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

BÁO CÁO ĐỒ ÁN
Môn : HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH
Lớp : 21HTTT1

Nhóm 7
20127200 - Nguyễn Nam Khang
21127194 - Đỗ Anh Tuấn
21127483 - Nguyễn Ngọc Vũ
21127719 - Nguyễn Minh Tuấn

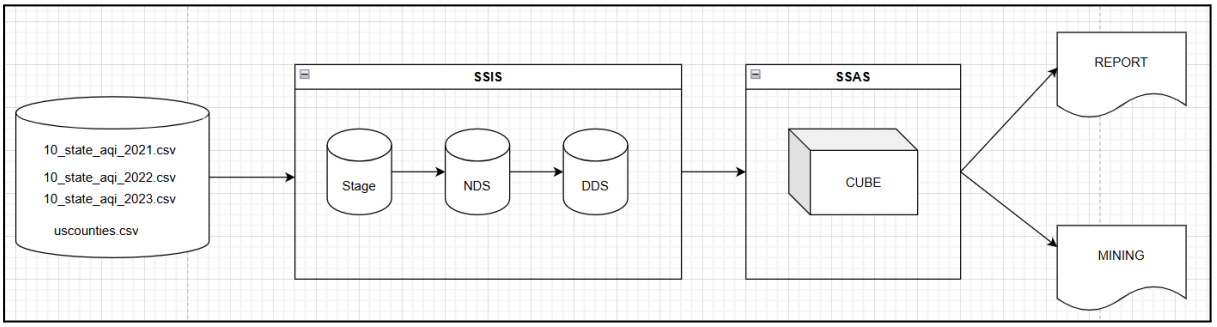
GVLT: Ths. Hồ Thị Hoàng Vy
GVTH: Ths. Tiết Gia Hồng
GVTG: Ths. Nguyễn Ngọc Minh Châu

HỌC KỲ I - NĂM HỌC 2024 - 2025

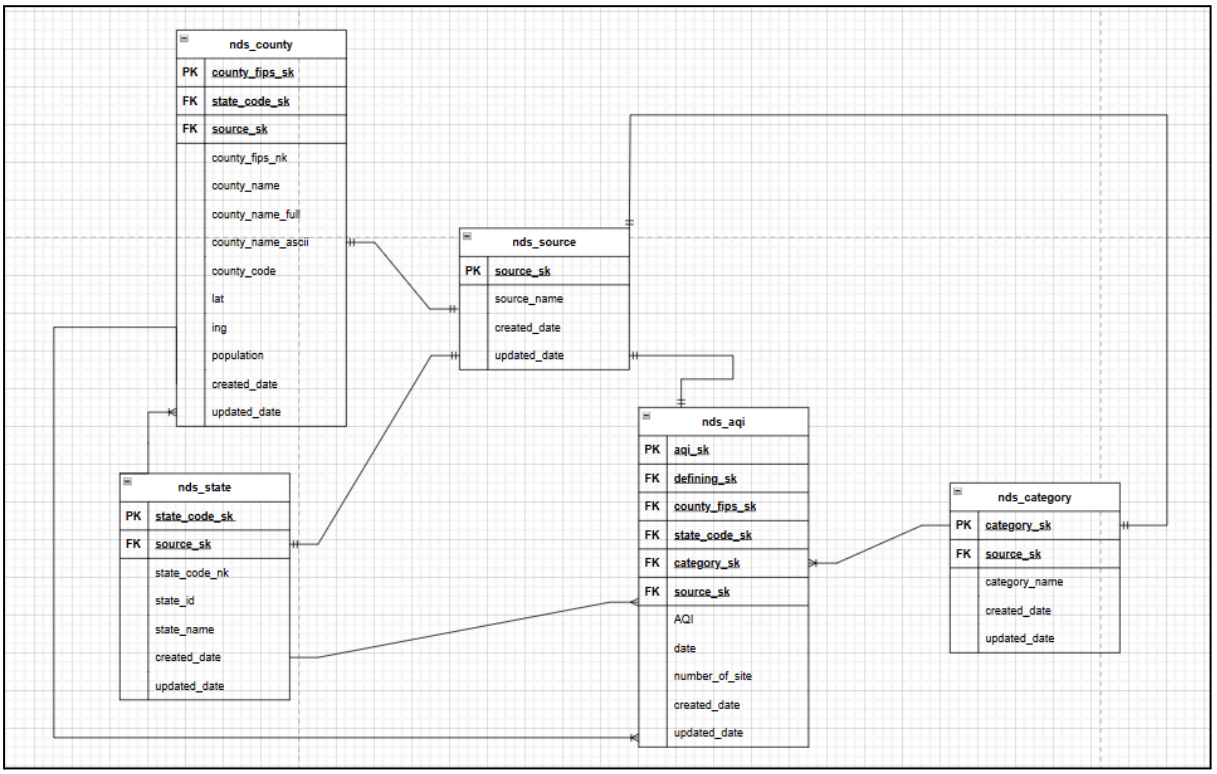
MỤC LỤC

1. Thiết kế Data Warehouse

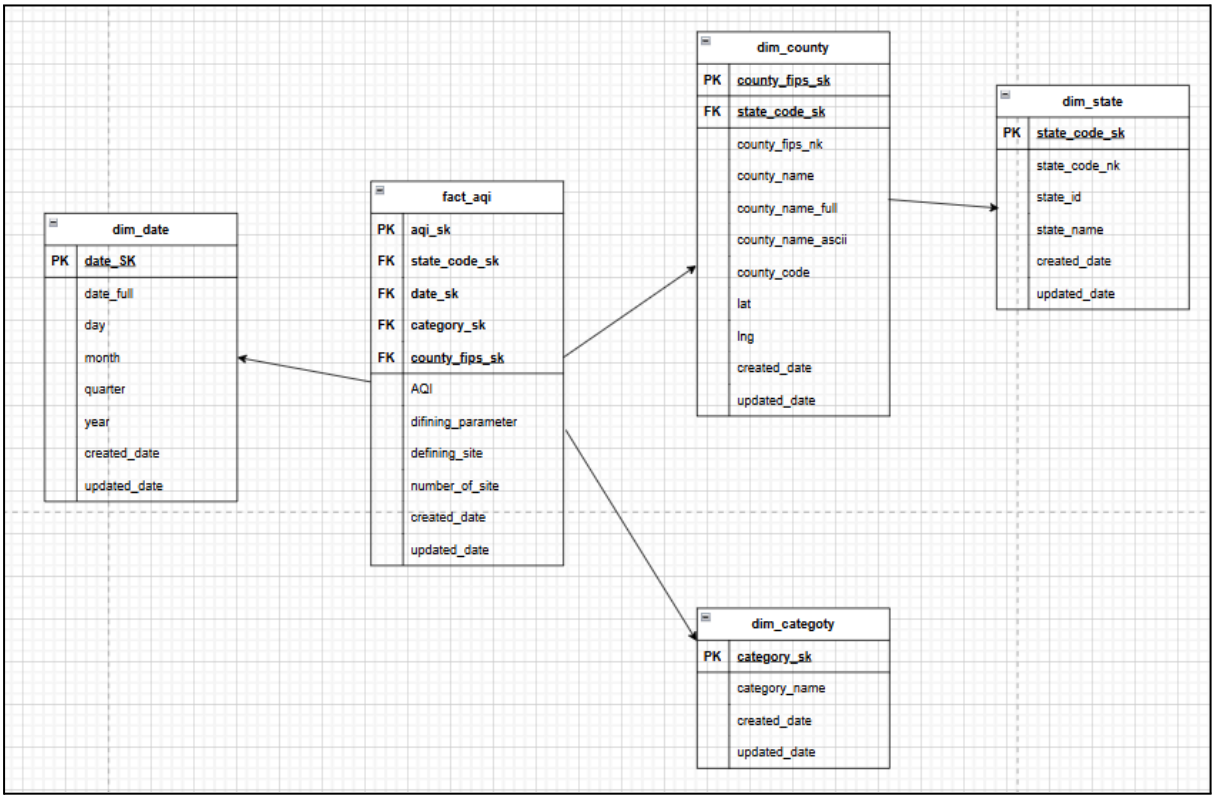
1.1. Thiết kế Data Flow



1.2. Thiết kế NDS



1.3. Thiết kế DDS



ProjectBI.drawio - draw.io

Sự kiện:

- Đo lường chất lượng không khí AQI tại các quận, bang ở Mỹ.

Bối cảnh sự kiện:

- **Ai**: Các bang và quận tại Mỹ.
- **Ở đâu**: Quận và bang cụ thể (State > County).
- **Cái gì**: Chỉ số AQI, thông số định nghĩa chất lượng không khí.
- **Khi nào**: Thời gian đo AQI (Năm > Quý > Tháng > Ngày).

Đo lường:

- Giá trị AQI (AQI), số lượng trạm đo (number_of_site), và thông số định nghĩa (defining_parameter).

Fact Table:

- **fact_aqi**: Chứa các giá trị đo lường AQI, bao gồm các thông số cần phân tích như min, max, mean, std của AQI.

Dimension Table:

- **dim_date**:
  - Phân chia chiều thời gian thành năm, quý, tháng, ngày.

    - Hỗ trợ phân tích xu hướng AQI qua các giai đoạn.
- **dim_state**:
    - Lưu thông tin về bang, cho phép phân tích AQI ở cấp độ bang.
- **dim_county**:
    - Lưu thông tin về quận và thuộc tính địa lý, hỗ trợ phân tích ở cấp quận.
- **dim_category**:
    - Xác định các hạng mục chất lượng không khí (Good, Moderate, etc.), phục vụ báo cáo số ngày thuộc từng hạng mục.

Phân chiều:

- **dim_date**: Năm > Quý > Tháng > Ngày.
- **dim_state** và **dim_county**: State > County.
- **dim_category**: Các hạng mục chất lượng không khí.

Độ chi tiết:

- Một dòng trong bảng fact tương ứng với một phép đo AQI tại một quận vào một ngày.

2. ETL

2.1. Tạo Metadata

2.1.1. Tạo Metadata cho quá trình từ Source to Stage

- Tại quá trình này, tiến hành tạo data_flow để có thể theo dõi và thực hiện incremental extract

```
create table data_flow (
id int not null identity(1,1),
name  varchar(20) not null,
LSET  datetime,
CET   datetime,
constraint pk_data_flow
primary key (id)
)
```

- Tiến hành insert các dòng lưu giá trị lset, cet của các stage mà mình đã thiết kế

```
insert into data_flow (name, LSET, CET) values ('10_state_AQI_2021','2007-12-01 03:00:00', null)
insert into data_flow (name, LSET, CET) values ('10_state_AQI_2022','2007-12-01 03:00:00', null)
insert into data_flow (name, LSET, CET) values ('10_state_AQI_2023','2007-12-01 03:00:00', null)
insert into data_flow (name, LSET, CET) values ('uscounties','2007-12-01 03:00:00', null)
```

2.1.2. Tạo Metadata cho quá trình từ NDS to DDS

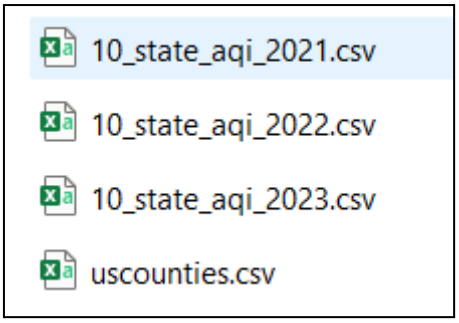
- Tại quá trình này, quá trình cũng như lần tại data_flow trước. Mục đích là để theo dõi quá trình từ NDS to DDS và thực hiện Incremental Extract từ NDS to DDS

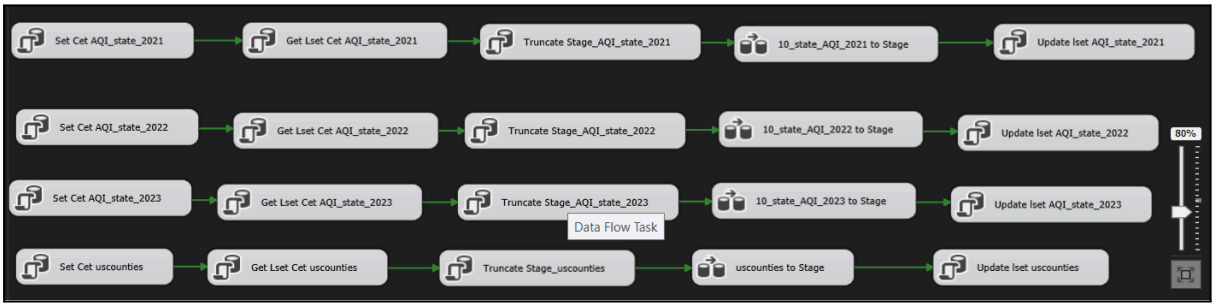
```
create table nds_data_flow (
id int not null identity(1,1),
table_name  varchar(20) not null,
LSET  datetime,
CET   datetime,
constraint pk_nds_data_flow
primary key (id)
)
```

- Tiến hành insert các dòng lưu giá trị lset, cet cho các bảng NDS mình đã thiết kế

```
insert into nds_data_flow (table_name, LSET, CET) values ('nds_source','2007-12-01 03:00:00', null)
insert into nds_data_flow (table_name, LSET, CET) values ('nds_state','2007-12-01 03:00:00', null)
insert into nds_data_flow (table_name, LSET, CET) values ('nds_county','2007-12-01 03:00:00', null)
insert into nds_data_flow (table_name, LSET, CET) values ('nds_aqi','2007-12-01 03:00:00', null)
insert into nds_data_flow (table_name, LSET, CET) values ('nds_category','2007-12-01 03:00:00', null)
```

2.2. Source to Stage

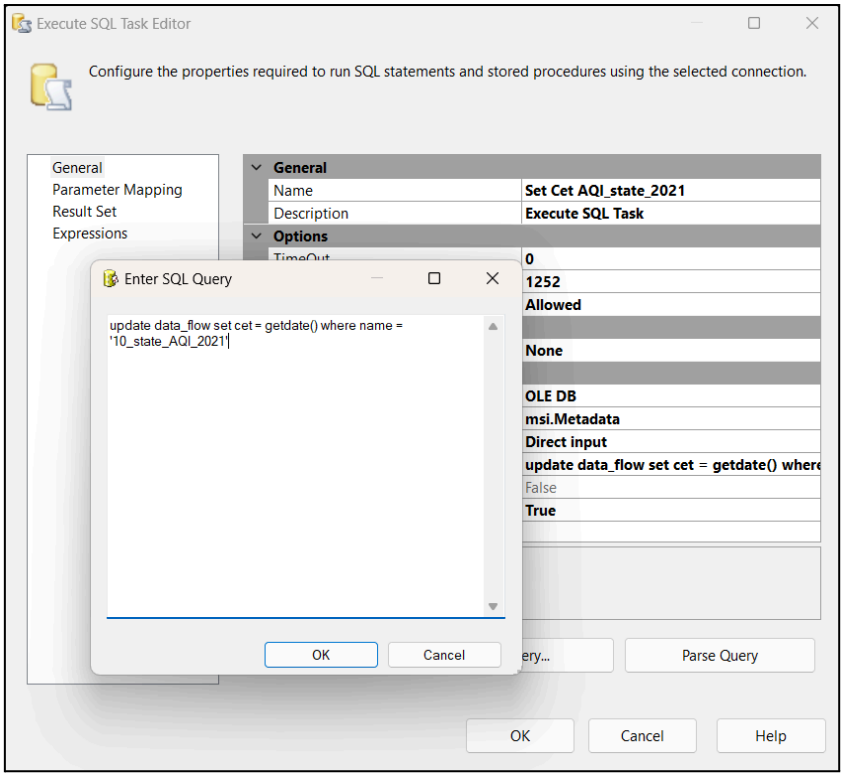




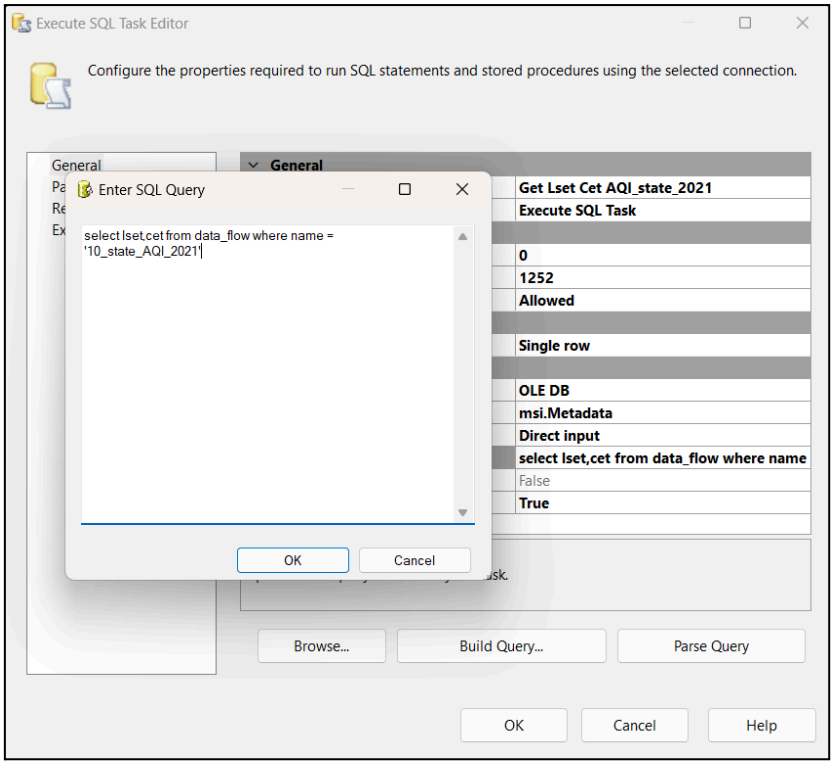
- Cách làm là chúng ta sẽ thực hiện đổ dữ liệu của từng source vào từng stage dành riêng cho source đó để đảm bảo quy tắc stage là bản lưu tạm của nguồn
    + Source 10_state_aqi_2021.csv đổ vào Stage_AQI_State_2021
    + Source 10_state_aqi_2022.csv đồ vào Stage_AQI_State_2022
    + Source 10_state_aqi_2023.csv đổ vào Stage_AQI_State_ 2023
    + Source uscounties.csv đổ vào Stage_uscounties

2.2.1. Xử lý dữ liệu từ nguồn 10_state_aqi_2021.csv váo stage

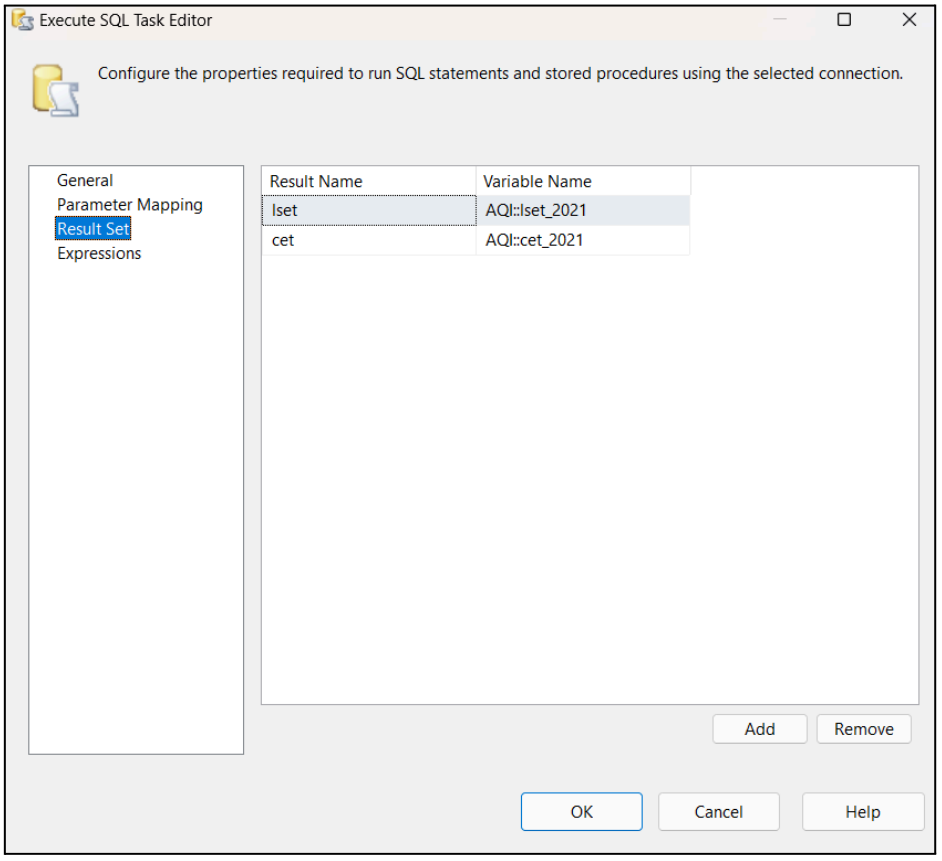
- Đầu tiên , update cet=getdate() trong metadata



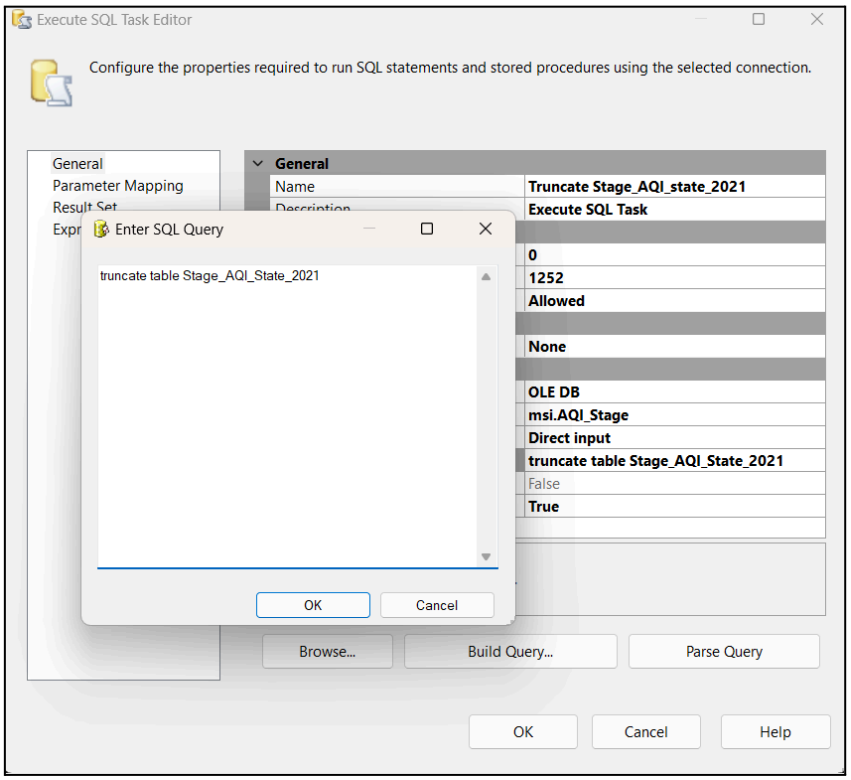
- Tiếp theo, Select lset cet để sử dụng cho việc thực hiện Incremental Extract



- Lưu trữ lset ,cet vào 2 variable. Ở đây tiến hành tạo tên biến là theo format là AQI::lset_2021, AQI::cet_2021 vì là từ source năm 2021. Tương tự như vậy cho các source sau

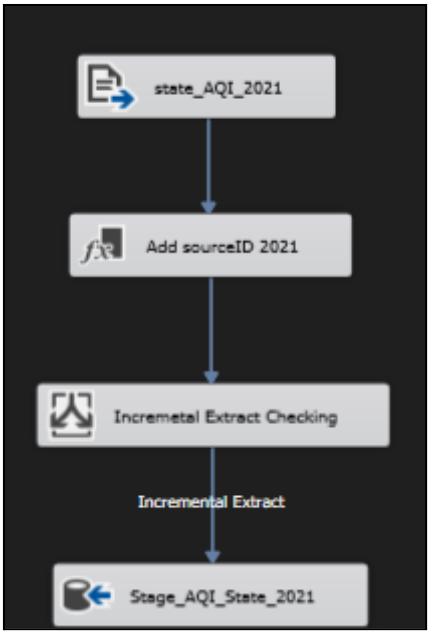


- Thực hiện truncate table để đảm bảo stage sẽ trống rỗng và không chứa dữ liệu lỗi hoặc cũ từ quá trình ETL trước đó



- Thực hiện quá trình đổ dữ liệu từ source 10_state_aqi_2021 vào stage

+ Kết nối dữ liệu với nguồn
+ Thêm cột sourceID (2021)
+ Thực hiện kiểm tra Incremental Extract
+ Mapping kết nối dữ liệu vào stage



- Cuối cùng là sẽ cập nhật lại giá trị cho lset cho bảng stage của source 10_state_aqi_2021 là getdate()

2.2.2. Xử lý dữ liệu từ nguồn 10_state_aqi_2022.csv váo Stage

- Quy trình nạp Source to Stage của nguồn 10_state_aqi_2022.csv giống như nguồn 10_state_aqi_2021.csv

2.2.3. Xử lý dữ liệu từ nguồn 10_state_aqi_2023.csv váo Stage

- Các bước như :
    + Set cet
    + Select lset, cet
    + Truncate table

  => Các bước này thực hiện y như lúc nạp từ source năm 2021
- Tới bước nạp dữ liệu vào Stage chúng ta sẽ gặp các vấn đề

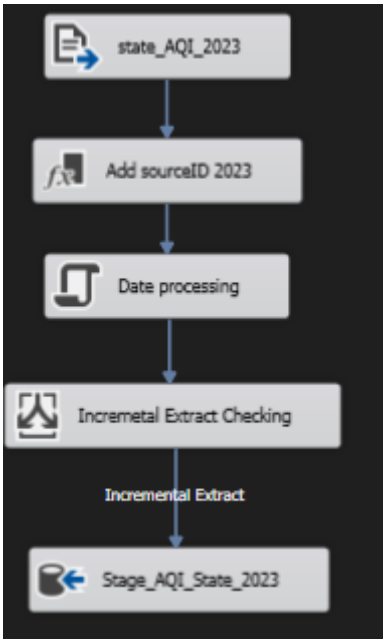
+ Giá trị ngày trong source năm 2023 không có chung định dạng

	State Nam	county Nar	State Code	County Co	Date	AQI	Category	Defining Pa	Defining Si	Number of	Created	Last Updated
2	Virginia	Fauquier	51	61	3/29/2022	7	Good	Ozone	51-061-00	1	3/29/2022 16:30	12/31/2022 23:54
3	Virginia	Fauquier	51	61	3/30/2022	19	Good	Ozone	51-061-00	1	3/30/2022 16:30	12/31/2022 23:54
4	Alabama	Sumter	1	119	6/12/2023	7	Good	PM2.5	01-119-00	1	12/6/2023 16:30	12/31/2023 23:42
5	Alabama	Sumter	1	119	7/12/2023	8	Good	PM2.5	01-119-00	1	12/7/2023 16:30	12/31/2023 23:42
6	Alabama	Sumter	1	119	8/12/2023	11	Good	PM2.5	01-119-00	1	12/8/2023 16:30	12/31/2023 23:42
7	Alabama	Sumter	1	119	9/12/2023	4	Good	PM2.5	01-119-00	1	12/9/2023 16:30	12/31/2023 23:42
8	Alabama	Sumter	1	119	10/12/2023	4	Good	PM2.5	01-119-00	1	12/10/2023 16:30	12/31/2023 23:42
9	Alabama	Sumter	1	119	11/12/2023	7	Good	PM2.5	01-119-00	1	12/11/2023 16:30	12/31/2023 23:42
10	Alabama	Sumter	1	119	12/12/2023	13	Good	PM2.5	01-119-00	1	12/12/2023 16:30	12/31/2023 23:42
11	Alabama	Sumter	1	119	13-12-2023	51	Good	PM2.5	01-119-00	1	12/13/2023 16:30	12/31/2023 23:42
12	Alabama	Sumter	1	119	14-12-2023	14	Good	PM2.5	01-119-00	1	12/14/2023 16:30	12/31/2023 23:42
13	Alabama	Sumter	1	119	15-12-2023	26	Good	PM2.5	01-119-00	1	12/15/2023 16:30	12/31/2023 23:42
14	Alabama	Sumter	1	119	16-12-2023	12	Good	PM2.5	01-119-00	1	12/16/2023 16:30	12/31/2023 23:42

```
1  State Name,county Name,State Code,County Code,Date,AQI,Category,Defining Parameter,Defining Site,Number of Sites Reporting,Created,Last Updated
2  Virginia,Fauquier,51,61,2022-03-29,7,Good,Ozone,51-061-0002,1,2022-03-29 16:30:00,2022-12-31 23:54:00
3  Virginia,Fauquier,51,61,2022-03-30,19,Good,Ozone,51-061-0002,1,2022-03-30 16:30:00,2022-12-31 23:54:00
4  Alabama,Sumter,1,119,06-12-2023,7,Good,PM2.5,01-119-0003,1,2023-12-06 16:30:00,2023-12-31 23:42:00
5  Alabama,Sumter,1,119,07-12-2023,8,Good,PM2.5,01-119-0003,1,2023-12-07 16:30:00,2023-12-31 23:42:00
6  Alabama,Sumter,1,119,08-12-2023,11,Good,PM2.5,01-119-0003,1,2023-12-08 16:30:00,2023-12-31 23:42:00
```

- Thực hiện quá trình đổ dữ liệu từ source 10_state_aqi_2021 vào stage

+ Kết nối dữ liệu với nguồn
+ Thêm cột sourceID (2021)
+ Tạo một Script Component để thực hiện chuyển cột Date trong Source về cùng một định dạng
+ Thực hiện kiểm tra Incremental Extract
+ Mapping kết nối dữ liệu vào Stage




```
2 references
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    string InputDate = Row.Date; // mm/dd/yyyy

    string[] dateParts = InputDate.Split('-');

    // Xu ly nam/thang/ngay =>
    if (dateParts[0].Length == 4 ) {
        string formattedYear = dateParts[0];
        string formattedMonth = dateParts[1];
        string formattedDate = dateParts[2];

        Row.ProcessedDate = formattedDate + '/' + formattedMonth + '/' + formattedYear;

        return;
    }

    // Xu ly ngay-thang-nam => thang/ngay/nam
    if (dateParts.Length == 3 ) {

        string formattedDate = dateParts[0].PadLeft(2, '0') ; // Day
        string formattedMonth = dateParts[1].PadLeft(2, '0');
        string formattedYear = dateParts[2];


        Row.ProcessedDate = formattedDate + '/' + formattedMonth + '/' + formattedYear;

    }
    else
    {
        Row.ProcessedDate = InputDate;
    }

}
```

- Cuối cùng là sẽ cập nhật lại giá trị cho lset cho bảng Stage của Source 10_state_aqi_2023 là getdate()

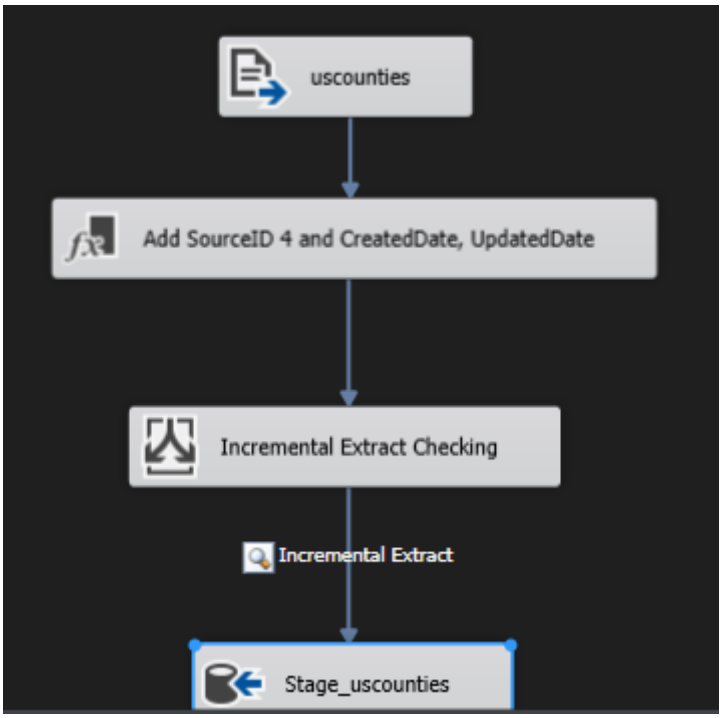
2.2.4. Xử lý dữ liệu từ nguồn uscounties.csv vào Stage

- Các bước như :
    + Các bước như :
    + Set cet
    + Select lset, cet
    + Truncate table

=> Các bước này thực hiện y như lúc nạp từ source năm 2021

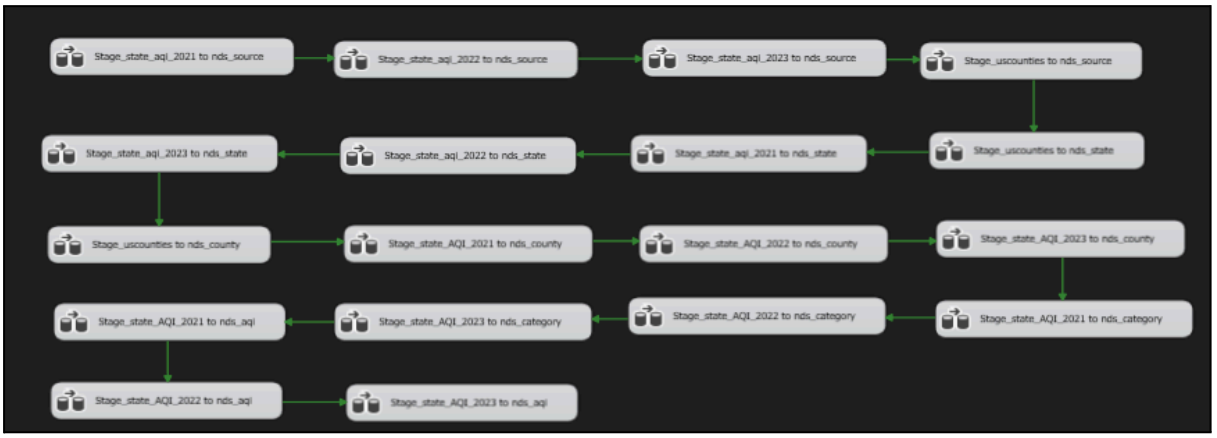
- Thực hiện quá trình đổ dữ liệu từ source uscounties vào stage

+ Kết nối dữ liệu với nguồn
+ Thêm cột sourceID (4)
+ Thêm cột createdDate và updatedDate vì 2 source uscounties không có cột createdData và updatedDate
+ Mapping kết nối dữ liệu vào Stage

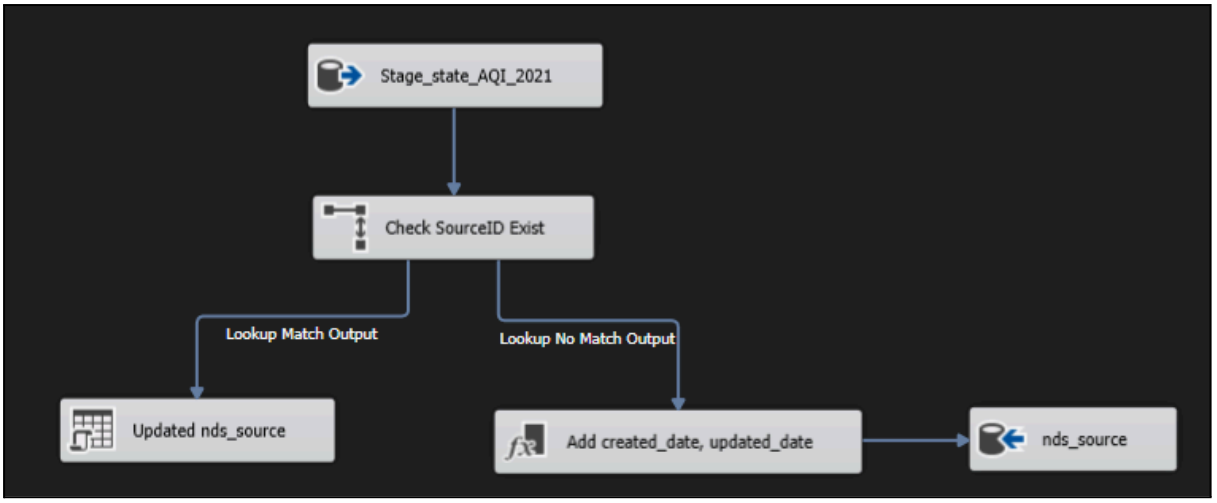


- Cuối cùng là sẽ cập nhật lại giá trị cho lset cho bảng stage của source uscounties là getdate()

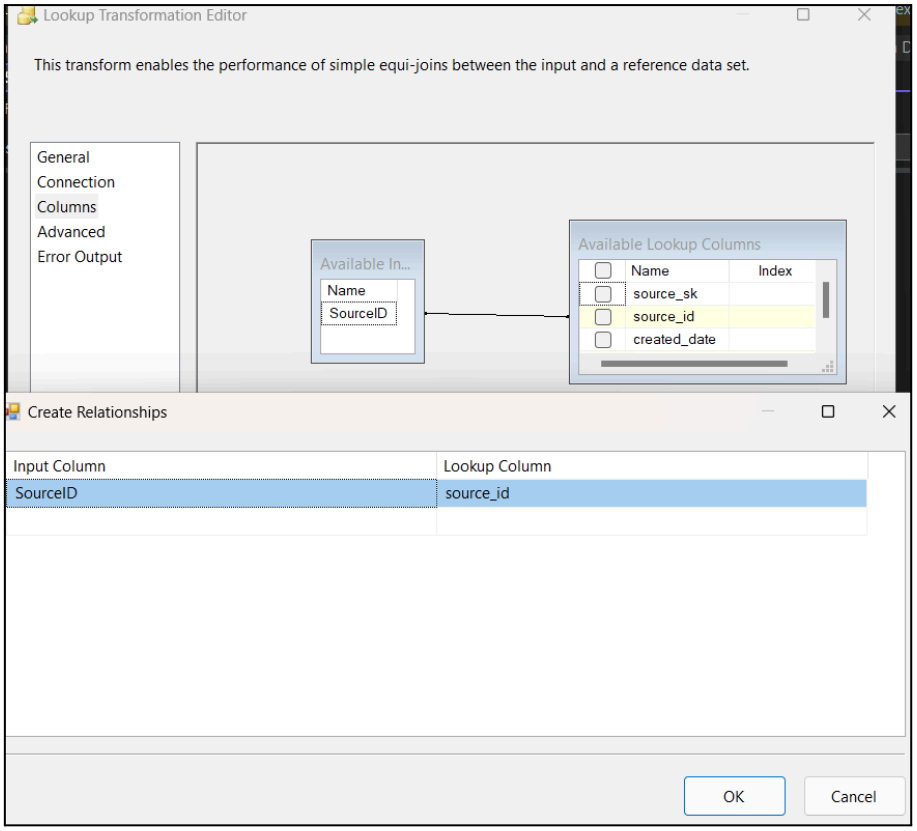
2.3. Stage to NDS



2.3.1. Đổ dữ liệu vào bảng nds_source

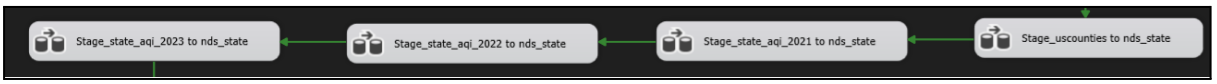


- Kết nối dữ liệu bảng Stage_state_AQI_2021
- Tiến hành kiểm tra xem đã tồn tại giá trị của sourceID chưa.

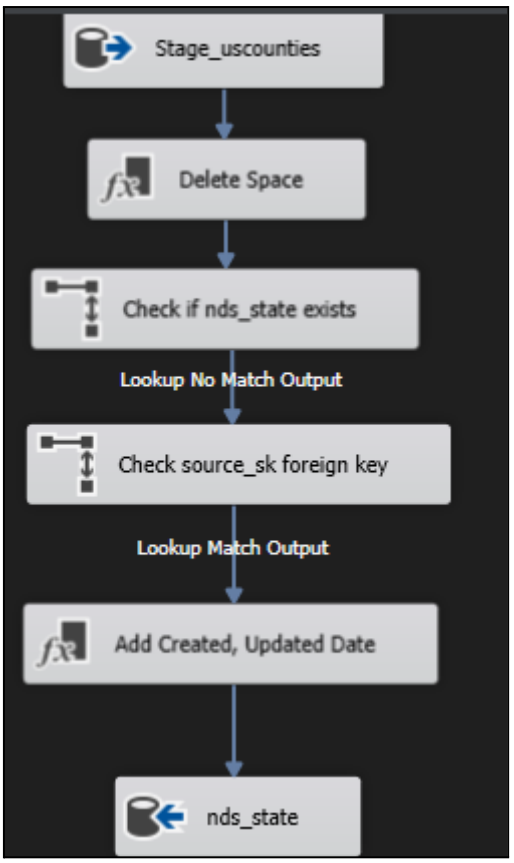


- Nếu có rồi thì cập nhật, chưa có thì tạo mới

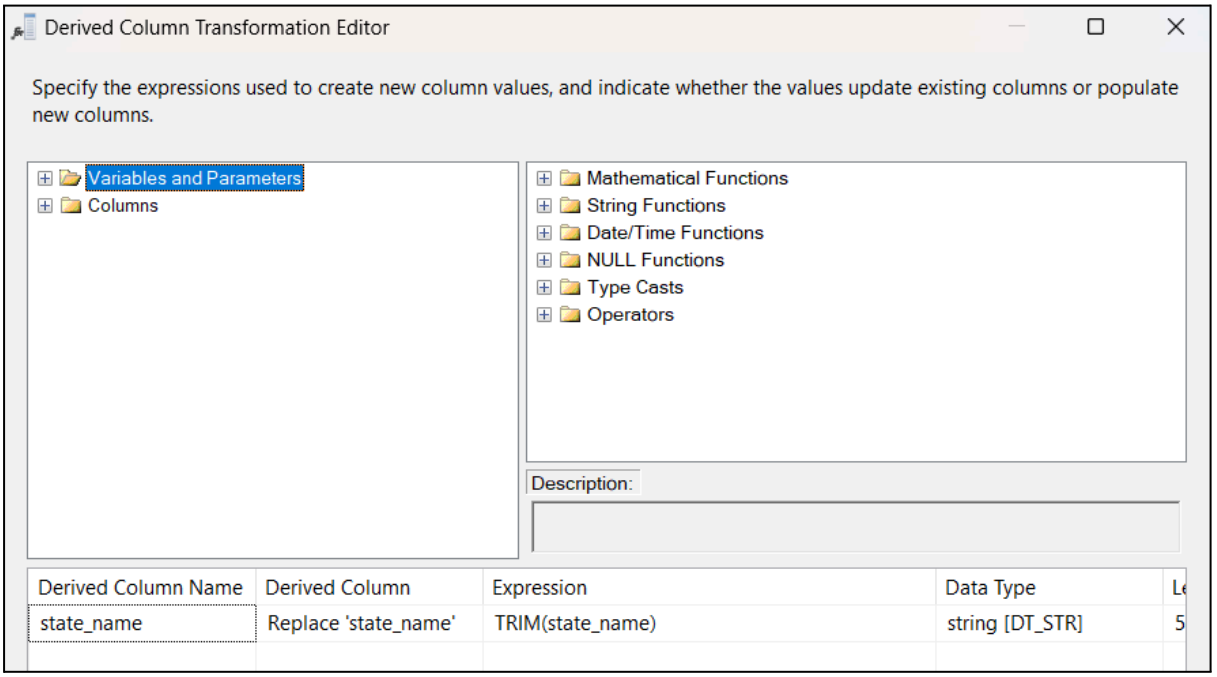
2.3.2. Đổ dữ liệu vào bảng nds_state

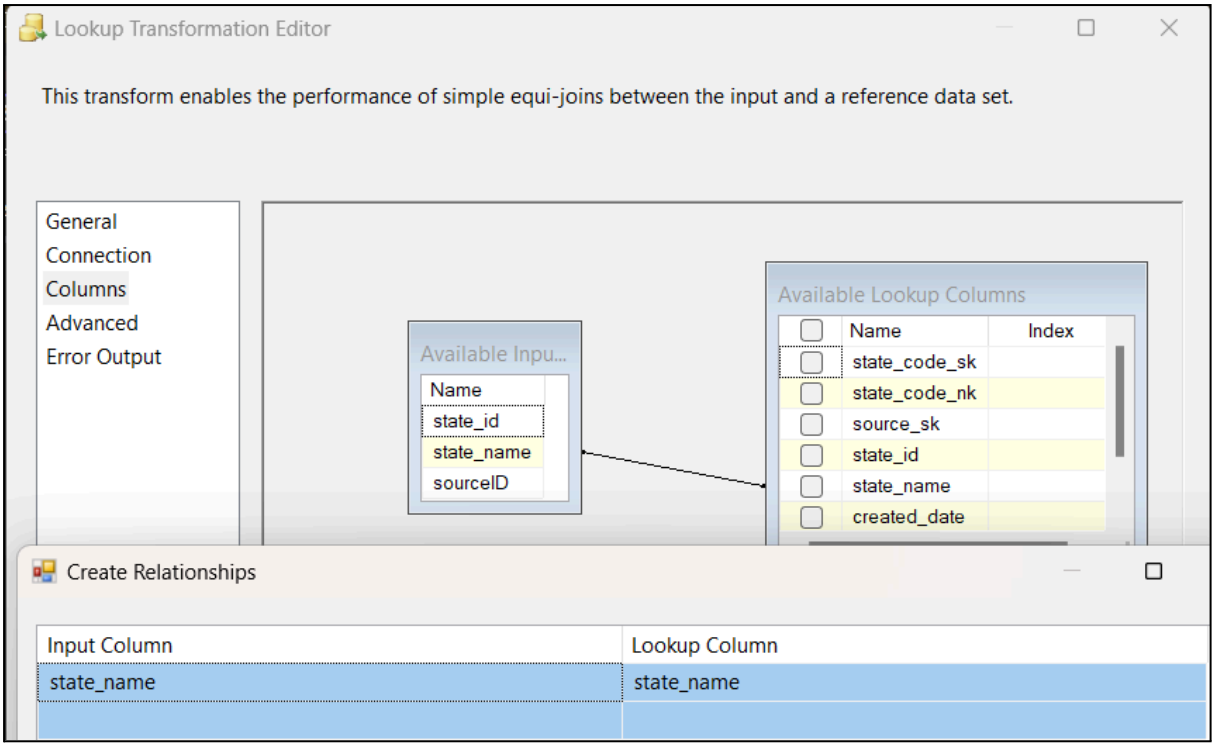


- Đầu tiên phải đổ dữ liệu từ uscounties vào trước vì uscounties có nhiều thông tin về state.

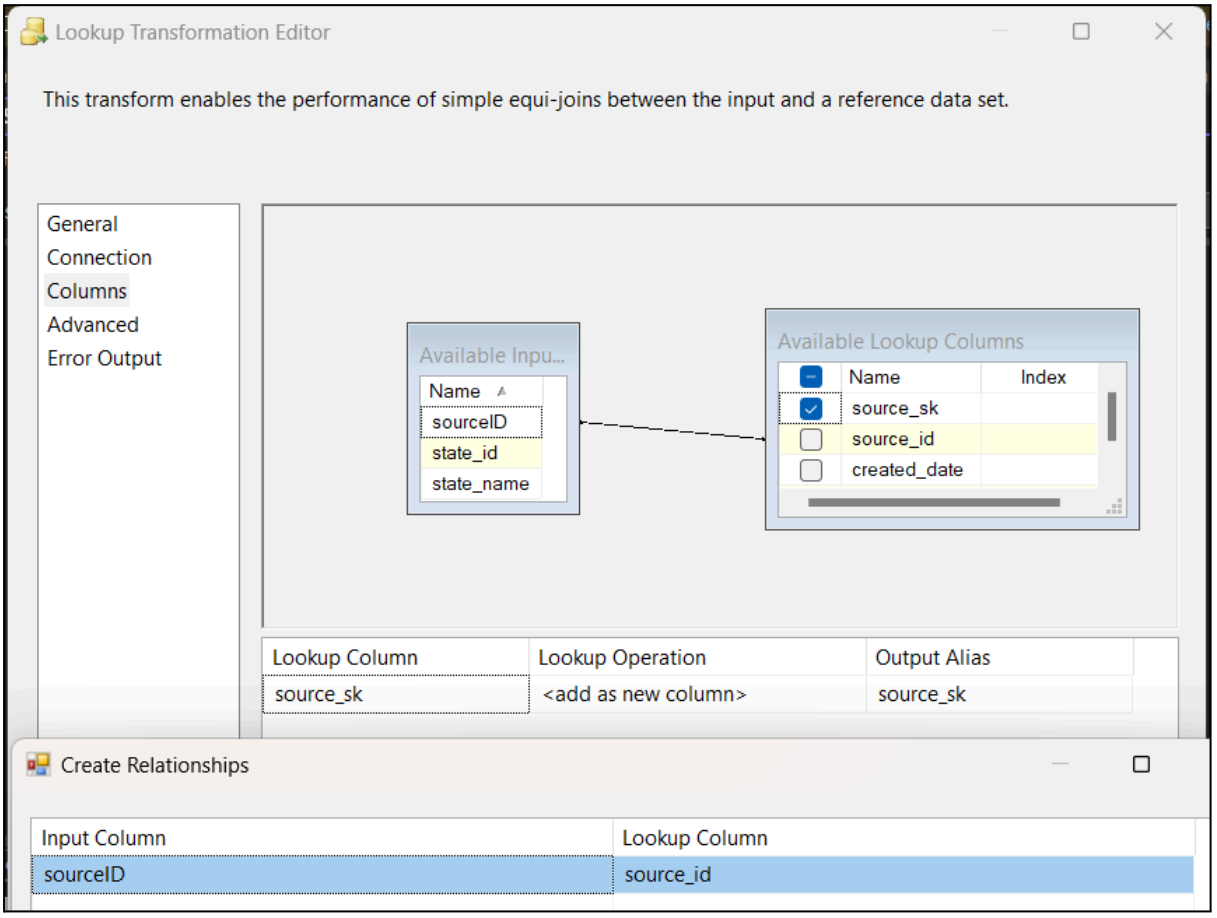


- Kết nối với lại Stage_uscounties
- Delete space là để xóa khoảng trắng của state_name vì sẽ tiến hành mapping dựa trên state_name để tiến hành kiểm tra xem đã tồn tại state đó trong nds_state hay chưa. Nếu chưa thì tiếp tục, nếu có rồi thì bỏ qua



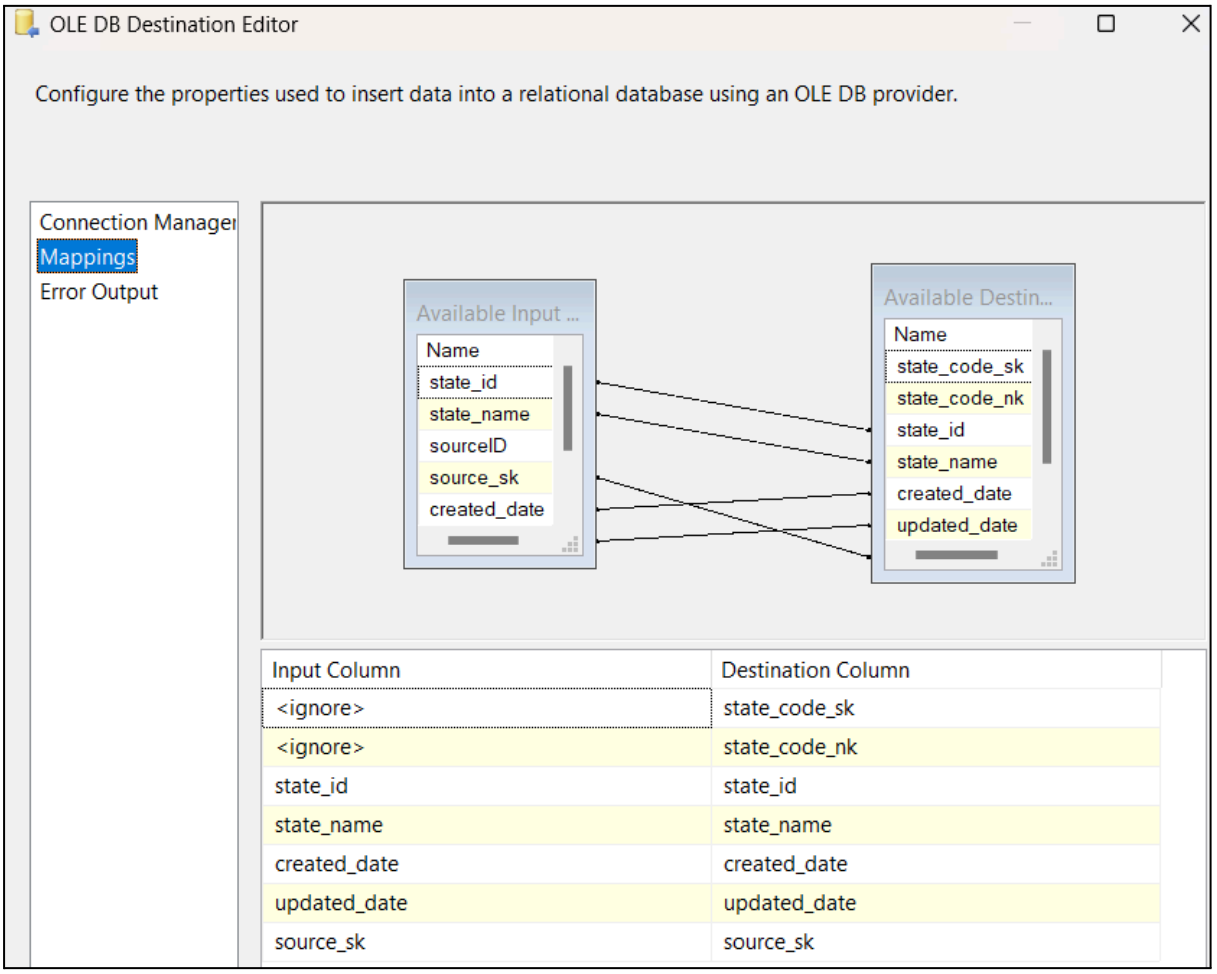


- Tiếp theo là mapping với lại nds_source dựa trên sourceID để lấy được surrogate key của bảng nds_source

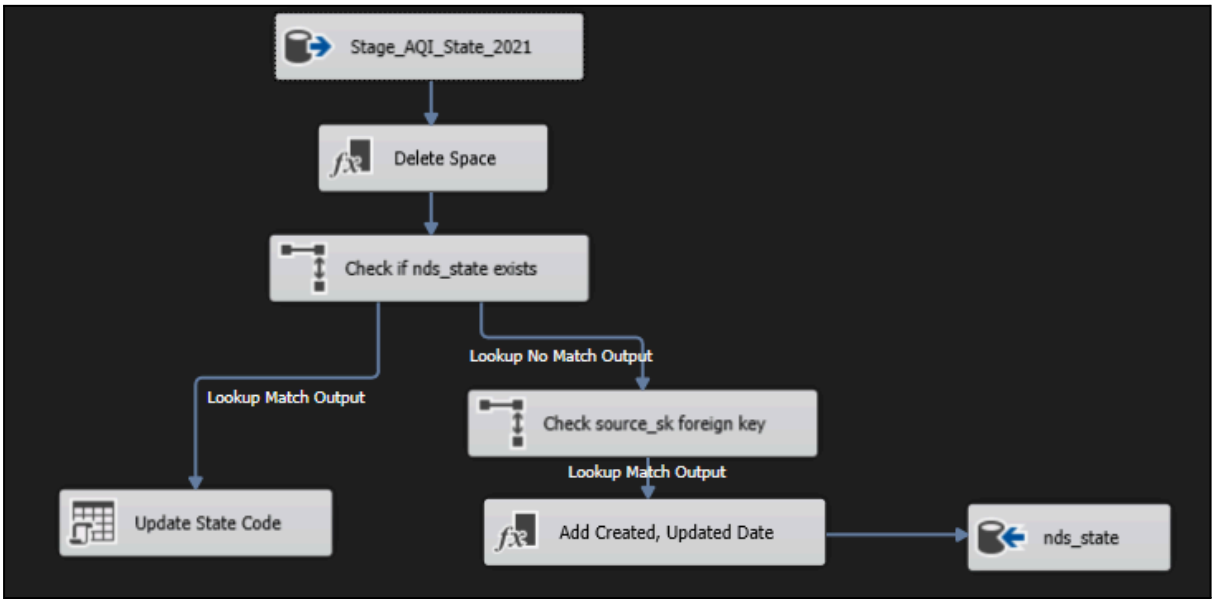


- Nếu có thì thêm tạo thêm cột created_date = getdate() và update_date = getdate()

- Mapping dữ liệu vào nds_state

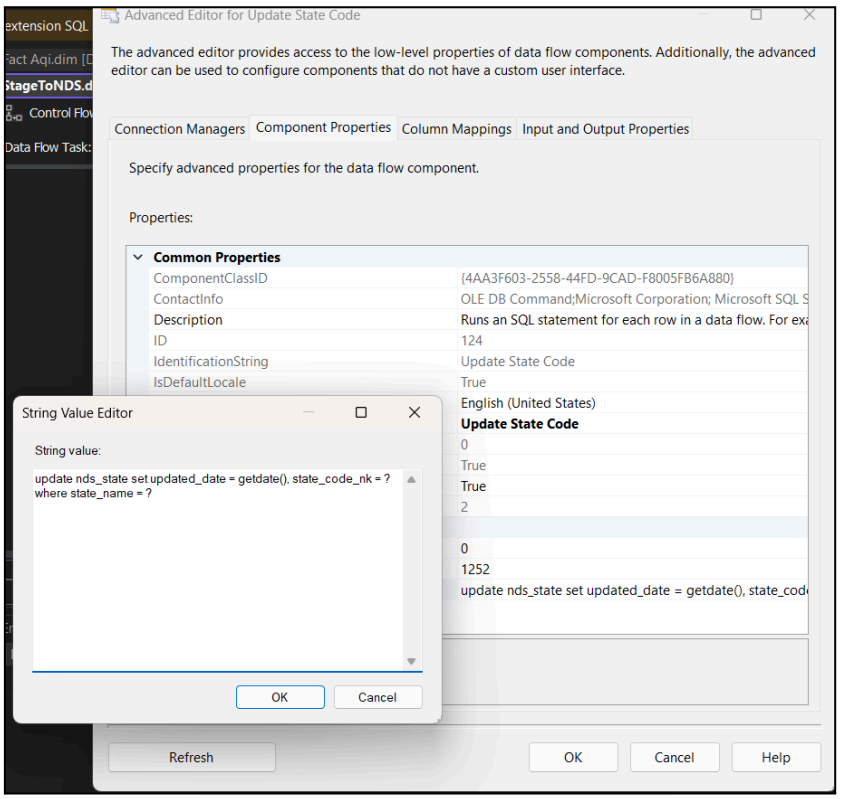


- Sau khi đổ dữ liệu từ Stage_uscounties xong thì tiếp theo sẽ đổ lần lượt từ các Stage_AQI_State_2021, Stage_AQI_State_2022, Stage_AQI_State_2023 (3 stage làm tương tự nhau)



- Kết nối dữ liệu tới Stage_AQI_State_2021

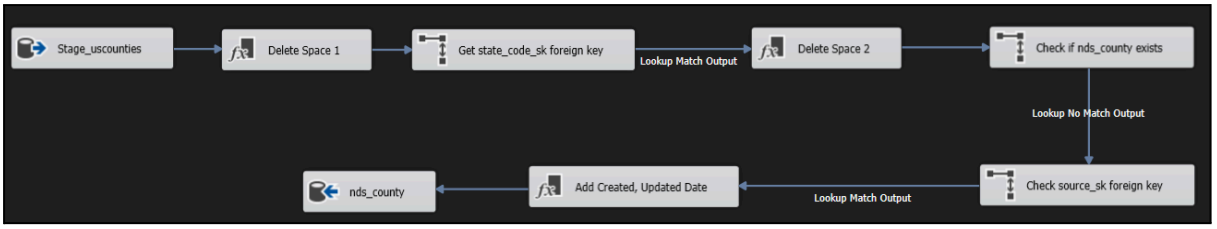
- Delete space là để xóa khoảng trắng của state_name vì sẽ tiến hành mapping dựa trên state_name để tiến hành kiểm tra xem đã tồn tại state đó trong nds_state hay chưa.
    + Nếu đã tồn tại thì cập nhật thêm state_code_nk



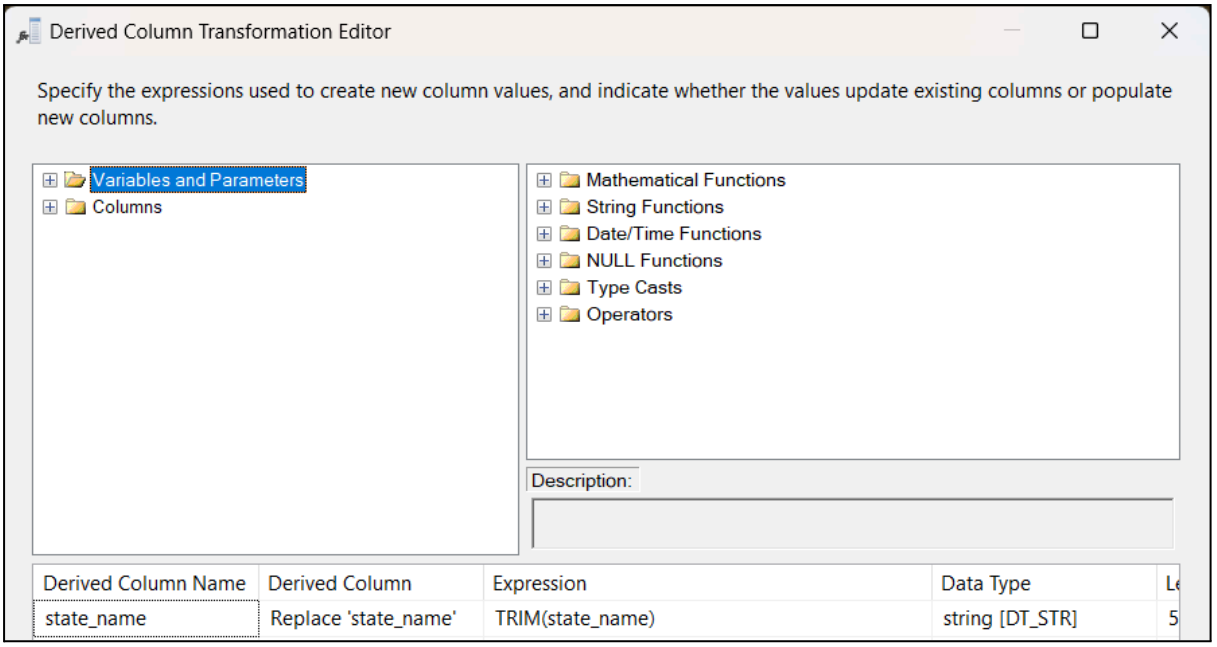
    + Nếu chưa tồn tại thì thì mapping với nds_source dựa trên sourceID để lấy surrogate_key khóa ngoại đến bảng nds_source. Sau đó thêm ngày tạo vá cập nhật rồi thêm mới vào nds_state

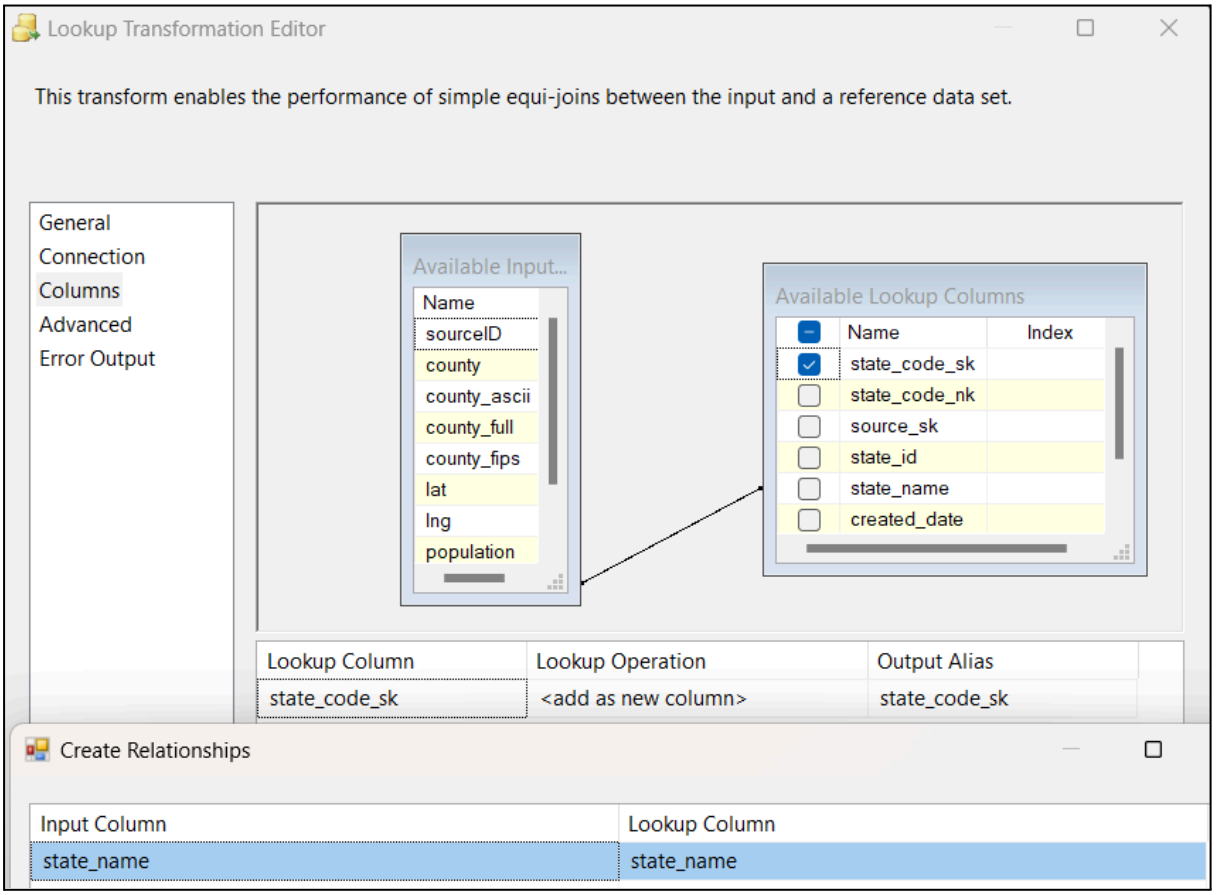
2.3.3. Đổ dữ liệu vào bảng nds_county

- Đầu tiên, chúng ta cũng tiến hành đổ từ bảng Stage_counties trước

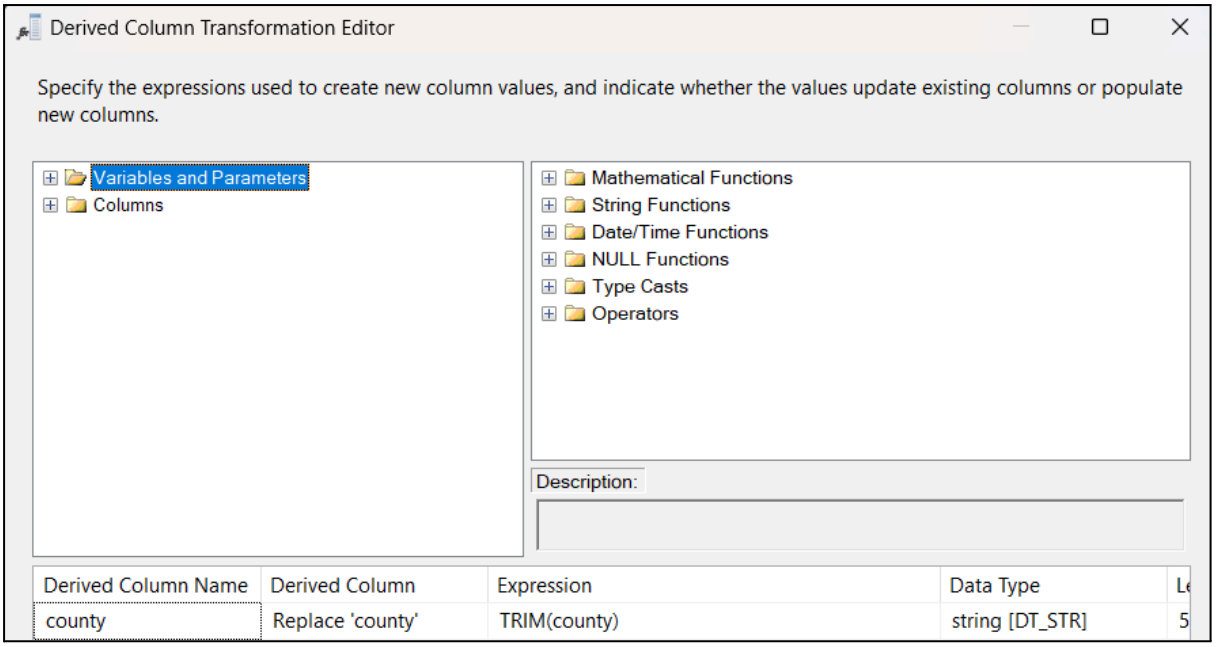


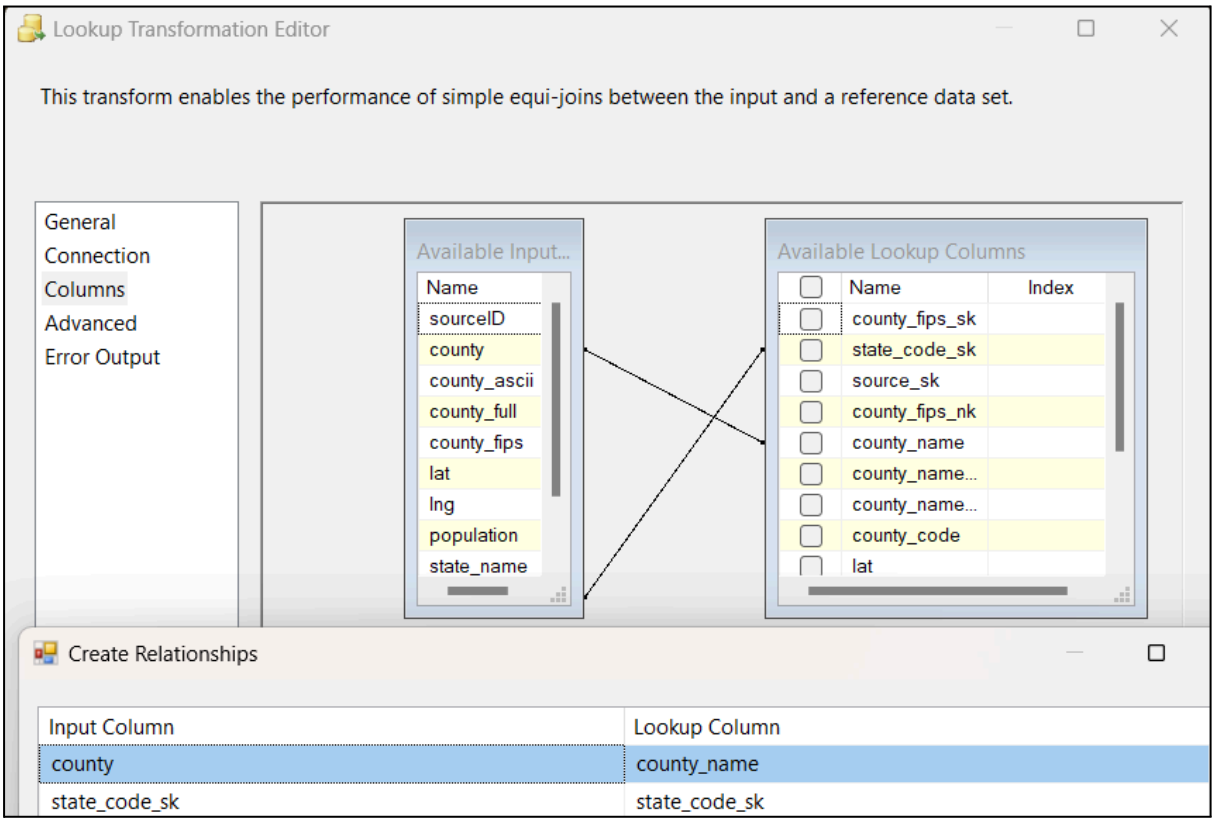
- Kết nối dữ liệu với lại Stage_uscounties
- Delete space 1 là để xóa khoảng trắng của state_name vì sẽ tiến hành mapping dựa trên state_name để lấy surrogate key của nds_state làm khóa ngoại trỏ đến bảng nds_state

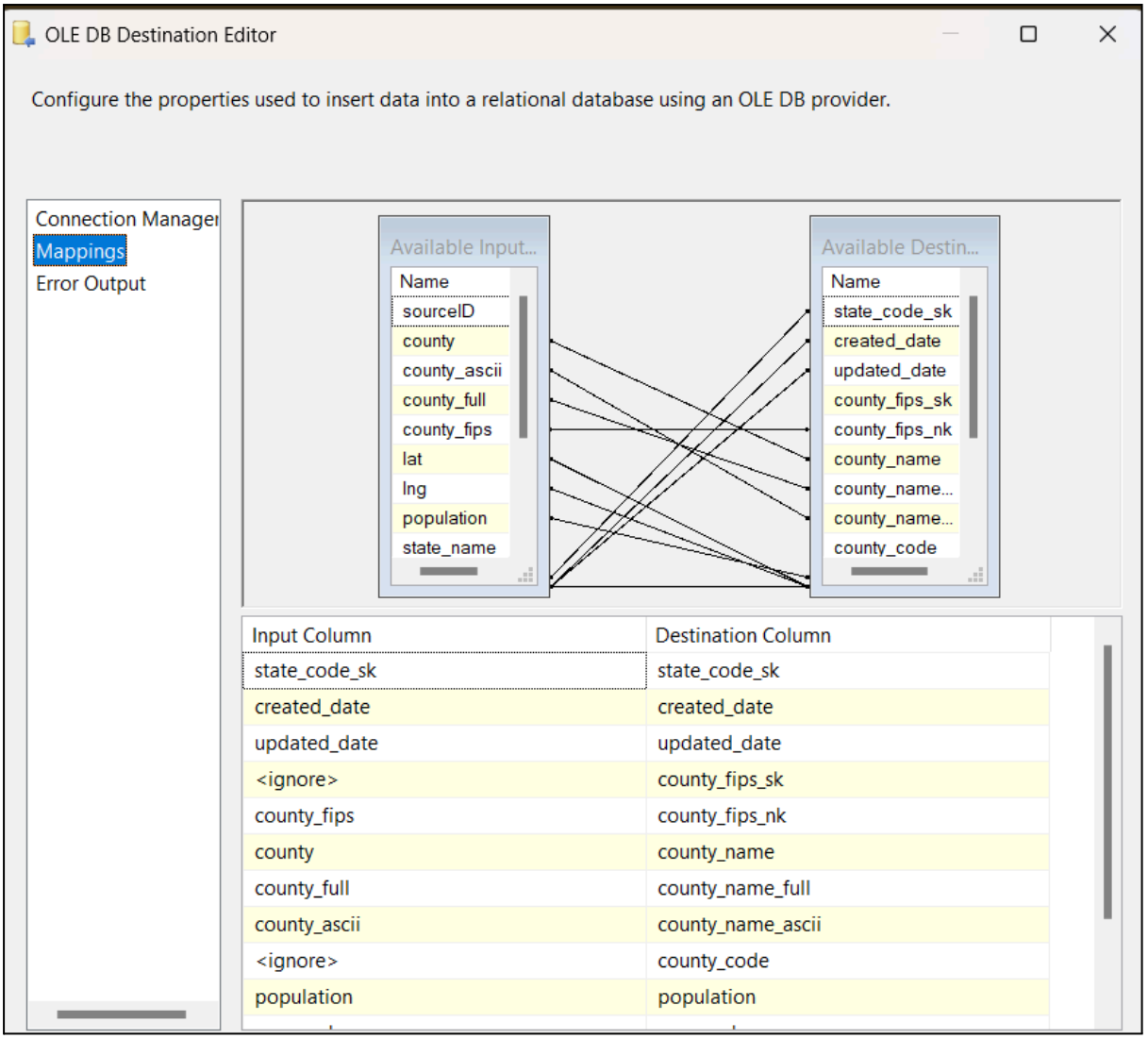




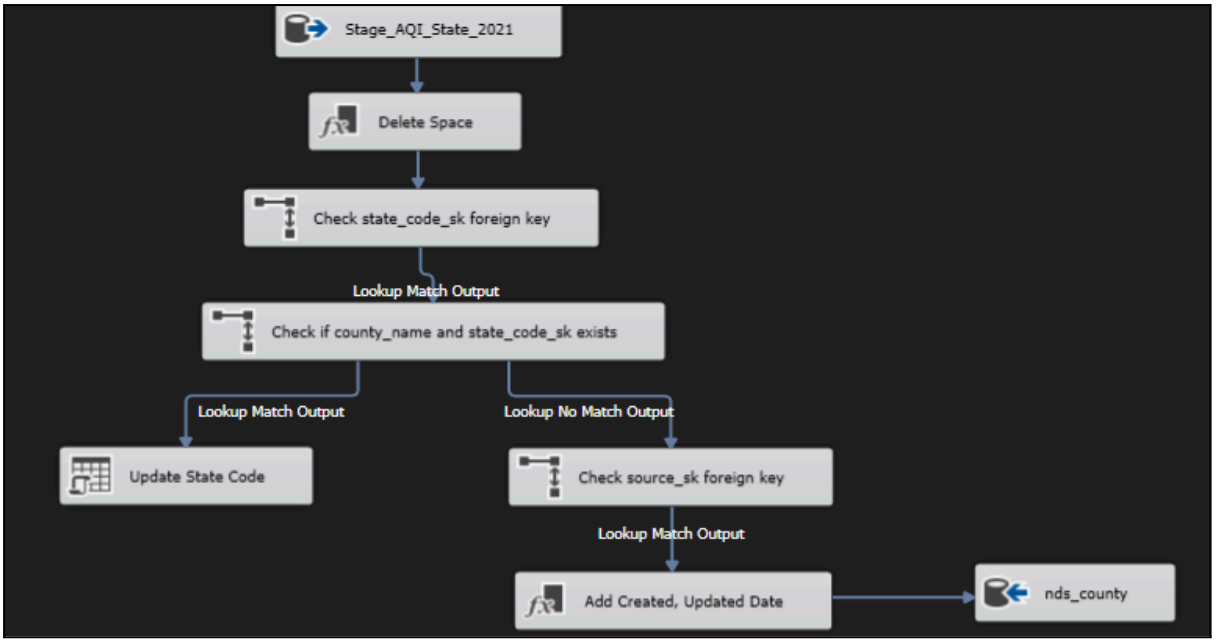
- Delete space 2 là để xóa khoảng trắng của county_name vì sẽ tiến hành mapping dựa trên county_name và state_code_sk (surrogate key của nds_state) để kiểm tra xem tại state đó, county đó đã tồn tại chưa.
    + Nếu có rồi thì bỏ qua
    + Nếu chưa tồn tại thì lấy khóa ngoại đến nds_source và thêm cột ngày tạo, ngày cập nhật và thêm mới dữ liệu


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
Columns
Mathematical Functions
String Functions
Date/Time Functions
NULL Functions
Type Casts
Operators
Description:
Derived Column Name
Derived Column
Expression
Data Type
county
Replace 'county'
TRIM(county)
string [DT_STR]

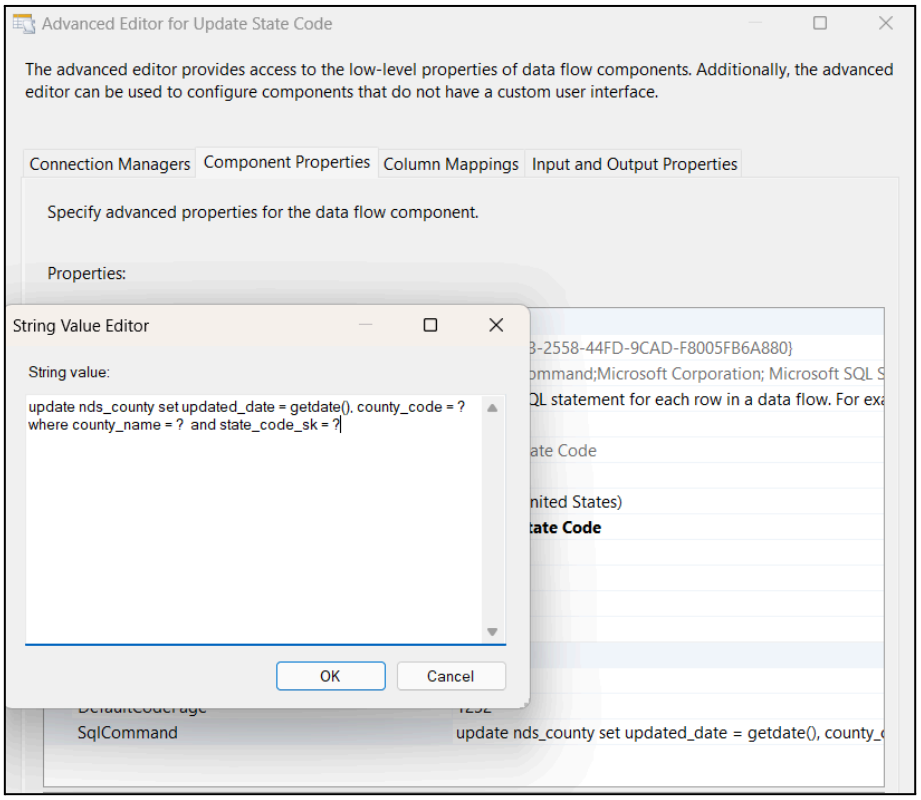

Lookup Transformation Editor
This transform enables the performance of simple equi-joins between the input and a reference data set.
General
Connection
Columns
Advanced
Error Output
Available Input...
Name
sourceID
county
county_ascii
county_full
county_fips
lat
lng
population
state_name
Available Lookup Columns
Name
Index
county_fips_sk
state_code_sk
source_sk
county_fips_nk
county_name
county_name...
county_name...
county_code
lat
Create Relationships
Input Column
Lookup Column
county
county_name
state_code_sk
state_code_sk



- Tiếp theo là sẽ lần lượt đổ dữ liệu từ Stage_AQI_State_2021, Stage_AQI_State_2022, Stage_AQI_State_2023 vào bảng nds_county (Các bảng này làm tương tự nhau)

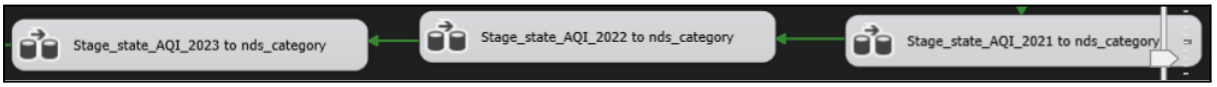


- Kết nối với lại Stage_AQI_State_2021
- Delete space tiến hành xóa khoảng trắng đầu và cuối của state_name và county_name
- Tiếp theo là sẽ lấy surrogate key của nds_state làm khóa ngoại của bản nds_county trỏ dến nds_state dựa trên mapping state_name
- Kiểm tra xem bang đó, county đó đã tồn tại chưa thông qua việc mapping county_name với lại state_code_sk
    + Nếu đã tồn tại rồi thì update county_code

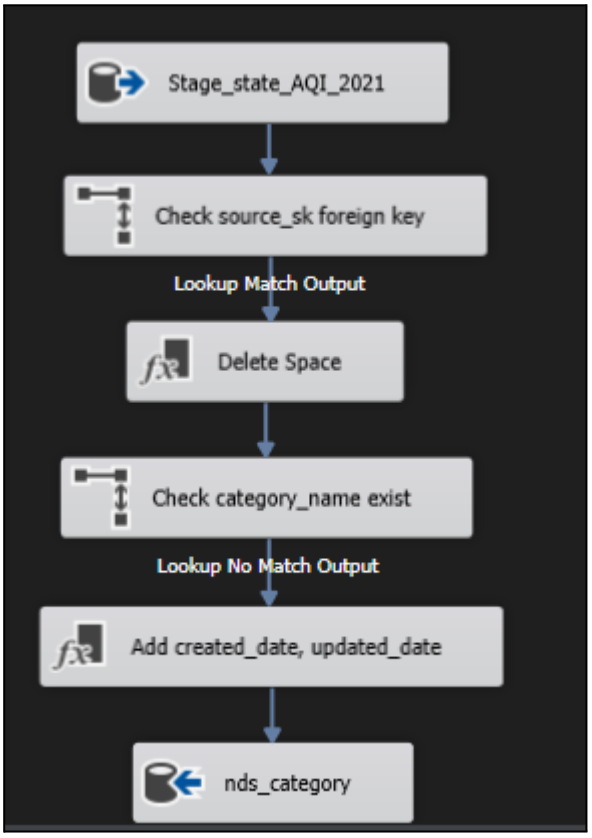


    + Nếu chưa tồn tại thì lấy khóa ngoại đến nds_source và thêm cột ngày tạo, ngày cập nhật và thêm mới dữ liệu

2.3.4. Đổ dữ liệu vào bảng nds_county



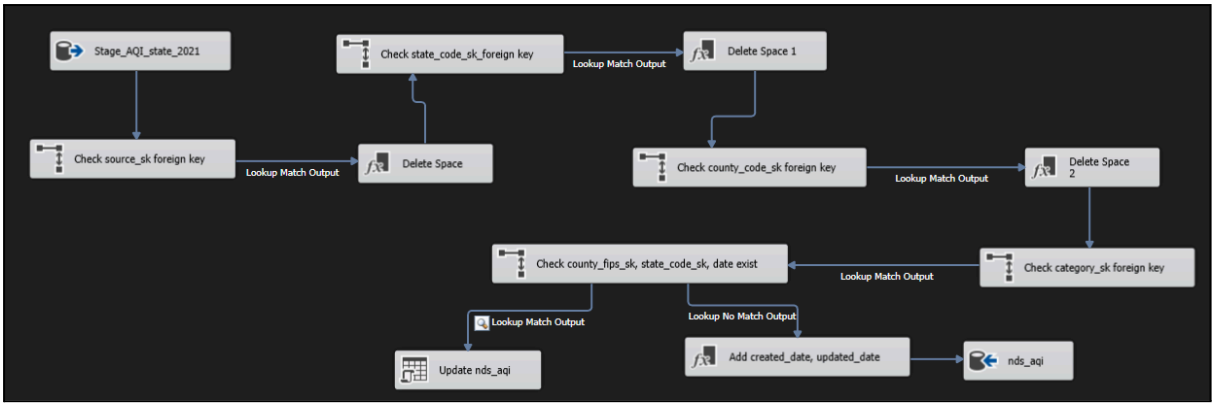
- Tiến hành đổ lần lượt dữ liệu từ Stage_AQI_State_2021, Stage_AQI_State_2022, Stage_AQI_State_2023 (3 bảng này làm tương tự nhau)



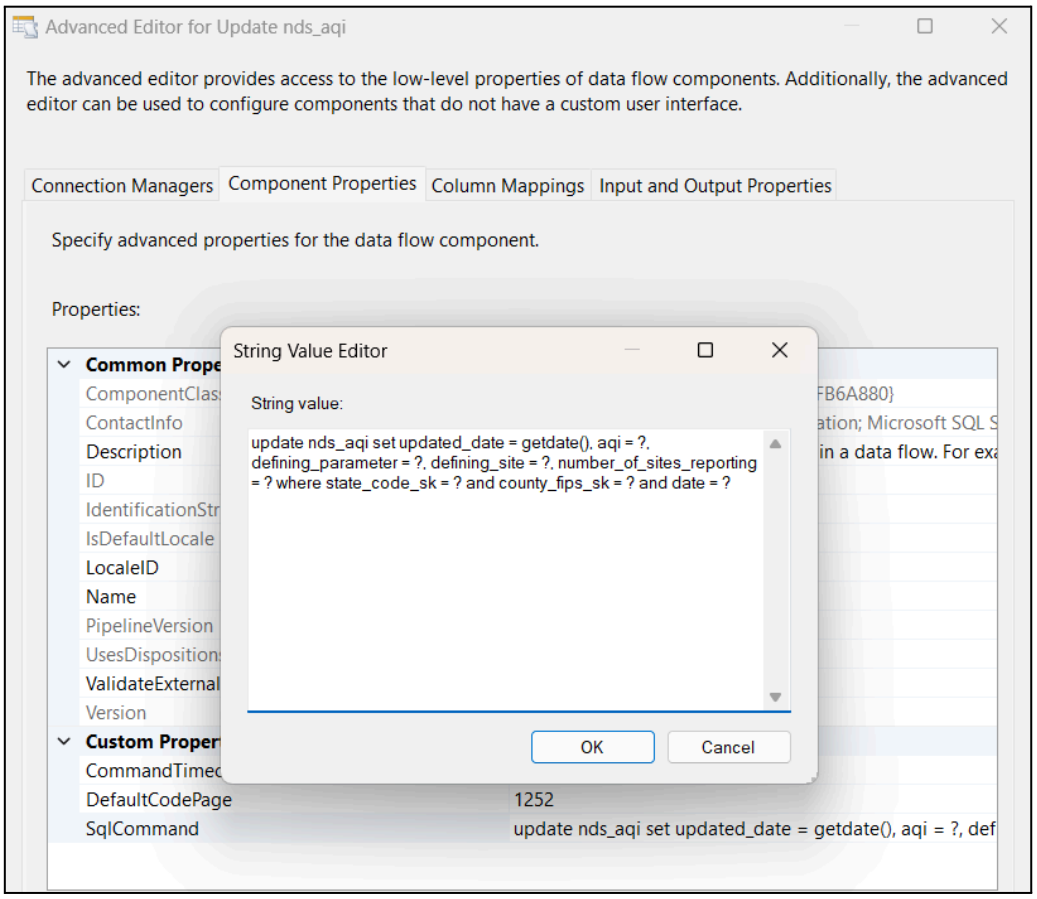
- Kết nối dữ liệu với lại Stage_state_AQI_2021
- Mapping dựa trên sourceID để lấy khóa ngoại đến nds_source
- Delete space để xóa khoảng trắng category name để tiến hành check xem tên loại category đó đã tồn tại chưa, nếu chưa thì thêm mới còn tồn tại rồi thì bỏ qua

2.3.5. Đổ dữ liệu vào bảng nds_aqi

- Tiến hành đổ lần lượt dữ liệu từ Stage_AQI_State_2021, Stage_AQI_State_2022, Stage_AQI_State_2023 (3 bảng này làm tương tự nhau)



- Tiến hành các thao tác lấy surrogate key từ các bảng nds_source, nds_state, nds_county, nds_category bằng cách
    + Mapping dựa trên sourceID để lấy được source_sk từ nds_source
    + Mapping dựa trên state_name để lấy được state_code_sk từ nds_state
    + Mapping dựa trên county_name để lấy được county_code_sk từ nds_county
    + Mapping dựa trên category_name để lấy được category_sk từ nds_category
- Sau đó tiến hành kiểm tra dữ liệu tại 1 ngày cụ thể của bang đó tại county đó đã tồn tại hay chưa (vì dữ liệu cấp thấp nhất là đơn vị ngày và không có trùng ngày trong dữ liệu)
    + Nếu đã tồn tại thì mình cập nhật

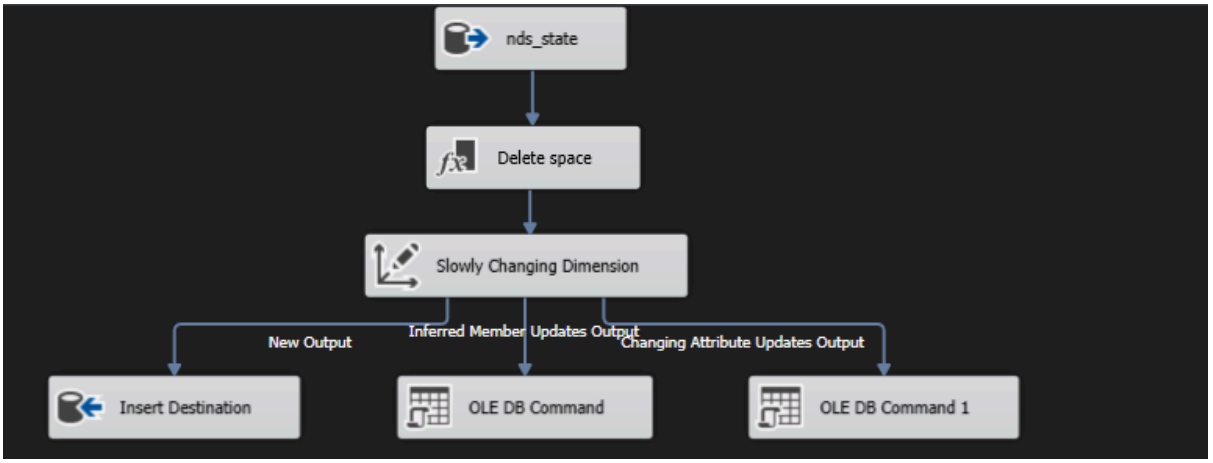


+ Nếu chưa tồn tại thì thêm mới

2.4. NDS to DDS

2.4.1. Đổ dữ liệu vào bảng dim_state

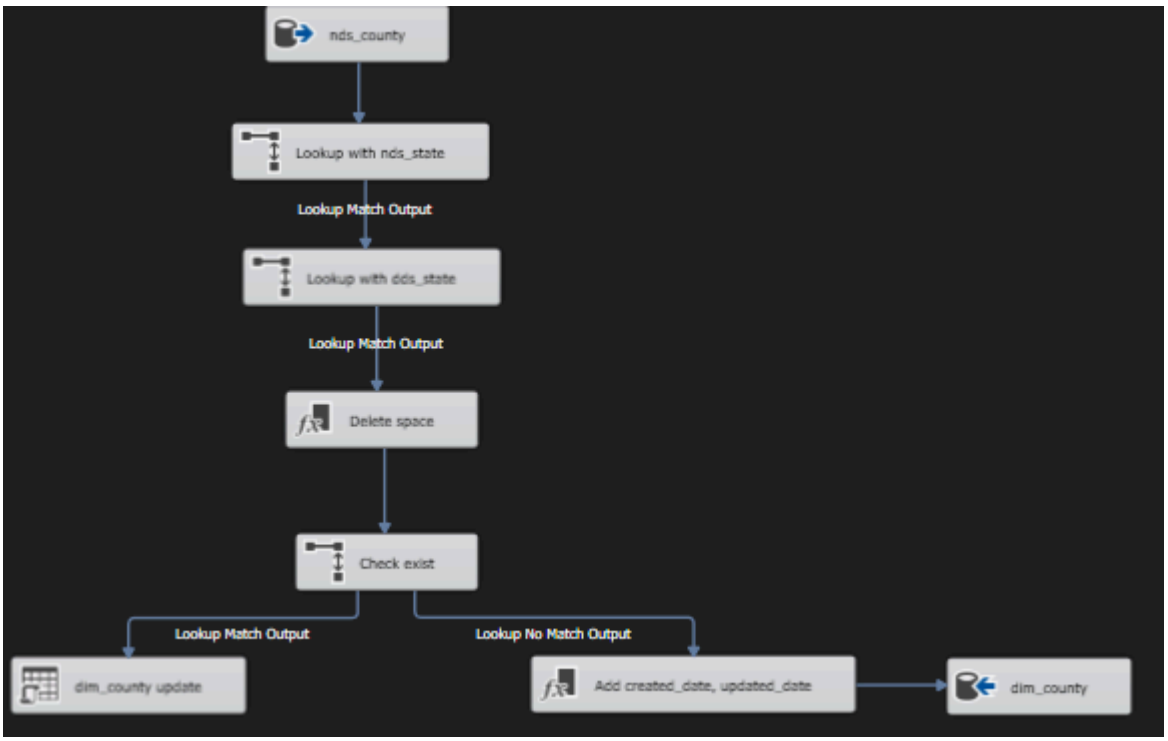
- Ở giai đoạn từ NDS to DDS chúng ta sẽ có bước thực hiện Incremental Extract dựa trên lset, cet mà chúng ta đã tạo ở bảng nds_data_flow trong metadata
- Bắt đầu cũng là việc cập nhật cet
- Sau đó lấy giá trị lset, cet để thực hiện Incremental Extract
- Thực hiện đổ dữ liệu vào dim_state



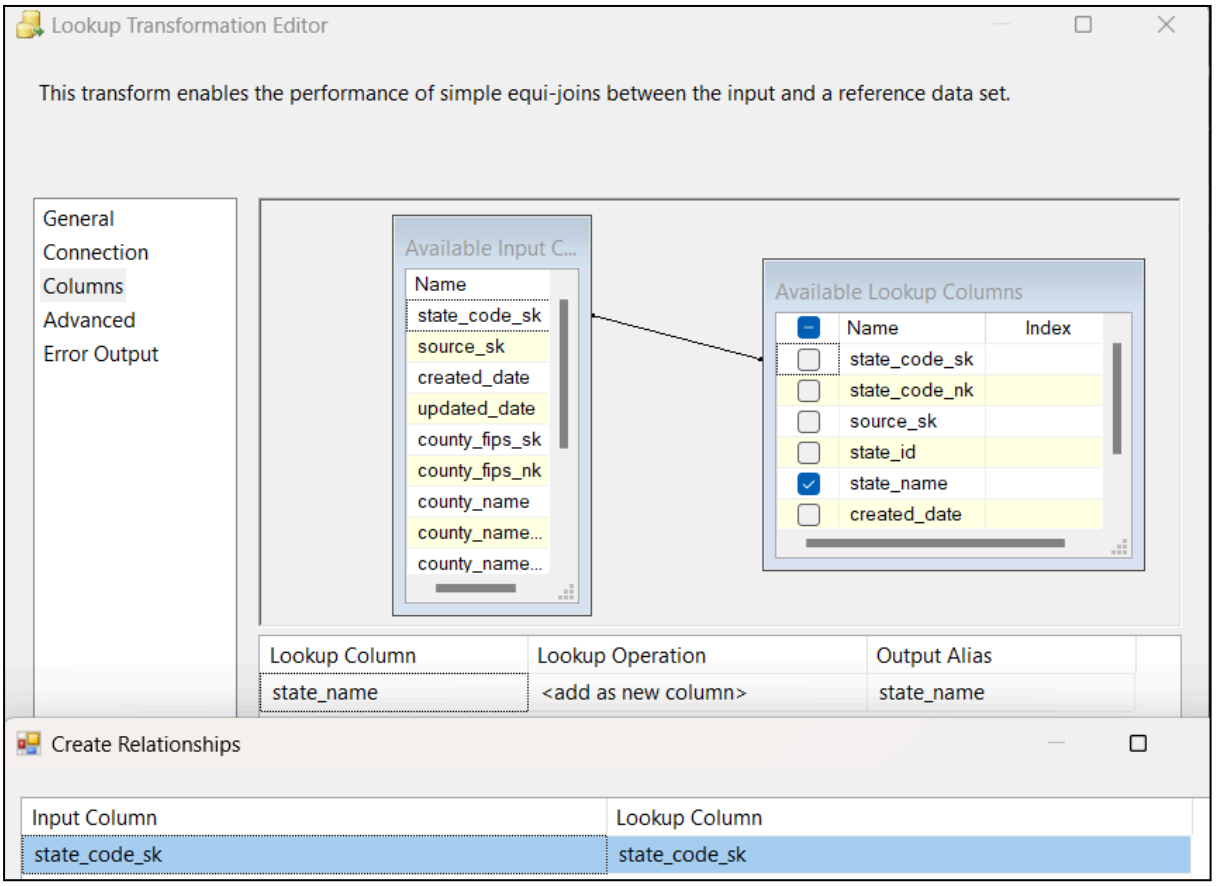
+ Kết nối với lại nds_state
+ Thực hiện đổ từ nds_state vào dim_state (Bước này cũng đơn giản vì nds_state và dim_state có cấu trúc y chang nhau)
+ Nếu đã tồn tại thì cập nhật, nếu chưa có thì thêm mới
- Cuối cùng là sẽ cập nhật lại giá trị cho lset là getdate()

2.4.2. Đổ dữ liệu vào bảng dim_county

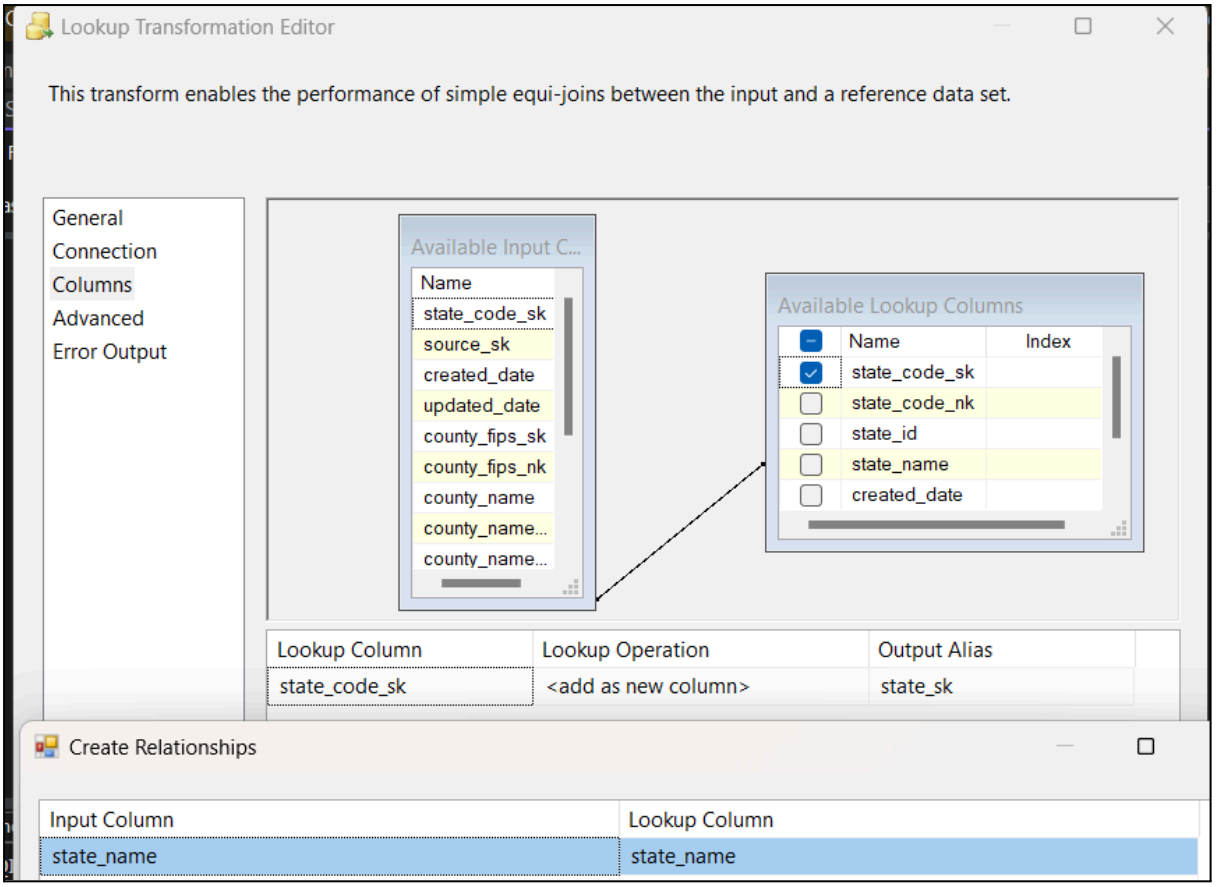
- Bắt đầu cũng là việc cập nhật cet
- Sau đó lấy giá trị lset, cet để thực hiện Incremental Extract



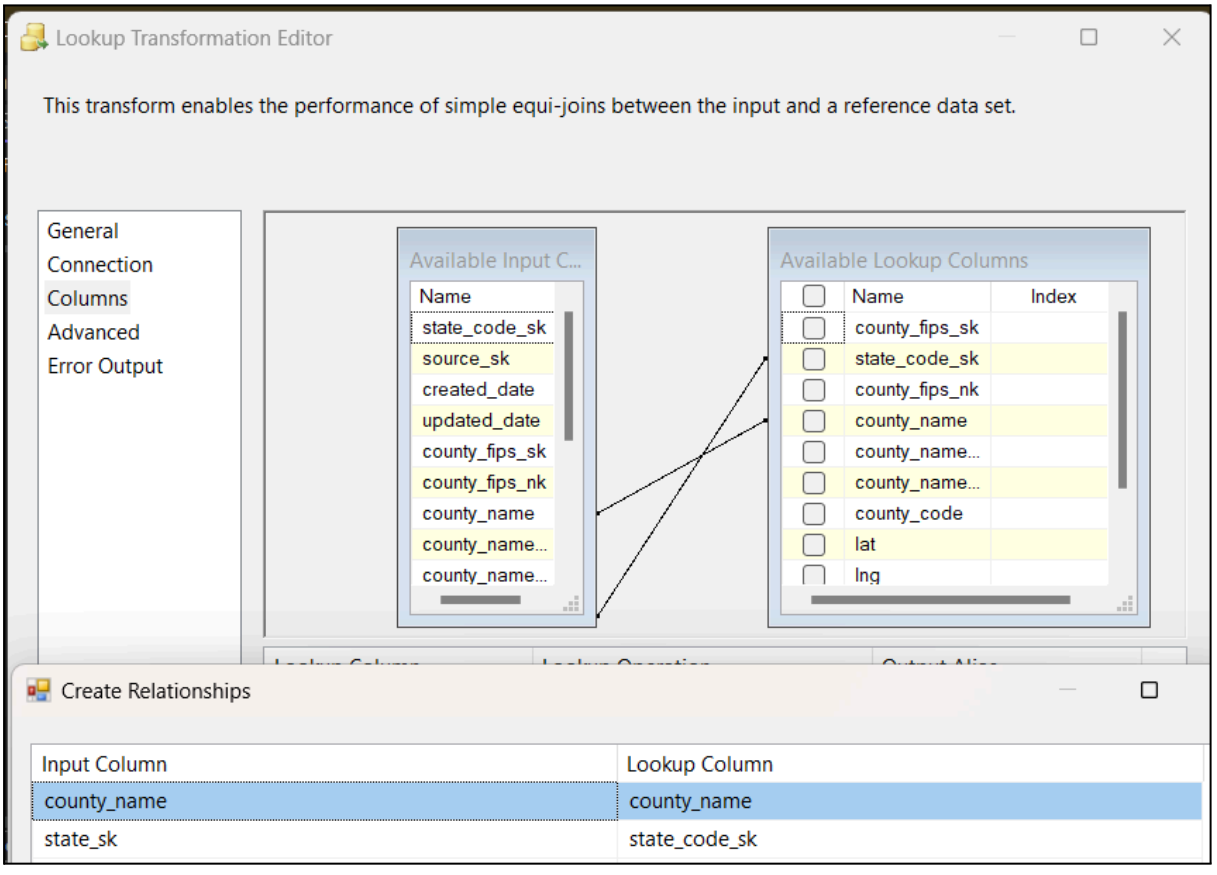
- Kết nối với lại nds_county
- Thực hiện mapping với lại nds_state dựa trên surrogate key của nds_state để lấy ra state_name



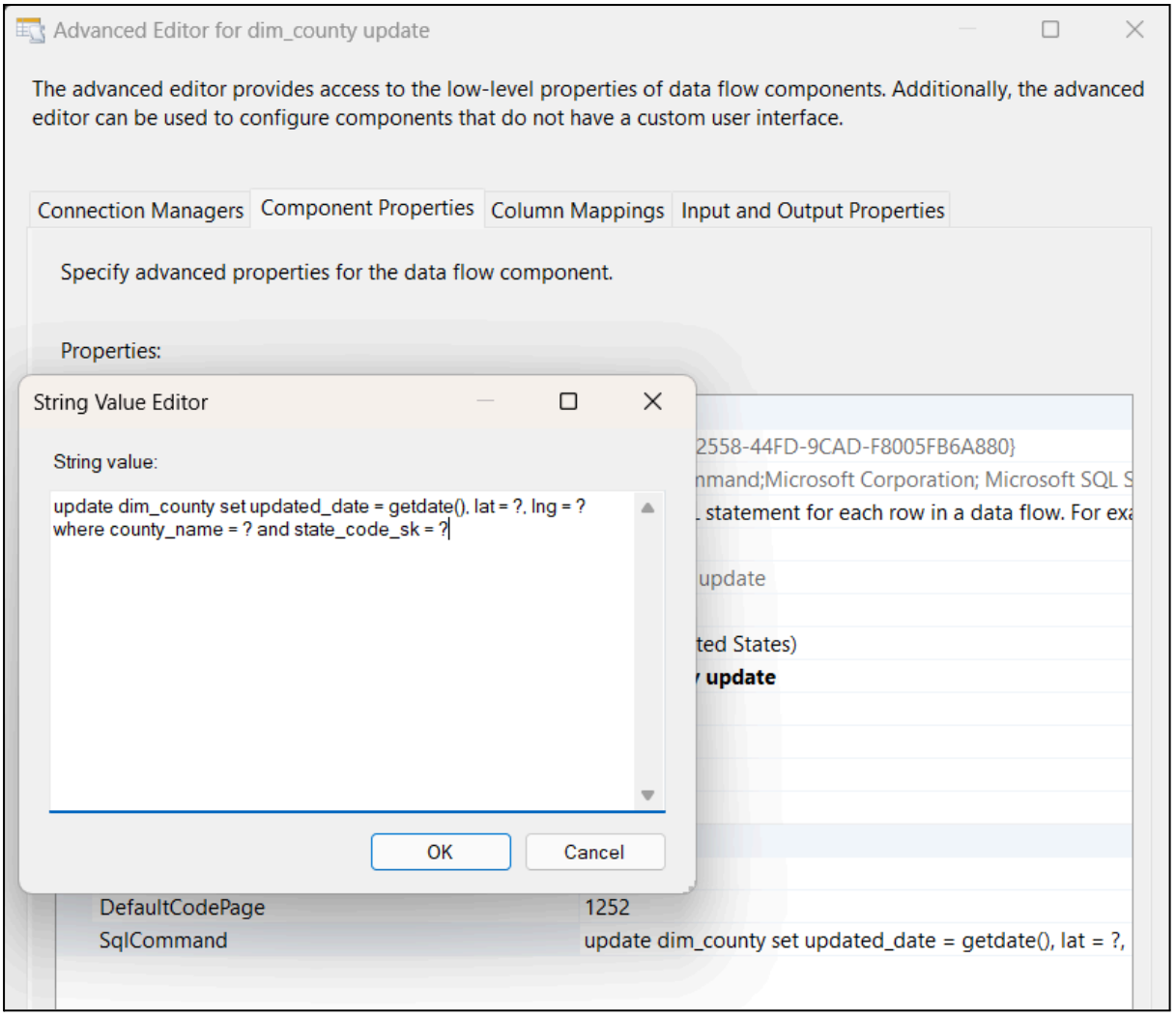
- Sau đó mapping với lại dds_state dựa trên state_name để có thể lấy được surrogate key của dds_state làm khóa ngoại trỏ đến dds_state



- Sau đó kiểm tra xem tại bang đó, county đó đã tồn tại chưa.



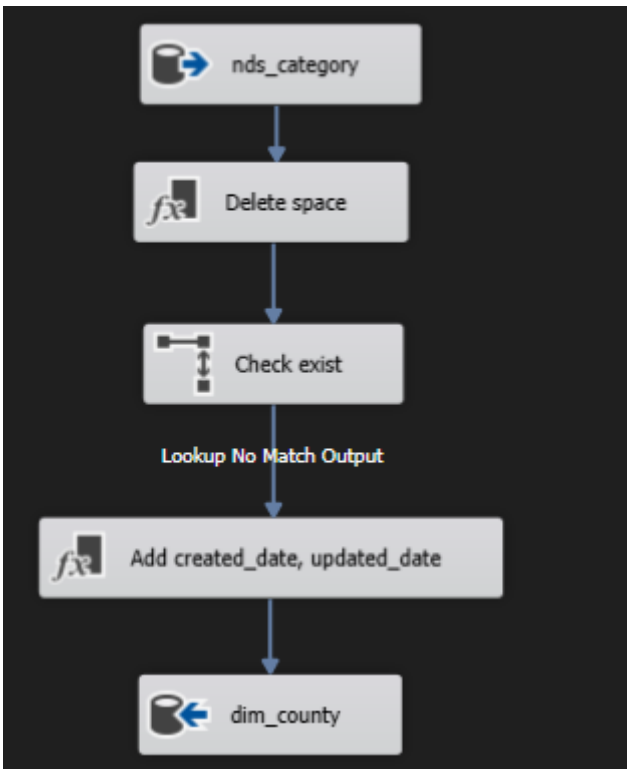
- Nếu đã tồn tại thì cập nhật



- Nếu chưa tồn tại thì thêm mới
- Cuối cùng là sẽ cập nhật lại giá trị cho lset là getdate()

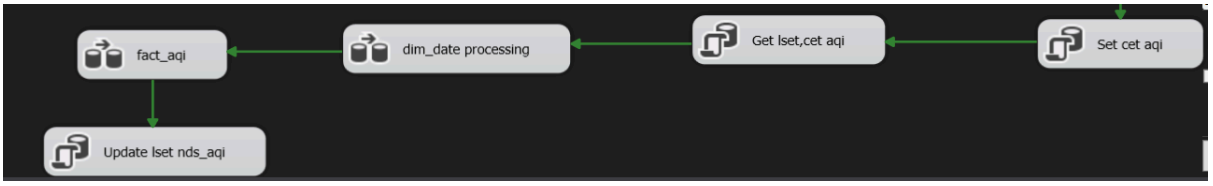
2.4.3. Đổ dữ liệu vào bảng dim_category

- Bắt đầu cũng là việc cập nhật cet
- Sau đó lấy giá trị lset, cet để thực hiện Incremental Extract



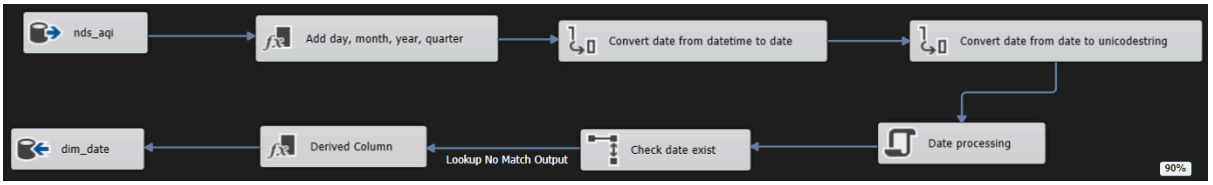
- Kết nối với lại nds_category (bước này đơn giản vì nds_category và dim_category có cấu trúc giống nhau)
- Thực hiện kiểm tra xem category_name đã tồn tại chưa, nếu chưa thì thêm mới, nếu có rồi thì bỏ qua
- Cuối cùng là sẽ cập nhật lại giá trị cho lset là getdate()

2.4.4. Đổ dữ liệu vào bảng dim_date và fact_aqi

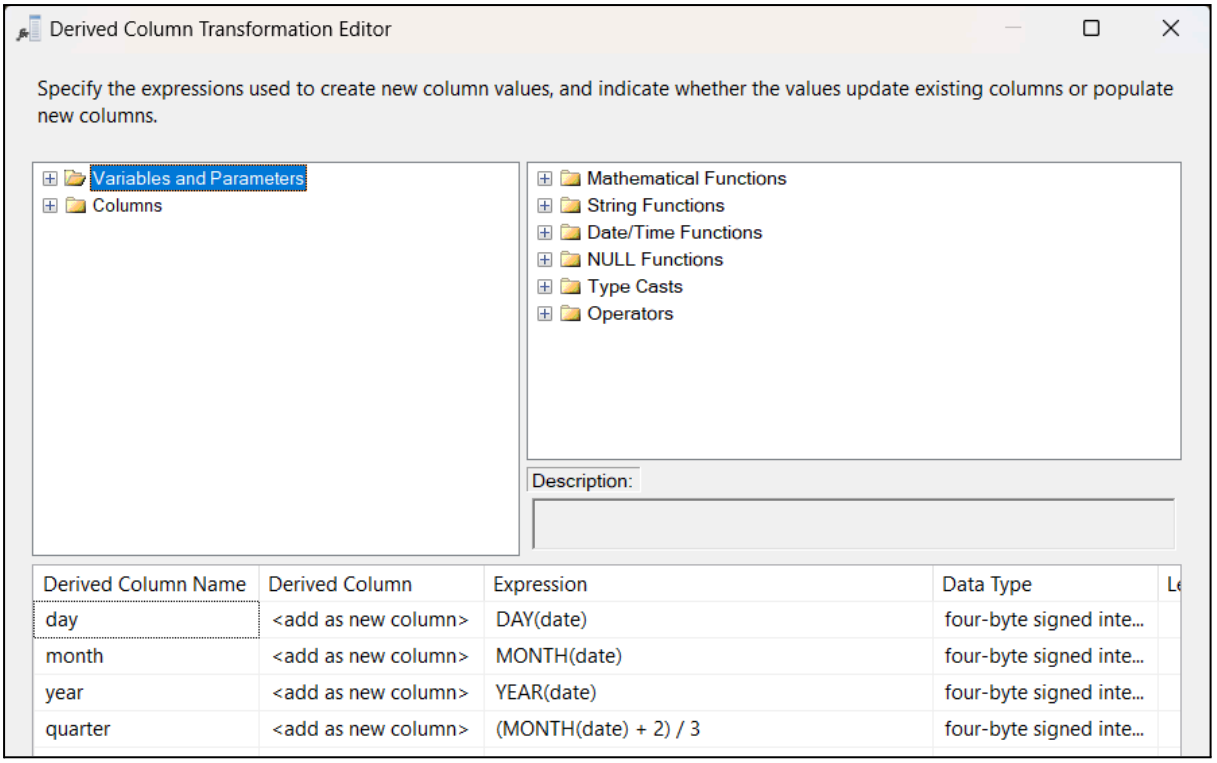


- Lý do gộp chung lại 2 bước này vì cột date và các giá trị aqi nằm chung 1 bảng và đồng thời dùng chung metadata. Chính vì thể phải đổ dim_date và fact_aqi hoàn tất mới cập nhật lset cho cả 2.

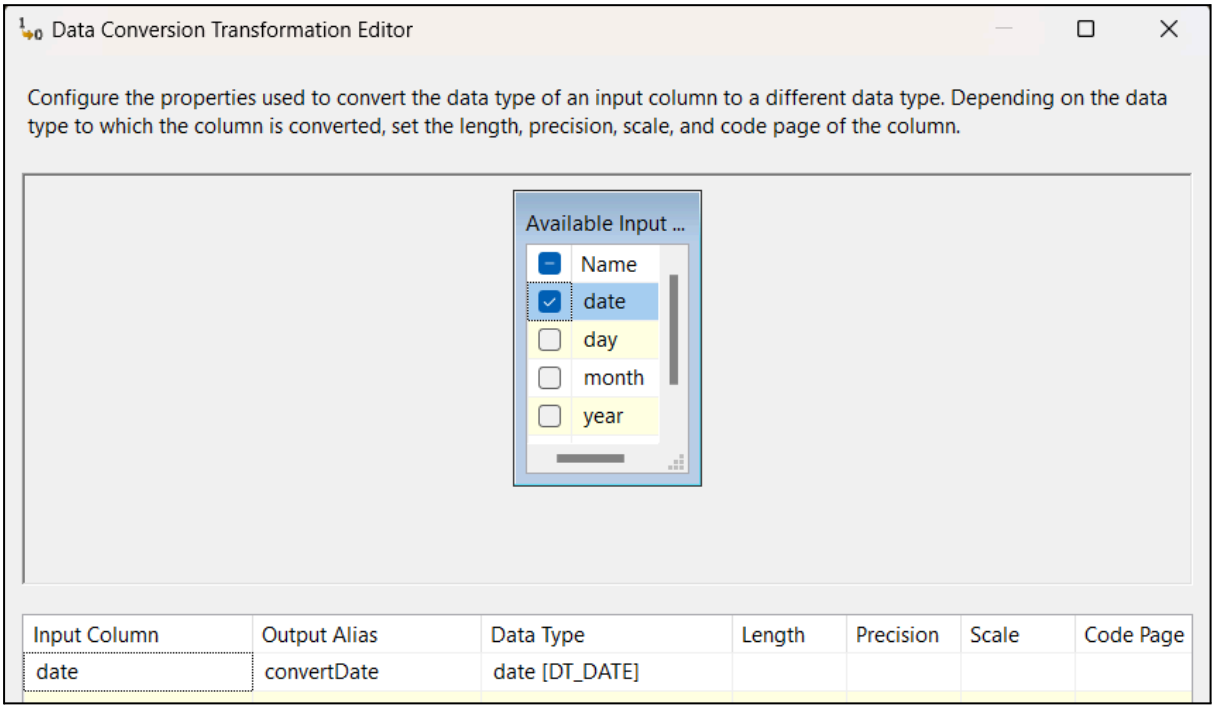
2.4.4.1. Đổ dữ liệu vào dim_date

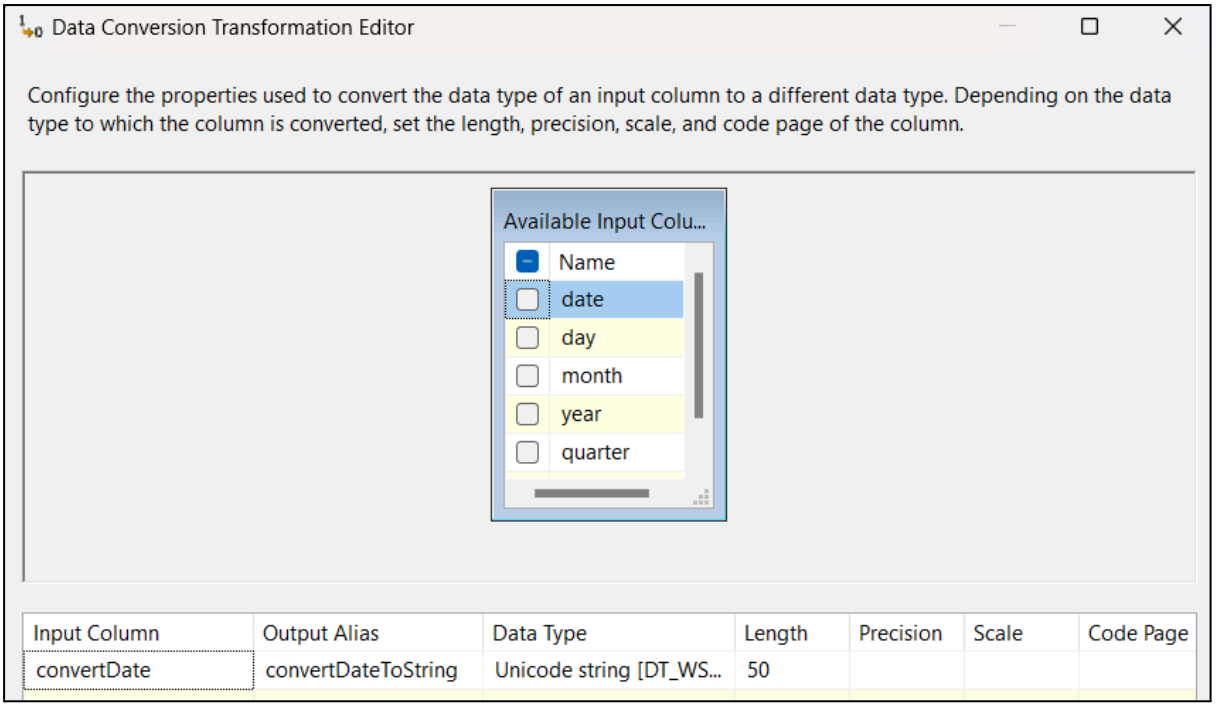


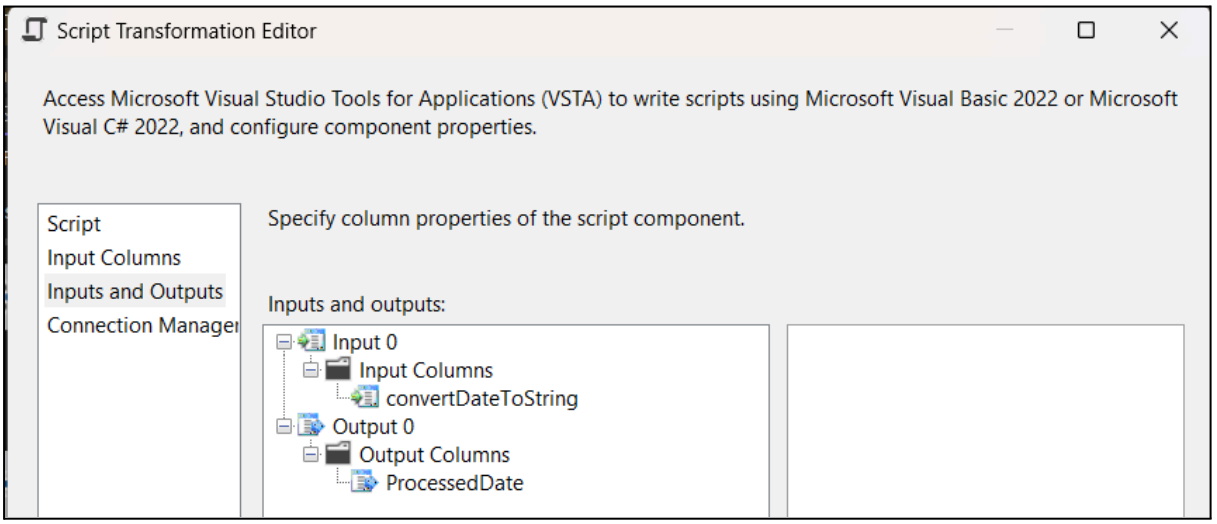
- Kết nối với lại nds_aqi
- Sau đó từ cột Date trong nds_aqi tách thêm thành 4 cột day, month, quarter, year



- Vì theo cấu trúc tạo trong dds thì cột date_full sẽ được thiết lập là dạng date với format là yyyy-mm-dd. Nên để có thể mapping thì phải đổi định dạng cột Date trong nds_aqi.
- Quy trình thực hiện là chuyển đổi Date trong nds_aqi từ datetime thành date. Xong tiếp tục thực hiện chuyển tiếp thành dạng string. Cuối cùng là dùng script component thực hiện việc chuyển đổi format về dạng yyyy-mm-dd



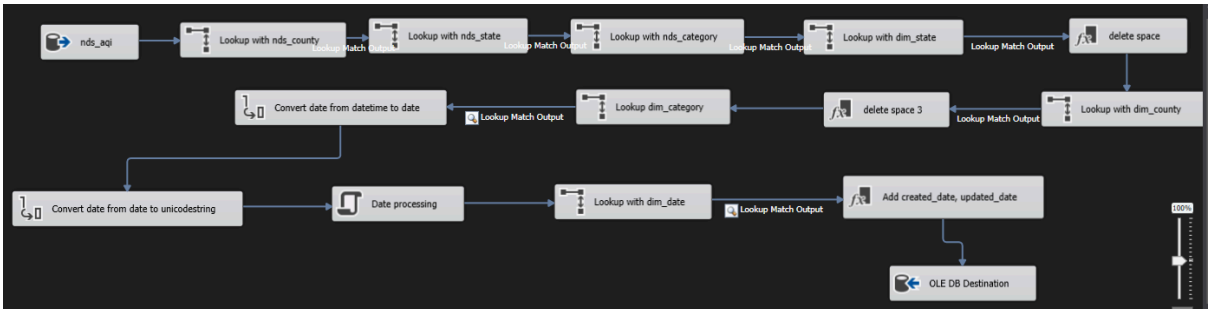

Data Conversion Transformation Editor
Configure the properties used to convert the data type of an input column to a different data type. Depending on the data type to which the column is converted, set the length, precision, scale, and code page of the column.
Available Input Colu...
Name
date
day
month
year
quarter
Input Column
Output Alias
Data Type
Length
Precision
Scale
Code Page
convertDate
convertDateToString
Unicode string [DT_WS...
50


Script Transformation Editor
Access Microsoft Visual Studio Tools for Applications (VSTA) to write scripts using Microsoft Visual Basic 2022 or Microsoft Visual C# 2022, and configure component properties.
Script
Input Columns
Inputs and Outputs
Connection Manager
Specify column properties of the script component.
Inputs and outputs:
Input 0
Input Columns
convertDateToString
Output 0
Output Columns
ProcessedDate

```
2 references
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    string InputDate = Row.convertDateToString; // mm/dd/yyyy

    string[] dateParts = InputDate.Split('/');



    // Xu ly thang-ngay-nam => nam-thang-ngay
    if (dateParts.Length == 3 ) {

        string formattedDate = dateParts[1].PadLeft(2, '0') ; // Day
        string formattedMonth = dateParts[0].PadLeft(2, '0');
        string formattedYear = dateParts[2];


        Row.ProcessedDate = formattedYear + '-' + formattedMonth + '-' + formattedDate;

    }
    else
    {
        Row.ProcessedDate = InputDate;
    }

}
```

- Sau đó kiểm tra ngày đó đã tồn tại chưa, nếu chưa thì thêm mới, nếu có rồi thì bỏ qua

2.4.4.2. Đổ dữ liệu vào fact_aqi



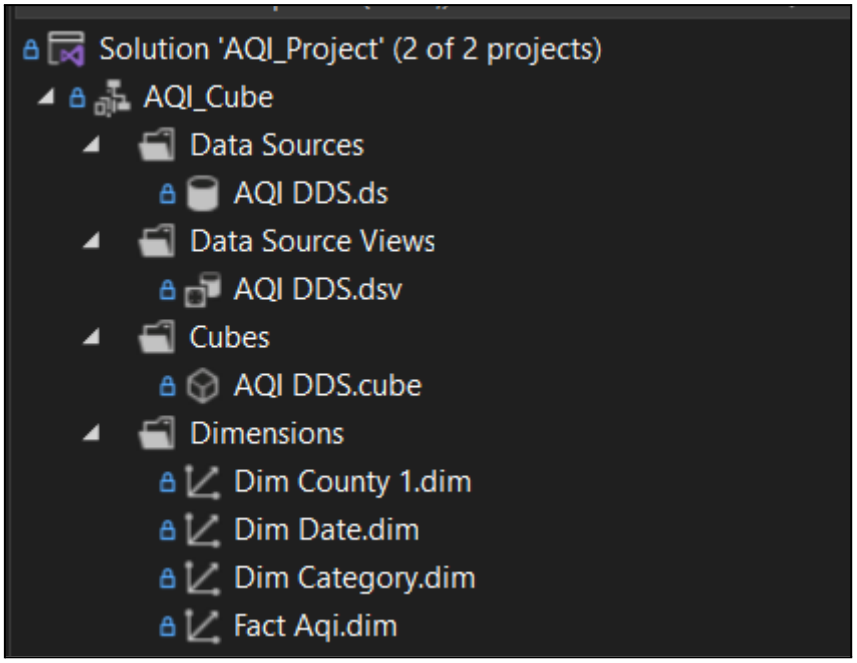
Các bước thực hiện đổ dữ liệu vào bảng fact_aqi như sau

- Kết nối với lại nds_aqi
- Tiến hành lấy các surrogate key từ các bảng chiều : dim_county, dim_date, dim_category
    + Mapping với nds_county để lấy được county_name. Sau đó mapping với lại nds_state để lấy state_name. Mục đích của việc lấy state_name là để có thể mapping với dim_state nhằm lấy ra được surrogate key của dim_state. Từ đây sẽ kiểm tra dựa trên county_name và state_code_sk của dds để có thể lấy ra được county_fips_sk của bảng dim_county

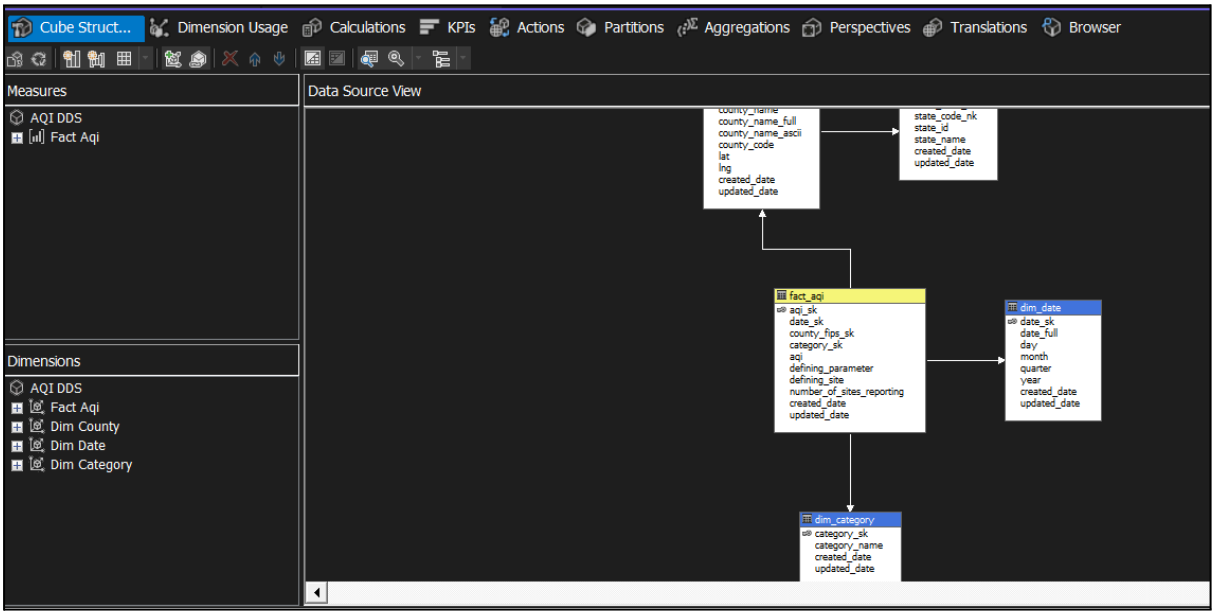
+ Mapping với nds_category để lấy được county_name. Sau đó mapping với lại dim_category để lấy được surrogate key của dim_category làm khóa ngoại
+ Thực hiện chuyển đổi cột date trong aqi về cùng định dạng và mapping với lại cột date_full trong dim_date để lấy được surrogate key của dim_date làm khóa ngoại
+ Mapping và đổ dữ liệu vào fact_aqi
+ Cuối cùng là sẽ cập nhật lại giá trị cho Iset là getdate()

3. CUBE

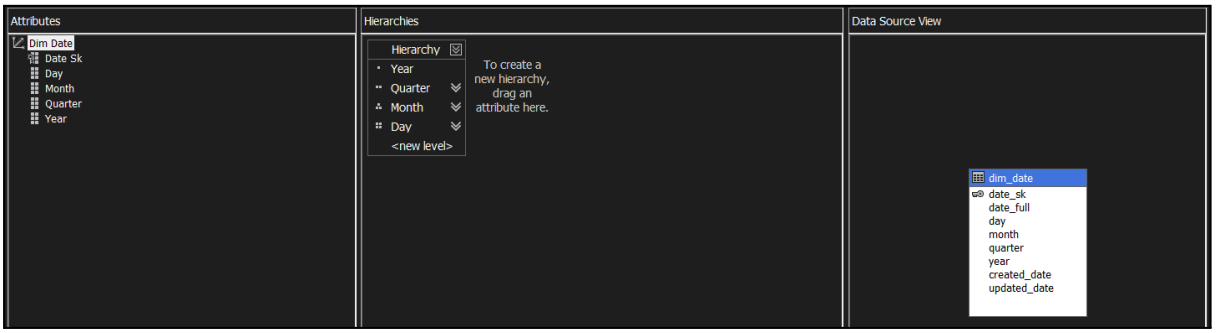
3.1. Tiến hành kết nối đến database DDS sau khi ETL

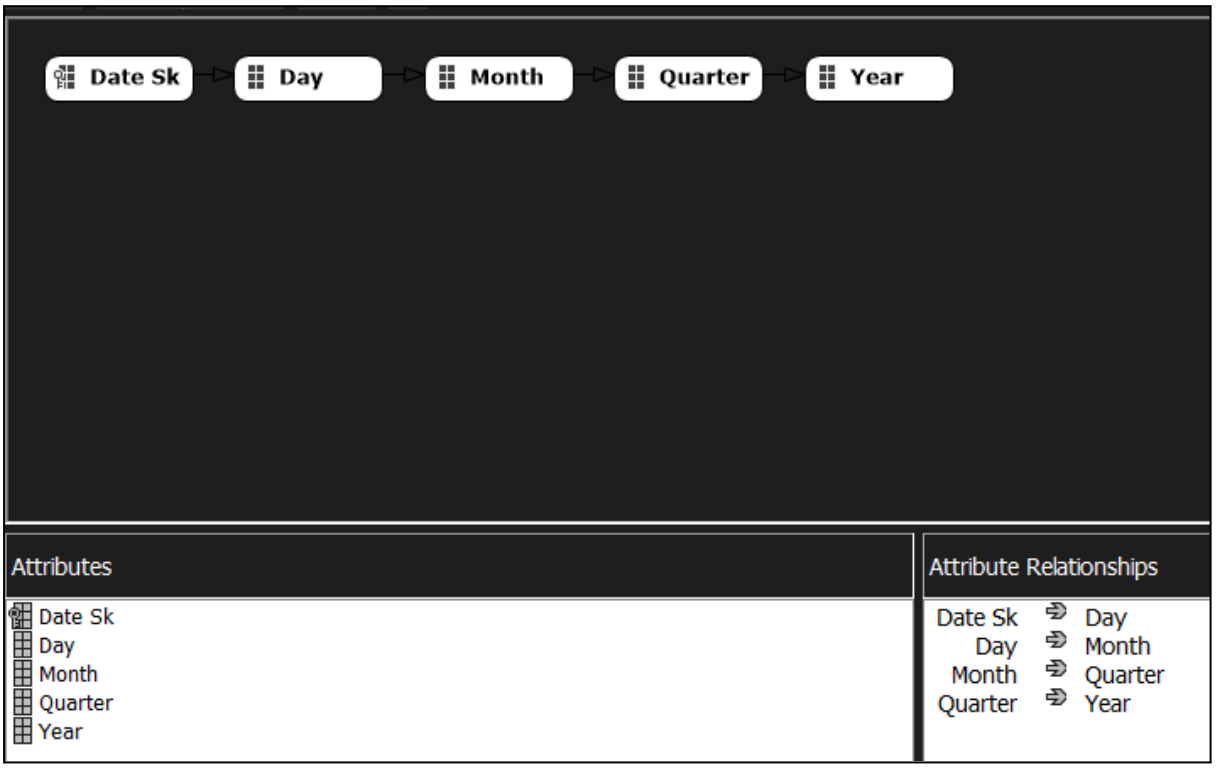


- Giao diện khi kết nối xong

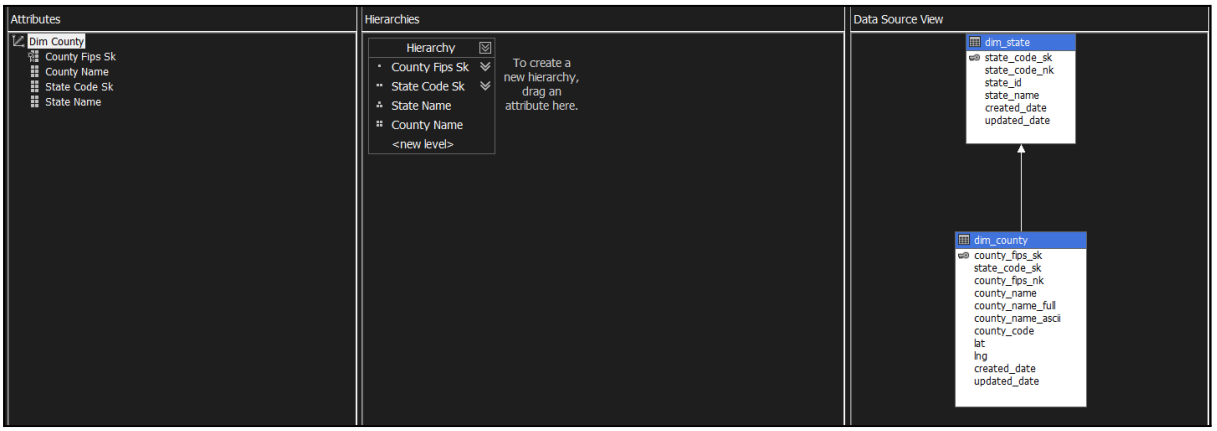


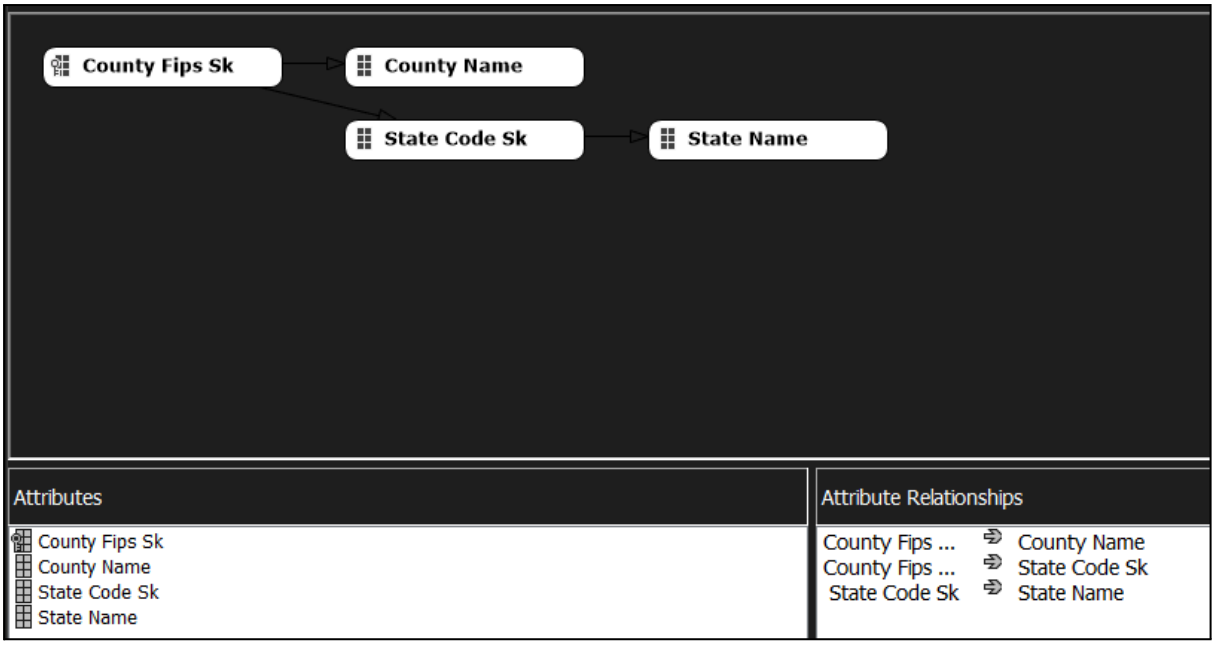
3.2. Tạo phân cấp chiều cho dim_date : year -> quarter -> month -> day





3.3. Tạo phân cấp chiều state -> county





4. OLAP, MDX

4.1. Báo cáo giá trị AQI nhỏ nhất và lớn nhất cho từng bang trong từng quý của các năm

```
-- Question 1
WITH
    MEMBER [Measures].[Max AQI] AS
        MAX(
            EXISTING [Dim Date].[Day].[Day].MEMBERS,
            [Measures].[Aqi]
        )
    MEMBER [Measures].[Min AQI] AS
        MIN(
            EXISTING [Dim Date].[Day].[Day].MEMBERS,
            [Measures].[Aqi]
        )
SELECT
    {
        [Measures].[Max AQI],
        [Measures].[Min AQI]
    } ON COLUMNS,
    NON EMPTY
    (
        [Dim County].[State Name].[State Name] *
        [Dim Date].[Year].[Year] *
        [Dim Date].[Quarter].[Quarter]
    ) ON ROWS
FROM [AQI DDS]
```

			Max AQI	Min AQI
Alabama	2022	3	51	4
Alabama	2022	4	46	3
Alabama	2023	1	46	2
Alabama	2023	2	132	4
Alabama	2023	3	86	10
Alabama	2023	4	129	7
Alaska	2021	1	631	96
Alaska	2021	2	281	83
Alaska	2021	3	312	54
Alaska	2021	4	385	66
Alaska	2022	1	290	87
Alaska	2022	2	336	73
Alaska	2022	3	436	84
Alaska	2022	4	451	57
Alaska	2023	1	313	61
Alaska	2023	2	318	69
Alaska	2023	3	260	66

4.2. Báo cáo giá trị trung bình (mean) và độ lệch chuẩn (standard deviation) của AQI cho từng bang trong từng quý của các năm.

```
WITH
    -- Tính giá trị trung bình AQI
    MEMBER [Measures].[Average AQI] AS
        AVG(
            EXISTING [Dim Date].[Day].[Day].MEMBERS,
            [Measures].[Aqi]
        )

    -- Tính độ lệch chuẩn AQI
    MEMBER [Measures].[Standard Deviation AQI] AS
        STDEV(
            EXISTING [Dim Date].[Day].[Day].MEMBERS,
            [Measures].[Aqi]
        )
```

```
SELECT
    {
        [Measures].[Average AQI],
        [Measures].[Standard Deviation AQI]
    } ON COLUMNS,

    FILTER(
        (
            [Dim County].[State Name].[State Name].Members,
            [Dim Date].[Year].[Year].Members,
            [Dim Date].[Quarter].[Quarter].Members
        ),
        NOT IsEmpty([Measures].[Average AQI])
        AND NOT IsEmpty([Measures].[Standard Deviation AQI])
    ) ON ROWS

FROM [AQI DDS]
```

			Average AQI	Standard Deviation AQI
Alabama	2022	3	15.9272727272727	9.49045240590178
Alabama	2022	4	13.734693877551	8.38593939829264
Alabama	2023	1	14.5111111111111	8.60395054865469
Alabama	2023	2	38.5054945054945	18.7512213684417
Alabama	2023	3	29.5869565217391	16.7942639173487
Alabama	2023	4	28.1086956521739	18.6879855167855
Alaska	2021	1	222.166666666667	75.4087364175391
Alaska	2021	2	144.43956043956	43.4432474720267
Alaska	2021	3	132.652173913043	53.2715281319007
Alaska	2021	4	196.141304347826	73.234469384111
Alaska	2022	1	146.088888888889	38.7497428608922
Alaska	2022	2	147.89010989011	52.4106542496531
Alaska	2022	3	198.04347826087	77.4749747676066
Alaska	2022	4	189.913043478261	86.0023833267904
Alaska	2023	1	151.588888888889	58.7487306952424
Alaska	2023	2	166.56043956044	63.6279478760375
Alaska	2023	3	145.945652173913	52.4011081441493

4.3. Báo cáo số ngày và giá trị AQI trung bình khi chất lượng không khí được xếp hạng là "rất không lành mạnh" (very unhealthy) hoặc tệ hơn cho từng bang và từng quận.

```
-- Question 3
WITH
MEMBER [Measures].[Unhealthy Days Count] AS
    SUM(
        FILTER(
            [Dim Category].[Category Name].MEMBERS,
            [Dim Category].[Category Name].CURRENTMEMBER.NAME = "Very Unhealthy"
            OR [Dim Category].[Category Name].CURRENTMEMBER.NAME = "Hazardous"
        ),
        [Measures].[Fact Aqi Count]
    )

MEMBER [Measures].[Average AQI] AS
    AVG(
        FILTER(
            [Dim Category].[Category Name].MEMBERS,
            [Dim Category].[Category Name].CURRENTMEMBER.NAME = "Very Unhealthy"
            OR [Dim Category].[Category Name].CURRENTMEMBER.NAME = "Hazardous"
        ),
        [Measures].[Aqi]
    )
```

```
SELECT
    {
        [Measures].[Unhealthy Days Count],
        [Measures].[Average AQI]
    } ON COLUMNS,
    NONEMPTY(
        CROSSJOIN(
            [Dim County].[State Name].[State Name].MEMBERS,
            [Dim County].[County Name].[County Name].MEMBERS
        ),
        {[Measures].[Unhealthy Days Count], [Measures].[Average AQI]}
    ) ON ROWS
FROM [AQI DDS]
WHERE
    ([Dim Date].[Year].MEMBERS);
```

		Unhealthy Days Count	Average AQI
Alaska	Fairbanks North Star	3	48.5
Arizona	Coconino	1	38
Arizona	Maricopa	79	16379
Arizona	Pima	1	313
Arizona	Pinal	7	227
California	Butte	3	733
California	Colusa	1	225
California	El Dorado	2	324.5
California	Fresno	2	501
California	Humboldt	4	431
California	Imperial	18	1082.5
California	Inyo	22	1547
California	Kern	3	313
California	Kings	1	500
California	Los Angeles	5	528
California	Mariposa	2	77
California	Mono	22	1198

4.4. Đối với 4 bang sau: Hawaii, Alaska, Illinois và Delaware, đếm số ngày trong từng hạng mục chất lượng không khí (Tốt, Trung bình, v.v.) theo từng quận.

```
-- Question 4
WITH
MEMBER [Measures].[Days Count] AS
    COUNT(
        NONEMPTY(
            [Dim Date].[Day].MEMBERS,
            [Measures].[Fact Aqi Count]
        )
    )

SELECT
    {
        [Measures].[Days Count]
    } ON COLUMNS,
    NONEMPTY(
        FILTER(
            CROSSJOIN(
                [Dim County].[State Name].[State Name].MEMBERS,
                [Dim County].[County Name].[County Name].MEMBERS,
                [Dim Category].[Category Name].[Category Name].MEMBERS
            ),
```

```
            [Dim County].[State Name].CURRENTMEMBER.NAME = "Hawaii"
            OR [Dim County].[State Name].CURRENTMEMBER.NAME = "Alaska"
            OR [Dim County].[State Name].CURRENTMEMBER.NAME = "Illinois"
            OR [Dim County].[State Name].CURRENTMEMBER.NAME = "Delaware"
        ),
        [Measures].[Fact Aqi Count]
    ) ON ROWS
FROM [AQI DDS];
```

			Days Count
Alaska	Aleutians East	Good	357
Alaska	Aleutians East	Moderate	5
Alaska	Anchorage	Good	911
Alaska	Anchorage	Moderate	183
Alaska	Anchorage	Unhealthy for Sensitive Groups	4
Alaska	Denali	Good	1039
Alaska	Denali	Moderate	14
Alaska	Fairbanks North Star	Good	649
Alaska	Fairbanks North Star	Hazardous	2
Alaska	Fairbanks North Star	Moderate	335
Alaska	Fairbanks North Star	Unhealthy	51
Alaska	Fairbanks North Star	Unhealthy for Sensitive Groups	61
Alaska	Fairbanks North Star	Very Unhealthy	3
Alaska	Juneau	Good	921
Alaska	Juneau	Moderate	144
Alaska	Kenai Peninsula	Good	360
Alaska	Kenai Peninsula	Moderate	2

4.5. Đối với 4 bang sau: Hawaii, Alaska, Illinois và Delaware, tính giá trị AQI trung bình theo từng quý.

```
WITH
-- Tính trung bình AQI
MEMBER [Measures].[Average AQI] AS
    AVG(
        NONEMPTY(
            [Dim Date].[Day].MEMBERS,
            [Measures].[Aqi]
        ),
        [Measures].[Aqi]
    )

SELECT
-- Tổ hợp State, Year và Quarter trên ROWS
    NONEMPTY(
        CROSSJOIN(
            {[Dim County].[State Name].[Hawaii],
             [Dim County].[State Name].[Alaska],
             [Dim County].[State Name].[Illinois],
             [Dim County].[State Name].[California]}, -- Chỉ chọn các bang cần
            [Dim Date].[Year].[Year].MEMBERS,
            [Dim Date].[Quarter].[Quarter].MEMBERS
        )
    ) ON ROWS,
-- Hiển thị Average AQI trên COLUMNS
   {[Measures].[Average AQI]} ON COLUMNS

FROM [AQI DDS];
```

			Average AQI
Hawaii	2021	1	88.2180656934307
Hawaii	2021	2	86.2381386861314
Hawaii	2021	3	85.8631386861314
Hawaii	2021	4	87.0894160583942
Hawaii	2022	1	87.4051094890511
Hawaii	2022	2	84.8156934306569
Hawaii	2022	3	83.698905109489
Hawaii	2022	4	86.3302919708029
Hawaii	2023	1	85.0155109489051
Hawaii	2023	2	83.6906934306569
Hawaii	2023	3	83.4534671532847
Hawaii	2023	4	83.7700729927007
Alaska	2021	1	188.579379562044
Alaska	2021	2	182.328467153285
Alaska	2021	3	181.470802919708
Alaska	2021	4	186.800182481752
Alaska	2022	1	182.332116788321

4.6. Báo cáo giá trị trung bình (mean), độ lệch chuẩn (standard deviation), giá trị nhỏ nhất (min) và lớn nhất (max) của AQI theo từng bang và quận trong mỗi quý của năm.

```
-- Question 9
WITH
MEMBER [Measures].[Mean AQI] AS
    AVG(
        NONEMPTY(
            [Dim Date].[Day].MEMBERS,
            [Measures].[Aqi]
        ),
        [Measures].[Aqi]
    )

MEMBER [Measures].[Standard Deviation AQI] AS
    STDDEV(
        NONEMPTY(
            [Dim Date].[Day].MEMBERS,
            [Measures].[Aqi]
        ),
        [Measures].[Aqi]
    )
```

```
MEMBER [Measures].[Min AQI] AS
    MIN(
        NONEMPTY(
            [Dim Date].[Day].MEMBERS,
            [Measures].[Aqi]
        ),
        [Measures].[Aqi]
    )

MEMBER [Measures].[Max AQI] AS
    MAX(
        NONEMPTY(
            [Dim Date].[Day].MEMBERS,
            [Measures].[Aqi]
        ),
        [Measures].[Aqi]
    )
```

```
SELECT
    {
        [Measures].[Mean AQI],
        [Measures].[Standard Deviation AQI],
        [Measures].[Min AQI],
        [Measures].[Max AQI]
    } ON COLUMNS,
    NONEMPTY(
        CROSSJOIN(
            [Dim Date].[Year].[Year].MEMBERS, -- Hiển thị Năm
            [Dim Date].[Quarter].[Quarter].MEMBERS, -- Hiển thị Quý
            [Dim County].[State Name].[State Name].MEMBERS,
            [Dim County].[County Name].[County Name].MEMBERS
        )
    ) ON ROWS
FROM [AQI DDS];
```

				Mean AQI	Standard Deviati
2021	1	Alaska	Aleutians East	30.6869806094183	31.97983864
2021	1	Alaska	Anchorage	37.5875912408759	106.3343893
2021	1	Alaska	Denali	43.31463878327	107.0537752
2021	1	Alaska	Fairbanks North Star	45.2381386861314	204.2810554
2021	1	Alaska	Juneau	23.4182330827068	78.940949185
2021	1	Alaska	Kenai Peninsula	20.3795013850416	24.10388444
2021	1	Alaska	Matanuska-Susitna	21.1603415559772	87.67373419
2021	1	Alaska	North Slope	13.9404761904762	18.55566932
2021	1	Arizona	Apache	29.0606617647059	35.60348427
2021	1	Arizona	Cochise	51.2609489051095	124.2966747
2021	1	Arizona	Coconino	53.0428832116788	112.9031386
2021	1	Arizona	Gila	58.3430656934307	118.9848269
2021	1	Arizona	La Paz	79.5423572744015	124.3662776
2021	1	Arizona	Maricopa	102.098540145985	368.6255148
2021	1	Arizona	Mohave	30.9310986964618	40.90393926
2021	1	Arizona	Navajo	54.2833638025594	109.8690985

4.7. Đếm số ngày theo từng bang và hạng mục (Category) trong mỗi tháng.

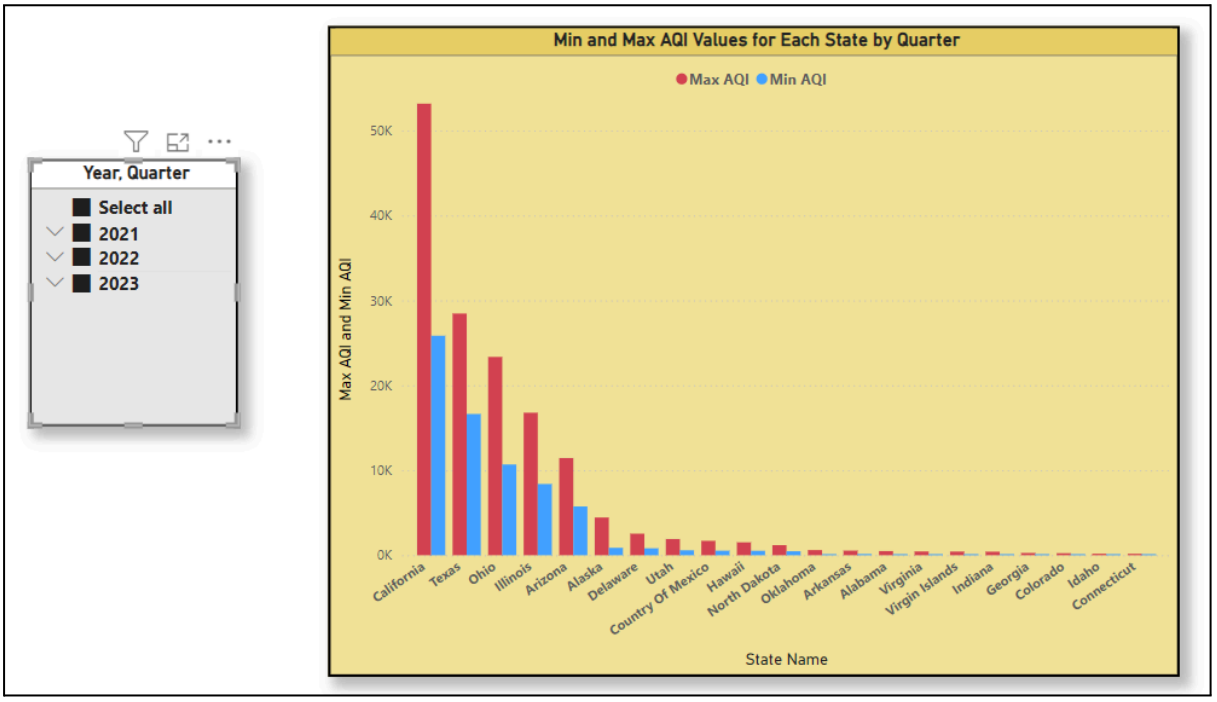
```
-- Question 11
WITH
MEMBER [Measures].[Days Count] AS
    SUM(
        [Dim County].[County Name].[County Name].MEMBERS, -- Lấy tất cả các County
        COUNT(
            NONEMPTY(
                [Dim Date].[Day].MEMBERS,
                ([Measures].[Fact Aqi Count],
                 [Dim County].[County Name].CURRENTMEMBER,
                 [Dim Category].[Category Name].CURRENTMEMBER,
                 [Dim Date].[Year].CURRENTMEMBER,
                 [Dim Date].[Month].CURRENTMEMBER)
            )
        )
    )
```

```
SELECT
    NONEMPTY(
        CROSSJOIN(
            [Dim County].[State Name].[State Name].MEMBERS, -- Hiển thị State
            [Dim Category].[Category Name].[Category Name].MEMBERS, -- Hiển thị Cate
            [Dim Date].[Year].[Year].MEMBERS, -- Hiển thị Year
            [Dim Date].[Month].[Month].MEMBERS -- Hiển thị Month
        )
    ) ON ROWS,
    [Measures].[Days Count] ON COLUMNS
FROM [AQI DDS];
```

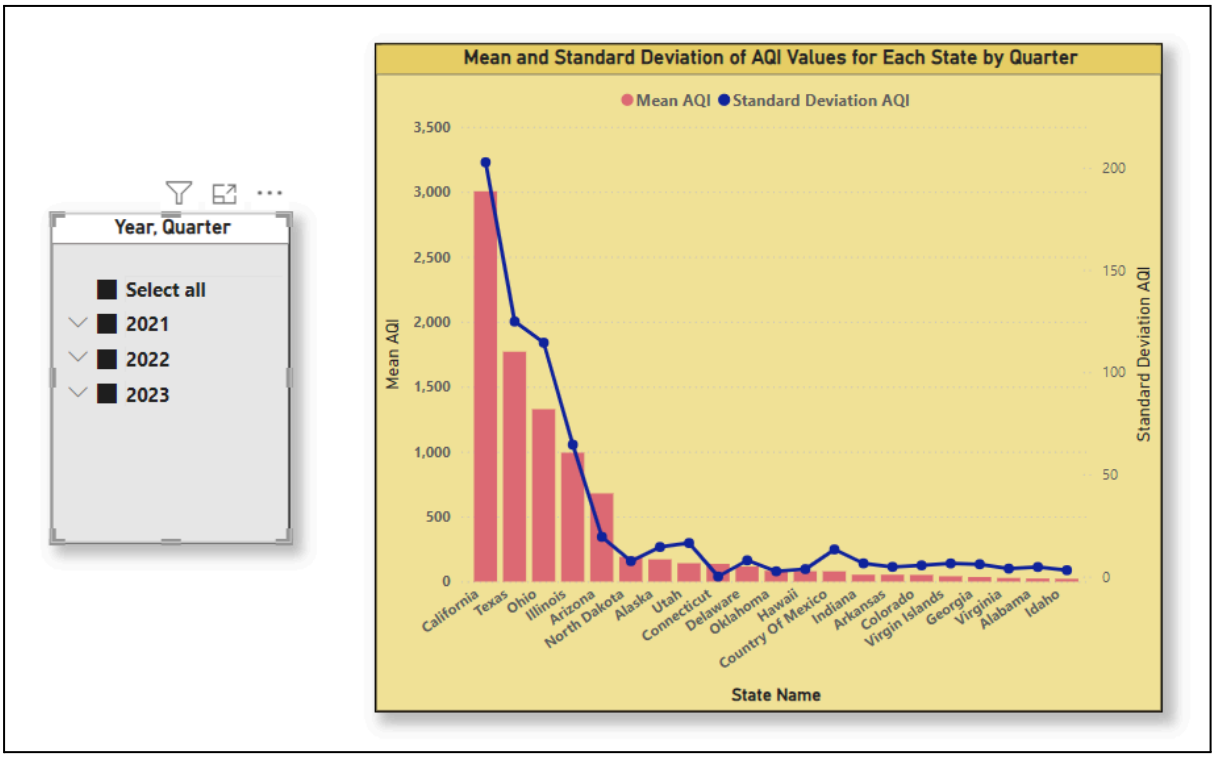
				Days Count
Alabama	Good	2022	10	26
Alabama	Good	2022	11	8
Alabama	Good	2022	12	10
Alabama	Good	2022	8	24
Alabama	Good	2022	9	27
Alabama	Good	2023	1	29
Alabama	Good	2023	10	28
Alabama	Good	2023	11	30
Alabama	Good	2023	12	59
Alabama	Good	2023	2	16
Alabama	Good	2023	3	18
Alabama	Good	2023	4	20
Alabama	Good	2023	5	9
Alabama	Good	2023	6	18
Alabama	Good	2023	7	23
Alabama	Good	2023	8	21
Alabama	Good	2023	9	23

5. Report

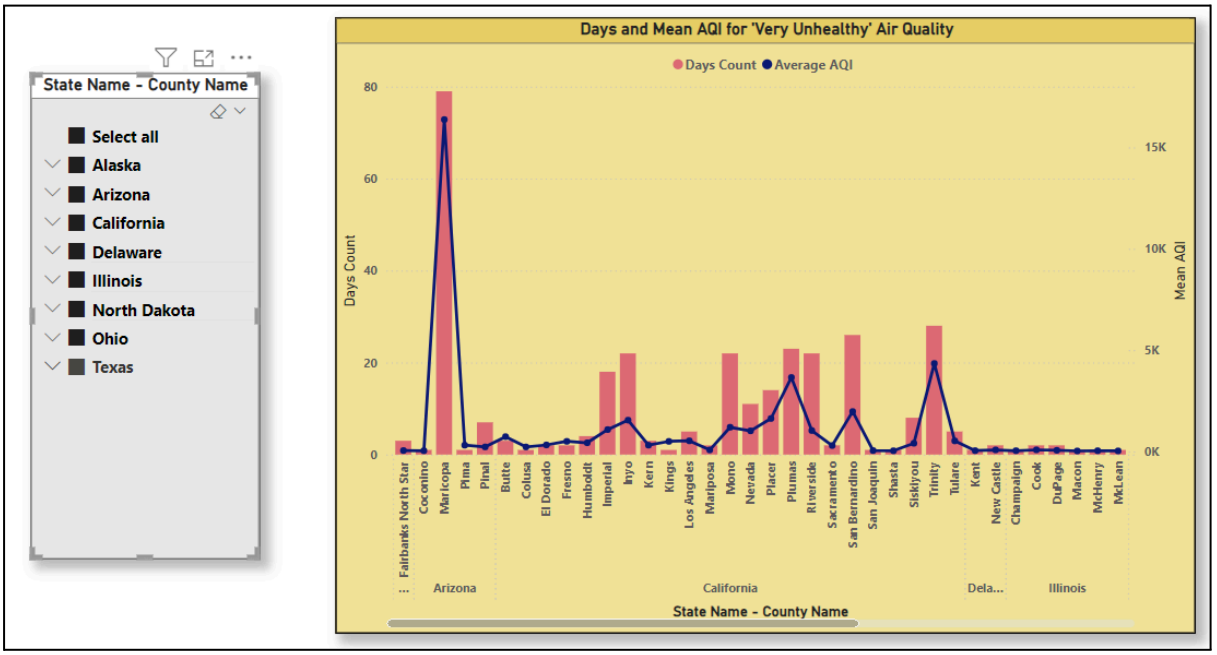
5.1. Báo cáo giá trị AQI nhỏ nhất và lớn nhất cho từng bang trong từng quý của các năm



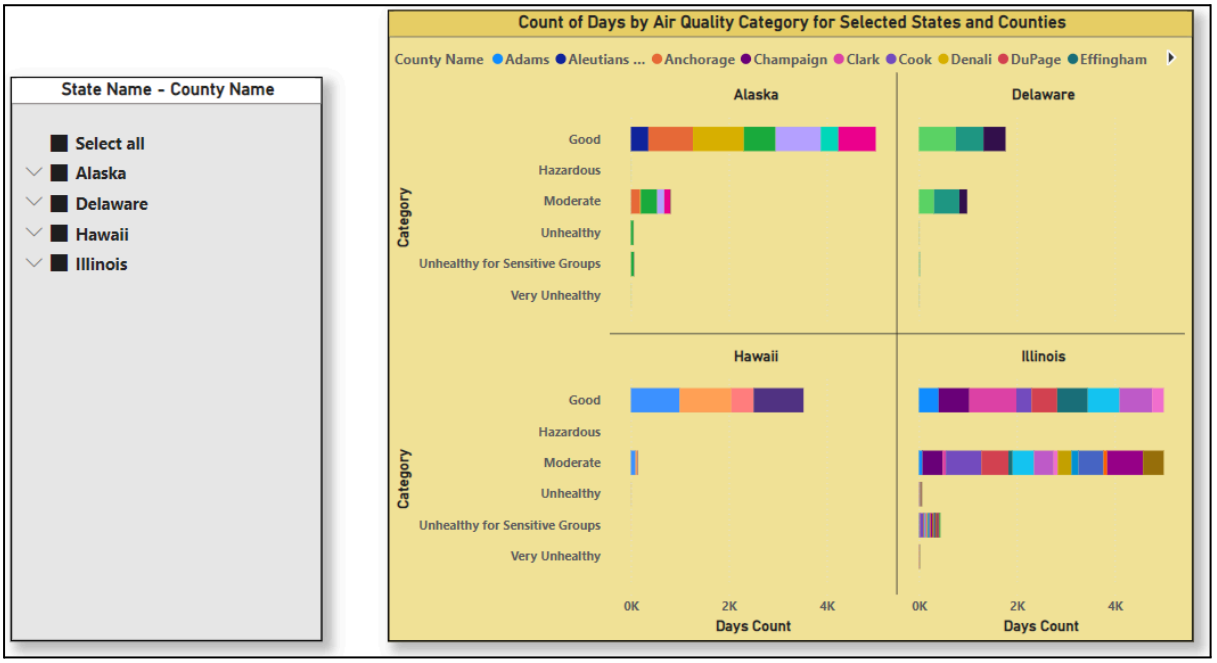
5.2. Báo cáo giá trị trung bình (mean) và độ lệch chuẩn (standard deviation) của AQI cho từng bang trong từng quý của các năm.



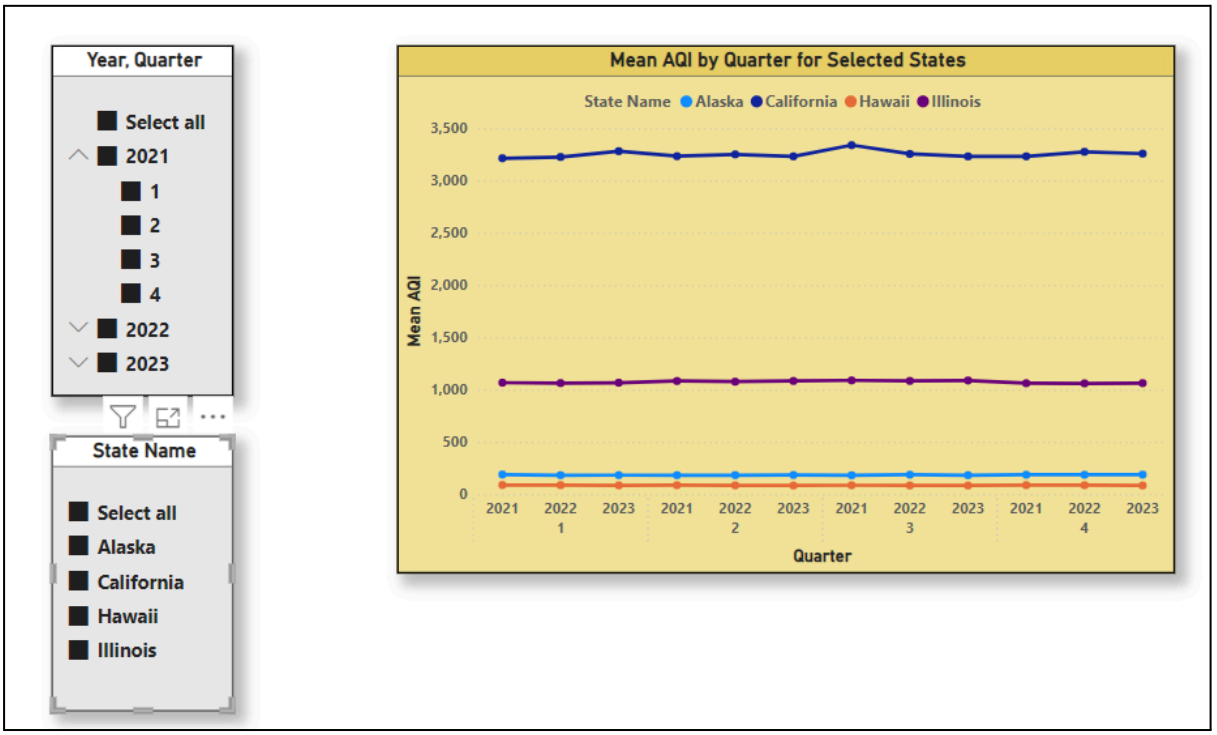
5.3. Báo cáo số ngày và giá trị AQI trung bình khi chất lượng không khí được xếp hạng là "rất không lành mạnh" (very unhealthy) hoặc tệ hơn cho từng bang và từng quận.



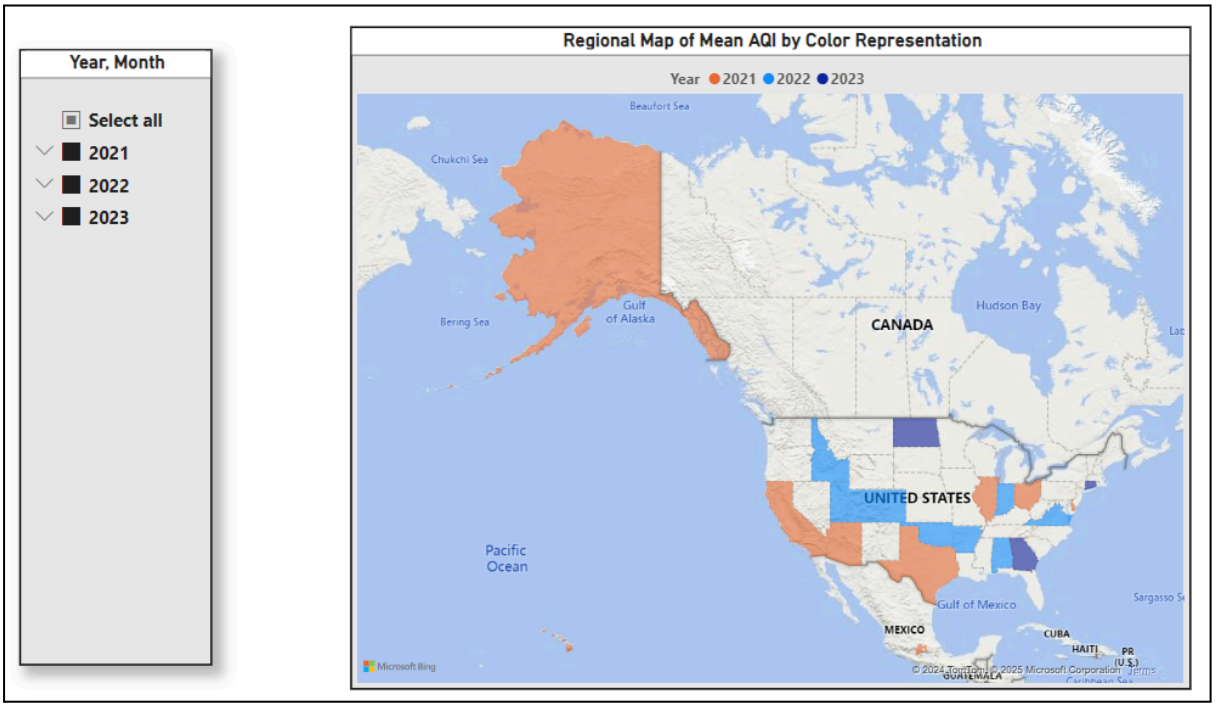
5.4. Đối với 4 bang sau: Hawaii, Alaska, Illinois và Delaware, đếm số ngày trong từng hạng mục chất lượng không khí (Tốt, Trung bình, v.v.) theo từng quận.



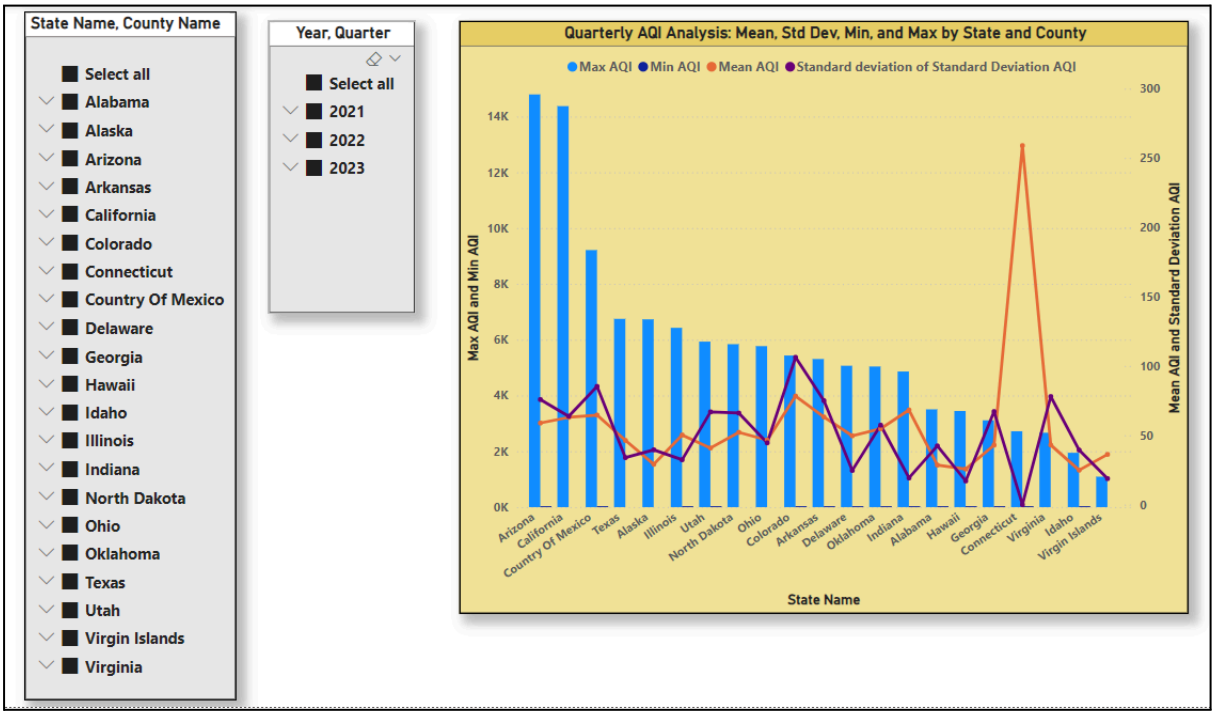
5.5. Đối với 4 bang sau: Hawaii, Alaska, Illinois và Delaware, tính giá trị AQI trung bình theo từng quý.



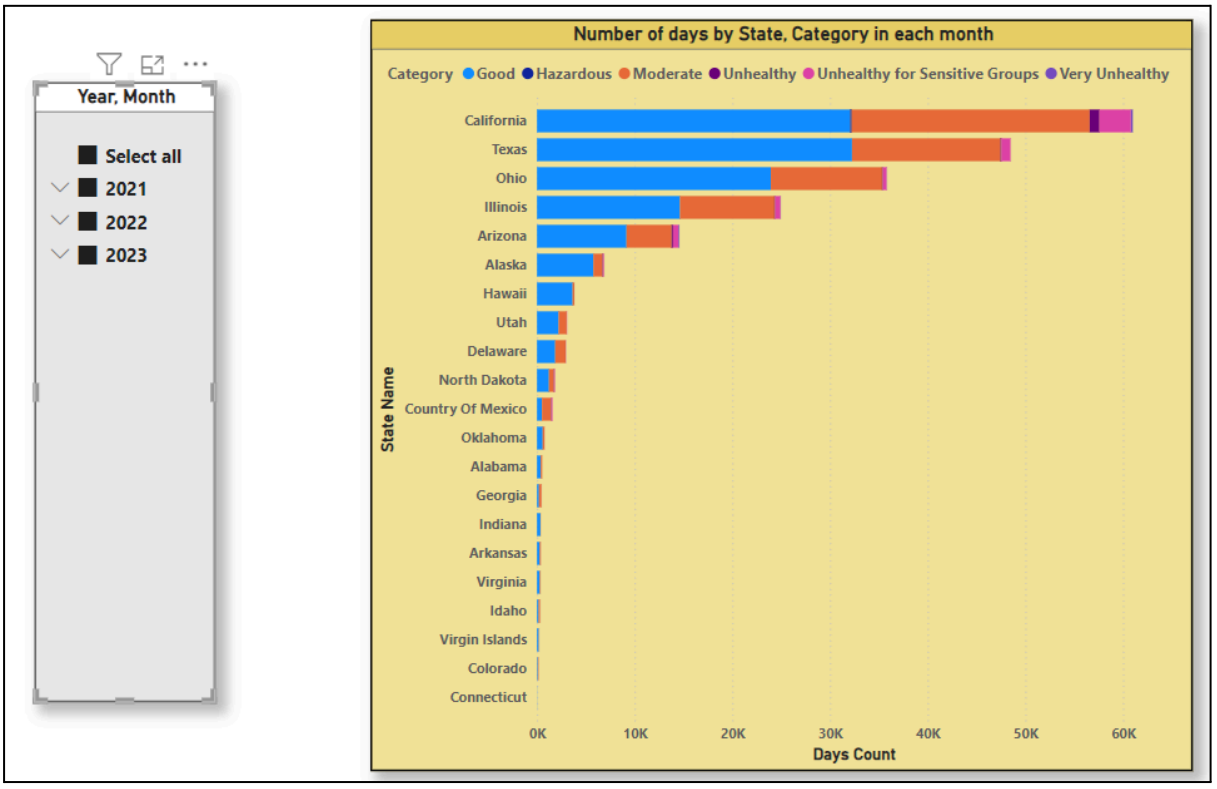
5.6. Sử dụng bản đồ khu vực để trực quan hóa (bằng màu sắc) giá trị trung bình AQI trong các khu vực trong một năm



5.7. Báo cáo giá trị trung bình (mean), độ lệch chuẩn (standard deviation), giá trị nhỏ nhất (min) và lớn nhất (max) của AQI theo từng bang và quận trong mỗi quý của năm.



5.8. Đếm số ngày theo từng bang và hạng mục (Category) trong mỗi tháng.



6. Mining

6.1. Mục tiêu

- Phân tích và dự đoán giá trị AQI trung bình trong tương lai, cụ thể:
  + Dự đoán giá trị AQI trung bình trong các tháng hoặc quý tiếp theo, ví dụ: Quý 1 năm 2024.
  + Đánh giá xu hướng AQI của một State cụ thể (ví dụ: Texas) để xác định các biện pháp quản lý không khí.

6.2. Đề xuất dùng phương pháp Time Series Analysis

- Vì nó bao gồm các quan sát có thứ tự theo thời gian. Do đó, Time Series Analysis là một lựa chọn tự nhiên và hiệu quả để dự đoán giá trị tương lai dựa trên các mẫu và xu hướng hiện tại.
- Hai mô hình sẽ được sử dụng là ARIMA và Prophet. Cả hai đều phù hợp với bài toán phân tích chuỗi thời gian, nhưng mỗi mô hình có cách tiếp cận và ưu điểm riêng:
  + ARIMA (AutoRegressive Integrated Moving Average) được biết đến với tính chính xác cao trong các dự báo có xu hướng và tính mùa vụ ổn định.
  + Prophet được phát triển bởi Facebook, nổi bật với khả năng tự động phát hiện và phân tích xu hướng, mùa vụ cũng như dễ triển khai.

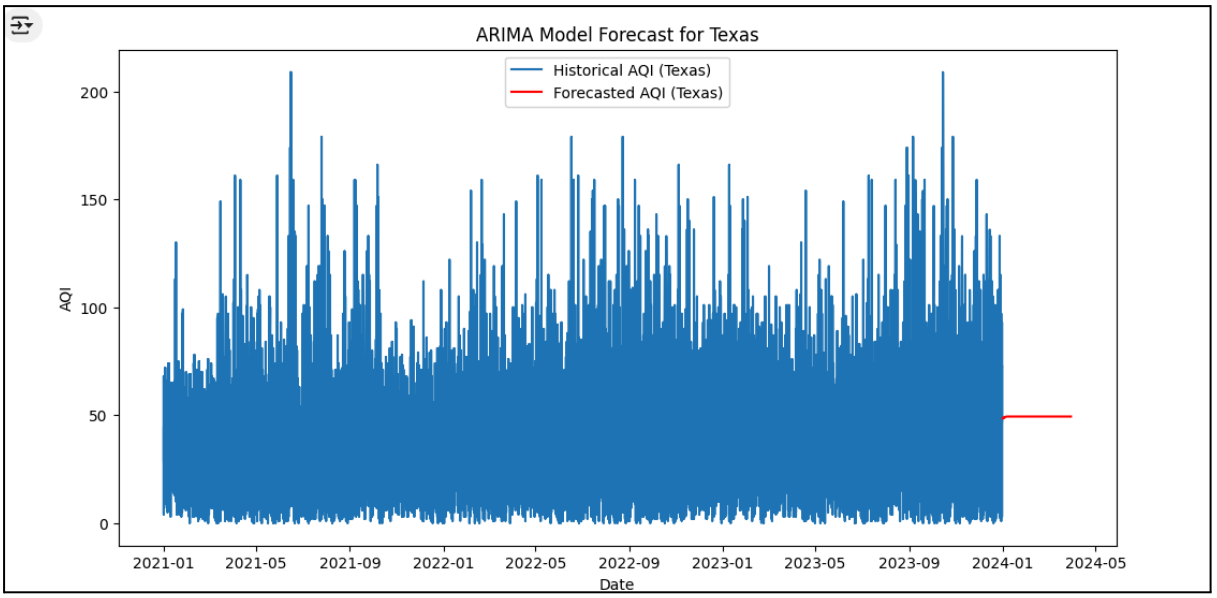
Dữ liệu cần chuẩn bị gồm:

  + date: Cột thời gian (kiểu DateTime).
  + AQI: Giá trị AQI trung bình.
  + state và county (tùy chọn): Dùng để phân tích theo khu vực.

1	1/1/2021	52	New Castl	Delaware
2	1/1/2021	26	Warren	Ohio
3	1/1/2021	34	Trumbull	Ohio
4	1/1/2021	44	Brewster	Texas
5	1/1/2021	33	Bell	Texas
6	1/1/2021	52	Sussex	Delaware
7	1/1/2021	31	Brazoria	Texas
8	1/1/2021	157	BAJA CALIF	Country Of Mexico

6.2.1. ARIMA

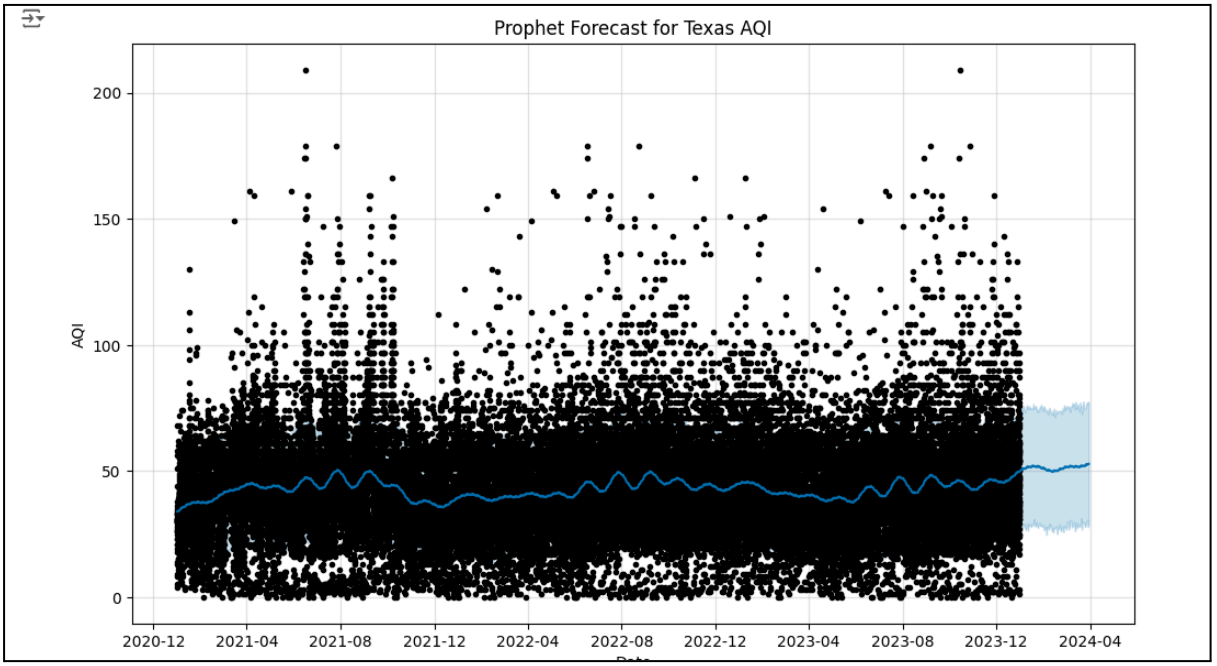
- Triển khai bằng Python

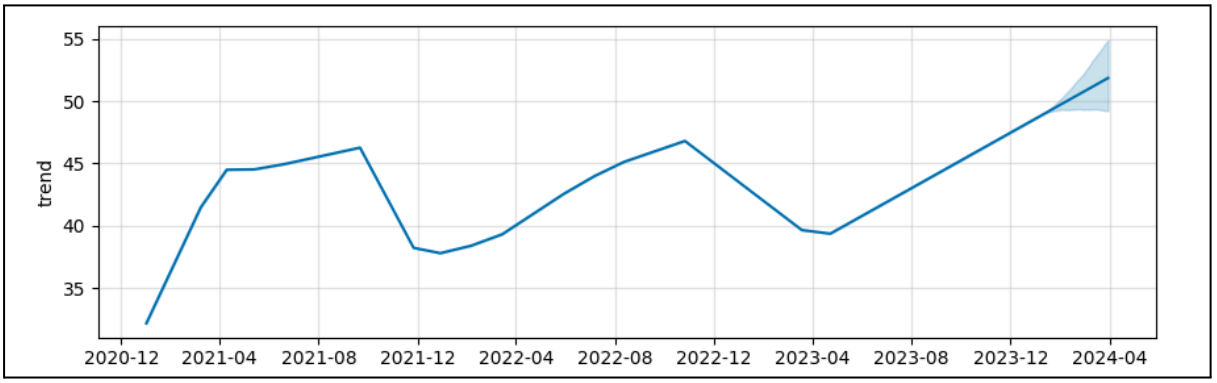


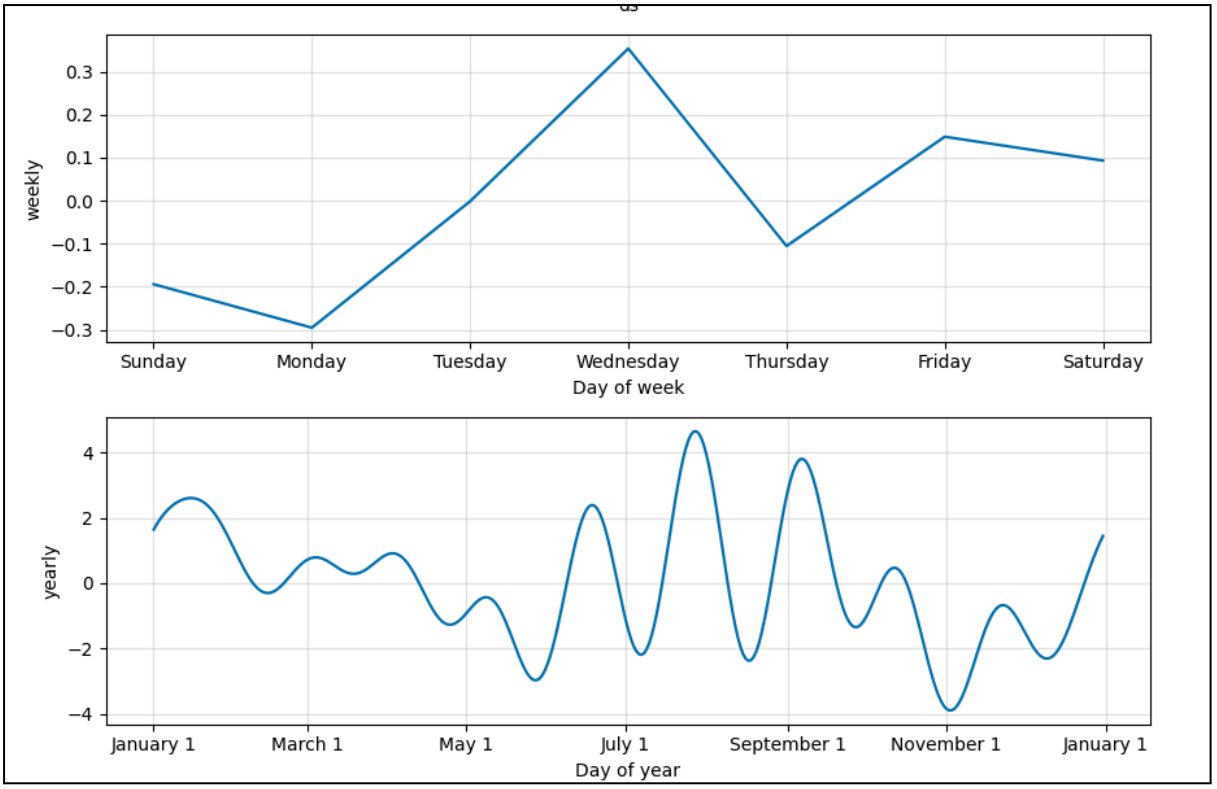
- Kết quả cho thấy ARIMA dự báo là một đường thẳng ngang, cho thấy mô hình chưa đánh giá được chính xác dự đoán xu hướng.

  => Mô hình chưa phù hợp cho yêu cầu

6.2.2. Prophet


Prophet Forecast for Texas AQI
AQI
200
150
100
50
0
2020-12
2021-04
2021-08
2021-12
2022-04
2022-08
2022-12
2023-04
2023-08
2023-12
2024-04


trend
55
50
45
40
35
2020-12
2021-04
2021-08
2021-12
2022-04
2022-08
2022-12
2023-04
2023-08
2023-12
2024-04



- Kết quả cho thấy dự đoán quý 1 năm 2024 của mô hình Prophet có chiều hướng đi lên, cho thấy mô hình có kết quả khả quan hơn khi có sự biến động cụ thể hơn là AMIRA.
- Dựa theo kết quả này ta có thể thấy dự báo chỉ số AQI của bang Texas có xu hướng gia tăng vào quý 1 năm 2024

https://github.com/20127200/SSIS_AQI.git