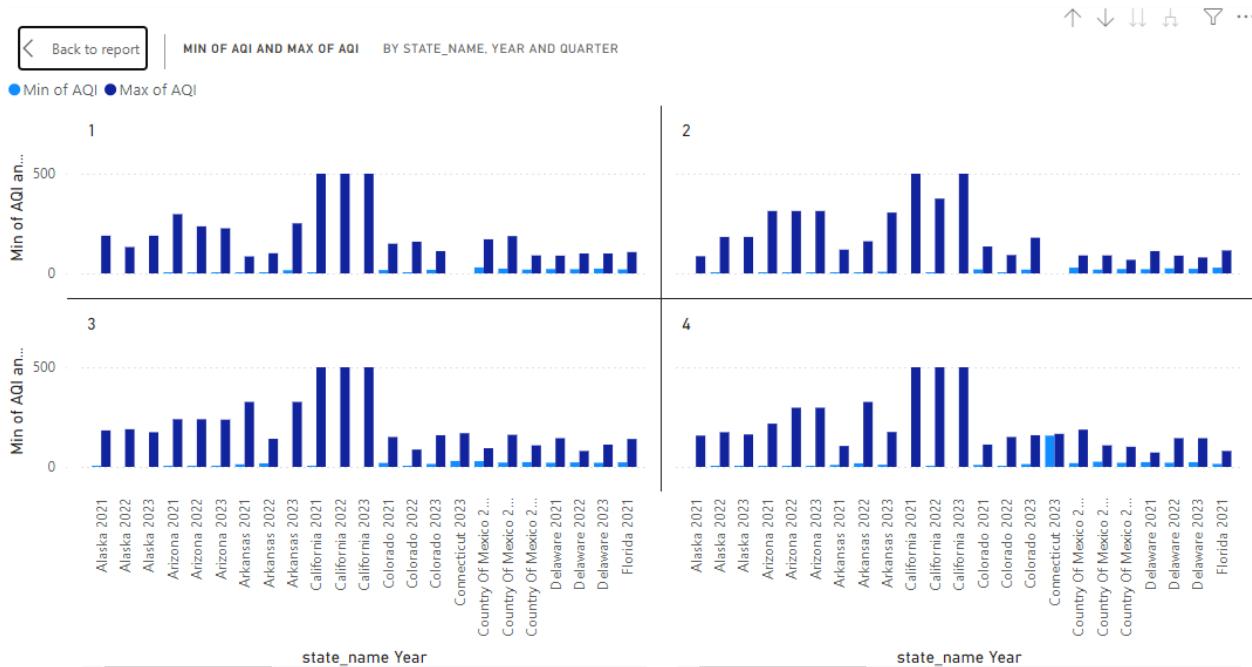
```
//1. Report the min and max of AQI value for each State during each quarter of years.
SELECT
    {[Measures].[Minimum AQI], [Measures].[Maximum AQI]} ON COLUMNS,
    NONEMPTY (
        [Dim Site].[State Name].[State Name] *
        [Dim Date].[Year].[Year] *
        [Dim Date].[Quarter].[Quarter]
    ) ON ROWS
FROM [AirQuality_OLAP];
```

- Kết quả thực thi:

121 %

		Minimum AQI	Maximum AQI
Alabama	2021 1	10	101
Alabama	2021 2	15	174
Alabama	2021 3	18	157
Alabama	2021 4	15	151
Alabama	2022 1	0	159
Alabama	2022 2	2	161
Alabama	2022 3	1	179
Alabama	2022 4	1	166
Alabama	2023 1	0	132
Alabama	2023 2	0	122
Alabama	2023 3	0	151
Alabama	2023 4	0	157
Alaska	2021 1	0	189
Alaska	2021 2	0	86
Alaska	2021 3	1	183
Alaska	2021 4	0	157
Alaska	2022 1	0	132
Alaska	2022 2	1	183
Alaska	2022 3	0	189
Alaska	2022 4	1	174

- Biểu đồ trực quan:



➔ Phân tích:

- Giá trị **Max của AQI** thường dao động ở mức cao hơn hẳn so với giá trị Min, cho thấy những thời điểm ô nhiễm cao trong năm.
- Một số bang như **California** và **Arkansas** có sự biến động mạnh (max AQI cao hơn nhiều so với các bang khác), đặc biệt trong các quý cuối năm.
- Giá trị **Min của AQI** ổn định hơn, thường không có nhiều thay đổi lớn giữa các bang.

➔ **Nhận xét:**

- California có sự biến động bất thường, thể hiện sự xuất hiện của các sự kiện ô nhiễm đặc biệt (cháy rừng hoặc công nghiệp).
- Những giá trị Max bất thường là yếu tố cần chú ý trong phân tích và quản lý chất lượng không khí.

6.2.2. Report the mean and the standard deviation of AQI value for each State during each quarter of years. *Analysis hints:* How do the AQI values fluctuate during the year? Pay attention to the values (mean, std, max, min). Are any unusually large or small?

• **Các bước thực hiện**

1. Mean là giá trị Measure đã có sẵn, Standard Deviation là độ lệch chuẩn có công thức để tính toán, mặc dù không nằm trong những lựa chọn có trong thiết lập sẵn của Measure trong SSDT, nhưng MDX có hỗ trợ hàm tính giá trị này.
2. Dùng hàm Stdev trong MDX để tính độ lệch chuẩn.
3. Thực hiện query MDX

```
//2. Report the mean and the standard deviation of AQI value for each
//State during each quarter of years.

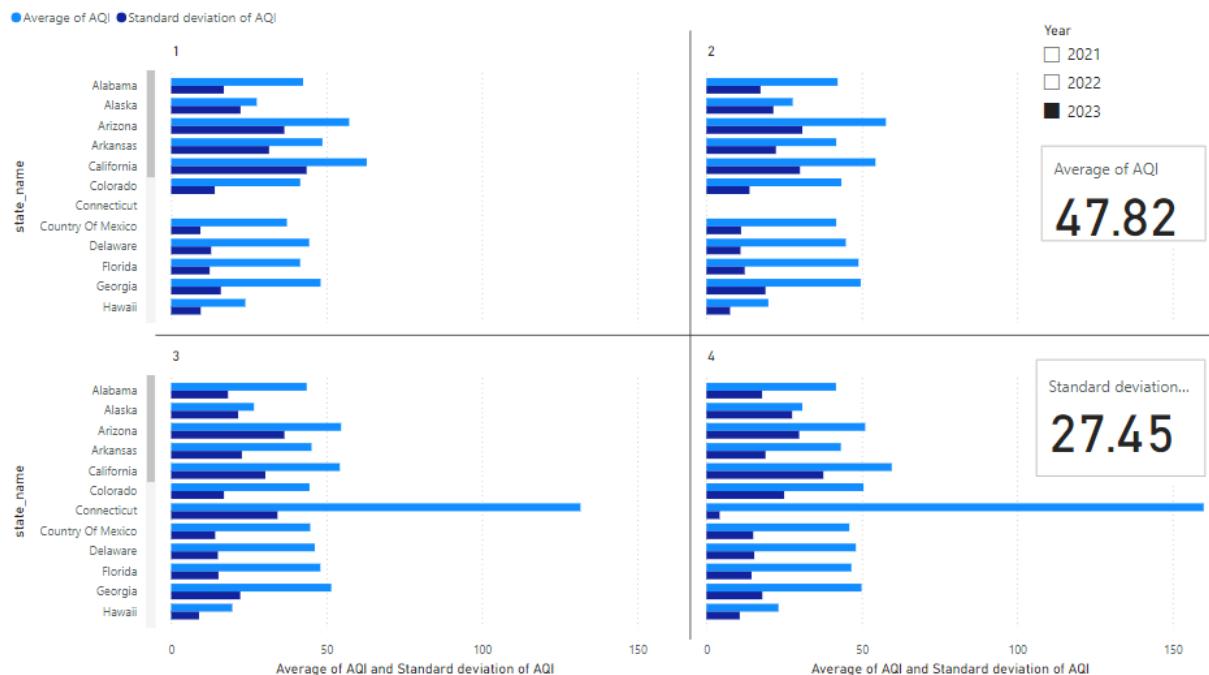
WITH
MEMBER [Measures].[Std] AS
  Stdev
  ([Dim Date].[Hierarchy Date].CurrentMember.Children,
  [Measures].[Mean])
-- Query Results
SELECT
  {[Measures].[Mean], [Measures].[Std]} ON COLUMNS,
  NON EMPTY
  [Dim Site].[State Name].[State Name] *
  FILTER(
    [Dim Date].[Year].MEMBERS,
    [Dim Date].[Year].CurrentMember.Name <> "All"
    AND [Dim Date].[Year].CurrentMember.Name <> "Unknown"
  ) *
  [Dim Date].[Hierarchy Date].[Quarter].MEMBERS
  ON ROWS
FROM [AirQuality_OLAP];
```

- Kết quả thực thi

		Messages	Results	Mean	Std
Florida	2022	3	38.57	3.4690060350535	
Florida	2022	4	40.18	3.77033749803848	
Florida	2023	1	41.51	0.246726974804844	
Florida	2023	2	48.93	2.53953007683281	
Florida	2023	3	48.01	3.61823229024274	
Florida	2023	4	46.64	1.8824027436798	
Georgia	2021	1	51.45	2.65154642037132	
Georgia	2021	2	54.37	5.27262137408719	
Georgia	2021	3	58.63	1.94357198745345	
Georgia	2021	4	50.47	1.32522942574232	
Georgia	2022	1	41.50	3.90477169294502	
Georgia	2022	2	43.98	1.16636039693087	
Georgia	2022	3	46.24	4.57777940289923	
Georgia	2022	4	38.73	7.34092409677133	
Georgia	2023	1	48.09	2.00225654461413	
Georgia	2023	2	49.64	5.83219016255383	
Georgia	2023	3	51.51	4.78796435335192	
Georgia	2023	4	49.92	4.76934712895411	
Hawaii	2021	1	28.39	0.486757741800129	
Hawaii	2021	2	22.12	3.10881183824186	
Hawaii	2021	3	20.76	0.859285578152716	

- Biểu đồ trực quan

Average of AQI and Standard deviation of AQI by state_name and Quarter



➔ Phân tích:

- Các bang như California và Arizona có độ lệch chuẩn (Standard Deviation) cao, điều này cho thấy sự biến động lớn của AQI trong năm.

- *Mean AQI* ở các bang thường ổn định ở mức trung bình, nhưng một số bang như Connecticut có giá trị *Mean AQI* cao hơn rõ rệt so với mức trung bình toàn quốc.

➔ **Nhận xét:**

- Những bang có độ lệch chuẩn cao (California, Arizona) cần được quan tâm đặc biệt do mức độ biến động về chất lượng không khí có thể gây ảnh hưởng tiêu cực đến sức khỏe.
- Mức trung bình cao của AQI là dấu hiệu của các hoạt động kinh tế hoặc khí hậu đặc biệt làm gia tăng ô nhiễm.

6.2.3. Report the number of days, and the mean AQI value where the air quality is rated as "very unhealthy" or worse for each State and County. Analysis hint: What is the AQI limit above which air quality is "very unhealthy" or worse?

• **Các bước thực hiện**

1. Mean là giá trị đã được tính toán, thuộc tính ‘Number of Days’ không sẵn có nhưng có thể sử dụng Fact AQI Count để suy ra. Ý nghĩa của thuộc tính Fact AQI count là đếm số dòng có trong bảng Fact dựa theo điều kiện, mà một dòng trong bảng Fact là thông tin về AQI của một trạm số liệu vào một ngày cụ thể, nên có thể đếm số dòng này như là số ngày ứng với các yêu cầu cần thiết.
2. Theo như bảng phân loại được cung cấp, dưới cấp ‘Very Unhealthy’ chỉ có ‘Harzardous’ và 2 thuộc tính này trong bảng Dim_Category có ID tương đương là 2 và 6 nên thực hiện tìm các dòng trong bảng Fact có FK_Category là 2 hoặc 6.
3. Thực hiện query MDX

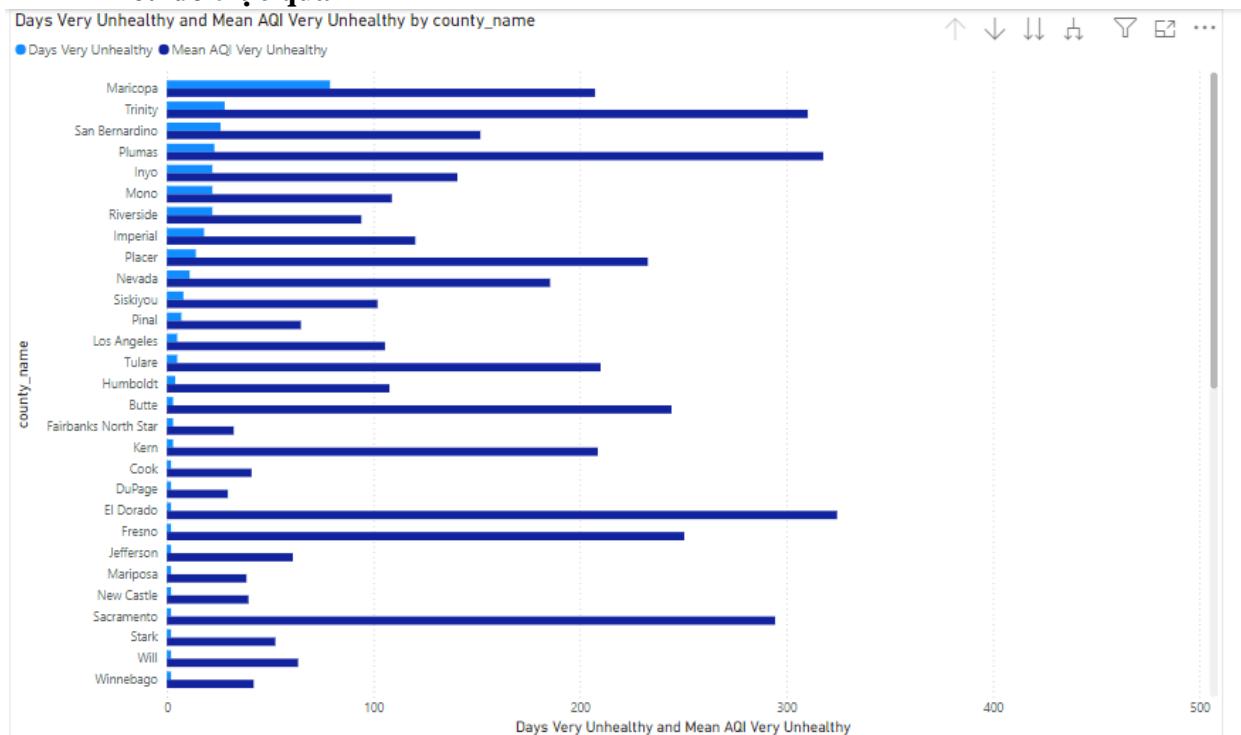
```
//3. Report the number of days, and the mean AQI value
//where the air quality is rated as "very unhealthy" or worse for each State and County.
SELECT
{[Measures].[Fact AQI Count], [Measures].[Mean]} ON COLUMNS,
NONEMPTYCROSSJOIN (
    [Dim Site].[State Name].[State Name],
    [Dim Site].[County Name].[County Name]
) ON ROWS
FROM [AirQuality_OLAP]
WHERE {[Dim Category].[SK Category].&[2], [Dim Category].[SK Category].&[6]};
```

- Kết quả thực thi

121 %

		Fact AQI Count	Mean
Alabama	Franklin	1	61.00
Alabama	Jefferson	2	61.00
Alabama	Macon	1	28.00
Alaska	Fairbanks North Star	3	32.33
Arizona	Coconino	1	38.00
Arizona	Maricopa	79	207.33
Arizona	Pima	1	313.00
Arizona	Pinal	7	64.86
Arkansas	Nevada	11	185.55
California	Butte	3	244.33
California	Colusa	1	225.00
California	El Dorado	2	324.50
California	Fresno	2	250.50
California	Humboldt	4	107.75
California	Imperial	18	120.28
California	Inyo	22	140.64
California	Kern	3	208.67
California	Kings	1	500.00
California	Los Angeles	5	105.60
California	Mariposa	2	38.50
California	Mono	22	108.91
California	Placer	14	232.86
California	Plumas	23	317.78
California	Riverside	22	94.14
California	Sacramento	2	294.50
California	San Bernardino	26	151.77
California	San Joaquin	1	47.00

- Biểu đồ trực quan



➔ Phân tích:

- Các hạt như **Maricopa, Trinity, San Bernardino** có số ngày "very unhealthy" cao hơn hẳn, cho thấy tình trạng ô nhiễm không khí nghiêm trọng.
- Giá trị *Mean AQI Very Unhealthy* của các hạt này cũng cao, tuy nhiên có những hạt có số ngày "very unhealthy" thấp nhưng lại có giá trị *Mean AQI Very Unhealthy* cực kì cao.

➔ **Nhận xét:**

- Những hạt có số ngày và AQI cao phản ánh sự ảnh hưởng mạnh mẽ từ hoạt động công nghiệp hoặc điều kiện khí hậu bất lợi.
- Cần có chính sách giảm thiểu ô nhiễm đặc thù, tập trung vào các khu vực chịu ảnh hưởng nghiêm trọng.

6.2.4. For the four following states: Hawaii, Alaska, Illinois and Delaware, count the number of days in each air quality Category (Good, Moderate,etc.) by County. Analysis hints: Comparing the data of the states and counties, focus on the distribution of the harmful air condition. What could you conclude about the differences?)

• **Các bước thực hiện**

1. Tạo measure mới để đếm số ngày, logic tương tự như câu 3 nhưng sử dụng hướng tiếp cận khác. Với mỗi giá trị mới của cột Day trong bảng Dim Date, nếu có dữ liệu AQI sẽ tính là 1 ngày.

2. Thực hiện query MDX

```
//4. For the four following states: Hawaii, Alaska, Illinois, and Delaware,
//count the number of days in each air quality Category (Good, Moderate, etc.) by County.
```

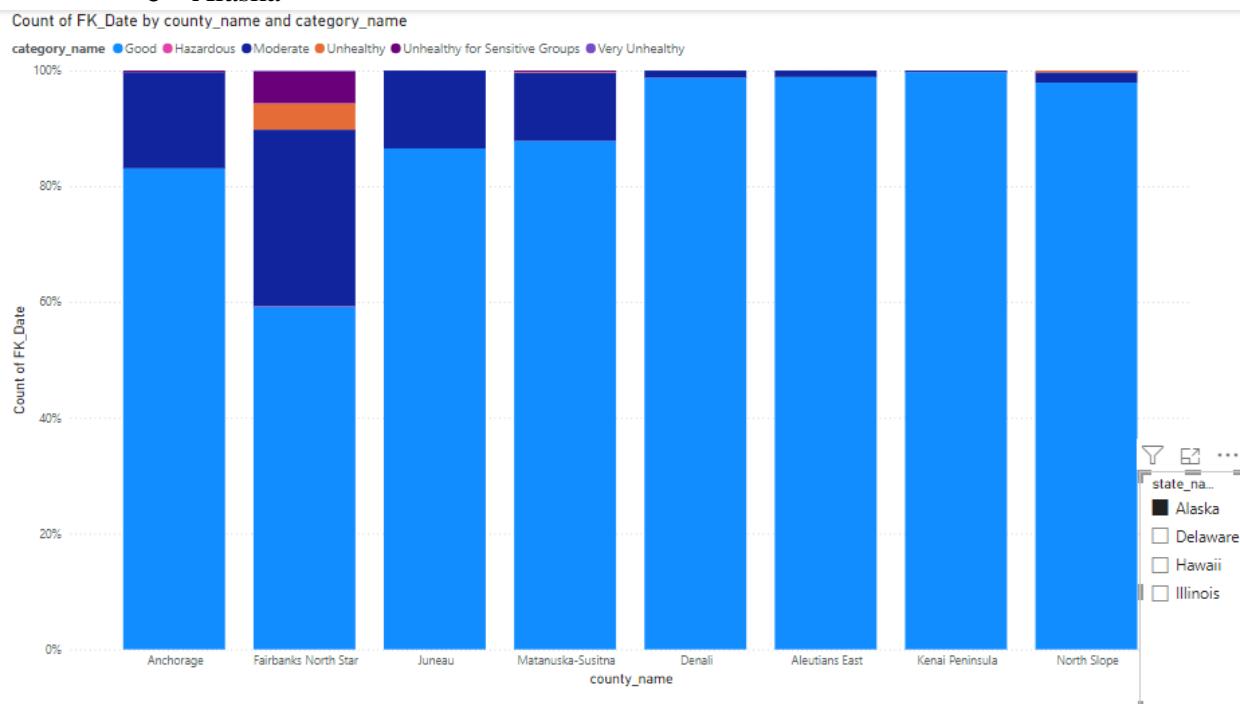
```
WITH
MEMBER [Measures].[NumOfDays4States] AS
COUNT(FILTER(
    [Dim Date].[Day].MEMBERS,
    [Measures].[AQI] > 0))
SELECT
{[Dim Category].[Category Name].[Category Name] * [Measures].[NumOfDays4States]}
ON COLUMNS,
NONEMPTYCROSSJOIN(
    {[Dim Site].[State Name].&[Hawaii],
    [Dim Site].[State Name].&[Alaska],
    [Dim Site].[State Name].&[Illinois],
    [Dim Site].[State Name].&[Delaware]},
    [Dim Site].[County Name].[County Name]
) ON ROWS
FROM [AirQuality_OLAP];
```

- Kết quả thực thi**

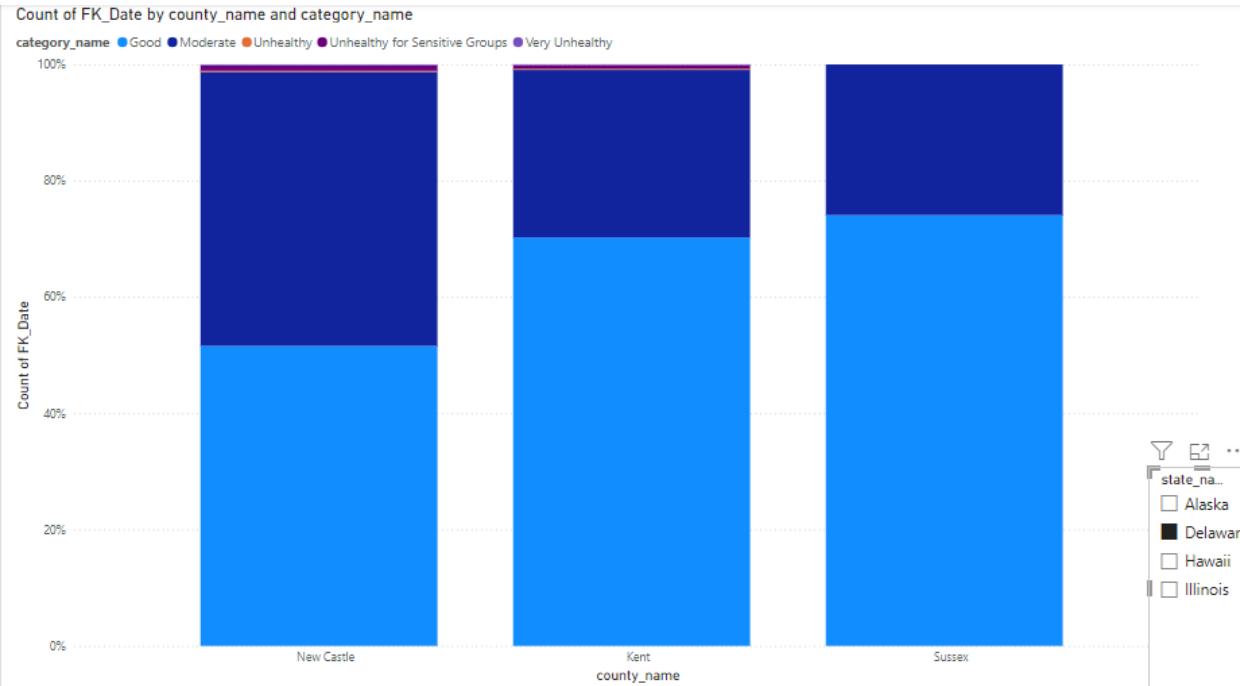
		Good NumOfDays4States	Hazardous NumOfDays4States	Moderate NumOfDays4States	Unhealthy NumOfDays4States	Unhealthy for Sensitive Groups NumOfDays4States	Very Unhealthy NumOfDays4States	Nu
Hawaii	Maui	1019	0	9	2	0	0	
Alaska	Aleutians East	357	0	5	0	0	0	
Alaska	Anchorage	911	0	183	0	4	0	
Alaska	Denali	1039	0	14	0	0	0	
Alaska	Fairbanks North Star	646	2	335	51	61	3	
Alaska	Juneau	911	0	144	0	0	0	
Alaska	Kenai Peninsula	357	0	2	0	0	0	
Alaska	Matanuska-Susitna	918	0	125	2	4	0	
Alaska	North Slope	329	0	7	2	0	0	
Illinois	Champaign	626	0	411	2	13	2	
Illinois	Clinton	625	0	99	0	6	0	
Illinois	DuPage	525	0	552	5	15	3	
Illinois	Jersey	665	0	402	2	25	0	
Illinois	Jo Daviess	967	0	84	2	9	0	
Illinois	Kane	556	0	279	6	21	0	
Illinois	Knox	620	0	83	0	3	0	
Illinois	Macoupin	1002	0	74	0	13	0	
Illinois	McHenry	578	0	491	5	22	2	
Illinois	McLean	583	0	498	3	14	2	
Illinois	Mercer	253	2	97	6	11	0	
Illinois	Peoria	550	0	449	2	23	0	
Illinois	Rock Island	584	0	481	4	20	2	
Illinois	Saint Clair	435	0	644	4	16	0	

- Biểu đồ trực quan**

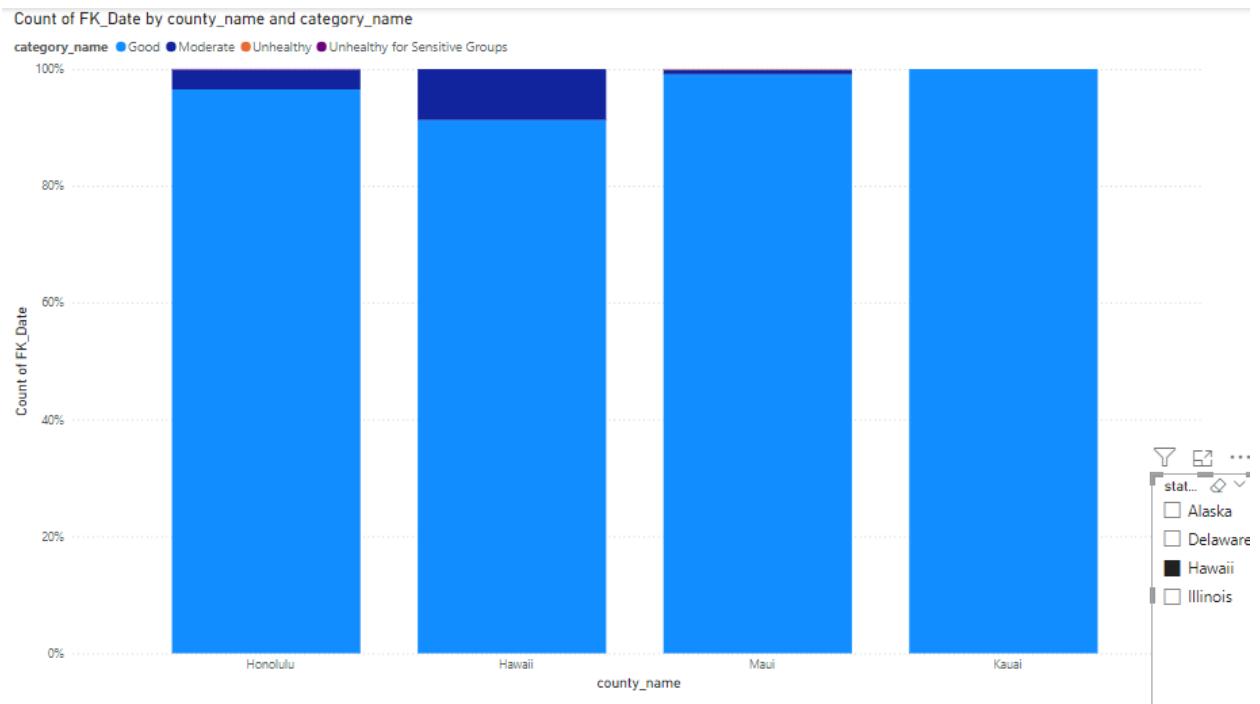
○ Alaska



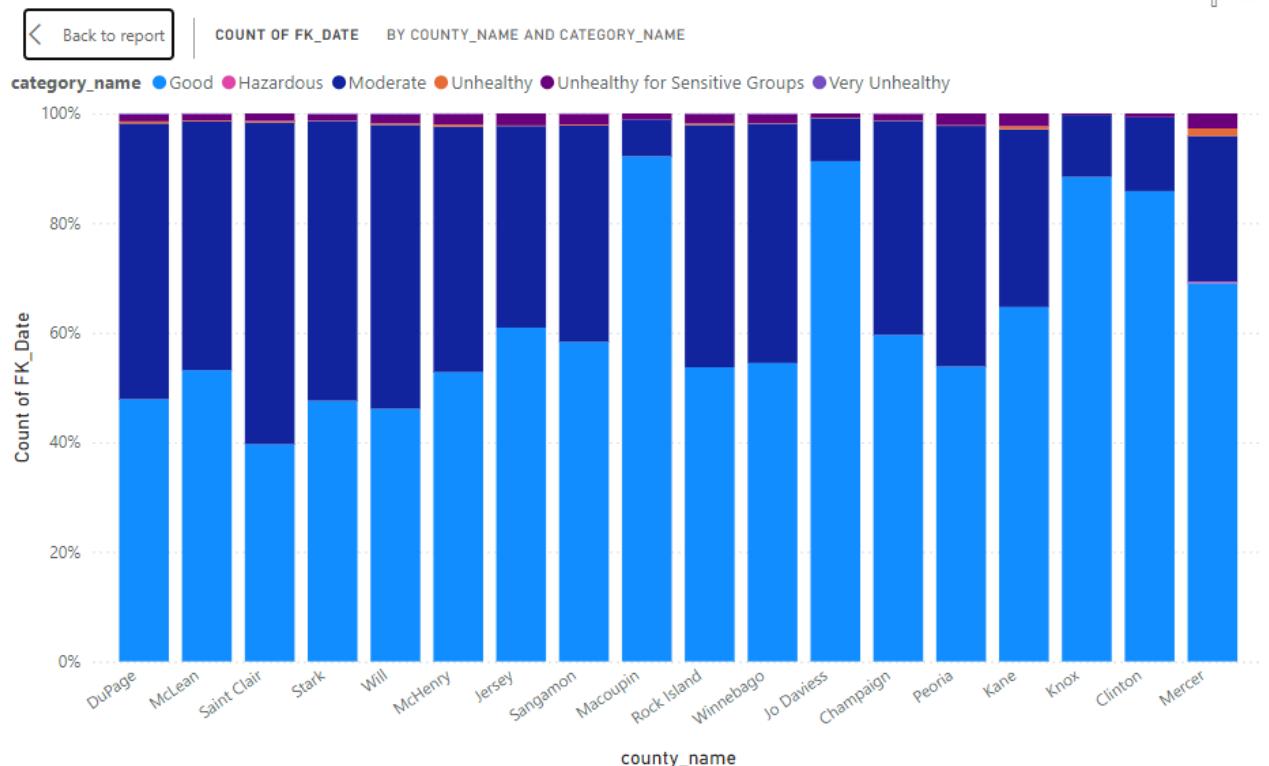
○ Delaware



○ Hawaii



- Illinois



➔ Phân tích:

1. Hawaii:

- Gần như toàn bộ các ngày có chất lượng không khí tốt (Good), chỉ một phần nhỏ thuộc loại Trung bình (Moderate) và gần như không có ngày nào ở mức độ xấu hơn. Điều này cho thấy môi trường ở Hawaii rất sạch sẽ và ít ô nhiễm.
- 2. Alaska:**
- Một số khu vực như Fairbanks North Star có sự hiện diện rõ rệt của các ngày thuộc loại Trung bình (Moderate), Không tốt cho nhóm nhạy cảm (Unhealthy for Sensitive Groups), thậm chí có không khí không lành mạnh (Unhealthy).
 - Các khu vực khác như Anchorage, Juneau, và North Slope hầu hết vẫn duy trì mức Good.
- 3. Delaware:**
- Số ngày chất lượng không khí Good chiếm phần lớn ở cả ba hạt (New Castle, Kent, Sussex), nhưng tỷ lệ các ngày thuộc nhóm Trung bình (Moderate) và Không tốt cho nhóm nhạy cảm cao hơn so với Hawaii và một số khu vực của Alaska.
- 4. Illinois:**
- Chất lượng không khí biến động nhiều hơn so với các bang khác. Một số hạt có tỷ lệ ngày Trung bình (Moderate), Không tốt cho nhóm nhạy cảm (Unhealthy for Sensitive Groups), và thậm chí Không lành mạnh (Unhealthy) cao hơn rõ rệt, chẳng hạn như ở Stark, McHenry, Saint Clair.

➔ **Nhận xét:**

- **So sánh giữa các bang:**
 - *Hawaii* có chất lượng không khí tốt nhất, phù hợp với môi trường tự nhiên sạch và ít ô nhiễm.
 - *Illinois* và *Alaska* có sự biến động lớn hơn trong chất lượng không khí, có thể do khí hậu hoặc các hoạt động công nghiệp. *Delaware* nằm giữa, với tình trạng ô nhiễm ở mức độ trung bình.
- **Khuyến nghị:**
 - Tăng cường theo dõi các khu vực có nhiều ngày Unhealthy như *Fairbanks North Star (Alaska)* và một số hạt tại *Illinois* để triển khai các biện pháp cải thiện chất lượng không khí.
 - Hạn chế các nguồn gây ô nhiễm, đặc biệt ở các khu vực có tỷ lệ ngày Trung bình (Moderate) và Không tốt cho nhóm nhạy cảm (Unhealthy for Sensitive Groups) cao.

6.2.5. For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the mean AQI value by quarters. Analysis hints: Comparing the data of the states over the year. What could you conclude about the fluctuations?

• **Các bước thực hiện**

1. Giá trị Mean đã được tính toán và lưu sẵn trong cube.
2. Thực hiện query MDX

//5. For the four following states: Hawaii, Alaska, Illinois, and Delaware,
 // compute the mean AQI value by quarters.

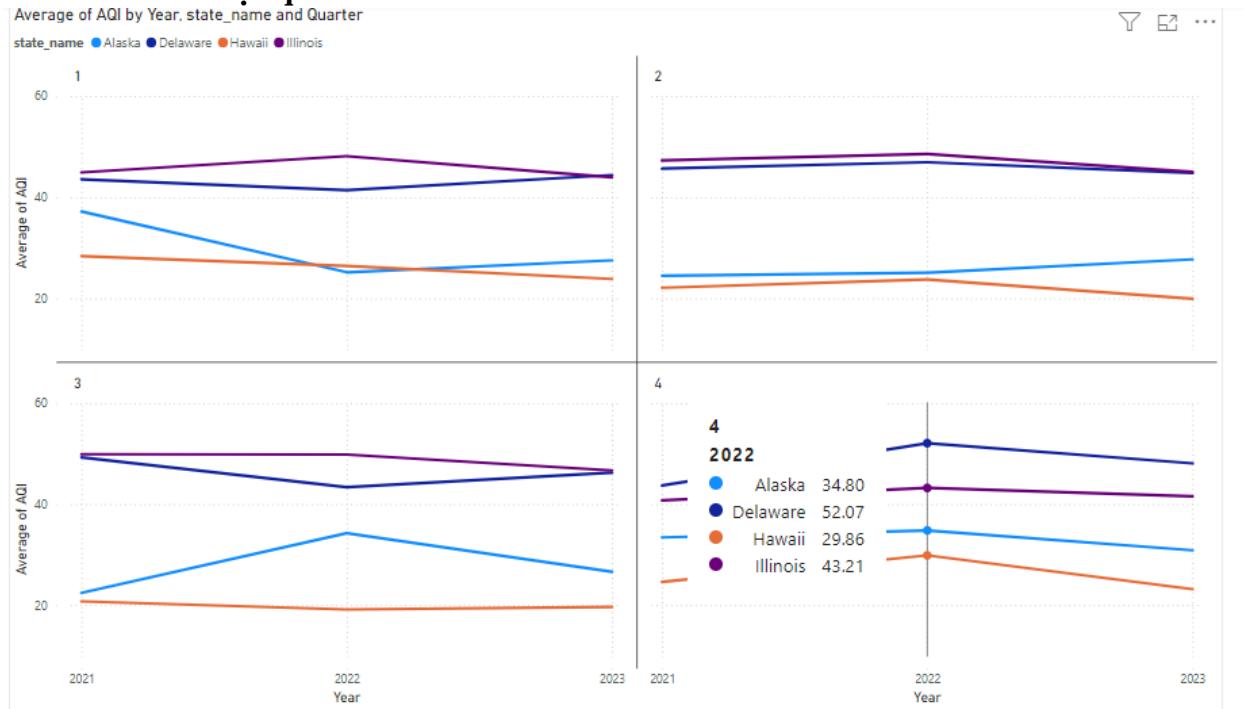
```

SELECT
  NON EMPTY([Dim Date].[Year].[Year] * [Dim Date].[Quarter].[Quarter]) ON COLUMNS,
  NON EMPTY(
    {[Dim Site].[State Name].&[Hawaii],
     [Dim Site].[State Name].&[Alaska],
     [Dim Site].[State Name].&[Illinois],
     [Dim Site].[State Name].&[Delaware]}
  ) ON ROWS
FROM [AirQuality_OLAP]
WHERE [Measures].[Mean];
  
```

- Kết quả thực thi

	2021	2021	2021	2021	2022	2022	2022	2022	2023	2023	2023	2023
	1	2	3	4	1	2	3	4	1	2	3	4
Hawaii	28.39	22.12	20.76	24.62	26.43	23.78	19.16	29.86	23.87	19.93	19.65	23.18
Alaska	37.23	24.52	22.43	33.42	25.19	25.11	34.25	34.80	27.54	27.75	26.64	30.87
Illinois	44.96	47.33	49.89	40.76	48.17	48.59	49.82	43.21	43.96	45.05	46.67	41.56
Delaware	43.59	45.71	49.26	43.69	41.45	46.97	43.34	52.07	44.42	44.85	46.22	48.08

- Biểu đồ trực quan



➔ Phân tích:

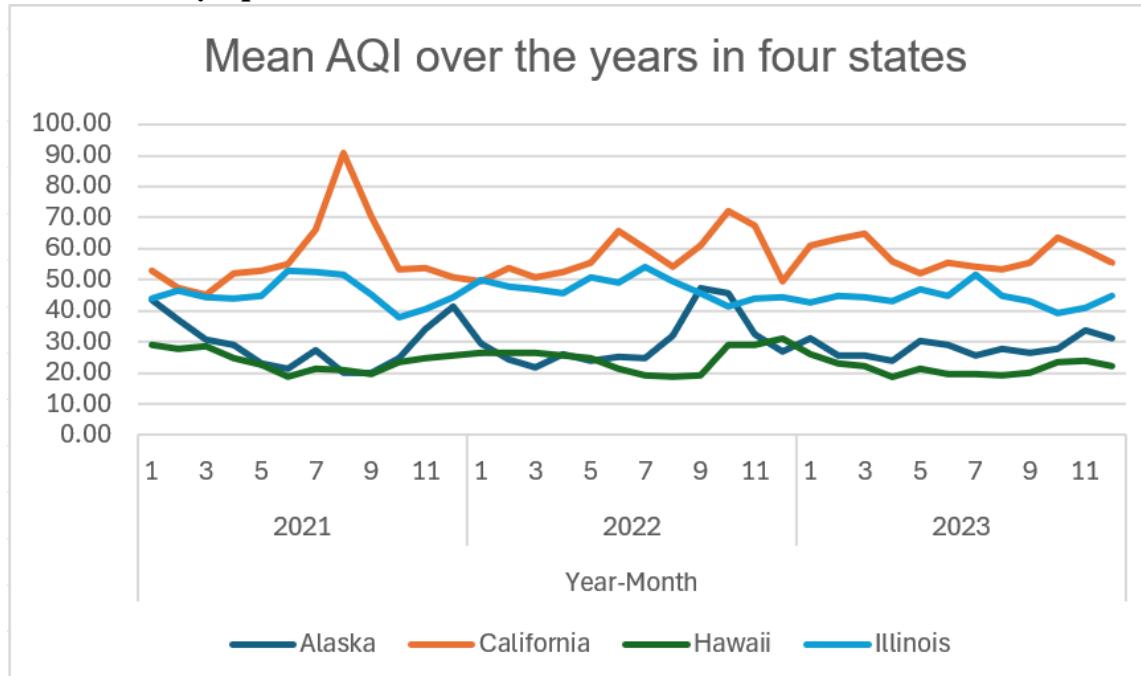
- **Hawaii:** AQI trung bình khá ổn định qua các quý với độ lệch chuẩn thấp, điều này cho thấy môi trường không khí ở bang này ít bị ảnh hưởng bởi các yếu tố biến động trong năm.
- **Alaska:** AQI biến động đáng kể giữa các quý, đặc biệt vào mùa xuân và thu, có thể do điều kiện khí hậu hoặc hiện tượng thiên nhiên như cháy rừng.
- **Delaware:** Có xu hướng AQI tăng nhẹ vào các quý mùa hè, có thể liên quan đến hoạt động kinh tế hoặc sự gia tăng ô nhiễm không khí do thời tiết nóng hơn.
- **Illinois:** AQI trung bình tương đối cao so với các bang khác trong phân tích, với mức độ biến động vừa phải, có thể liên quan đến hoạt động giao thông hoặc công nghiệp địa phương.

➔ Nhận xét:

- Những bang có sự biến động lớn như Alaska cần được theo dõi sát sao, vì điều này có thể chỉ ra các yếu tố rủi ro đặc biệt theo mùa.
- Các bang có mức AQI cao trung bình như Illinois cần thực hiện các chính sách giảm thiểu ô nhiễm, đặc biệt trong các mùa có hoạt động kinh tế sôi động.

6.2.6. Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California. Analysis hint: Give your opinion about the fluctuations of AQI value.

• Biểu đồ trực quan



➔ Phân tích:

- **California:** Có độ lệch chuẩn rất cao, đặc biệt trong các tháng mùa hè, cho thấy các đợt cháy rừng là nguyên nhân chính làm tăng độ biến AQI. Đây là bang có mức AQI trung bình cao nhất, vượt trội so với các bang khác.
- **Alaska:** Biến động AQI khá lớn, đặc biệt trong các tháng lạnh, cho thấy tác động của điều kiện khí hậu đặc thù và hiện tượng tự nhiên.
- **Hawaii:** Có xu hướng AQI ổn định nhất, với mức trung bình thấp nhất so với các bang còn lại, điều này phản ánh môi trường trong lành và ít bị tác động bởi hoạt động công nghiệp.
- **Illinois:** Xu hướng AQI có sự thay đổi nhẹ theo mùa, nhưng nhìn chung vẫn duy trì ở mức trung bình, không có sự đột biến rõ ràng.

➔ **Nhận xét:**

- *California* cần có các biện pháp khắc phục dài hạn để kiểm soát AQI trong các mùa dễ xảy ra cháy rừng, như kiểm soát nguồn lửa và bảo vệ rừng.
- *Hawaii* duy trì được chất lượng không khí tốt, là mô hình lý tưởng để các bang khác học hỏi, đặc biệt trong việc quản lý môi trường và tài nguyên.
- *Illinois* và *Alaska* cần có chính sách cụ thể để kiểm soát AQI trong các mùa có sự tăng cao, nhằm bảo vệ sức khỏe người dân.

6.2.7. Build graphs/charts for the above reports

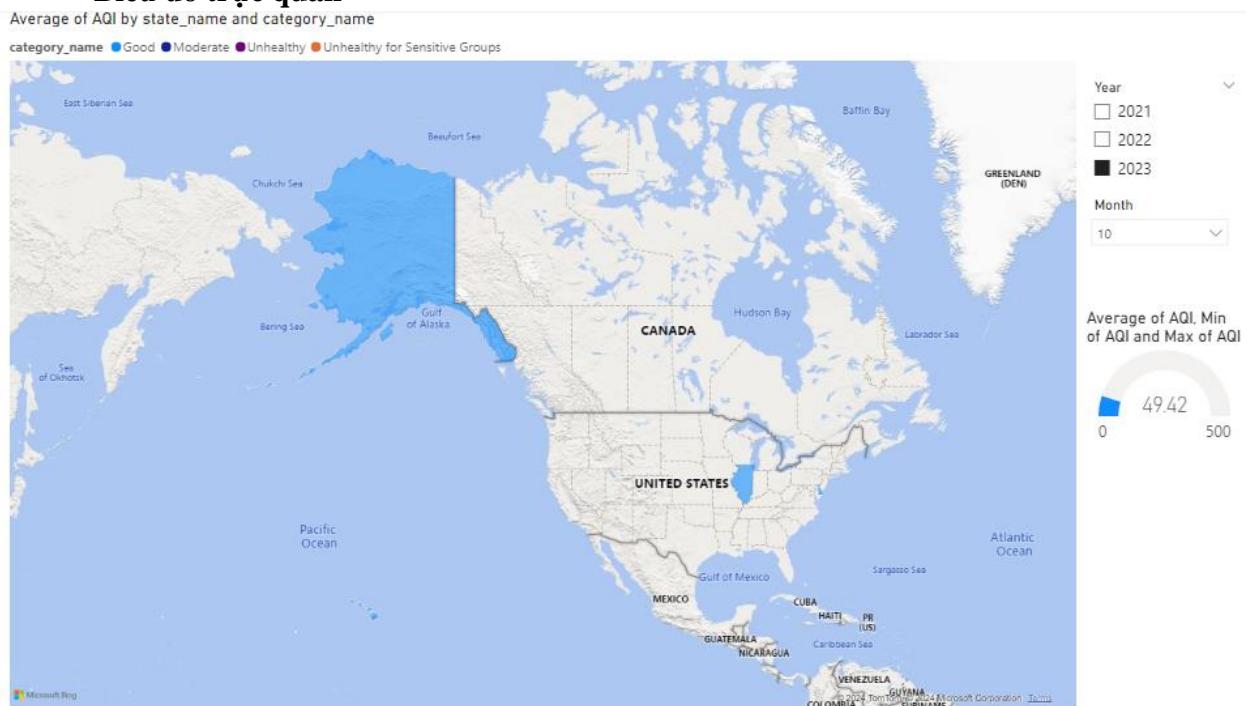
- Đã gắn trực tiếp vào từng câu hỏi ở trên để dễ dàng phân tích và trực quan hơn.

6.2.8. Use a regional map to visually represent (by color) the mean AQI value in regions during a year. Example:

US mean AQI of four states: Alaska, Delaware, Hawaii, Illinois over the year 2023

Month	Alaska	Delaware	Hawaii	Illinois
2023-01	40.339	43.151	38.581	44.517
2023-02	28.032	47.893	34.405	40.057
2023-03	29.077	46.570	29.600	45.179
2023-04	24.994	53.278	25.500	48.929
2023-05	25.632	50.699	28.364	65.065
2023-06	20.050	82.956	24.435	90.900
2023-07	22.762	56.355	24.462	55.857
2023-08	30.117	53.000	26.544	50.501
2023-09	17.956	47.822	25.256	48.688
2023-10	29.303	41.032	18.419	40.935
2023-11	30.620	46.012	24.156	41.243
2023-12	36.558	43.284	23.593	35.737

- **Biểu đồ trực quan**



➔ Phân tích:

- Trên bản đồ vùng, các bang như *Delaware* nổi bật với mức AQI trung bình cao (do cháy rừng và hoạt động kinh tế), trong khi *Hawaii* giữ mức thấp nhất, phản ánh môi trường trong sạch.
- Các bang khác như *Alaska* và *Illinois* có AQI trung bình gần sát nhau, nhưng Alaska có độ biến động cao hơn.

➔ Nhận xét:

- Các khu vực có mức AQI trung bình cao nên được ưu tiên trong các kế hoạch giảm thiểu ô nhiễm không khí.
- Hawaii là minh chứng cho việc duy trì môi trường sống sạch, nên khuyến khích các bang khác học hỏi phương pháp quản lý khí thải và bảo vệ môi trường.

6.2.9. Report the mean, the standard deviation, min and max of AQI value group by State and County during each quarter of the year. Analysis hints: Pay attention to the values (mean, std, max, min). Are any unusually large or small? Compare the standard deviation values between question 1 and 2, explain.

- **Các bước thực hiện**

1. Tạo Measure để tính Standard Deviation (Std) của AQI qua việc sử dụng hàm tích hợp trong MDX để tính độ lệch chuẩn (Standard Deviation).
2. Thực hiện query MDX

```
//9. Report the mean, the standard deviation, min and max of AQI value group by
//State and County during each quarter of the year.

WITH
MEMBER [Measures].[Std] AS
    Stdev
    ([Dim Date].[Hierarchy Date].CurrentMember.Children,
    [Measures].[Mean])

SELECT {[Measures].[Minimum AQI], [Measures].[Maximum AQI], [Measures].[Mean], [Measures].[Std]} ON COLUMNS,
NON EMPTY
[Dim Site].[State Name].[State Name]
* [Dim Site].[County Name].[County Name]
* [Dim Date].[Year].[Year]
* [Dim Date].[Quarter].[Quarter]
ON ROWS
FROM [OLAP];
```

- Kết quả**

Messages Results

				Minimum AQI	Maximum AQI	Mean	Std
Alabama	Butler	2021	1	16	101	53.04	8.17527921171505
Alabama	Butler	2021	2	30	101	57.10	3.2344095554028
Alabama	Butler	2021	3	28	116	60.78	7.0989182978085
Alabama	Butler	2021	4	26	85	50.16	2.46705107422065
Alabama	Butler	2022	1	3	76	19.52	2.71854380831953
Alabama	Butler	2022	2	2	77	19.11	2.47631078032252
Alabama	Butler	2022	3	1	76	23.83	6.32939079263653
Alabama	Butler	2022	4	26	101	55.78	2.60508853956915
Alabama	Butler	2023	1	17	71	38.87	2.00623245820028
Alabama	Butler	2023	2	17	71	36.76	2.36646586965055
Alabama	Butler	2023	3	16	77	36.75	5.20361810145789
Alabama	Butler	2023	4	16	90	41.96	3.13730446527265
Alabama	Butler	Unknown	Unknown	(null)	(null)	(null)	-nan(ind)
Alabama	Dallas	2021	1	14	71	44.20	2.58143889003318
Alabama	Dallas	2021	2	19	174	50.59	8.04279830762304
Alabama	Dallas	2021	3	21	119	58.67	7.11028742230653
Alabama	Dallas	2021	4	25	97	45.84	4.77801539847969
Alabama	Dallas	2022	1	17	159	42.46	7.82832332437258
Alabama	Dallas	2022	2	13	85	33.70	5.5315444683562
Alabama	Dallas	2022	3	11	74	42.67	2.43143705502347
Alabama	Dallas	2022	4	17	73	40.51	1.87174169301032
Alabama	Dallas	2023	1	6	85	47.71	4.46471907118204
Alabama	Dallas	2023	2	10	92	45.52	4.2026507224793
Alabama	Dallas	2023	3	14	77	37.66	5.41435608909862
Alabama	Dallas	2023	4	21	92	46.72	0.951501319825171
Alabama	Dallas	Unknown	Unknown	(null)	(null)	(null)	-nan(ind)
Alabama	Fayette	2021	1	16	58	34.25	7.99788827051707
Alabama	Fayette	2021	2	24	90	43.74	0.952968072995119
Alabama	Fayette	2021	3	18	54	37.81	2.01387521920085
Alabama	Fayette	2021	4	16	46	28.84	3.53951584661124
Alabama	Fayette	2022	1	19	77	44.69	5.681175632192
Alabama	Fayette	2022	2	27	67	41.46	7.07950611111002

➔ Phân tích: Lấy đại diện thông tin của New Mexico để phân tích

- Có thể thấy các mục Good và Moderate chiếm phần lớn, áp đảo so với các mục khác. Ở đa số các tháng, số ngày mà chất lượng không khí Good hoặc Moderate nhiều hơn 28 ngày).

➔ Nhận xét:

- Nhìn chung, chất lượng không khí ở New Mexico là rất tốt, có thể là vùng với bầu không khí trong lành, thích hợp để sinh sống.

6.2.10. Create a new attribute, DayLightSaving, in a suitable table. DayLightSaving may have two values: True: Between March 12, 2023, and November 5, 2023; False: Otherwise. Report the mean AQI value by State, Category, DayLightSaving over years. Analysis hint: Is there any notable difference on the air quality during the DaylightSaving period compared to the other?

- **Các bước thực hiện**

- Thực thi câu lệnh sau trong SQL Server

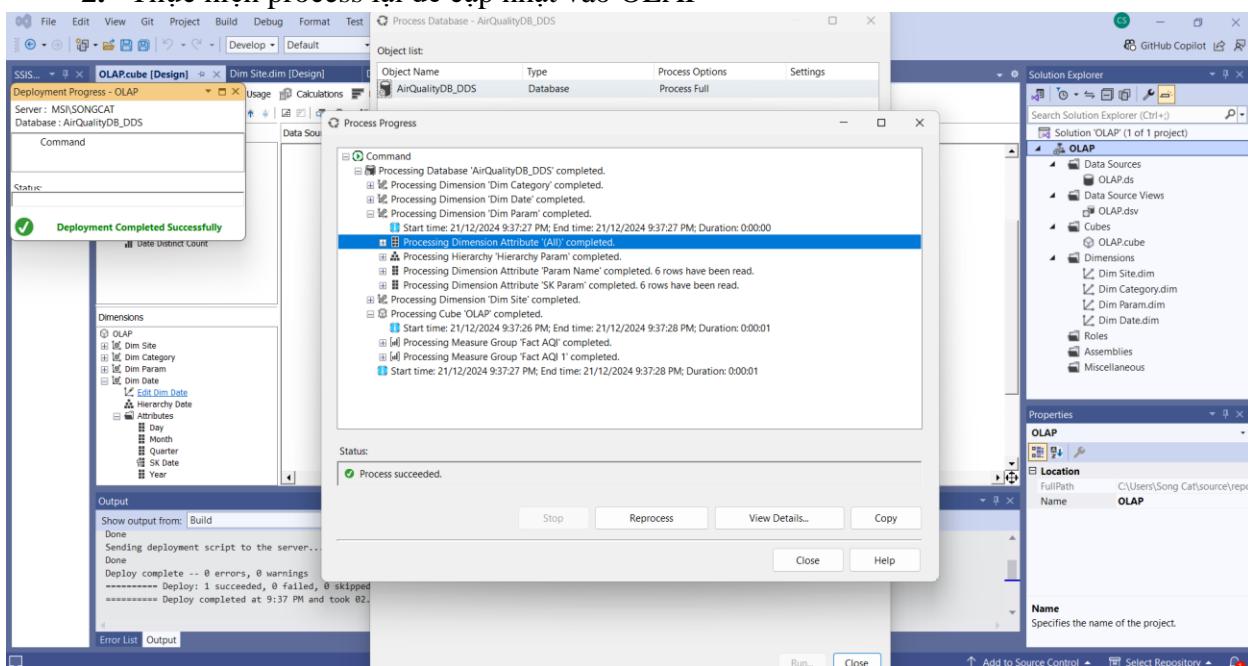
```
USE AirQualityDB_DDS;
```

```
GO
```

```
ALTER TABLE Dim_Date
ADD DayLightSaving bit;
GO
```

```
UPDATE Dim_Date
SET DayLightSaving =
CASE
    WHEN ((Month = 3 AND Day >= 12)
        OR (MONTH > 3 AND MONTH < 11)
        OR (MONTH = 11 AND DAY <=5) ) AND YEAR = 2023
    THEN 1
    ELSE 0
END;
```

- Thực hiện process lại để cập nhật vào OLAP



- Thực thi câu query MDX

```
//10. Create a new attribute, DayLightSaving, in a suitable table. DayLightSaving may
//have two values:
//True: Between March 12, 2023, and November 5, 2023
//False: Otherwise
//Report the mean AQI value by State, Category, DayLightSaving over years.
```

```
SELECT
    NON EMPTY
        [Dim Site].[State Name].[State Name] *
        [Dim Date].[Year].[Year] *
        [Dim Date].[Day Light Saving].[Day Light Saving] ON ROWS,
    NON EMPTY
        [Dim Category].[Category Name].[Category Name] ON COLUMNS
FROM [OLAP]
WHERE [Measures].[Mean];
```

- Kết quả

	Messages	Results					
		Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Indiana	2022	0	37.5677083333333	(null)	62.36666666666667	(null)	(null)
North Dakota	2023	0	29.8461538461538	(null)	60.5824175824176	(null)	113.333333333333
North Dakota	2023	1	30.5	(null)	62.2853470437018	(null)	103.75
Ohio	2021	0	34.3025355596784	(null)	61.5768525792092	157	110.25
Ohio	2022	0	34.1261153282727	(null)	61.5572477563231	157	110.261904761905
Ohio	2023	0	34.956064947469	(null)	61.5086021505376	157	112
Ohio	2023	1	34.939151813153	(null)	61.5634167385677	(null)	110.285714285714
Oklahoma	2022	0	36.6024305555556	(null)	62.0845070422535	(null)	110
Texas	2021	0	33.1213980789755	(null)	61.5453575240128	162.4375	115.930555555556
Texas	2022	0	33.7998175552402	(null)	61.5032822757112	161.6666666666667	115.778625954198
Texas	2023	0	34.3154305200341	(null)	62.0541494997057	155	117.964912280702
Texas	2023	1	33.8250716332378	(null)	61.2143874643875	163.5	114.597701149425
Utah	2022	0	28.2641509433962	(null)	61.8479020979021	174	117.461538461538
Utah	2023	0	18.8588235294118	(null)	62.2950819672131	(null)	(null)
Utah	2023	1	21.9448979591837	(null)	61.7094017094017	(null)	(null)
Virginia	2022	0	23.3098591549296	(null)	59.25	(null)	(null)
Virginia	2023	0	26.4827586206897	(null)	63.625	(null)	(null)
Virginia	2023	1	22.66666666666667	(null)	57.0625	(null)	(null)

➔ Phân tích:

- Có thể thấy dataset bị thiếu nhiều dòng dữ liệu để có thể so sánh, một số vùng dữ liệu không trọn vẹn cả năm. Với những vùng có đầy đủ dòng dữ liệu thì mặc dù là ngày Day Light Saving thì chất lượng không khí vẫn không có chênh lệch quá nhiều.

➔ Nhận xét:

- Nhìn chung, chất lượng không khí ở New Mexico là rất tốt, có thể là vùng với bầu không khí trong lành, thích hợp để sinh sống.

6.2.11. Count the number of days by State, Category in each month.

Be caution: The Category in the data set is calculated for each County, not State.

- Các bước thực hiện

- Tạo measure Date Distinct Count bằng hàm Count Distinct các FK_Date trong bảng Fact AQI.
- Thực hiện query MDX

//11. Count the number of days by State, Category in each month.
 //Be caution: The Category in the data set is calculated for each County, not State.

```

SELECT
  NON EMPTY
    [Dim Site].[State Name].[State Name] *
    [Dim Date].[Year].[Year] *
    [Dim Date].[Month].[Month] ON ROWS,
  NON EMPTY
    [Dim Category].[Category Name].[Category Name] ON COLUMNS
FROM [OLAP]
WHERE [Measures].[Date Distinct Count];
  
```

- Kết quả

			Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
New Mexico	2021	1	18	(null)	13	(null)	(null)	(null)
New Mexico	2021	10	17	(null)	14	(null)	(null)	(null)
New Mexico	2021	11	27	(null)	3	(null)	(null)	(null)
New Mexico	2021	12	11	(null)	19	(null)	1	(null)
New Mexico	2021	2	11	(null)	15	(null)	(null)	(null)
New Mexico	2021	3	10	(null)	21	(null)	(null)	(null)
New Mexico	2021	4	10	(null)	20	(null)	(null)	(null)
New Mexico	2021	5	14	(null)	17	(null)	(null)	(null)
New Mexico	2021	6	16	(null)	14	(null)	(null)	(null)
New Mexico	2021	7	13	(null)	18	(null)	(null)	(null)
New Mexico	2021	8	11	(null)	20	(null)	(null)	(null)
New Mexico	2021	9	11	(null)	18	(null)	1	(null)
New Mexico	2022	1	15	(null)	16	(null)	(null)	(null)
New Mexico	2022	10	21	(null)	10	(null)	(null)	(null)
New Mexico	2022	11	28	(null)	2	(null)	(null)	(null)
New Mexico	2022	12	21	(null)	10	(null)	(null)	(null)
New Mexico	2022	2	15	(null)	13	(null)	(null)	(null)
New Mexico	2022	3	13	(null)	18	(null)	(null)	(null)
New Mexico	2022	4	9	(null)	20	(null)	1	(null)
New Mexico	2022	5	8	(null)	23	(null)	(null)	(null)
New Mexico	2022	6	13	(null)	15	(null)	2	(null)
New Mexico	2022	7	13	(null)	18	(null)	(null)	(null)
New Mexico	2022	8	20	(null)	11	(null)	(null)	(null)
New Mexico	2022	9	21	(null)	9	(null)	(null)	(null)
New Mexico	2023	1	20	(null)	11	(null)	(null)	(null)
New Mexico	2023	10	21	(null)	9	(null)	(null)	(null)
New Mexico	2023	11	18	(null)	11	(null)	(null)	(null)
New Mexico	2023	12	19	(null)	11	(null)	(null)	(null)
New Mexico	2023	2	12	(null)	16	(null)	(null)	(null)
New Mexico	2023	3	15	(null)	16	(null)	(null)	(null)
New Mexico	2023	4	12	(null)	18	(null)	(null)	(null)
New Mexico	2023	5	21	(null)	10	(null)	(null)	(null)
New Mexico	2023	6	17	(null)	9	(null)	(null)	(null)
New Mexico	2023	7	12	(null)	19	(null)	(null)	(null)
New Mexico	2023	8	17	(null)	14	(null)	(null)	(null)
New Mexico	2023	9	20	(null)	11	(null)	(null)	(null)

➔ **Phân tích: Lấy đại diện thông tin của New Mexico để phân tích**

- Có thể thấy các mục Good và Moderate chiếm phần lớn, áp đảo so với các mục khác. Ở đa số các tháng, số ngày mà chất lượng không khí Good hoặc Moderate nhiều hơn 28 ngày).

➔ **Nhận xét:**

- Nhìn chung, chất lượng không khí ở New Mexico là rất tốt, có thể là vùng với bầu không khí trong lành, thích hợp để sinh sống.

6.2.12. Report the number of days by Category and Defining Parameter. Analysis hints:

What is your opinion on the pollution situation in the United States as a whole?

Additionally, please identify the primary factors that the country should consider in order to enhance air quality

- Các bước thực hiện**

- Thực hiện query MDX

//12. Report the number of days by Category and Defining Parameter.

```
SELECT
```

```
NON EMPTY [Dim Category].[Category Name].[Category Name] ON ROWS,
NON EMPTY
    [Dim Param].[Param Name].[Param Name] ON COLUMNS
FROM [OLAP]
WHERE [Measures].[Fact AQI Count];
```

- Kết quả thực thi**

The screenshot shows a results grid from SSMS. The columns are labeled CO, NO2, Ozone, PM10, and PM2.5. The rows represent air quality categories: Good, Hazardous, Moderate, Unhealthy, Unhealthy for Sensitive Groups, and Very Unhealthy. The data shows that 'Good' is the most frequent category across all pollutants.

	CO	NO2	Ozone	PM10	PM2.5
Good	232	3571	73225	4320	42314
Hazardous	(null)	(null)	(null)	67	41
Moderate	2	89	18378	2385	44269
Unhealthy	(null)	(null)	582	71	537
Unhealthy for Sensitive Groups	(null)	9	3192	263	1179
Very Unhealthy	(null)	(null)	113	26	106

➔ **Phân tích:**

- Chất lượng không khí "Good" chiếm ưu thế (đặc biệt với Ozone và PM2.5), cho thấy tình hình chung khá tích cực.
- PM2.5 và Ozone là hai thông số chính gây ra các ngày ô nhiễm cao, đặc biệt trong nhóm "Moderate" và "Unhealthy for Sensitive Groups".
- Số ngày "Hazardous" và "Very Unhealthy" rất thấp, nhưng vẫn có nguy cơ ô nhiễm cục bộ nghiêm trọng.

➔ **Nhận xét:**

- Mục tiêu soát tốt ô nhiễm không khí, nhưng cần tập trung giảm thiểu PM2.5 và Ozone ở các khu vực đông dân cư.
- Đề xuất: Thắt chặt tiêu chuẩn khí thải, chuyển đổi năng lượng sạch, và quản lý tốt hơn hiện tượng tự nhiên như cháy rừng.

7. Data Mining

7.1. Tổng quan

- **Mục đích:** Dự đoán chất lượng không khí và phân loại chất lượng không khí trong các khoảng thời gian sắp tới (ví dụ: Quý 1-2024, Tháng 01-2024) bằng cách sử dụng các mô hình phân tích dữ liệu.
- **Dữ liệu:** Chất lượng không khí (AQI) thu thập theo từng ngày, tháng, quý từ các trạng thái, quận và các yếu tố ảnh hưởng khác.

7.2. Mô hình và Thuật toán sử dụng

7.2.1. Thuật toán đề xuất

- **SARIMA:** Dự đoán giá trị đo đạc AQI theo chuỗi thời gian.
- **Random Forest:** Dự đoán giá trị đo đạc AQI dựa trên tập hợp các biến độc lập (nhiệt độ, độ ẩm, nồng độ các chất ô nhiễm, v.v.).
- **Naive Bayes:** Phân loại nhãn chất lượng không khí ("Good", "Hazardous", "Moderate", "Unhealthy", "Unhealthy for Sensitive Groups", "Very Unhealthy").

7.2.2. Lý do chọn thuật toán

- **SARIMA:**
 - Xử lý hiệu quả chuỗi thời gian có xu hướng và tính chu kỳ.
 - Có khả năng biểu diễn mối quan hệ giữa thời gian và giá trị AQI.
- **Random Forest:**
 - Mạnh mẽ trong xử lý dữ liệu phi tuyến tính.
 - Giảm nguy cơ overfitting nhờ việc kết hợp nhiều cây quyết định.
 - Khả năng dự đoán chính xác khi có nhiều yếu tố ảnh hưởng.
- **Naïve Bayes:**
 - Đơn giản, nhanh và hiệu quả cho bài toán phân loại.
 - Hoạt động tốt khi dữ liệu phân phối không đồng nhất.

7.3. Các bước thực hiện

1. Chuẩn bị môi trường và dữ liệu

- Nhóm sử dụng ngôn ngữ **Python** để thực hiện khai thác dữ liệu. Các thư viện sử dụng chủ yếu gồm: **pandas**, **sqlalchemy** và **scikit-learn**.
- Thiết lập kết nối cần thiết đến cơ sở dữ liệu DDS đã xây dựng trước đó.
- Truy vấn các dữ liệu cần thiết và lưu vào một Pandas DataFrame. Câu lệnh truy vấn như sau:

```
select AQI, FK_Param, FK_Category, FK_County,
       FK_Site, Day, Month, Year, Quarter, lat,
       lng from Fact_AQI
join Dim_Category on SK_Category = FK_Category
join Dim_Date on SK_Date = FK_Date
join Dim_Param on SK_Param = FK_Param
join Dim_Site on SK_Site = FK_Site
join Dim_County on SK_County = FK_County
join Dim_State on SK_State = FK_State
order by SK_AQI ASC
```

- Sau khi kết nối và truy vấn dữ liệu thành công, đồng thời lượt bỏ những giá trị *missing value* ta sẽ được một bảng dữ liệu như sau:

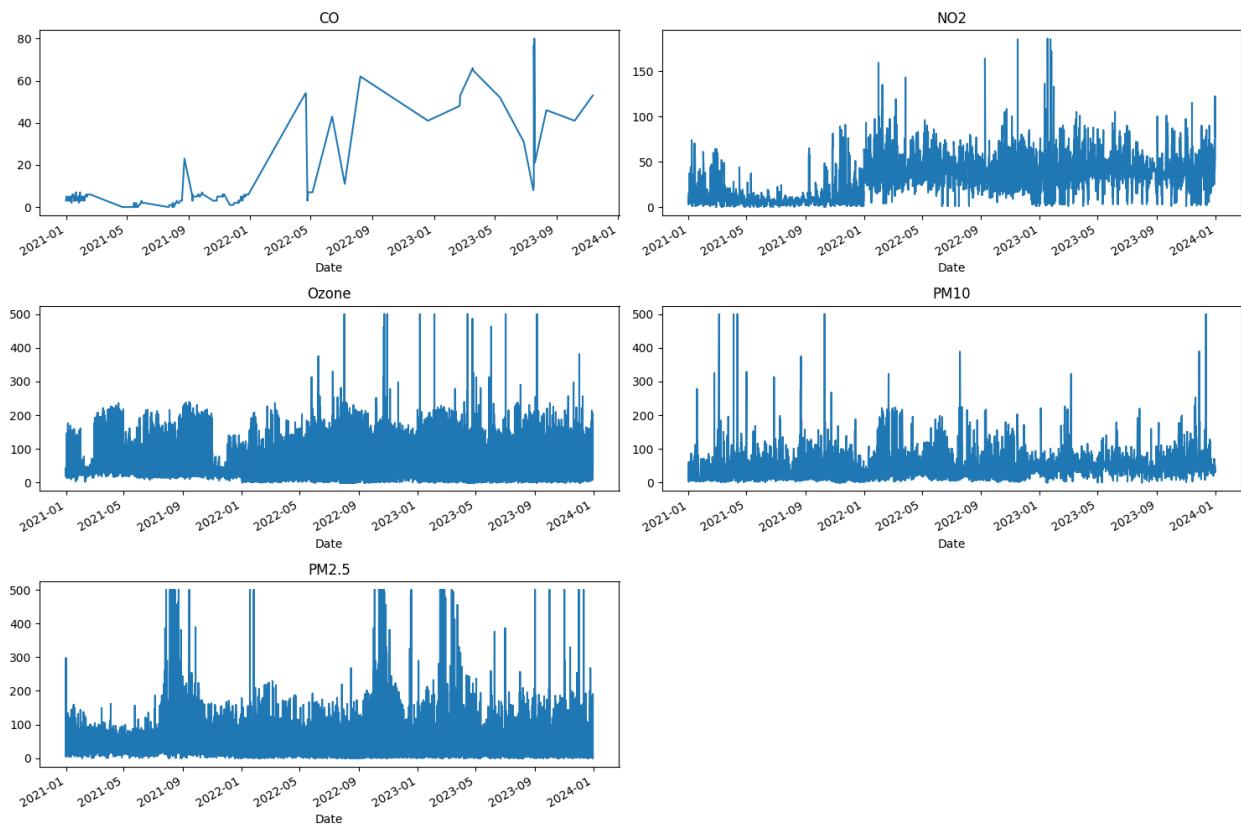
	AQI	FK_Param	FK_Category	FK_County	FK_Site	Day	Month	Year	Quarter	lat	lng
0	57	3	1	860	462	10	4	2023	2	38.7787	-120.5247
1	71	3	1	860	462	10	5	2023	2	38.7787	-120.5247
2	125	3	3	860	462	10	6	2023	2	38.7787	-120.5247
3	123	3	3	860	462	10	7	2023	3	38.7787	-120.5247
4	93	3	3	860	101	10	8	2023	3	38.7787	-120.5247
...
194966	49	3	1	340	560	19	4	2022	2	38.2046	-120.5541
194967	80	3	1	340	560	20	4	2022	2	38.2046	-120.5541
194968	90	3	1	340	560	21	4	2022	2	38.2046	-120.5541
194969	115	3	1	340	560	22	4	2022	2	38.2046	-120.5541
194970	87	3	1	340	560	23	4	2022	2	38.2046	-120.5541

185770 rows × 11 columns

2. Tiền xử lý dữ liệu

- Đầu tiên, ta biết rằng mỗi dòng trong bảng Fact sẽ là thông tin đo đặc chất lượng không khí tại một trạm của một quận thuộc một bang ở một thời điểm cụ thể với **một chỉ số đo đặc cụ thể**.
- Dữ liệu tồn tại tổng cộng 5 loại chỉ số đo đặc lần lượt là: '**CO**', '**NO2**', '**Ozone**', '**PM10**', '**PM2.5**'. Tương ứng với mỗi chỉ số đo đặc, AQI sẽ có một miền giá trị tương đối khác

nhau, chênh lệch nhiều và phân bố cũng khác nhau. Chi tiết xin xem trong loạt biểu đồ bên dưới:



- **Nhận xét:** Ta có thể nhận thấy biên độ giá trị AQI tương ứng với các chỉ số khác nhau sẽ có chênh lệch nhiều, trong khi chỉ số CO có khoảng giá trị 0 – 80, NO2 từ 0-186, thì ba chỉ số Ozone, PM10 và PM5 lại có khoảng giá trị từ 0-500. Do đó, để có thể mô hình hoá một cách hiệu quả, ta cần tách dữ liệu ra thành 5 phần tương ứng với 5 chỉ số và huấn luyện lần lượt.
- Kế đến, ta cần thực hiện chuẩn hoá dữ liệu do để cải thiện hiệu suất mô hình, đặc biệt là khi miền giá trị của dữ liệu đang khá rộng.
- Đối với mô hình phân loại (classification) mà ở đây chính là mô hình Naïve Bayes, ta còn cần thực hiện lấy mẫu quá mức (oversampling) để cân bằng các nhãn dữ liệu, vì phân bố các nhãn hiện tại đang quá mất cân bằng.
- Việc xử lí missing value trong bối cảnh hiện tại là không quá cần thiết vì dữ liệu từ DDS đã đảm bảo chất lượng, không tồn tại các dòng bị mất giá trị tại các cột quan trọng.

7.4. Kết quả và Đánh giá

7.4.1. Kết quả dự đoán

7.4.1.1. Mô hình Naïve Bayes

param_name	accuracy	precision	recall	f1
CO	1.000	1.000	1.000	1.000
NO2	0.858	0.859	0.858	0.857

Ozone	0.521	0.516	0.522	0.489
PM10	0.391	0.428	0.389	0.390
PM2.5	0.432	0.431	0.432	0.405

- Trung bình các chỉ số:

- **accuracy**: 64.04%.
- **precision**: 64.68%.
- **recall**: 64.02%.
- **F1**: 62.82%.

7.4.1.2. Mô hình SARIMA

param_name	MAE	MSE	RMSE
CO	14.035	341.037	18.467
NO2	16.998	397.382	19.934
Ozone	20.505	633.470	25.169
PM10	14.408	434.234	20.838
PM2.5	21.571	829.164	28.795

- Trung bình các chỉ số:

- **MAE (Mean Absolute Error)**: 17.5034
- **MSE (Mean Squared Error)**: 527.0574
- **RMSE (Root Mean Squared Error)**: 22.6406

7.4.1.3. Mô hình Random Forest

param_name	MAE	MSE	RMSE
CO	3.079	38.379	6.195
NO2	17.567	435.635	20.872
Ozone	17.154	662.805	25.745
PM10	18.306	531.226	23.048
PM2.5	16.730	662.320	25.736

- Trung bình các chỉ số:

- **MAE (Mean Absolute Error)**: 14.5674
- **MSE (Mean Squared Error)**: 466.0731
- **RMSE (Root Mean Squared Error)**: 20.3192

7.4.2. Dánh giá mô hình

7.4.2.1. Mô hình Naïve Bayes

- Mô hình hoạt động tốt với tham số CO, cho thấy khả năng phân biệt chính xác và đầy đủ. Tuy nhiên, có thể mô hình đang gặp tình trạng “**overfitting**”. Tình trạng này có thể khiến mô hình hoạt động rất tốt với dữ liệu hiện tại nhưng sẽ không tốt cho các trường hợp tổng quát hơn.

- **NO2** cũng đạt hiệu suất khá cao, cho thấy khả năng dự đoán tương đối ổn định. Có thể thấy, mô hình được huấn luyện trên dữ liệu của tham số NO là đạt hiệu suất tốt nhất trong cả 5 tham số.
- Các tham số **Ozone**, **PM10**, và **PM2.5** có độ chính xác và điểm F1 thấp, cho thấy mô hình cần cải thiện với các tham số này.

7.4.2.2. Mô hình SARIMA

- **CO**: Sai số trung bình nhỏ nhất trong tất cả các tham số, cho thấy mô hình hoạt động hiệu quả nhất với tham số này.
- **NO2**: Sai số tương đối nhỏ, hiệu suất khá tốt, gần với CO.
- **Ozone**: Sai số bắt đầu tăng lên, đặc biệt với RMSE cao hơn đáng kể so với CO và NO2.
- **PM10**: Mặc dù MAE gần với CO, nhưng MSE và RMSE cao hơn, cho thấy một số điểm dữ liệu có sai số lớn.
- **PM2.5**: Sai số cao nhất trong tất cả các tham số, đặc biệt RMSE lớn, chỉ ra rằng dự đoán PM2.5 gặp khó khăn nhất.
- **Kết luận chung:**
 - **CO và NO2**: Mô hình hoạt động tốt với các tham số này, sai số nhỏ và ổn định.
 - **Ozone, PM10, và PM2.5**: Sai số lớn hơn, cần cải thiện đặc biệt với PM2.5.
 - Lý giải cho tình trạng các mô hình Ozone, PM10 và PM5 có sai số lớn là do một số nguyên nhân như số lượng dòng dữ liệu thuộc các chỉ số này lớn hơn nhiều so với các chỉ số khác.Thêm vào đó, giá trị của các dòng cũng biến động theo vị trí địa lý và khó có thể huấn luyện một cách trọn tru nếu ta không phân dữ liệu theo từng vùng cụ thể hơn.

7.4.2.3. Mô hình Random Forest

- **CO**: Sai số nhỏ nhất trong tất cả các tham số, cho thấy mô hình Random Forest dự đoán **CO** hiệu quả. Đây có thể là do dữ liệu CO dễ nắm bắt với các mô hình phi tuyến tính như Random Forest.
- **NO2**: Hiệu suất khá tốt, tương đương với CO, cho thấy Random Forest có khả năng dự đoán chính xác NO2.
- **Ozone**: Sai số bắt đầu tăng lên. Mô hình gặp khó khăn hơn với Ozone, có thể do Ozone có sự phụ thuộc phức tạp vào thời gian và điều kiện môi trường.
- **PM10**: Sai số ở mức trung bình. Mặc dù MAE thấp (gần với CO), RMSE cao hơn, cho thấy có một số điểm dữ liệu với sai số lớn.
- **PM2.5**: Sai số cao nhất trong tất cả các tham số. Đặc biệt RMSE lớn cho thấy mô hình không dự đoán tốt PM2.5. Điều này có thể do PM2.5 phụ thuộc nhiều yếu tố phi tuyến phức tạp.
- **Kết luận chung:**
 - Random Forest hoạt động tốt với **CO** và **NO2**, cho thấy khả năng học các đặc trưng phi tuyến và biến động phức tạp trong dữ liệu.
 - Dự đoán chưa tốt đối với **PM2.5** và **Ozone**, đặc biệt với RMSE cao, cho thấy sai số lớn từ một số dự đoán.

- Mô hình gặp khó khăn trong việc nắm bắt đặc điểm của các thông số phức tạp hơn như PM2.5. Điều này có thể được giải thích do số lượng dòng dữ liệu quá lớn và được đo đạc tại những địa điểm cách rất xa nhau, dẫn đến dữ liệu trở nên phức tạp và khó để nắm bắt được chu kỳ biến động.

8. Tài liệu tham khảo

- [1] [SSIS Tutorial](#)
- [2] Slide hướng dẫn seminar#1, 2
- [3] [SSIS Tutorial — from Basics to Advanced Development](#)
- [4] [Dimensional Data Modeling - GeeksforGeeks](#)
- [5] [What is Dimensional Modeling in Data Warehouse? Learn Types](#)
- [6] Slide ENG và VIET của lớp lý thuyết
- [7] Các video demo của các anh chị khóa trước
- [8] <https://www.codeproject.com/Articles/710387/Learn-to-Write-Custom-MDX-Query-First-Time>
- [9] Slide hướng dẫn thực hành
- [10] Slide bài giảng môn “Nhập môn Khoa học Dữ liệu” – TS. Nguyễn Ngọc Thảo