

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 CUỐI KỲ

Môn học: Hệ thống thông tin phục vụ Trí tuệ Kinh doanh

Nhóm CQ.BI.2425.06:

Giảng viên hướng dẫn:

21120394 - Nguyễn Hoàng Ngọc Bảo

Hồ Thị Hoàng Vy

21120405 - Trần Minh Triết

Tiết Gia Hồng

21120424 - Nguyễn Đình Phương Đại

Nguyễn Ngọc Minh Châu

21120433 - Nguyễn Quang Định



fit@hcmus

MỤC LỤC

I. Thông tin thành viên.....	2
II. Phân tích cơ sở dữ liệu.....	2
1. Stage.....	2
2. NDS.....	5
3. DDS.....	8
4. Metadata.....	13
III. ETL.....	15
1. Source to Stage.....	15
2. Stage to NDS.....	24
3. NDS to DDS.....	35
IV. OLAP.....	45
V. MDX.....	54
1. Report the min and max of AQI value for each State during each quarter of years.....	54
2. Report the mean and the standard deviation of AQI value for each State during each quarter of years.....	57
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.....	63
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.....	66
5. For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the mean AQI value by quarters.....	69
6. Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California.....	71
7. Build graphs/charts for the above reports.....	74
8. Use a regional map to visually represent (by color) the mean AQI value in regions during a year.....	74
9. Report the mean, the standard deviation, min and max of AQI value group by State and County during each quarter of the year.....	77
10. Report the mean AQI value by State, Category, DayLightSaving over	

years.....	80
11. Count the number of days by State, Category in each month.....	83
12. Report the number of days by Category and Defining Parameter.....	87
VI. Mining.....	89
1. Mô hình ARIMA.....	89
a. Tổng quan về ARIMA.....	89
Các tham số trong ARIMA:.....	89
b. Tính dừng:.....	90
c. Lựa chọn tham số ARIMA (p, d, q).....	90
2. Mining data.....	91
Tham Khảo.....	98

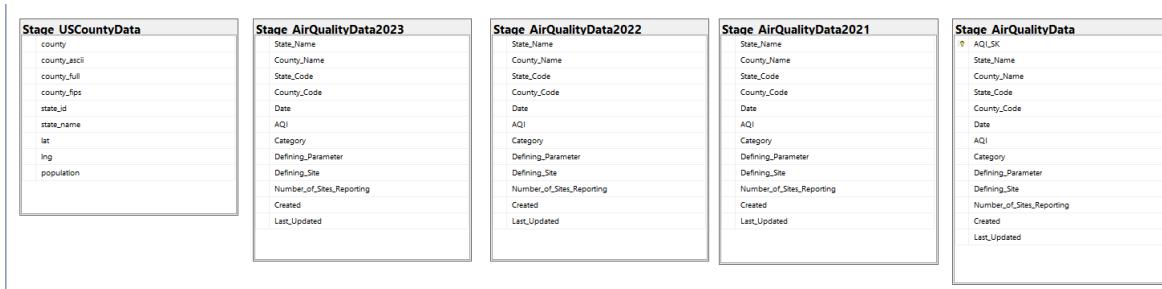
I. Thông tin thành viên

Nhóm: CQ.BI.2425.06

MSSV	Họ và tên	Phân công	Đánh giá
21120394	Nguyễn Hoàng Ngọc Bảo	Project Manager, ETL, MDX + tổng hợp MDX, làm Dashboard, tổng hợp và viết báo cáo.	100%
21120405	Trần Minh Triết	Tham gia làm ETL, MDX, Data Mining, tổng hợp và viết báo cáo.	100%
21120424	Nguyễn Đình Phương Đại	Làm video, tham gia làm ETL, MDX, tổng hợp và sửa chữa báo cáo.	100%
21120433	Nguyễn Quang Định	Tham gia làm ETL, MDX, Data Mining tổng hợp và sửa chữa báo cáo..	100%

II. Phân tích cơ sở dữ liệu

1. Stage



Stage bao gồm:

- Stage_USCountyData: Dùng để chứa dữ liệu của file Excel (2B)uscounties.csv

```
DROP TABLE IF EXISTS PROJECT_STAGE.dbo.Stage_USCountyData;
CREATE TABLE PROJECT_STAGE.dbo.Stage_USCountyData (
    county VARCHAR(50),
    county_ascii VARCHAR(50),
    county_full VARCHAR(100),
    county_fips CHAR(5),
    state_id CHAR(2),
    state_name VARCHAR(50),
    lat DECIMAL(8, 5),
    lng DECIMAL(8, 5),
    population INT
);
GO
```

- Stage_AirQualityData2021, Stage_AirQualityData2022, Stage_AirQualityData2023: Dùng để chứa dữ liệu tương ứng từ 3 file Excel 10_state_aqi_2021.csv, 10_state_aqi_2022.csv, 10_state_aqi_2023.csv

```
DROP TABLE IF EXISTS PROJECT_STAGE.dbo.Stage_AirQualityData2021;
CREATE TABLE PROJECT_STAGE.dbo.Stage_AirQualityData2021 (
    State_Name VARCHAR(50),
    County_Name VARCHAR(50),
    State_Code INT,
    County_Code INT,
    Date DATE,
    AQI INT,
    Category VARCHAR(50),
    Defining_Parameter VARCHAR(50),
    Defining_Site VARCHAR(50),
    Number_of_Sites_Reported INT,
    Created DATETIME,
    Last_Updated DATETIME
);
```

```

DROP TABLE IF EXISTS PROJECT_STAGE.dbo.Stage_AirQualityData2022;
CREATE TABLE PROJECT_STAGE.dbo.Stage_AirQualityData2022 (
    State_Name VARCHAR(50),
    County_Name VARCHAR(50),
    State_Code INT,
    County_Code INT,
    Date DATE,
    AQI INT,
    Category VARCHAR(50),
    Defining_Parameter VARCHAR(50),
    Defining_Site VARCHAR(50),
    Number_of_Sites_Reported INT,
    Created DATETIME,
    Last_Updated DATETIME
);
GO

DROP TABLE IF EXISTS PROJECT_STAGE.dbo.Stage_AirQualityData2023;
CREATE TABLE PROJECT_STAGE.dbo.Stage_AirQualityData2023 (
    State_Name VARCHAR(50),
    County_Name VARCHAR(50),
    State_Code INT,
    County_Code INT,
    Date DATE,
    AQI INT,
    Category VARCHAR(50),
    Defining_Parameter VARCHAR(50),
    Defining_Site VARCHAR(50),
    Number_of_Sites_Reported INT,
    Created DATETIME,
    Last_Updated DATETIME
);
GO

```

- Stage_AirQualityData: Dùng để tổng hợp toàn bộ dữ liệu của 3 table trên.

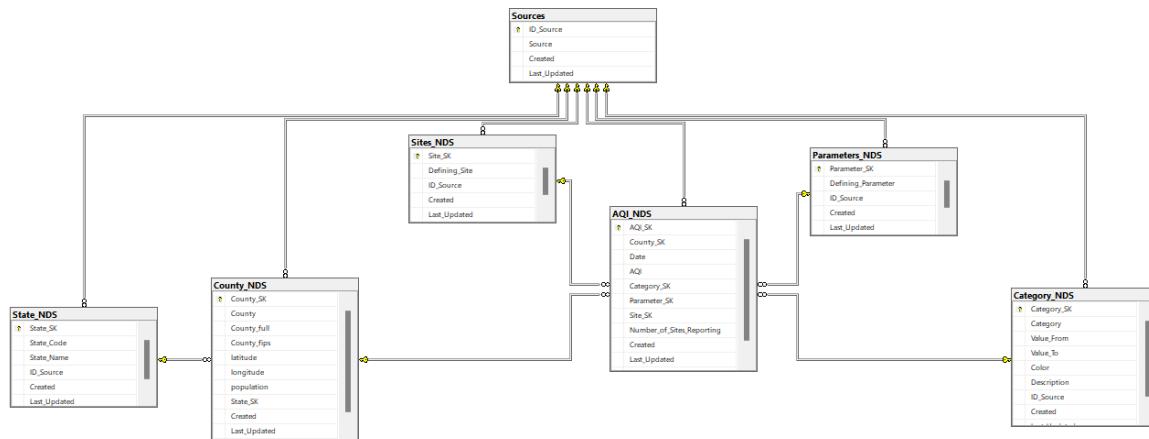
```

DROP TABLE IF EXISTS PROJECT_STAGE.dbo.Stage_AirQualityData;
CREATE TABLE PROJECT_STAGE.dbo.Stage_AirQualityData (
    AQI_SK INT IDENTITY(1,1) PRIMARY KEY,
    State_Name VARCHAR(50),
    County_Name VARCHAR(50),
    State_Code INT,
    County_Code INT,
    Date DATE,
    AQI INT,
    Category VARCHAR(50),
    Defining_Parameter VARCHAR(50),
    Defining_Site VARCHAR(50),
    Number_of_Sites_Reported INT,
    Created DATETIME,
    Last_Updated DATETIME,
    UNIQUE(State_Code,County_Code,Date, Defining_Parameter)
);
GO

```

2. NDS

Cấu trúc của NDS như sau:



- Sources: bảng lưu thông tin về nguồn dữ liệu, bao gồm: khóa SK tự tăng, tên nguồn, ngày tạo và ngày update.

```
[CREATE TABLE PROJECT_NDS.dbo.Sources (
    ID_Source INT IDENTITY(1,1) PRIMARY KEY,
    Source VARCHAR(255),
    Created DATETIME,
    Last_Updated DATETIME
);
```

- State_NDS: bảng lưu thông tin về tiểu bang, gồm: khóa SK tự tăng, code bang, tên bang, ID nguồn, ngày tạo và ngày update.

```
[CREATE TABLE PROJECT_NDS.dbo.State_NDS (
    State_SK INT IDENTITY(1,1) PRIMARY KEY,
    State_Code VARCHAR(50) UNIQUE,
    State_Name VARCHAR(50),
    ID_Source INT FOREIGN KEY REFERENCES Sources(ID_Source),
    Created DATETIME,
    Last_Updated DATETIME
);
```

- County_NDS: bảng lưu thông tin về hạt, gồm: khóa SK tự tăng, tên hạt, tên đầy đủ, kinh độ, vĩ độ, dân số, tiểu bang trực thuộc, nguồn, ngày tạo và ngày update.

```
[CREATE TABLE PROJECT_NDS.dbo.County_NDS (
    County_SK INT IDENTITY(1,1) PRIMARY KEY,
    County VARCHAR(50),
    County_full VARCHAR(100),
    County_fips VARCHAR(50) UNIQUE,
    latitude FLOAT,
    longitude FLOAT,
    population INT,
    State_SK INT FOREIGN KEY REFERENCES State_NDS(State_SK),
    Created DATETIME,
    Last_Updated DATETIME,
    ID_Source INT FOREIGN KEY REFERENCES Sources(ID_Source),
);
```

- Category_NDS: lưu thông tin về phân loại chất lượng không khí, gồm: khóa SK tự tăng, tên, giới hạn trái và phải của miền giá trị, màu hiển thị, mô tả, nguồn,

và ngày tạo, update.

```
CREATE TABLE PROJECT_NDS.dbo.Category_NDS (
    Category_SK INT IDENTITY(1,1) PRIMARY KEY,
    Category VARCHAR(50) NOT NULL,
    Value_From INT NOT NULL,
    Value_To INT NOT NULL,
    Color VARCHAR(20) NOT NULL,
    Description VARCHAR(500),
    ID_Source INT FOREIGN KEY REFERENCES Sources(ID_Source),
    Created DATETIME,
    Last_Updated DATETIME
);
```

- Parameter_NDS: lưu thông tin về thang đo, bao gồm khóa SK tự tăng, tên thang đo, nguồn dữ liệu và ngày tạo, update.

```
CREATE TABLE PROJECT_NDS.dbo.Parameters_NDS (
    Parameter_SK INT IDENTITY(1,1) PRIMARY KEY,
    Defining_Parameter VARCHAR(50) UNIQUE,
    ID_Source INT FOREIGN KEY REFERENCES Sources(ID_Source),
    Created DATETIME,
    Last_Updated DATETIME
);
```

- Sites_NDS: lưu thông tin về trạm đo lường, gồm khóa SK tự tăng, mã của trạm đo lường, nguồn, và ngày tạo, update.

```
CREATE TABLE PROJECT_NDS.dbo.Sites_NDS (
    Site_SK INT IDENTITY(1,1) PRIMARY KEY,
    Defining_Site VARCHAR(50) UNIQUE,
    ID_Source INT FOREIGN KEY REFERENCES Sources(ID_Source),
    Created DATETIME,
    Last_Updated DATETIME
);
```

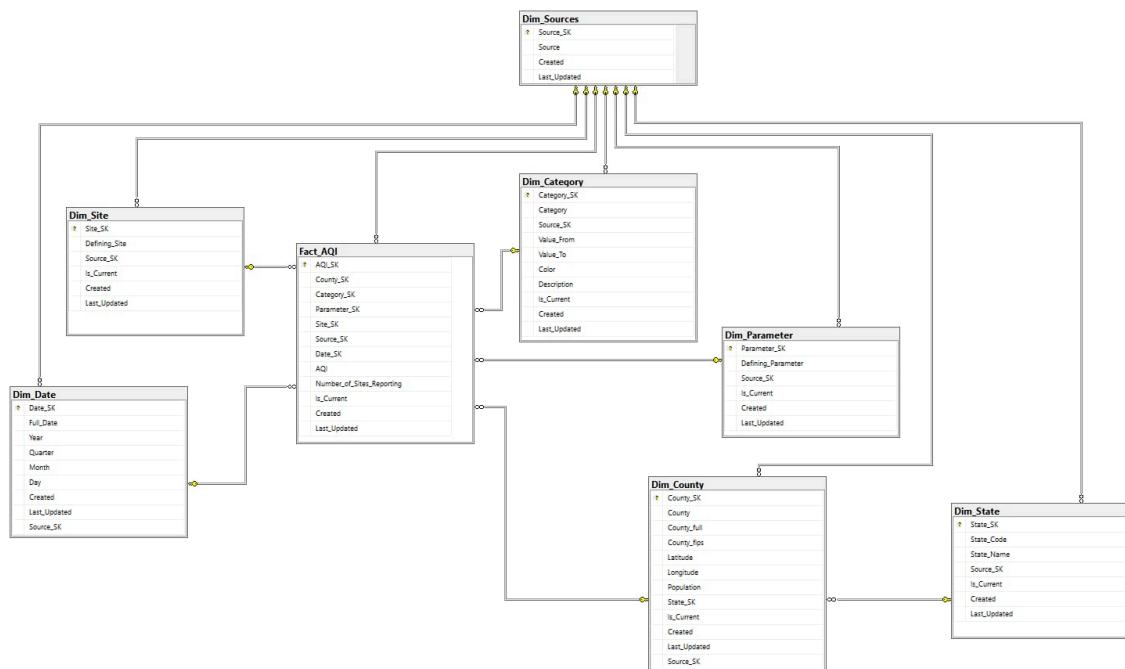
- AQI_NDS: lưu các thông tin về chất lượng không khí, bao gồm khóa SK tự tăng, tên hạt, ngày đo lường, phân loại chất lượng, thang đo, trạm đo, số lượng trạm

báo cáo, ngày tạo, ngày update, nguồn dữ liệu.

```
-- Table: AQI_NDS
DROP TABLE IF EXISTS PROJECT_NDS.dbo.AQI_NDS;
CREATE TABLE PROJECT_NDS.dbo.AQI_NDS (
    AQI_SK INT IDENTITY(1,1) PRIMARY KEY,
    County_SK INT FOREIGN KEY REFERENCES County_NDS(County_SK),
    Date DATE,
    AQI INT,
    Category_SK INT FOREIGN KEY REFERENCES Category_NDS(Category_SK),
    Parameter_SK INT FOREIGN KEY REFERENCES Parameters_NDS(Parameter_SK),
    Site_SK INT FOREIGN KEY REFERENCES Sites_NDS(Site_SK),
    Number_of_Sites_Reported INT,
    Created DATETIME,
    Last_Updated DATETIME,
    ID_Source INT FOREIGN KEY REFERENCES Sources(ID_Source),
    UNIQUE (County_SK, Date, Parameter_SK)
);
```

3. DDS

Cấu trúc của DDS như sau:



Trong đó bao gồm các bảng:

- Fact_AQI: Dùng để chứa dữ liệu của chất lượng không khí cho việc phân tích: bao gồm các thuộc tính khóa của các bảng chiều, các thuộc tính đo lường. Ngoài ra còn lưu thời gian được tạo và lần update gần nhất.

```

CREATE TABLE PROJECT_DDS.dbo.Fact_AQI (
    AQI_SK INT PRIMARY KEY,
    County_SK INT FOREIGN KEY REFERENCES Dim_County(County_SK),
    Category_SK INT FOREIGN KEY REFERENCES Dim_Category(Category_SK),
    Parameter_SK INT FOREIGN KEY REFERENCES Dim_Parameter(Parameter_SK),
    Site_SK INT FOREIGN KEY REFERENCES Dim_Site(Site_SK),
    Source_SK INT FOREIGN KEY REFERENCES Dim_Sources(Source_SK),
    Date_SK INT FOREIGN KEY REFERENCES Dim_Date(Date_SK),
    AQI INT,
    Number_of_Sites_Reported INT,
    Is_Current BIT,
    Created DATETIME,
    Last_Updated DATETIME
);

```

- Bảng Dim_Source: dùng để lưu thông tin của các nguồn dữ liệu, bao gồm: surrogate key, tên nguồn, thời gian tạo và update gần nhất.

```

CREATE TABLE PROJECT_DDS.dbo.Dim_Sources (
    Source_SK INT PRIMARY KEY,
    Source VARCHAR(255),
    Created DATETIME,
    Last_Updated DATETIME
);

```

- Dim_Date: dùng để lưu thông tin cho chiều ngày, bao gồm các thông tin như: surgerate key, ngày tạo và ngày cập nhật, ngày đầy đủ và các trường chia nhỏ từ ngày đầy đủ ngày, tháng, quý, năm. Khóa ngoại để quản lý nguồn dữ liệu. Do ngày đầy đủ được chia thành các thành phần riêng lẻ, hai procedure được tạo để thêm và cập nhật dữ liệu vào bảng này.

```

CREATE TABLE PROJECT_dds.dbo.Dim_Date (
    Date_SK INT PRIMARY KEY,
    Full_Date DATE NOT NULL,
    Year INT NOT NULL,
    Quarter INT NOT NULL,
    Month INT NOT NULL,
    Day INT NOT NULL,
    Created DATETIME,
    Last_Updated DATETIME,
    Source_SK INT FOREIGN KEY REFERENCES Dim_Sources(Source_SK)
);

CREATE OR ALTER PROCEDURE Add_Dim_Date
    @InputDate DATE,          -- Ngày cần thêm
    @Source_SK INT,           -- Khóa ngoại nguồn dữ liệu
    @Created DATE,
    @Updated DATE
AS
BEGIN
    SET NOCOUNT ON;

    -- Kiểm tra xem ngày đã tồn tại trong Dim_Date chưa
    IF NOT EXISTS (SELECT 1 FROM PROJECT_dds.dbo.Dim_Date WHERE Full_Date = @InputDate)
    BEGIN
        INSERT INTO PROJECT_dds.dbo.Dim_Date (
            Date_SK,
            Full_Date,
            Year,
            Quarter,
            Month,
            Day,
            Created,
            Last_Updated,
            Source_SK
        )
        VALUES (
            CONVERT(INT, FORMAT(@InputDate, 'yyyyMMdd')), -- Tạo Date_SK từ ngày
            @InputDate,                                     -- Năm
            YEAR(@InputDate),                             -- Quý
            DATEPART(QUARTER, @InputDate),                -- Tháng
            MONTH(@InputDate),                            -- Ngày
            DAY(@InputDate),                             -- Ngày tạo
            @Created,                                    -- Ngày cập nhật
            @Source_SK                                    -- Khóa ngoại nguồn
        );
    END
END

```

```

]CREATE OR ALTER PROCEDURE Update_Dim_Date
    @InputDate DATE,          -- Ngày cần cập nhật
    @Source_SK INT,           -- Khóa ngoại nguồn dữ liệu
    @Update_Date DATE
AS
BEGIN
    SET NOCOUNT ON;

    -- Kiểm tra xem ngày đã tồn tại trong Dim_Date chưa
    IF EXISTS (SELECT 1 FROM PROJECT_DDS.dbo.Dim_Date WHERE Full_Date = @InputDate)
    BEGIN
        UPDATE PROJECT_DDS.dbo.Dim_Date
        SET
            Year = YEAR(@InputDate),          -- Cập nhật Năm
            Quarter = DATEPART(QUARTER, @InputDate), -- Cập nhật Quý
            Month = MONTH(@InputDate),        -- Cập nhật Tháng
            Day = DAY(@InputDate),           -- Cập nhật Ngày
            Last_Updated = @Update_Date,     -- Cập nhật Ngày cập nhật
            Source_SK = @Source_SK          -- Cập nhật Khóa ngoại nguồn
        WHERE Full_Date = @InputDate;
    END
    ELSE
    BEGIN
        PRINT 'Ngày này không tồn tại trong Dim_Date để cập nhật.';
    END
END;
GO

```

- Dim_State: dùng để lưu thông tin cho chiều tiểu bang, gồm các thông tin: surrogate key, mã code và tên của tiểu bang, nguồn dữ liệu này được nạp vào, thời gian tạo và cập nhật. Thuộc tính IS_CURRENT là thuộc tính cờ dùng để xác định liệu dòng này có đang được sử dụng để ghi dữ liệu lịch sử khi thực hiện Slowly Changing Dimension hay không.

```

]CREATE TABLE PROJECT_DDS.dbo.Dim_State (
    State_SK INT PRIMARY KEY,
    State_Code VARCHAR(50) UNIQUE,
    State_Name VARCHAR(50),
    Source_SK INT FOREIGN KEY REFERENCES Dim_Sources(Source_SK),
    Is_Current BIT,
    Created DATETIME,
    Last_Updated DATETIME
);
GO

```

- Dim_County: dùng để lưu thông tin cho chiều hạt, gồm các thông tin như khóa SK, mã hạt, tên đầy đủ, kinh tuyến, vĩ tuyến, dân số, tiểu bang trực thuộc, biến cờ để xác định có đang được sử dụng, ngày tạo và update.

```

CREATE TABLE PROJECT_DDS.dbo.Dim_County (
    County_SK INT PRIMARY KEY,
    County VARCHAR(50),
    County_full VARCHAR(100),
    County_fips VARCHAR(50) UNIQUE,
    Latitude FLOAT,
    Longitude FLOAT,
    Population INT,
    State_SK INT FOREIGN KEY REFERENCES Dim_State(State_SK),
    Is_Current BIT,
    Created DATETIME,
    Last_Updated DATETIME,
    Source_SK INT FOREIGN KEY REFERENCES Dim_Sources(Source_SK)
);

```

- Dim_Category: dùng để lưu thông tin cho chiều phân loại, bao gồm: khóa SK, tên phân loại, nguồn dữ liệu, giới hạn trái và phải của miền giá trị, màu, mô tả, thuộc tính cờ và ngày tạo, cập nhật.

```

DROP TABLE IF EXISTS PROJECT_DDS.dbo.Dim_Category;
CREATE TABLE PROJECT_DDS.dbo.Dim_Category (
    Category_SK INT PRIMARY KEY,
    Category VARCHAR(50) UNIQUE,
    Source_SK INT FOREIGN KEY REFERENCES Dim_Sources(Source_SK),
    Value_From INT NOT NULL,
    Value_To INT NOT NULL,
    Color VARCHAR(20) NOT NULL,
    Description VARCHAR(500),
    Is_Current BIT,
    Created DATETIME,
    Last_Updated DATETIME
);

```

- Dim_Parameter: dùng để lưu thông tin cho chiều đơn vị đo, bao gồm: khóa SK, tên parameter, nguồn dữ liệu, thuộc tính cờ và ngày tạo, cập nhật.

```

|DROP TABLE IF EXISTS PROJECT_dds.dbo.Dim_Parameter;
|CREATE TABLE PROJECT_dds.dbo.Dim_Parameter (
    Parameter_SK INT PRIMARY KEY,
    Defining_Parameter VARCHAR(50) UNIQUE,
    Source_SK INT FOREIGN KEY REFERENCES Dim_Sources(Source_SK),
    Is_Current BIT,
    Created DATETIME,
    Last_Updated DATETIME
);

```

- Dim_Sites:

```

|CREATE TABLE PROJECT_dds.dbo.Dim_Site (
    Site_SK INT PRIMARY KEY,
    Defining_Site VARCHAR(50) UNIQUE,
    Source_SK INT FOREIGN KEY REFERENCES Dim_Sources(Source_SK),
    Is_Current BIT,
    Created DATETIME,
    Last_Updated DATETIME
);

```

4. Metadata

PROJECT_METADATA: chứa các metadata về data warehouse.

Hiện tại, CSDL PROJECT_METADATA chỉ chứa bảng data_flow (**ID**, name, LSET, CET) dùng để lưu các thông tin về quá trình SSIS với LSET là lần cuối SSIS thành công và CET là lần SSIS gần nhất của bảng ‘name’.

```

|DROP TABLE IF EXISTS PROJECT_metadata.dbo.data_flow;
|CREATE TABLE PROJECT_metadata.dbo.data_flow (
    id INT IDENTITY(1,1),
    name VARCHAR(50) NOT NULL UNIQUE,
    LSET DATETIME, -- Thời gian SSIS thành công gần nhất
    CET DATETIME, -- Thời gian SSIS hiện tại
    CONSTRAINT pk_data_flow PRIMARY KEY CLUSTERED (id)
);

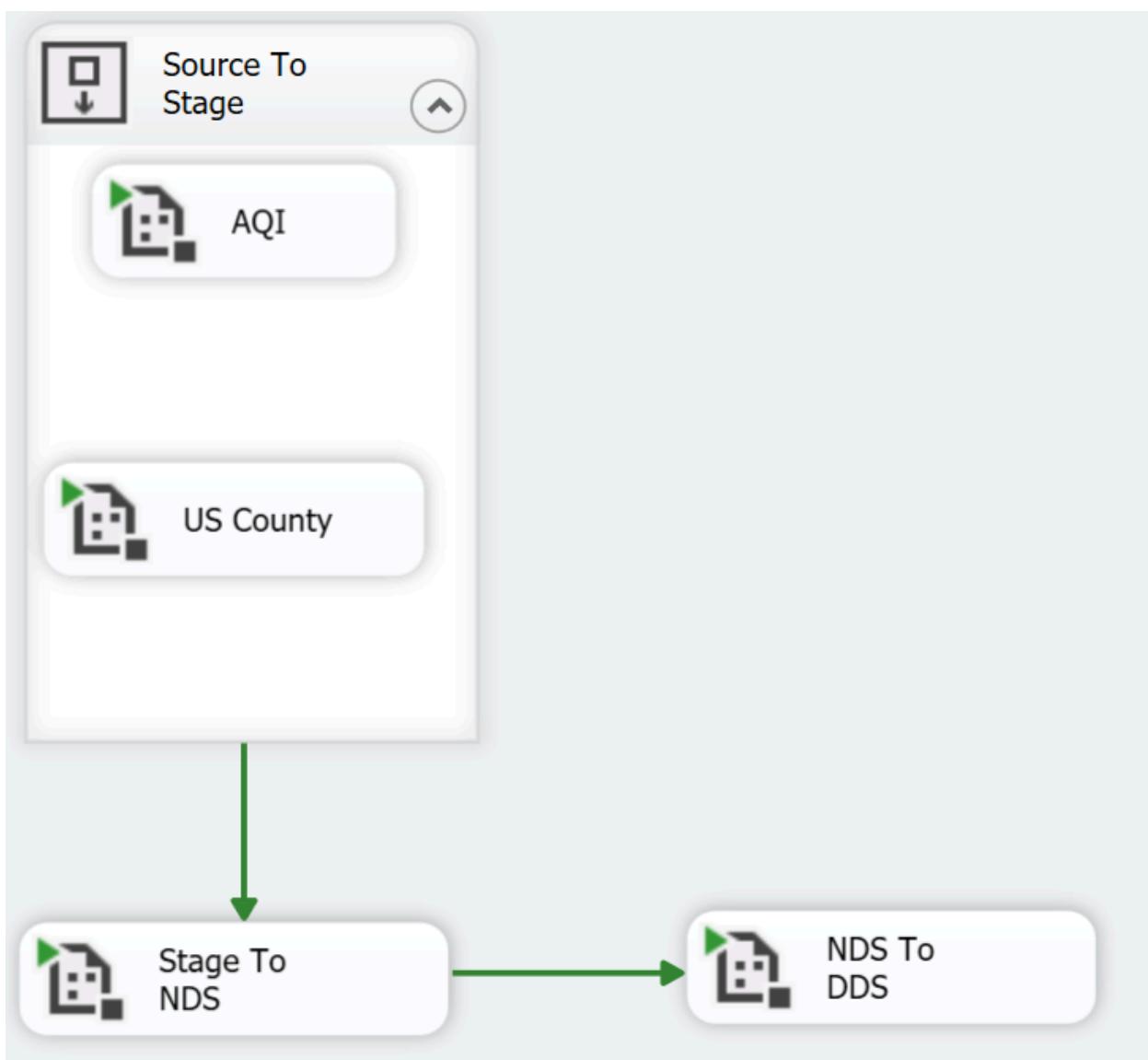
```

Sau đó insert vào data_flow các dòng dữ liệu đại diện cho giá trị CET và LSET của các bảng.

	id	name	LSET	CET
	1	Stage_AirQualityData2021	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	2	Stage_AirQualityData2022	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	3	Stage_AirQualityData2023	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	4	Stage_AirQualityData	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	5	Stage_USCountyData	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	6	Dim_Sources	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	7	Dim_Date	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	8	Dim_State	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	9	Dim_County	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	10	Dim_Category	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	11	Dim_Parameter	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
	12	Dim_Site	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
*	13	Fact_AQI	2020-01-01 00:00:00.000	2020-01-01 00:00:00.000
*	NULL	NULL	NULL	NULL

Nhóm có cung cấp file **setup.sql** để thực hiện cài đặt bước này.

III. ETL

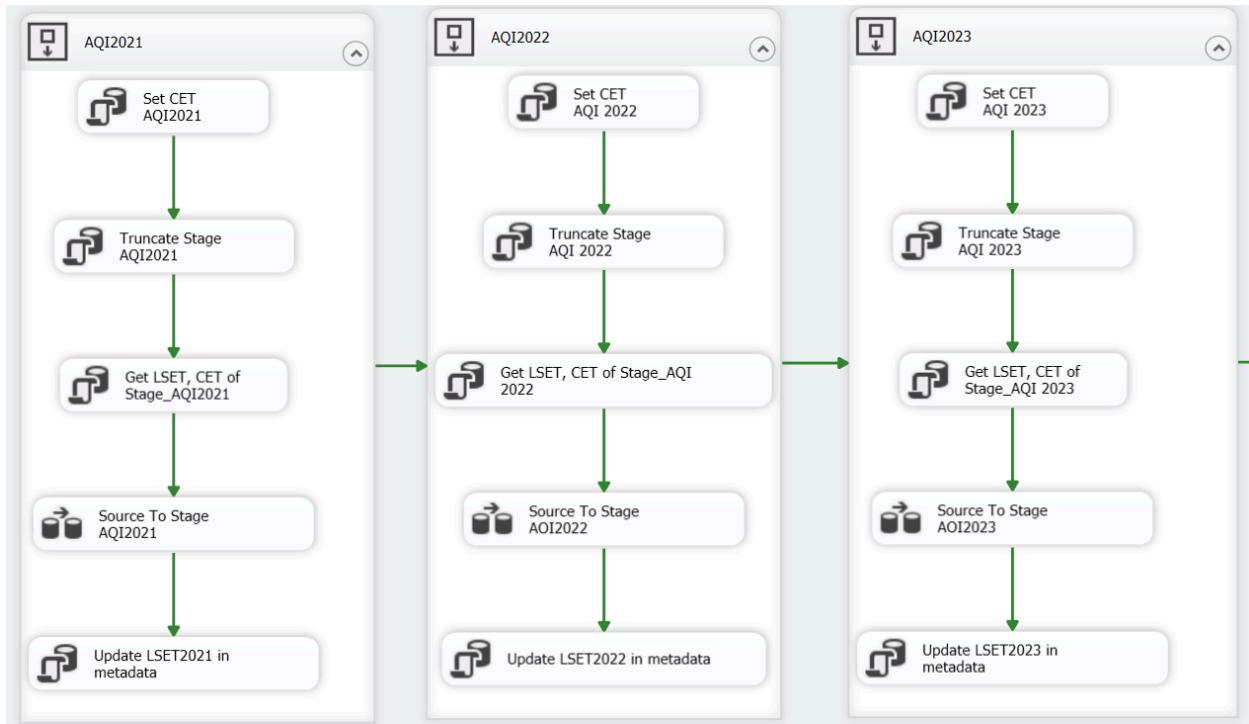


Workflow của quá trình nạp dữ liệu từ nguồn đến DDS

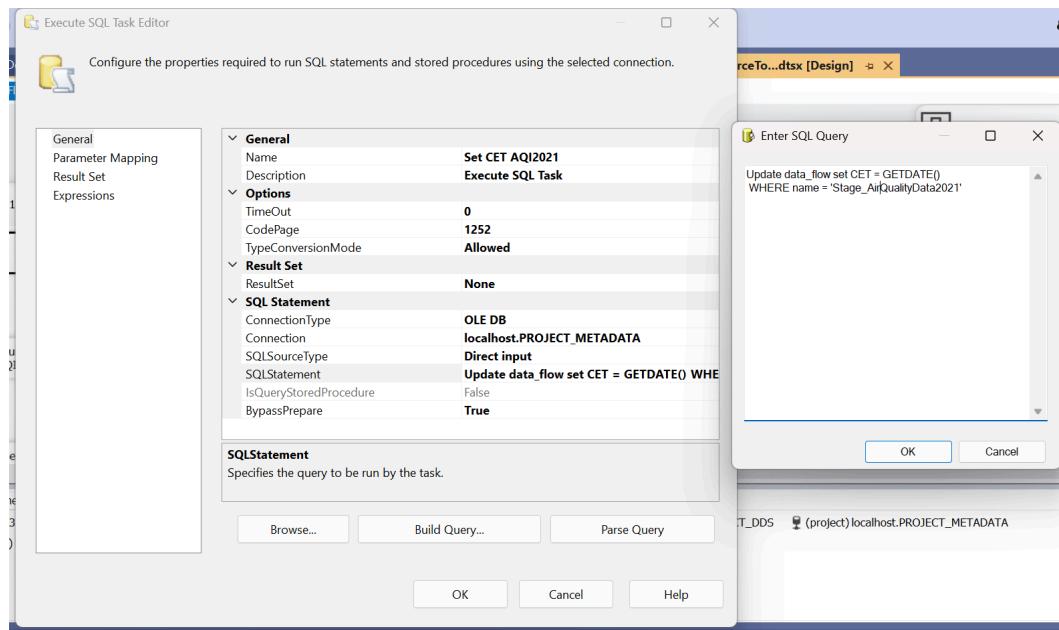
1. Source to Stage

Nhóm sẽ thực hiện load dữ liệu từ các file csv tương ứng vào các bảng Stage tương ứng. Đồng thời lưu trữ, cập nhật các dữ liệu từ metadata để thực hiện incremental loading.

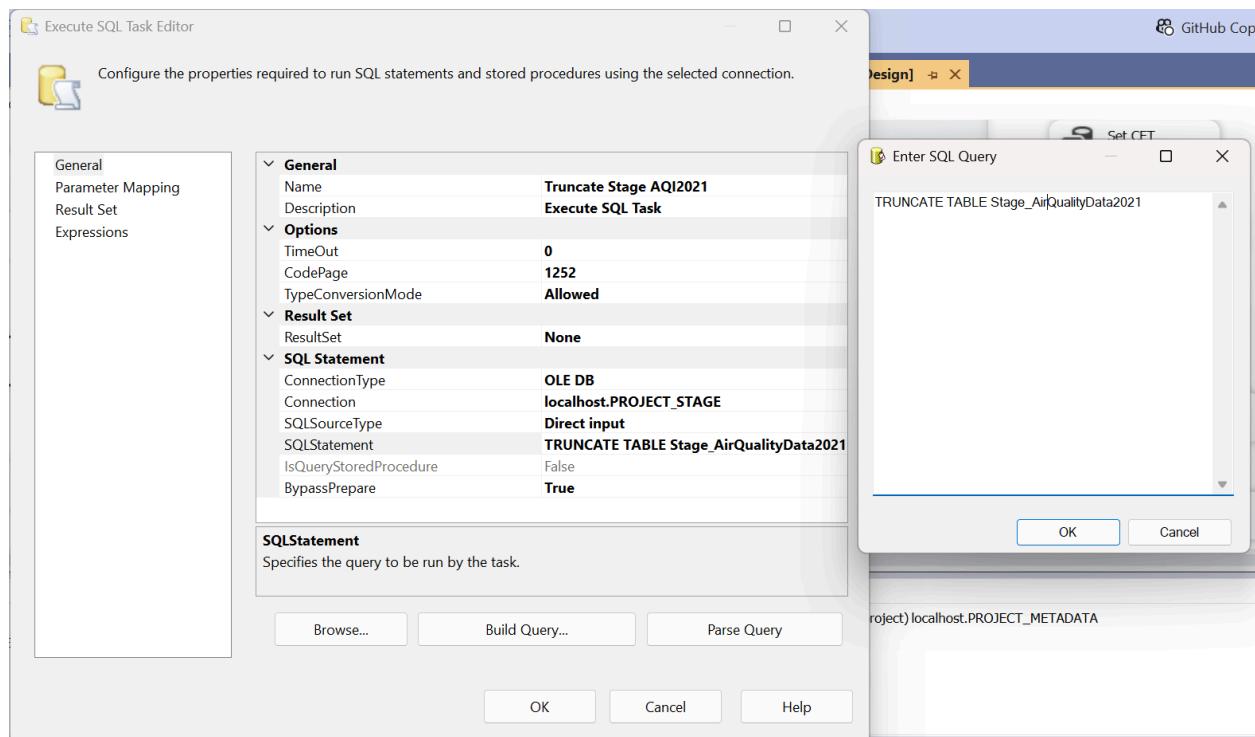
Đầu tiên, nhóm quyết định thực hiện load 3 bảng stage tương ứng với 3 file csv chất lượng không khí 2021, 2022, 2023 vì nếu thực hiện gộp thẳng 3 file csv vào bảng stage thì trong bảng stage chung sẽ tồn tại dữ liệu bị trùng nhau tồn tại trong cả 2-3 file csv và khi sử dụng để nạp vào NDS sẽ dễ gây ra lỗi.



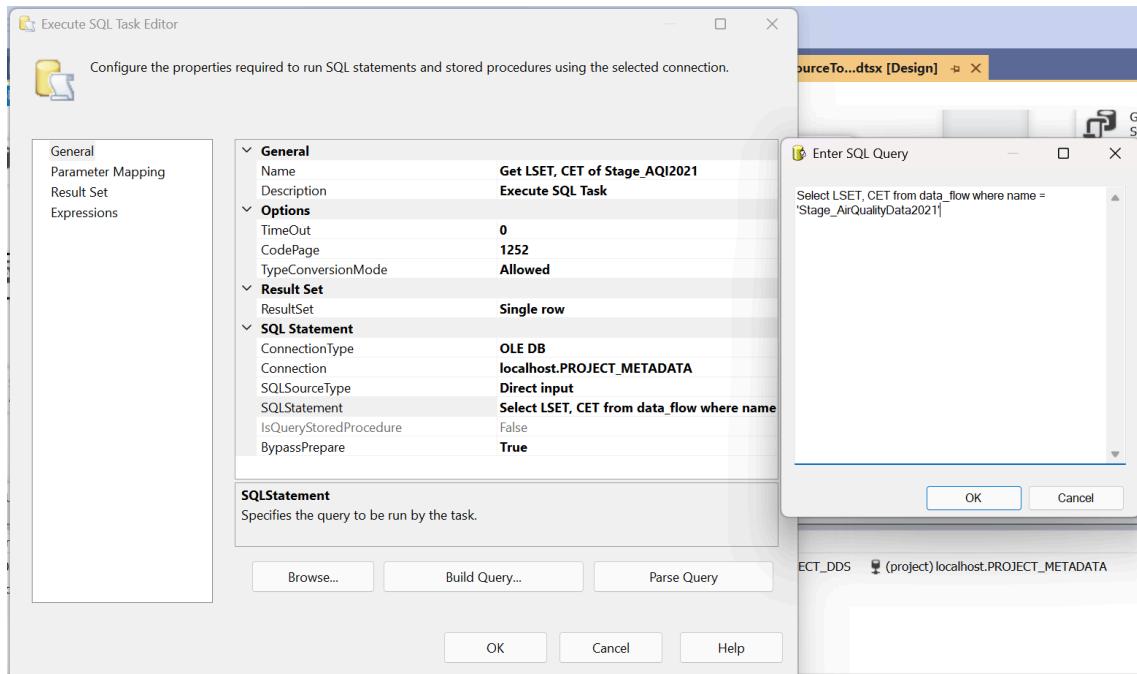
Đầu tiên ta set lại giá trị CET trong Metadata:



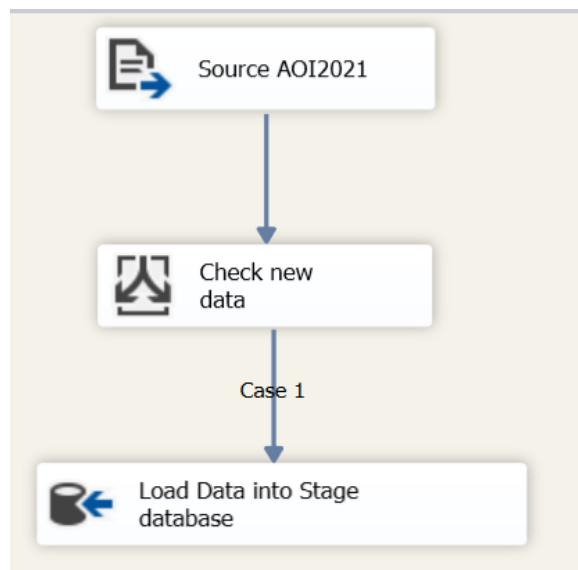
Sau đó ta truncate các giá trị tồn đong có sẵn trong Stage:



Tiếp theo ta lấy giá trị LSET và CET trong Metadata nhằm phục vụ khả năng incremental Update:



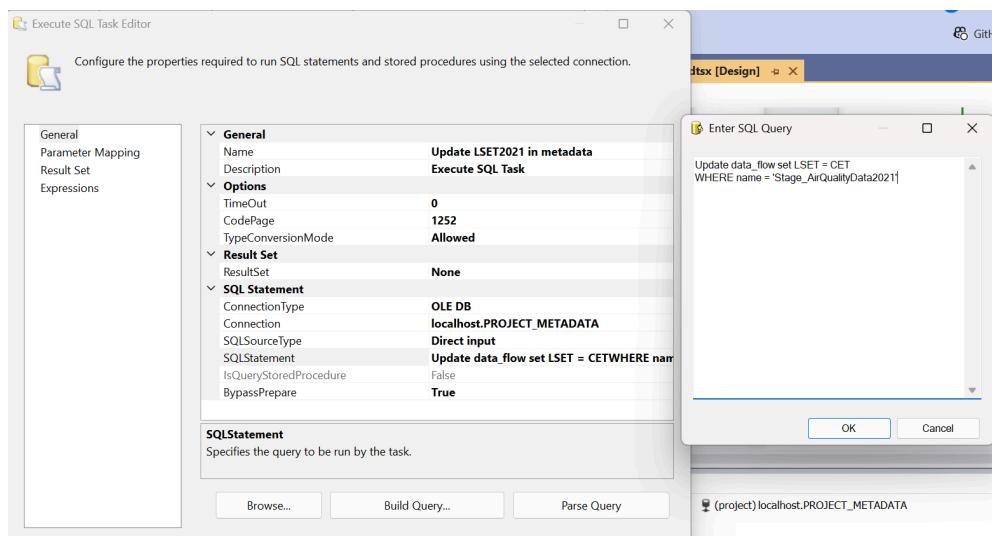
Trong Data Flow từ Source vào Stage của từng AQI thì nhóm có sử dụng Conditional Split:



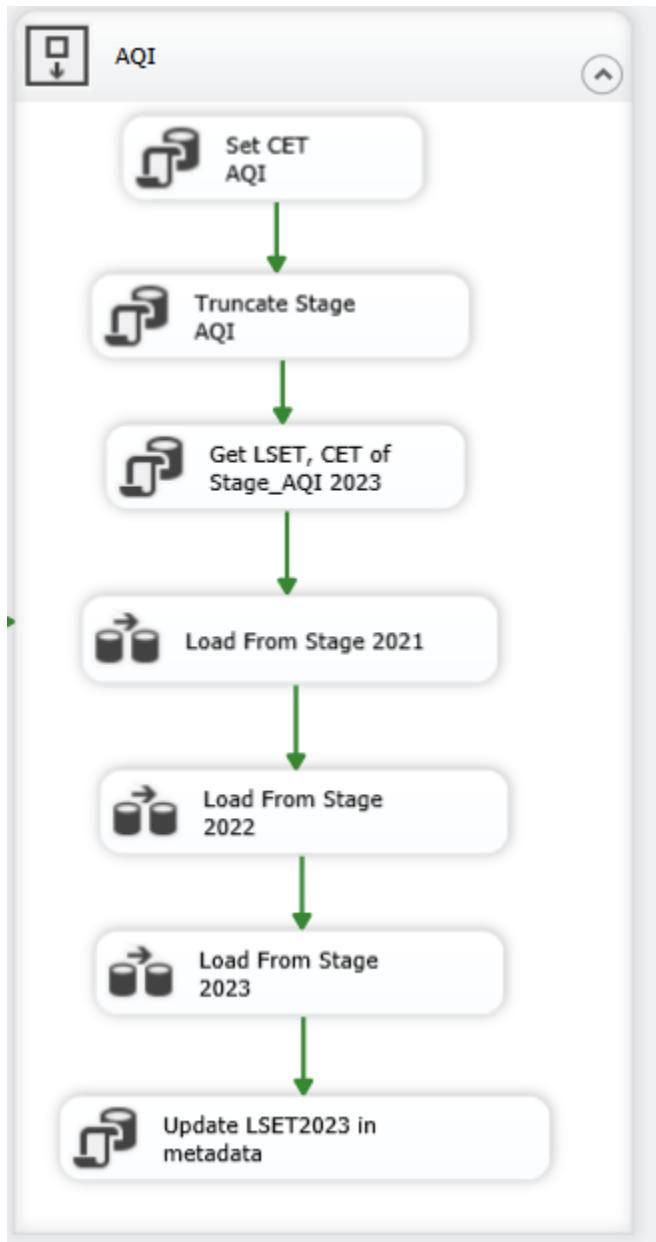
Order	Output Name	Condition
1	Case 1	(Created >= @[User::LSET2021] && Created < @[User::CET2021]) ([Last Updated] >= @[User::LSET2021] && [Last Updated] < @[User::CET2021])

Đây là điều kiện được sử dụng để có thể thực hiện incremental update: Các dữ liệu được lấy là các dữ liệu mới được thêm vào hoặc chỉnh sửa từ lần cuối extract cho đến thời điểm hiện tại.

Cuối cùng ta update giá trị LSET trong Metadata để phục vụ lần incremental update sau:



Sau khi nạp thành công dữ liệu vào 3 bảng tạm của riêng các năm, thực hiện việc incremental extract để gộp 3 bảng dữ liệu này vào một bảng chính, Stage_AirQualityData.

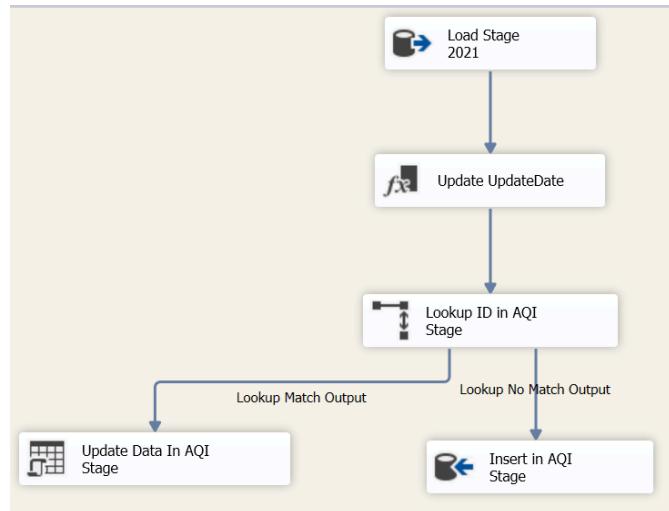


Source to Stage của các file chất lượng không khí

Tại đây, ta thực hiện các transform như:

- Derived Column:
 - Sửa lại Last_update thành thời gian đưa dữ liệu vào bảng.
 - Đ Đồng nhất cách gọi trong County_Name: Saint Clair, St. Clair -> St. Clair,....

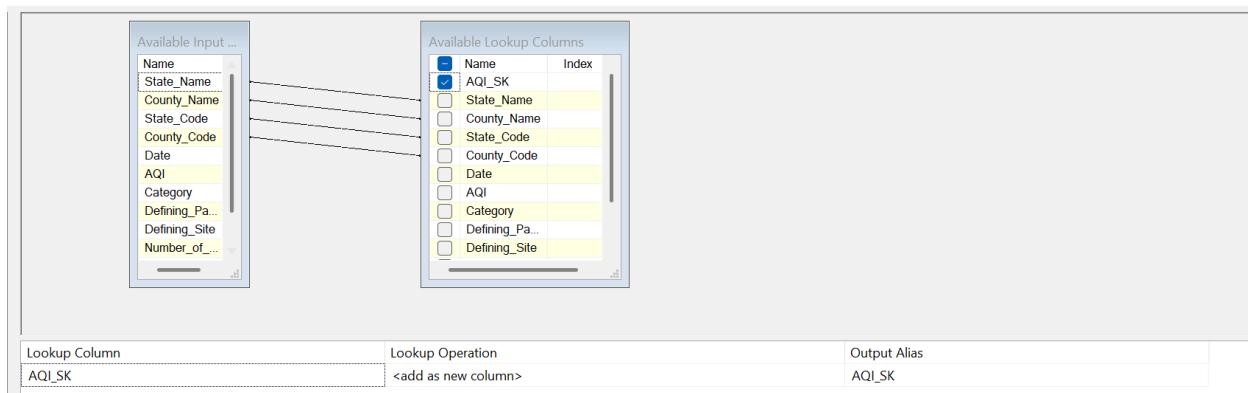
- Lookup: Dùng để kiểm tra xem đã tồn tại dữ liệu trong Stage_AirQualityData từ trước hay chưa. Từ đó sẽ xem xét việc đưa dữ liệu vào là insert dữ liệu mới hay update dữ liệu đã có.



Load from Stage AQI2021 to Stage_AirQualityData (Tương tự đối với 2022, 2023)

Derived Column Name	Derived Column	Expression	Data Type	Length	Precision	Scale	Code Page
Last_Updated	Replace 'Last_Updated'	@[User:=CET2021]	database timestamp [...]				
County_Name	Replace 'County_Name'	TRIM(REPLACE(REPLACE(REPLACE(County_Name,"Saint ","S t"), "St ","St "), " City",""))	string [DT_STR]	50			1252 (ANSI - Latin I)

Update UpdateDate

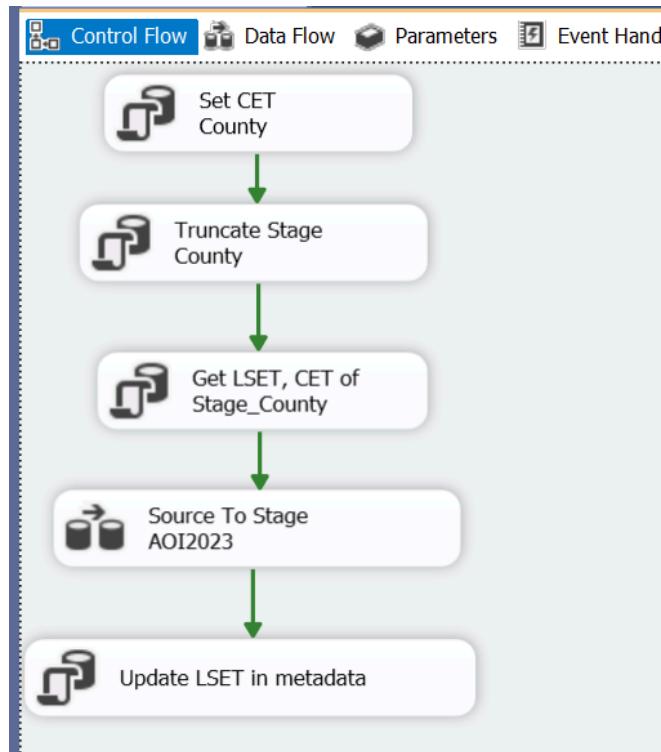


Lookup ID in AQI Stage

Cuối cùng sẽ là quá trình ETL file dữ liệu (2B)uscounties.csv, tuy nhiên trong quá trình làm nhóm sinh viên phát hiện file được cung cấp bị lỗi định dạng. Nhóm đã

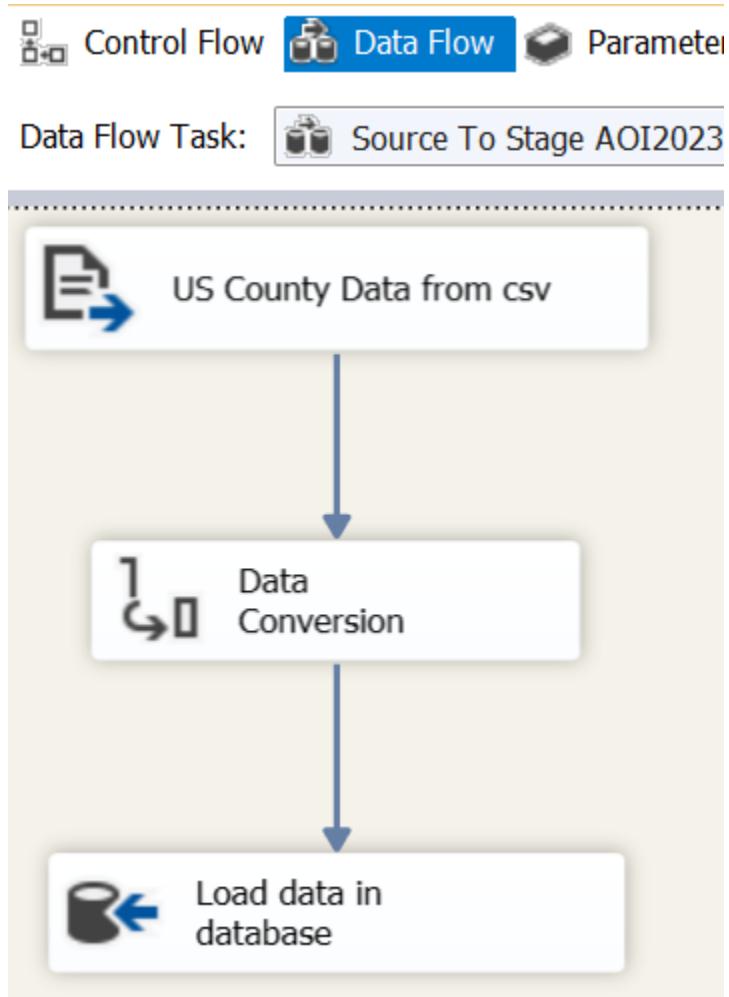
quyết định sửa lỗi bằng cách mở excel và chọn option convert file mà excel cung cấp (popup khi mở bằng file Excel) để fix và load được file trên SSIS.

Nhóm thực hiện incremental extract đối với (2B)uscounties.csv sau chỉnh sửa:



Source to Stage của file US county

Tuy nhiên trong quá trình chuyển dữ liệu, ta cần phải sử dụng Data Conversion:



Source To US_County Stage Data Flow

Input Column	Output Alias	Data Type	Length	Precision	Scale	Code Page
state_id	state_id	string [DT_STR]	2			1252 (ANSI)
county_fips	county_fips	string [DT_STR]	5			1252 (ANSI)

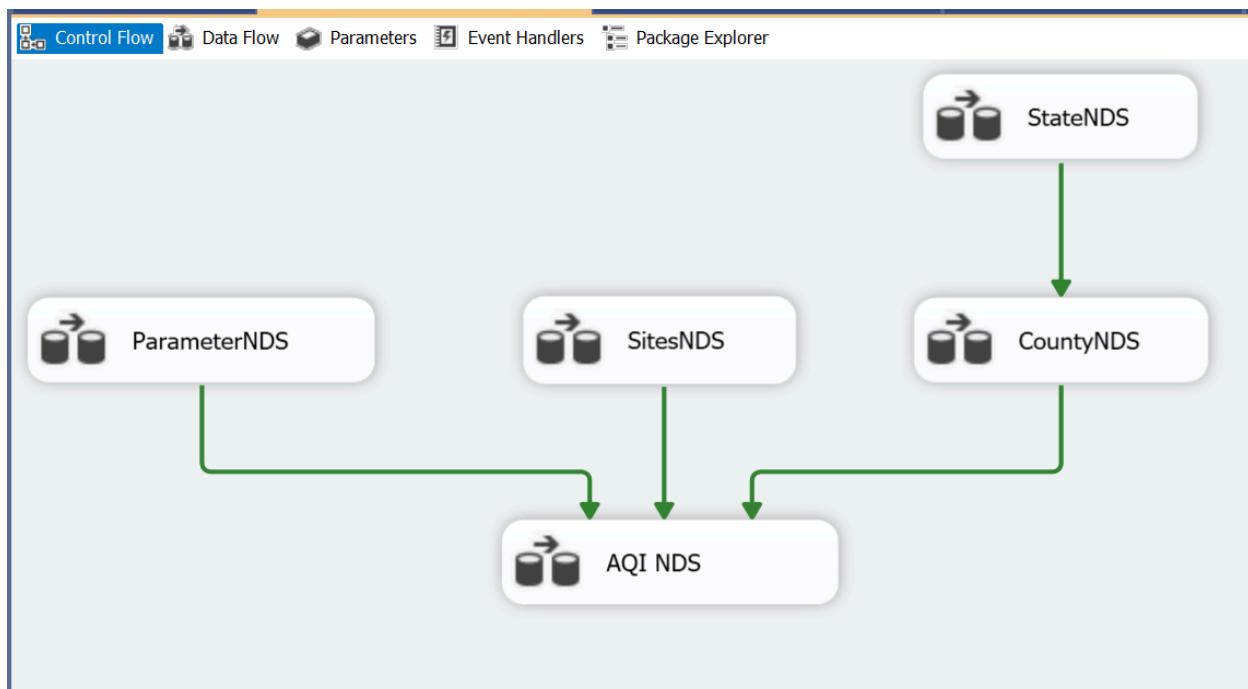
Data Conversion

Việc này là cần thiết, vì ta cần phải chuyển state_id thành kiểu varchar(2) và county_fips thành varchar(5) để có thể đưa vào Stage.

2. Stage to NDS

Vì yêu cầu về khóa ngoại, nhóm quyết định thực hiện ETL từ Stage sang NDS lần lượt nhau như: StateNDS xong rồi mới thực hiện trên CountyNDS, từ CountyNDS, SitesNDS và ParameterNDS xong rồi mới thực hiện trên AQINDS

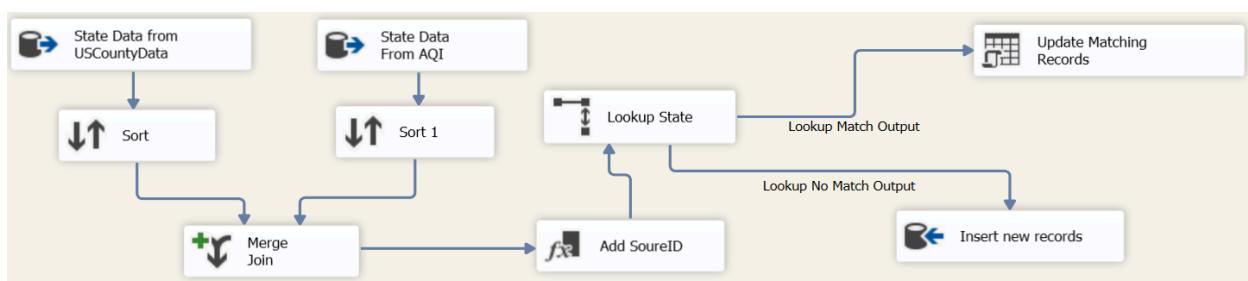
Quá trình xử lý dữ liệu từ Stage sang NDS như sau:



Control Flow quá trình ETL Stage -> NDS

Chi tiết từng DataFlow được thực hiện bao gồm:

- Trong StateNDS:



Data Flow Stage -> State NDS

- Merge Join: Vì dữ liệu trong State NDS có trong cả US_County Stage và AQI Stage nên cần phải join lại thông qua StateName.
- Sort: Trước khi merge join thì phải sort hai nguồn dữ liệu dựa trên điều kiện join là StateName.
- Devired Column: Thêm cột ID_Source vào dữ liệu.
- Lookup: Kiểm tra dữ liệu đã tồn tại trong cơ sở dữ liệu hay chưa. Từ đó xác định insert dữ liệu mới hay update dữ liệu đã tồn tại.

Available Input Columns

Name	Pass T...
state_id	<input checked="" type="checkbox"/>
state_name	<input type="checkbox"/>

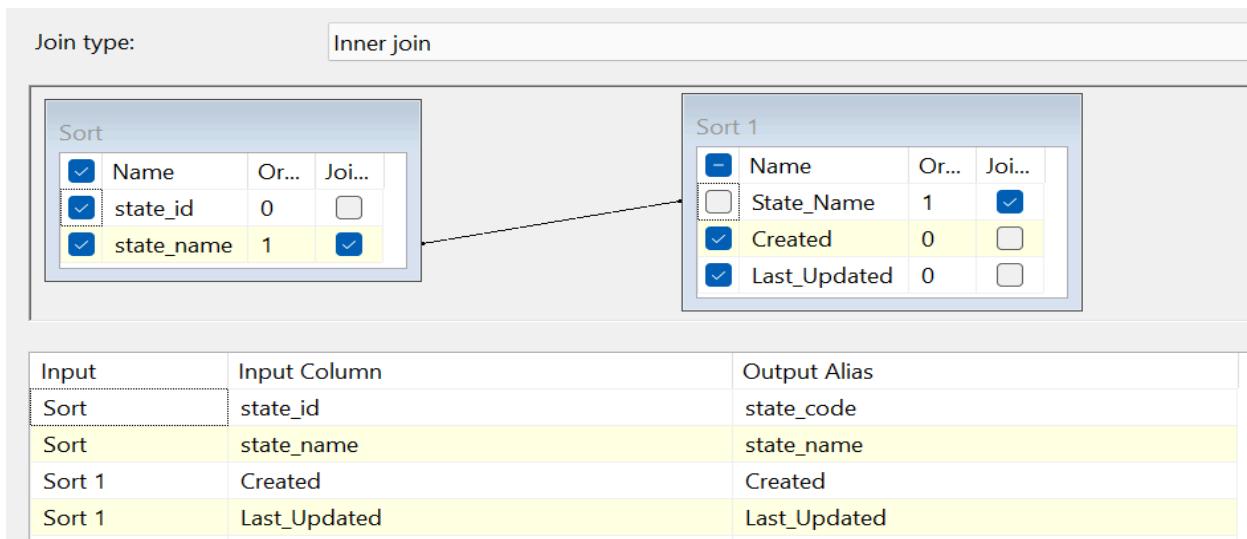
Input Column	Output Alias	Sort Type	Sort Order	Comparison Flags
state_name	state_name	ascending	1	

Available Input Columns

Name	Pass T...
State_Name	<input checked="" type="checkbox"/>
Created	<input type="checkbox"/>
Last_Updated	<input type="checkbox"/>

Input Column	Output Alias	Sort Type	Sort Order	Comparison Flags
State_Name	State_Name	ascending	1	

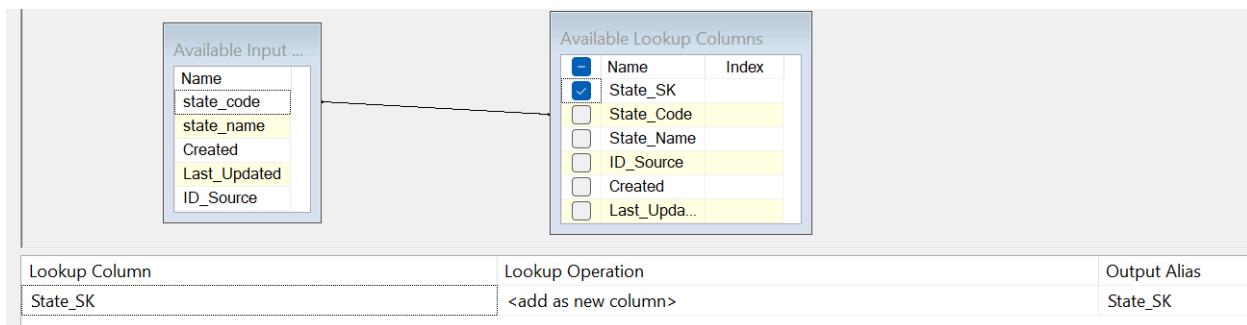
Sort, Sort1



Merge Join - Merge 2 bảng và chỉ lấy những thông tin cần thiết

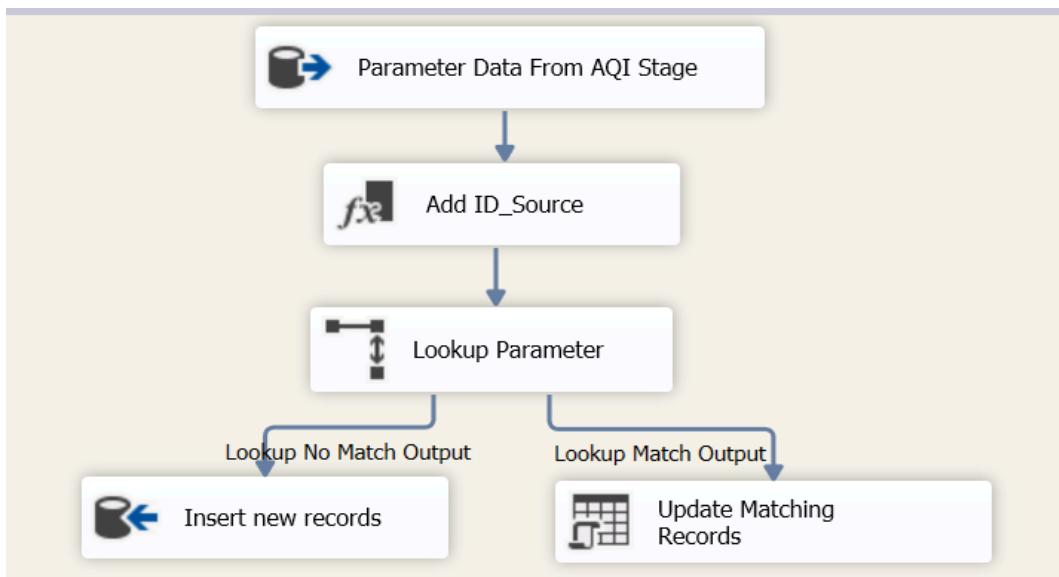
Derived Column Name	Derived Column	Expression	Data Type	Length	Precision	Scale	Code Page
ID_Source	<add as new column>	1	four-byte signed integer				

Add SourceID



Lookup State

- Trong ParameterNDS:

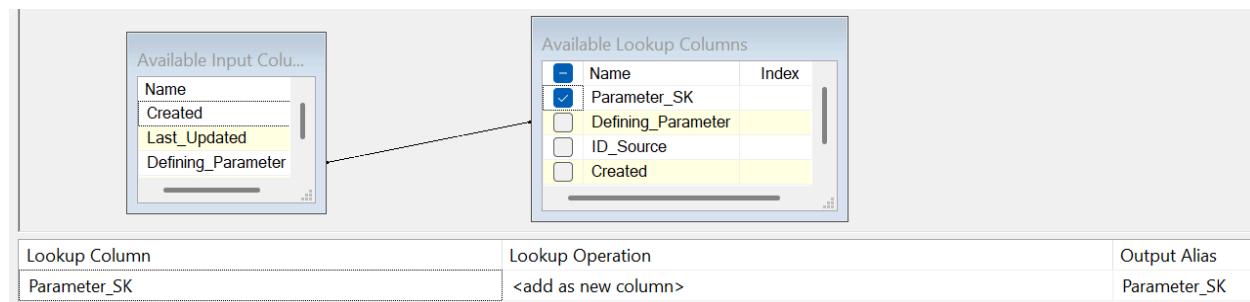


Data Flow Stage -> Parameter NDS

- Derived Column: Thêm ID source.
- Lookup: Kiểm tra dữ liệu đã tồn tại trong cơ sở dữ liệu hay chưa. Từ đó xác định insert dữ liệu mới hay update dữ liệu đã tồn tại.

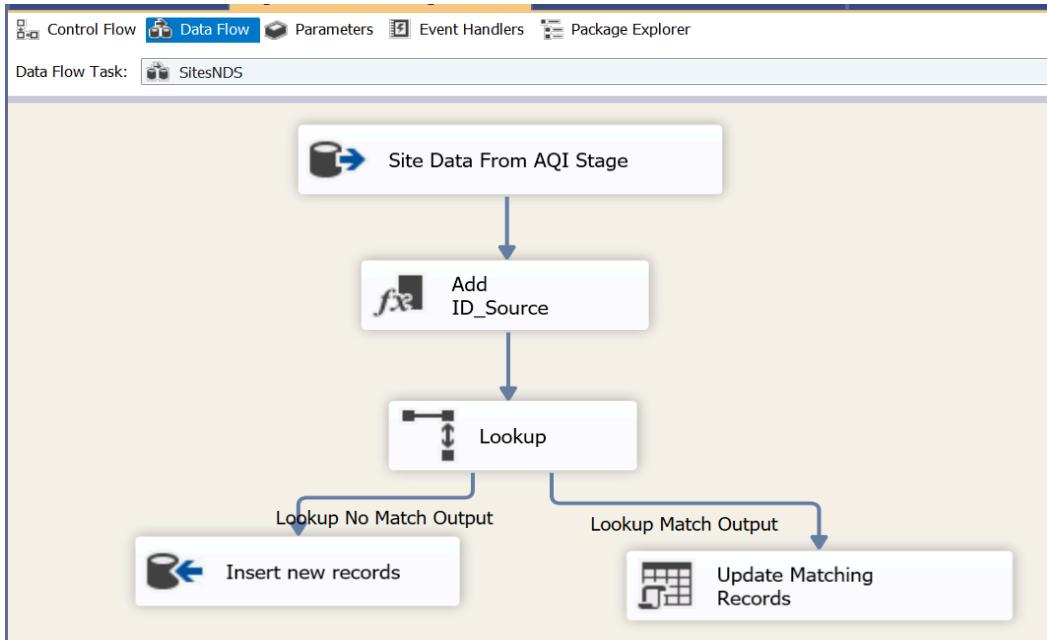
Derived Column Name	Derived Column	Expression	Data Type	Length	Precision	Scale	Code Page
ID_Source	<add as new column>	1	four-byte signed integer				

Add ID_Source



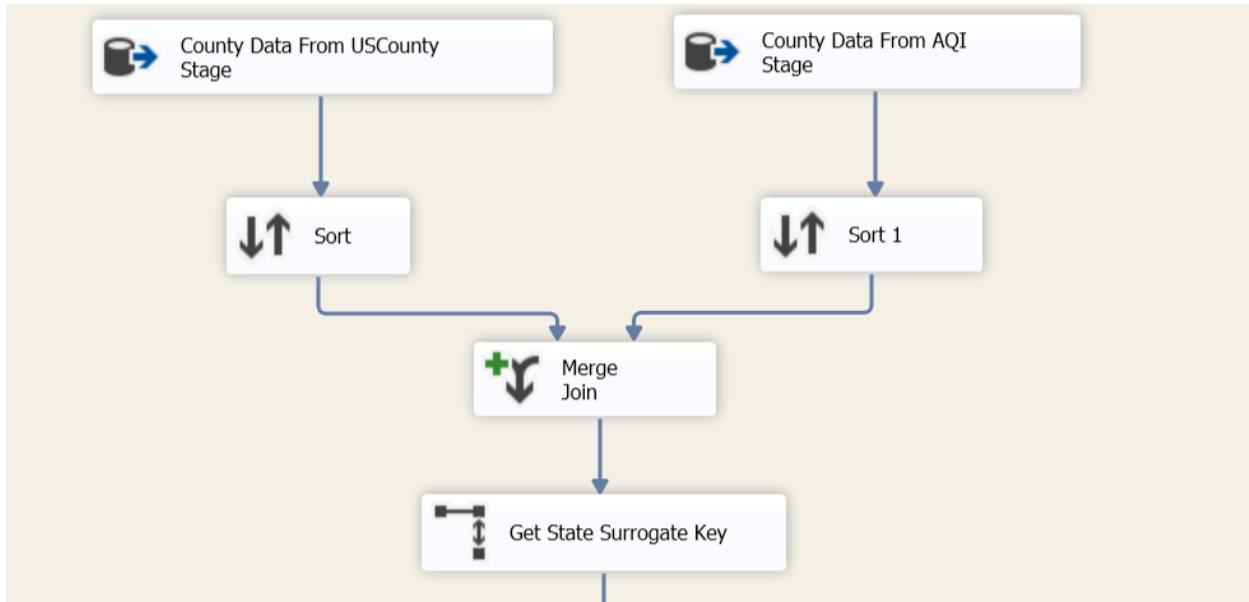
Lookup Parameter

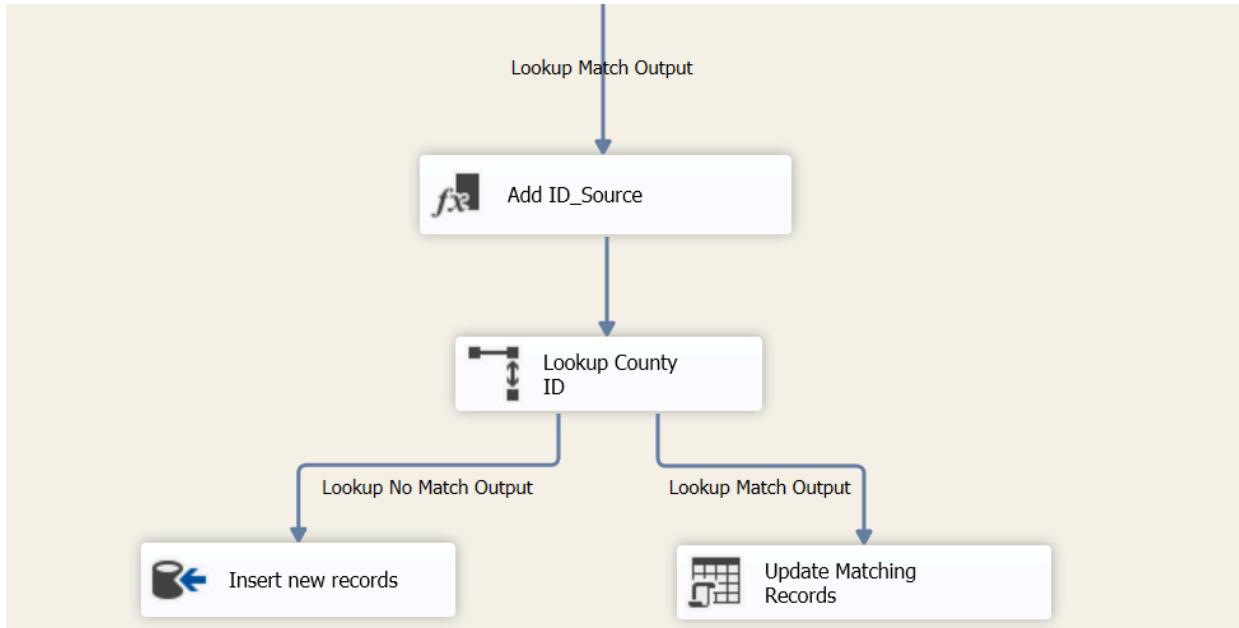
- Trong SitesNDS: tương tự với ParameterNDS



Data flow Stage -> Sites NDS

- Đối với County, các bước làm cũng tương tự với State, ngoài ra còn có truy xuất giá trị SK của State được lưu trong NDS ở bước trên để ghi vào khóa ngoại của County.





Data Flow Stage -> Country NDS

- Merge Join: Vì dữ liệu trong Country NDS có trong cả US_County Stage và AQI Stage nên cần phải join lại thông qua StateName.
- Sort: Trước khi merge join thì phải sort hai nguồn dữ liệu dựa trên điều kiện join là CountryName và StateName.
- Lookup: Kiểm tra xem có tồn tại State tương ứng để làm khoái ngoại cho County không, nếu không tồn tại thì ta bỏ luôn.
- Derived Column: Thêm ID source.
- Lookup: Kiểm tra dữ liệu đã tồn tại trong cơ sở dữ liệu hay chưa. Từ đó xác định insert dữ liệu mới hay update dữ liệu đã tồn tại.

Available Input Columns

Name	Pass T...
<input checked="" type="checkbox"/> county	<input type="button" value="–"/>
<input type="checkbox"/> county_full	<input checked="" type="checkbox"/>
<input type="checkbox"/> county_fips	<input checked="" type="checkbox"/>
<input type="checkbox"/> state_id	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/> state_name	<input type="button" value="–"/>
<input type="checkbox"/> lat	<input checked="" type="checkbox"/>
<input type="checkbox"/> lng	<input checked="" type="checkbox"/>

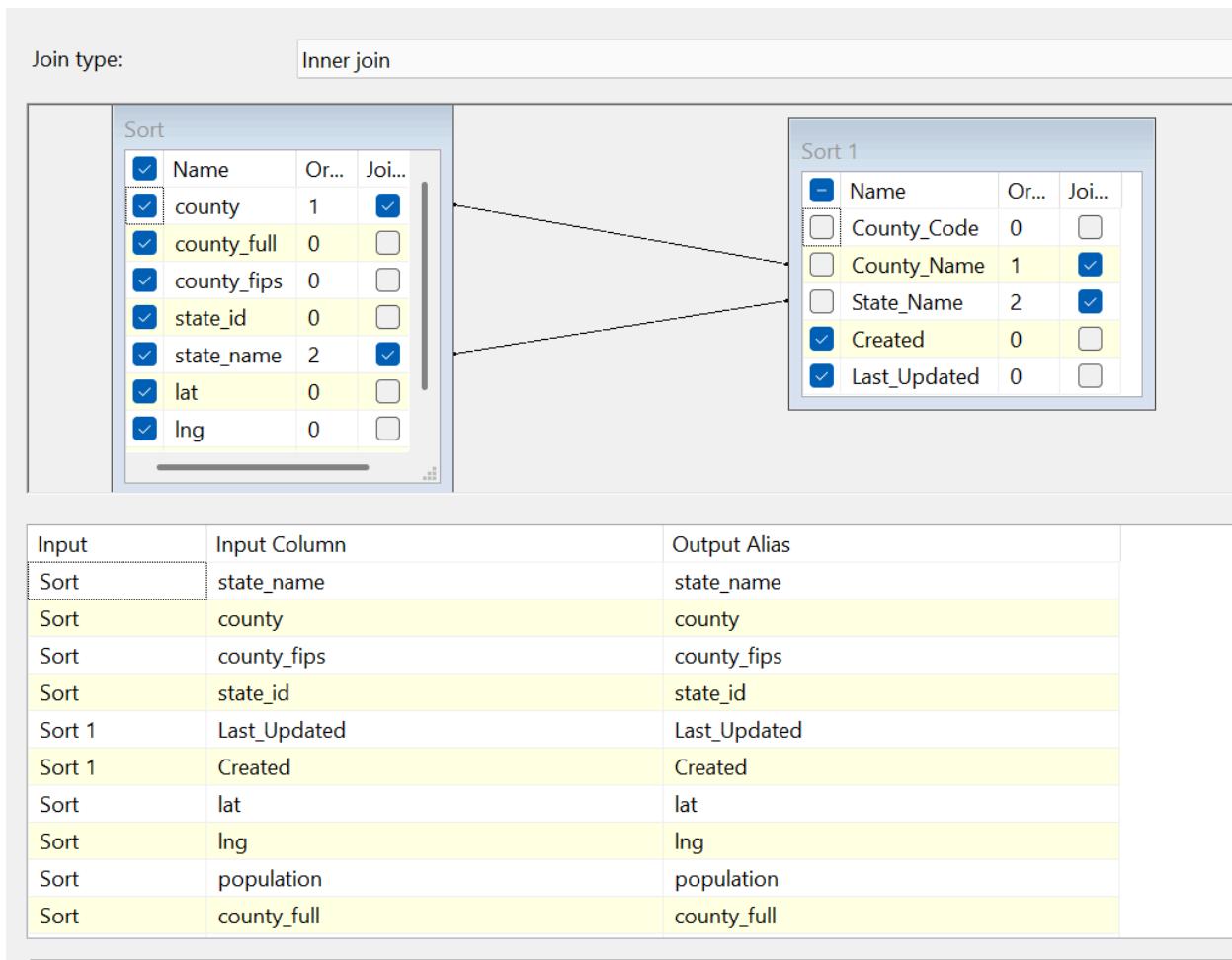
Input Column	Output Alias	Sort Type	Sort Order	Comparison Flags
county	county	ascending	1	
state_name	state_name	ascending	2	

Available Input Columns

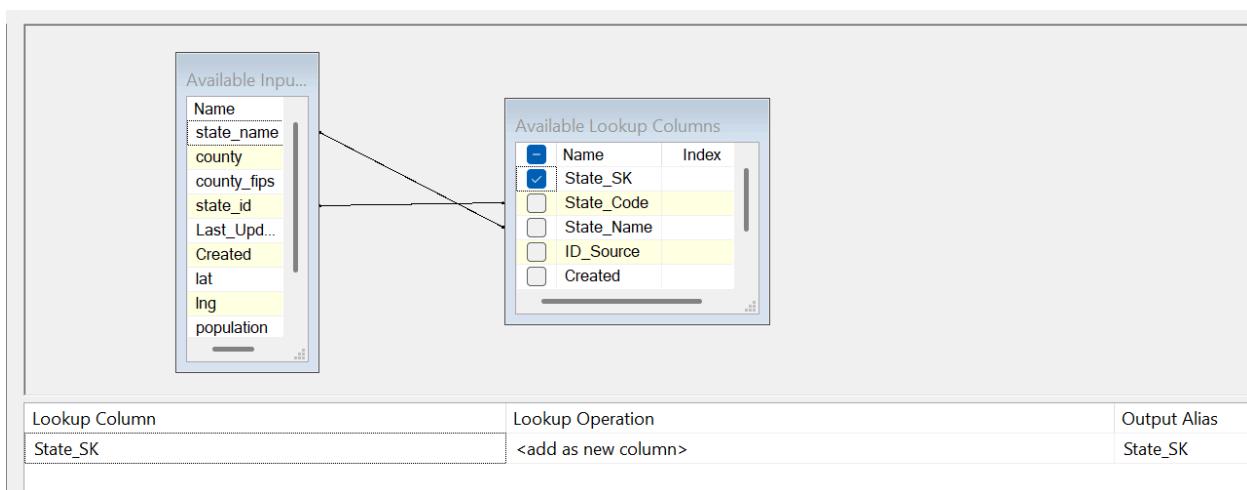
Name	Pass T...
<input type="checkbox"/> County_Code	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/> County_Name	<input type="button" value="–"/>
<input checked="" type="checkbox"/> State_Name	<input type="button" value="–"/>
<input type="checkbox"/> Created	<input checked="" type="checkbox"/>
<input type="checkbox"/> Last_Updated	<input checked="" type="checkbox"/>

Input Column	Output Alias	Sort Type	Sort Order	Comparison Flags
County_Name	County_Name	ascending	1	
State_Name	State_Name	ascending	2	

Sort, Sort 1



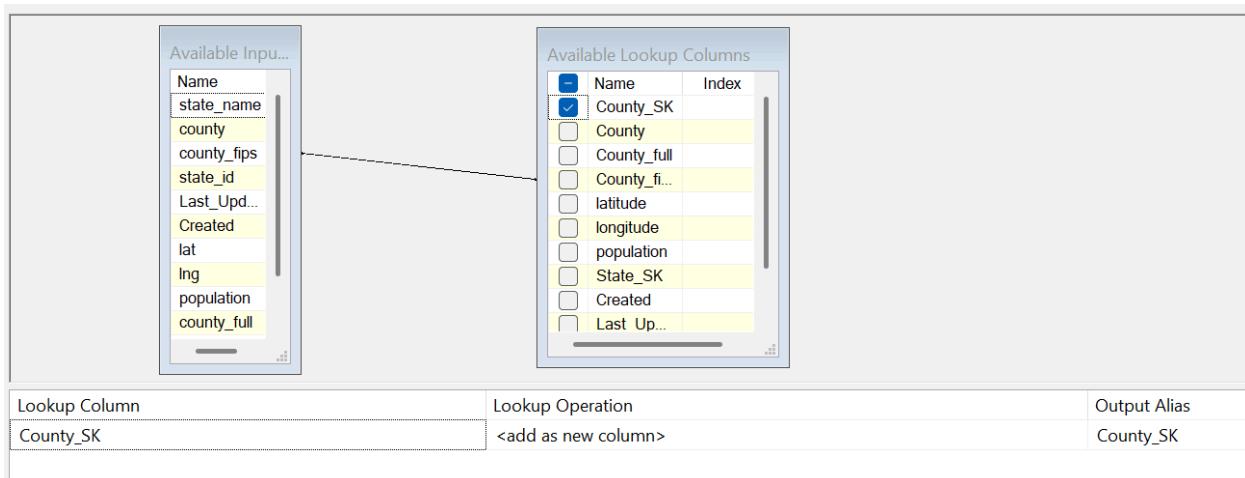
Merge join



Get State Surrogate Key - Lấy State_SK làm khóa ngoại

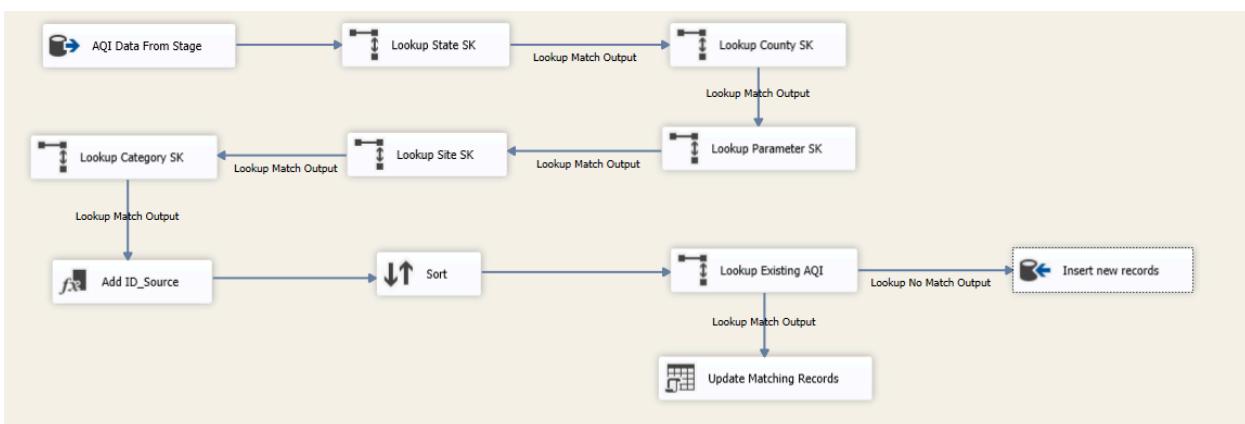
Derived Column Name	Derived Column	Expression	Data Type	Length
ID_Source	<add as new column>	1	four-byte signed integer	4

Add ID_Source



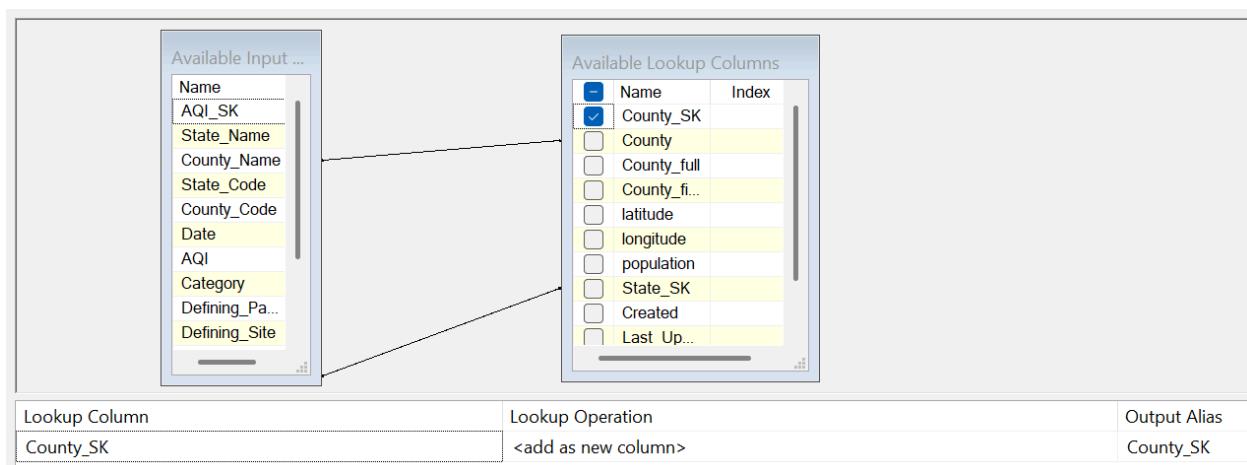
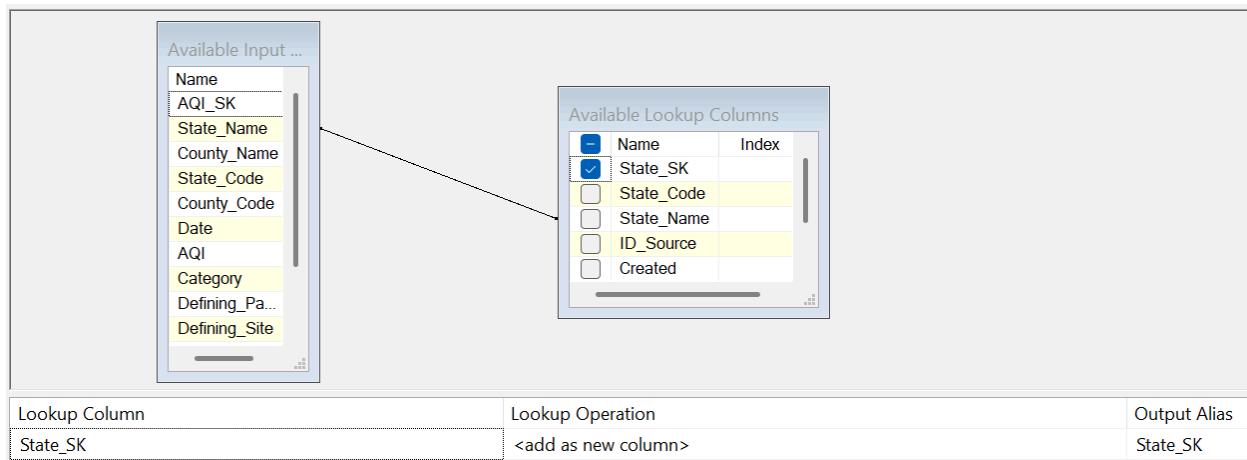
Lookup County ID

- Sau khi có đủ các thông tin về Sites, Parameter, County, State thì ta thực hiện load dữ liệu chất lượng không khí, sử dụng các lookup transformation để chắc chắn rằng những thông tin về site, state, county, parameter đã có trong NDS và lưu vào bảng AQI_NDS.



Data Flow Stage -> AQI

- Lookup: Tìm kiếm các khóa ngoại tương ứng về State, Conutry, Parameters, Site và Category.
- Derived Column: Thêm ID source
- Sort: Sort lại AQI dựa trên AQI_SK
- Lookup: Kiểm tra dữ liệu đã tồn tại trong cơ sở dữ liệu hay chưa. Từ đó xác định insert dữ liệu mới hay update dữ liệu đã tồn tại.



Available Input ...

Name
AQI_SK
State_Name
County_Name
State_Code
County_Code
Date
AQI
Category
Defining_Pa...
Defining_Site

Available Lookup Columns

Name	Index
<input checked="" type="checkbox"/> Parameter_SK	
<input type="checkbox"/> Defining_Parameter	
<input type="checkbox"/> ID_Source	
<input type="checkbox"/> Created	

Lookup Column	Lookup Operation	Output Alias
Parameter_SK	<add as new column>	Parameter_SK

Available Input ...

Name
AQI_SK
State_Name
County_Name
State_Code
County_Code
Date
AQI
Category
Defining_Pa...
Defining_Site

Available Lookup Columns

Name	Index
<input checked="" type="checkbox"/> Site_SK	
<input type="checkbox"/> Defining_Site	
<input type="checkbox"/> ID_Source	
<input type="checkbox"/> Created	

Lookup Column	Lookup Operation	Output Alias
Site_SK	<add as new column>	Site_SK

Available Input ...

Name
AQI_SK
State_Name
County_Name
State_Code
County_Code
Date
AQI
Category
Defining_Pa...
Defining_Site

Available Lookup Columns

Name	Index
<input checked="" type="checkbox"/> Category_SK	
<input type="checkbox"/> Category	
<input type="checkbox"/> Value_From	
<input type="checkbox"/> Value_To	
<input type="checkbox"/> Color	
<input type="checkbox"/> Description	
<input type="checkbox"/> ID_Source	
<input type="checkbox"/> Created	

Lookup Column	Lookup Operation	Output Alias
Category_SK	<add as new column>	Category_SK

Derived Column Name	Derived Column	Expression	Data Type	Length
ID_Source	<add as new column>	1	four-byte signed integer	4

Available Input Columns

	Name	Pass T...
<input checked="" type="checkbox"/>	AQI_SK	<input type="button" value="-"/>
<input type="checkbox"/>	State_Name	<input checked="" type="checkbox"/>
<input type="checkbox"/>	County_Name	<input checked="" type="checkbox"/>
<input type="checkbox"/>	State_Code	<input checked="" type="checkbox"/>
<input type="checkbox"/>	County_Code	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>
<input type="checkbox"/>	AQI	<input checked="" type="checkbox"/>

Input Column	Output Alias	Sort Type	Sort Order	Com
AQI_SK	AQI_SK	ascending	1	

Available Input ...

Name
AQI_SK
State_Name
County_Name
State_Code
County_Code
Date
AQI
Category
Defining_Pa...
Defining_Site
Number_of_...
Created
Last_Updated
State_SK
County_SK
Parameter_...
Site_SK
Category_SK
ID_Source

Available Lookup Columns

Name	Index
<input checked="" type="checkbox"/> AQI_SK	
<input type="checkbox"/> County_SK	
<input type="checkbox"/> Date	
<input type="checkbox"/> AQI	
<input type="checkbox"/> Category...	
<input type="checkbox"/> Parameter...	
<input type="checkbox"/> Site_SK	
<input type="checkbox"/> Number_...	
<input type="checkbox"/> Created	
<input type="checkbox"/> Last_Up...	
<input type="checkbox"/> ID_Source	

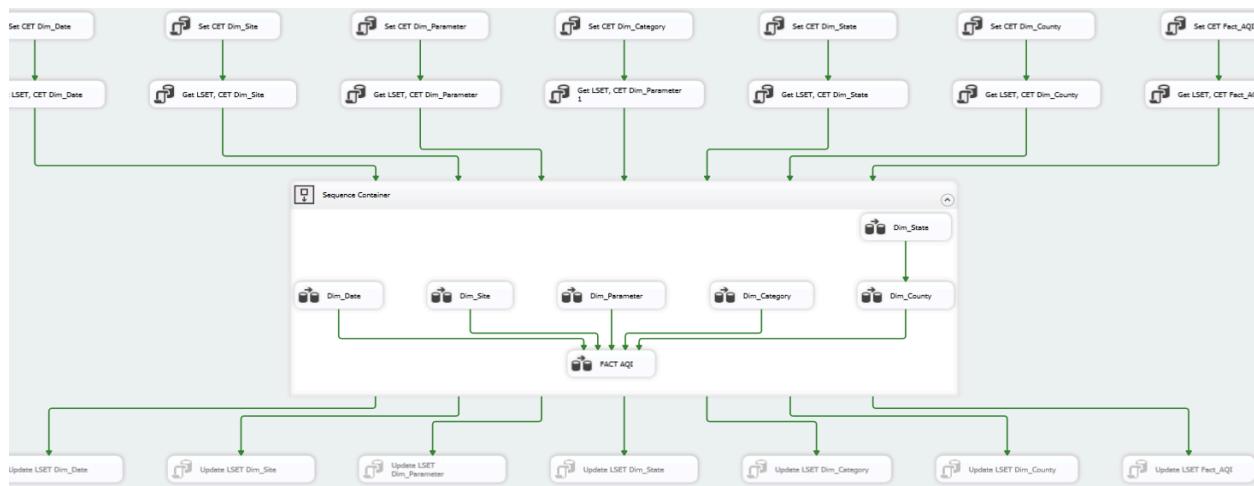
Lookup Column	Lookup Operation	Output Alias
AQI_SK	<add as new column>	AQI_SK

3. NDS to DDS

Note từ nhóm: Ở quá trình này nhóm sử dụng Slowly Changing Dimension Wizard và công cụ này đôi lúc sẽ bị lỗi. Nếu chạy ETL bị lỗi ở bước này, chỉ cần chạy lại từng

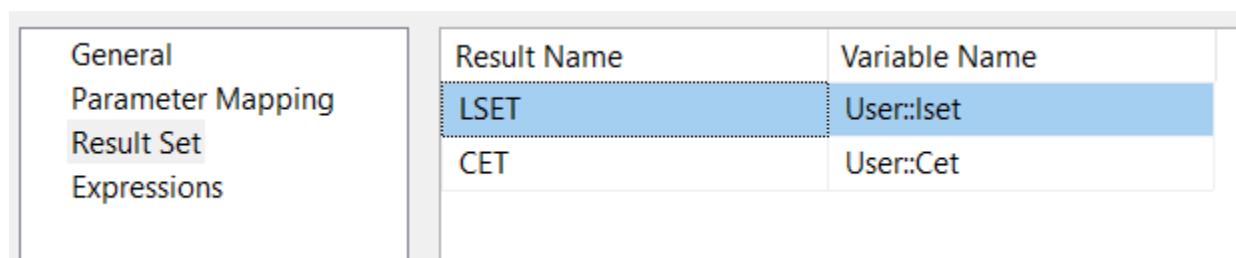
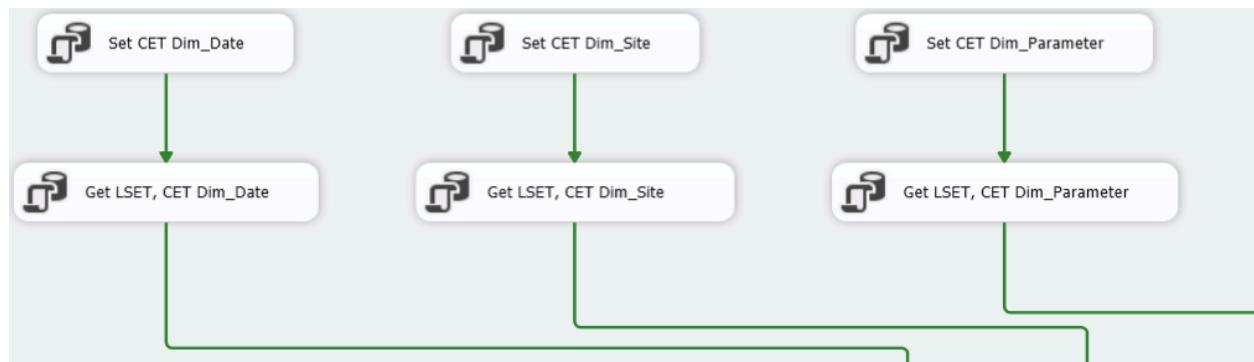
Slowly Changing Dimension Wizard trong package (nhấp đúp và next theo cài đặt nhóm đã thực hiện).

Giai đoạn nạp dữ liệu từ NDS vào DDS được thực hiện theo workflow như sau:

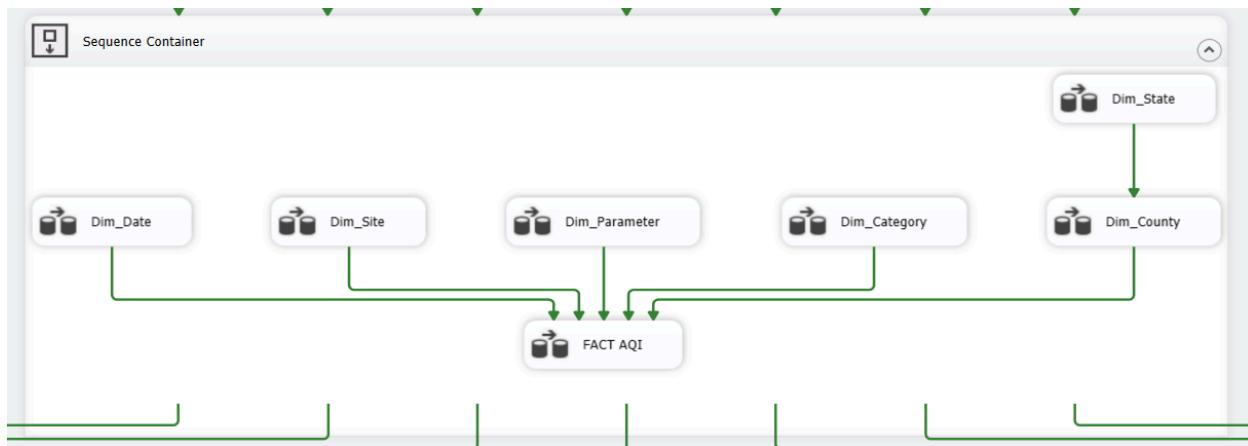


Trong bước đầu tiên, thực hiện việc cập nhật CET với giá trị là ngày hệ thống trong METADATA cho tất cả những bản ghi tương ứng với các bảng chiều.

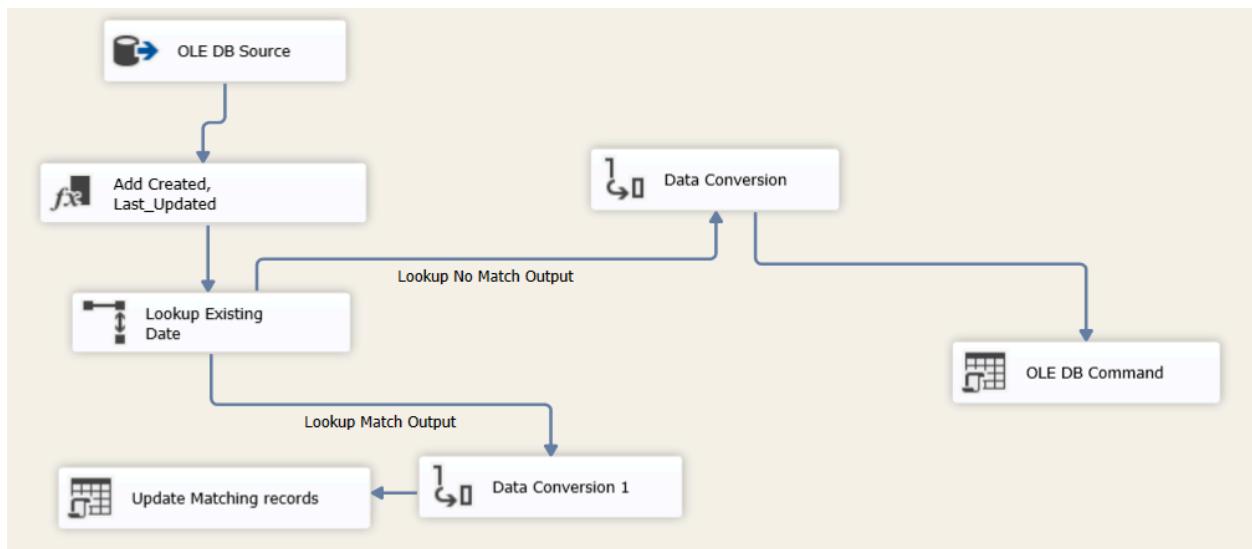
Sau đó, lấy giá trị CET, LSET của những bảng chiều trong METADATA và lưu vào một cặp biến dành riêng cho từng bảng.



Sau khi thực hiện việc truy xuất giá trị LSET và CET của tất cả bảng chiều trong METADATA, bắt đầu thực hiện việc load dữ liệu vào các bảng chiều và bảng fact. Quá trình thực hiện như sau:



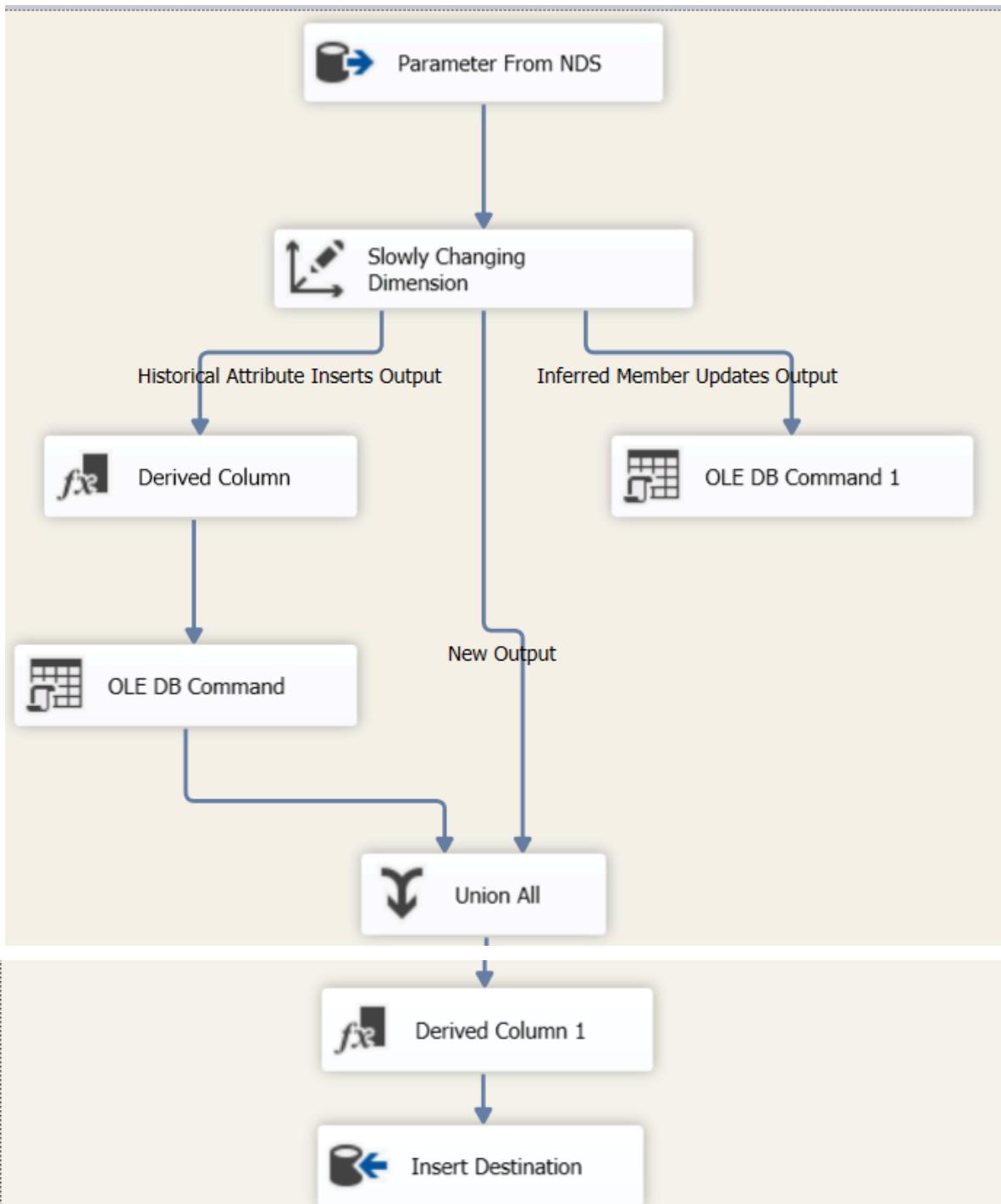
- Dim_Date: đọc dữ liệu ngày và mã nguồn từ bảng AQI_NDS và thực hiện incremental extract.



Trong đó bao gồm:

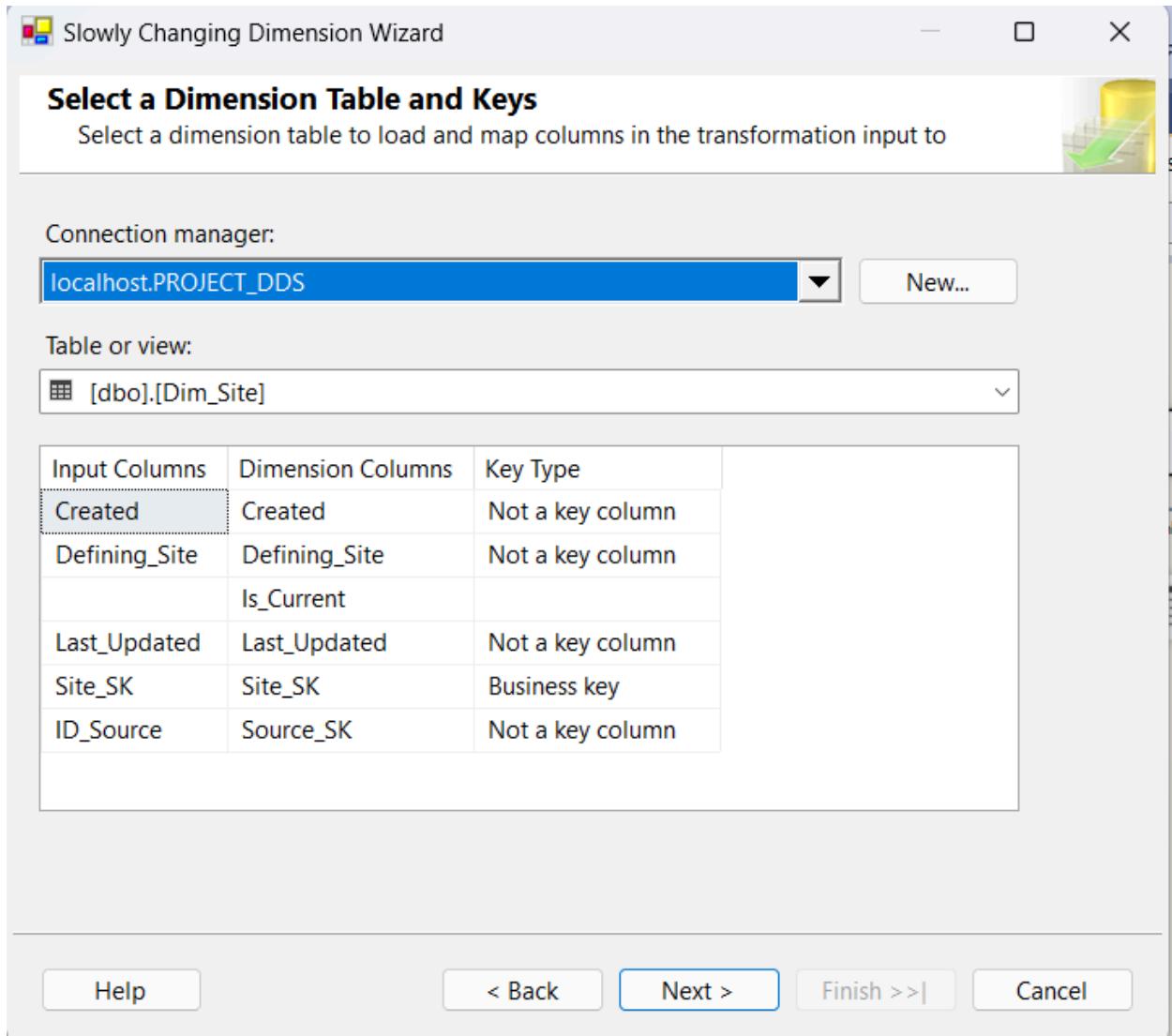
- Derived Column: thêm ngày tạo và ngày cập nhật lần cuối.
- Lookup: Kiểm tra xem Date đó đã tồn tại hay chưa. Nếu chưa tồn tại thì thêm vào, còn đã tồn tại thì update lại dữ liệu mới.

- Data Conversion: Ta cần chuyển dữ liệu Date, CreatedDate và UpdateDate về đúng định dạng để có thể sử dụng procedure được. Vì các dữ liệu khi được đưa vào SSIS thì sẽ tự động chuyển sang WSTRING.
- Đối với các bảng Dim_[Site, Parameter, Category, State, County] có cách xử lý giống nhau, là dùng transformation Slowly Changing Dimension để thêm vào bảng đích tương ứng. Tại transformation này, cấu hình như sau:

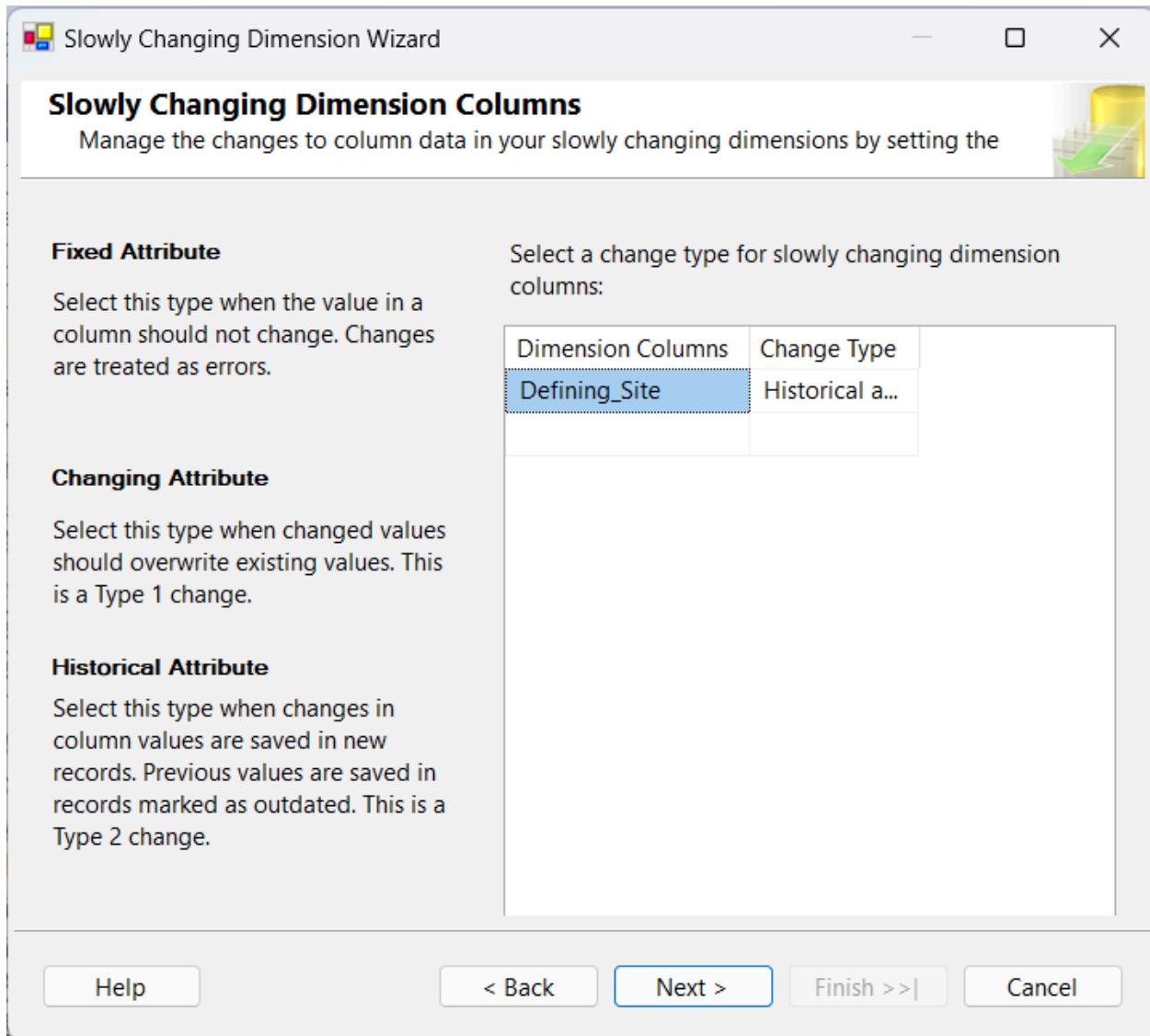


Trong đó Slowly Changing Dimension (SCD), ta cần xác định các cài đặt cơ bản sau:

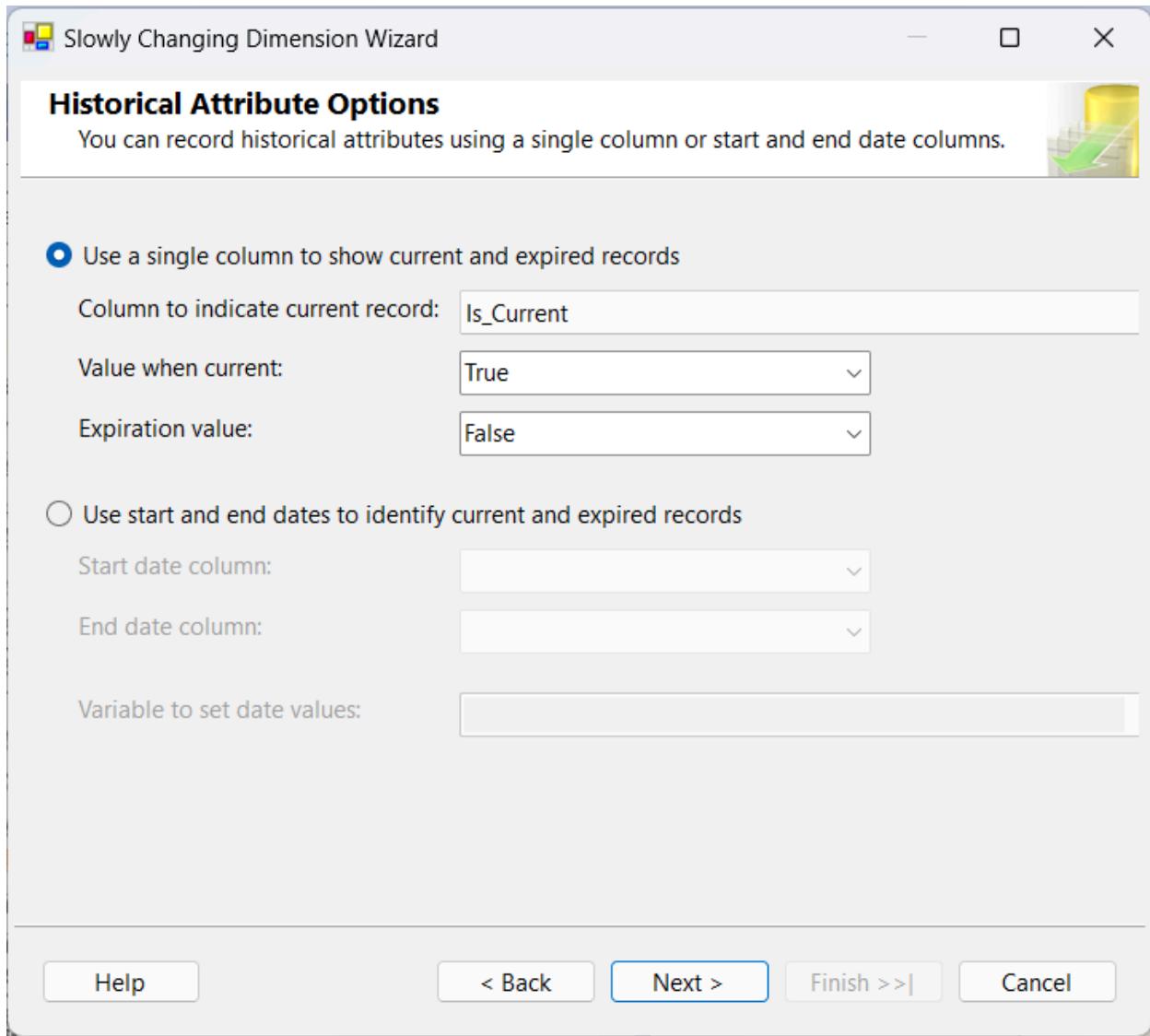
- Xác định khóa của bảng



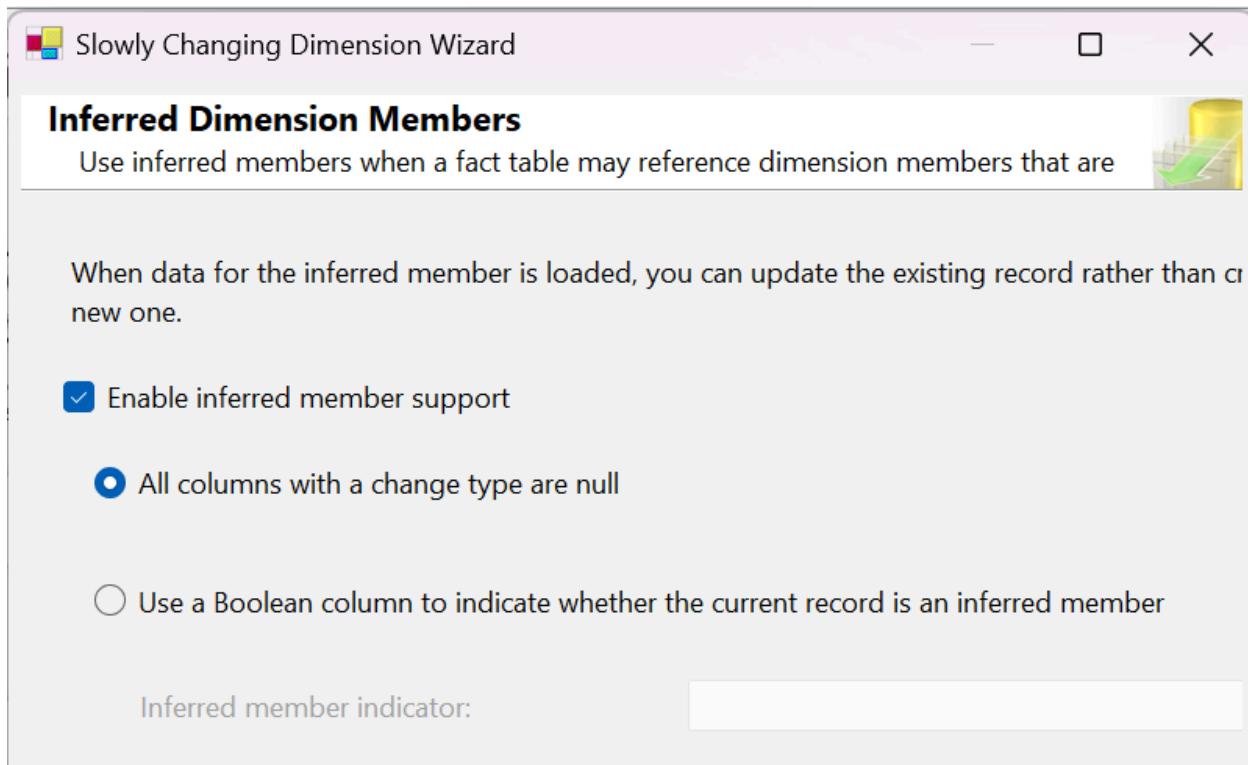
- Chọn thuộc tính làm thuộc tính lịch sử, dùng để lưu giá trị của những thuộc tính thay đổi theo thời gian.



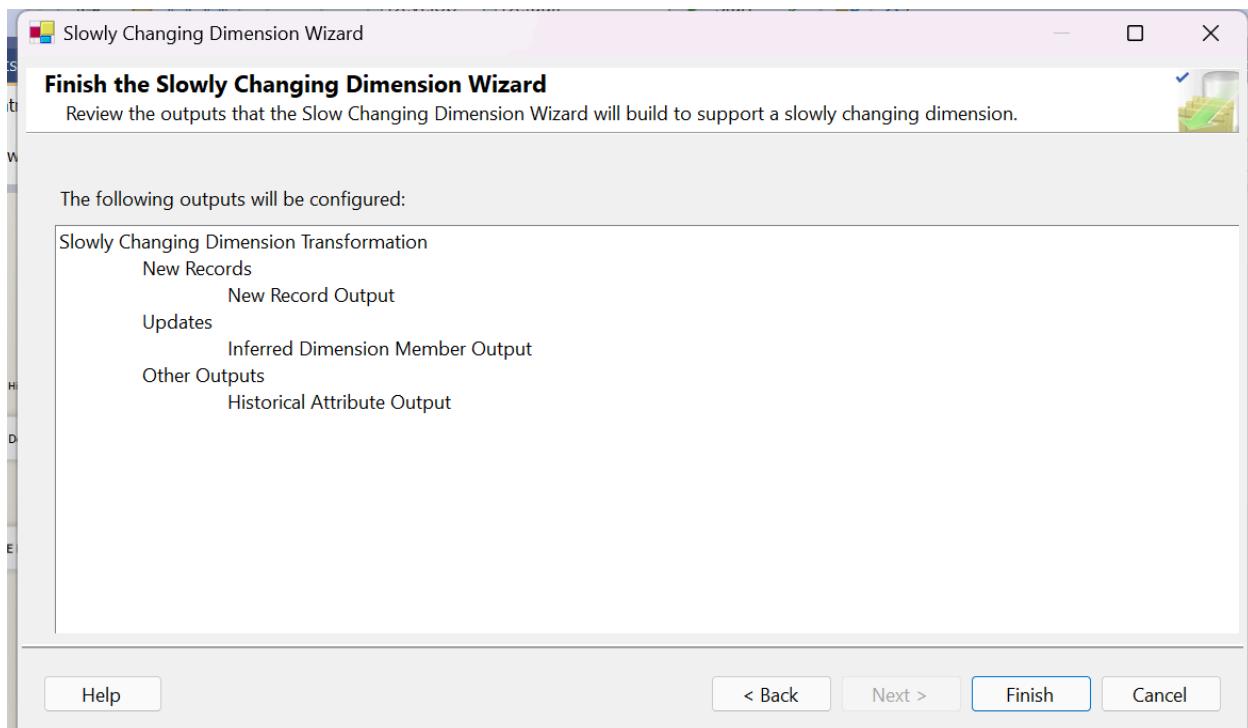
- Dùng thuộc tính cờ đã tạo trong NDS để đánh dấu record có hết hạn hay chưa.



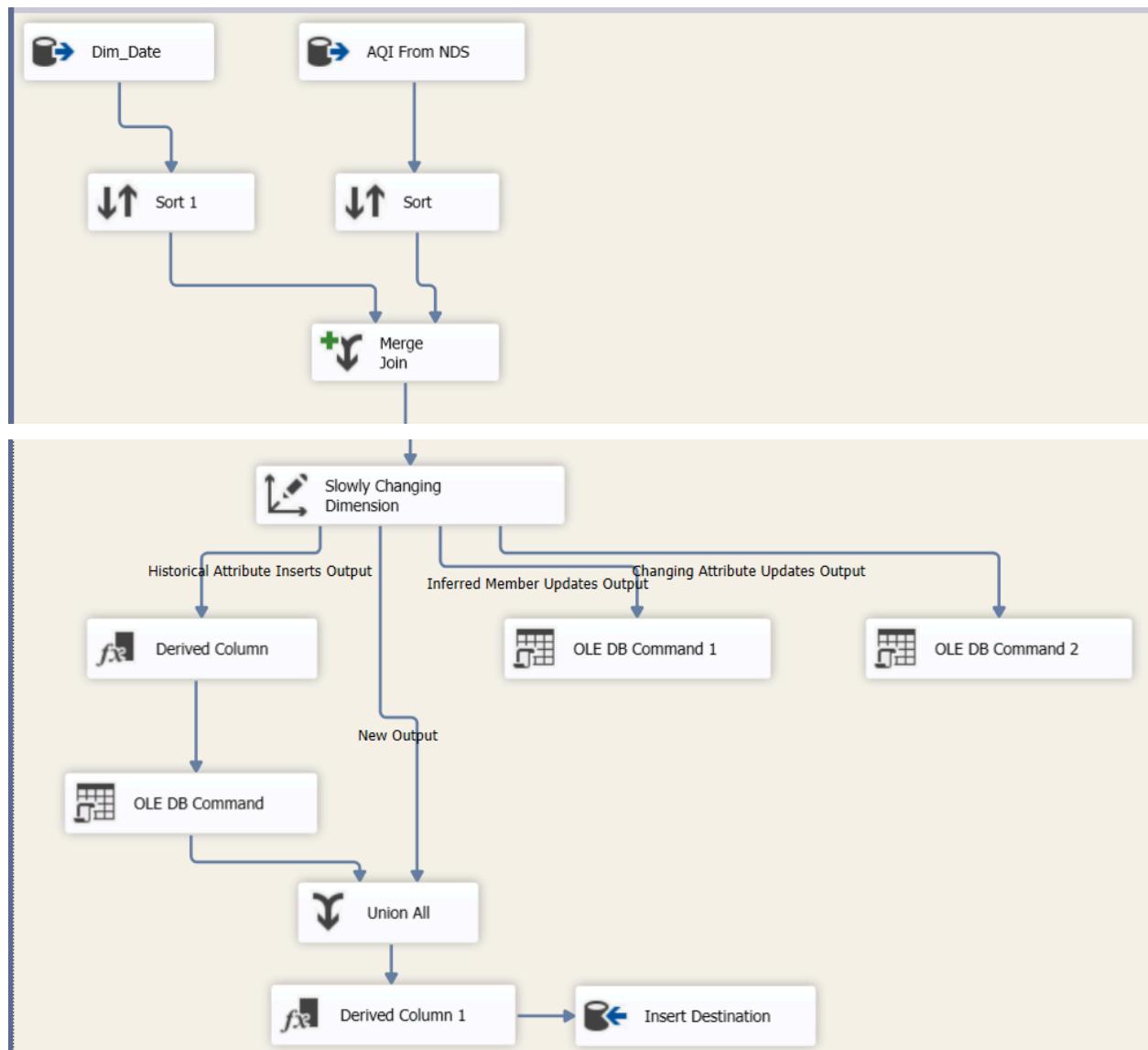
- Chọn Inferred Column sang Null để tránh xuất hiện tình trạng quá trình tải dữ liệu bị gián đoạn khi hệ thống dữ liệu không đồng bộ. Lúc đó các giá trị Null sẽ được đưa vào.



- Cuối cùng kiểm tra lại rồi chọn finish.



- Các transform khác là các bước SCD tự động tạo như:
 - Derived column: Thêm cột Is_Current với giá trị 0 hoặc 1.
 - Union All: Bỏ giá trị cũ, lấy giá trị mới.
- Trong Fact_AQI thì ta có thêm một bước đó là: Merge Join từ AQI NDS và Dim_Date (Phải có Sort đi kèm):



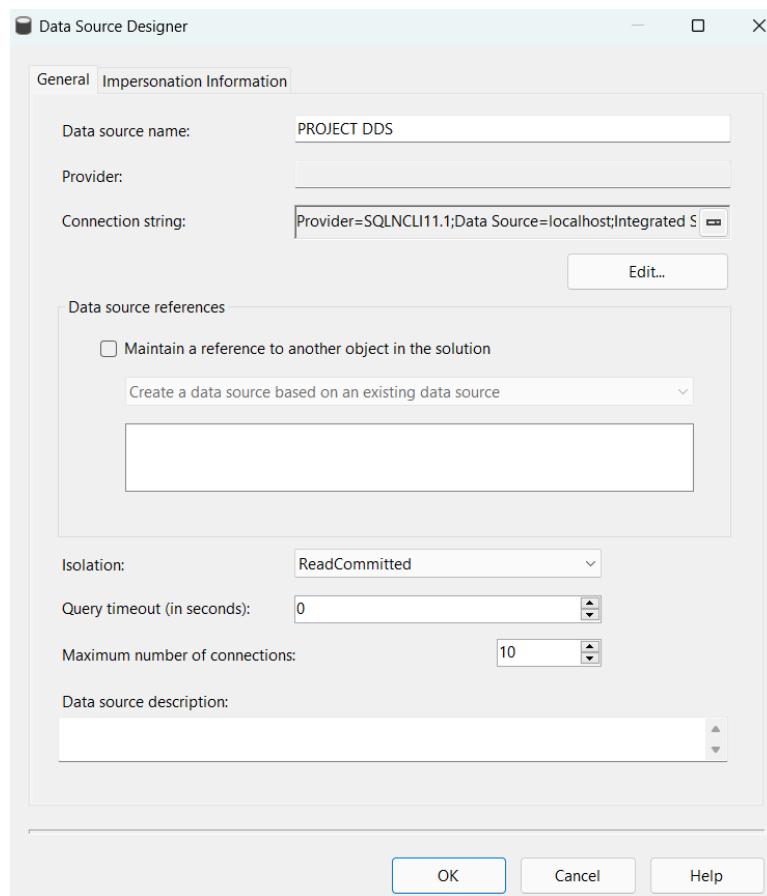
- Cuối cùng, ta update giá trị LSET trong Metadata:



IV. OLAP

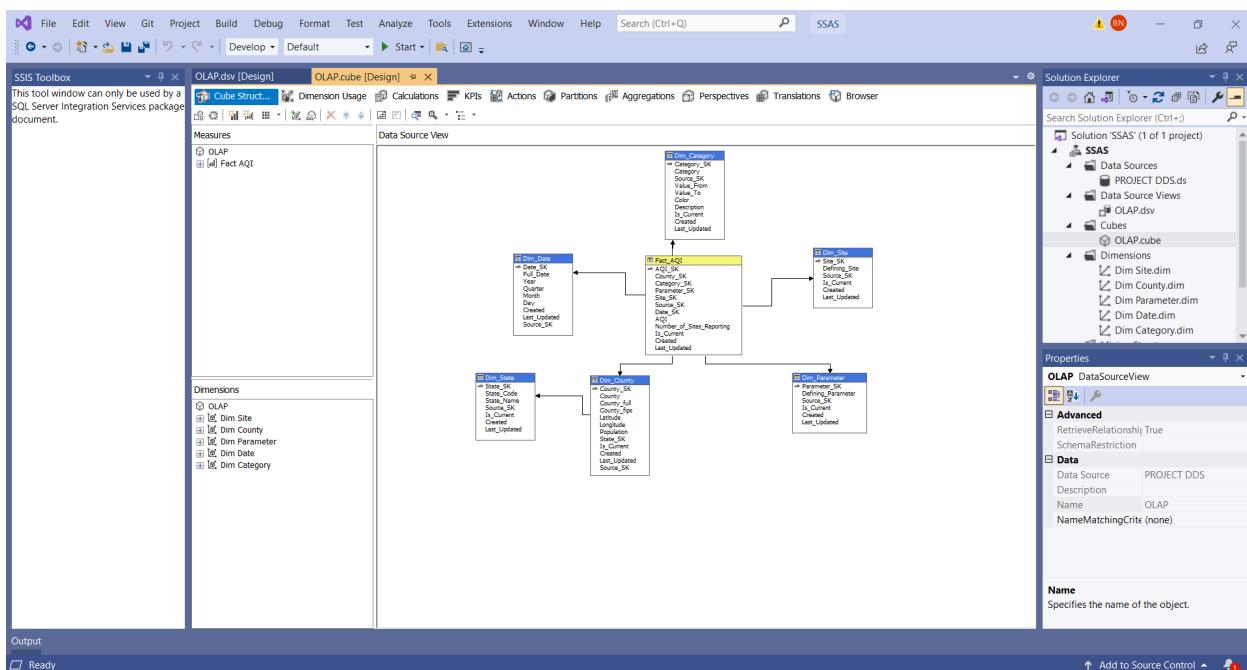
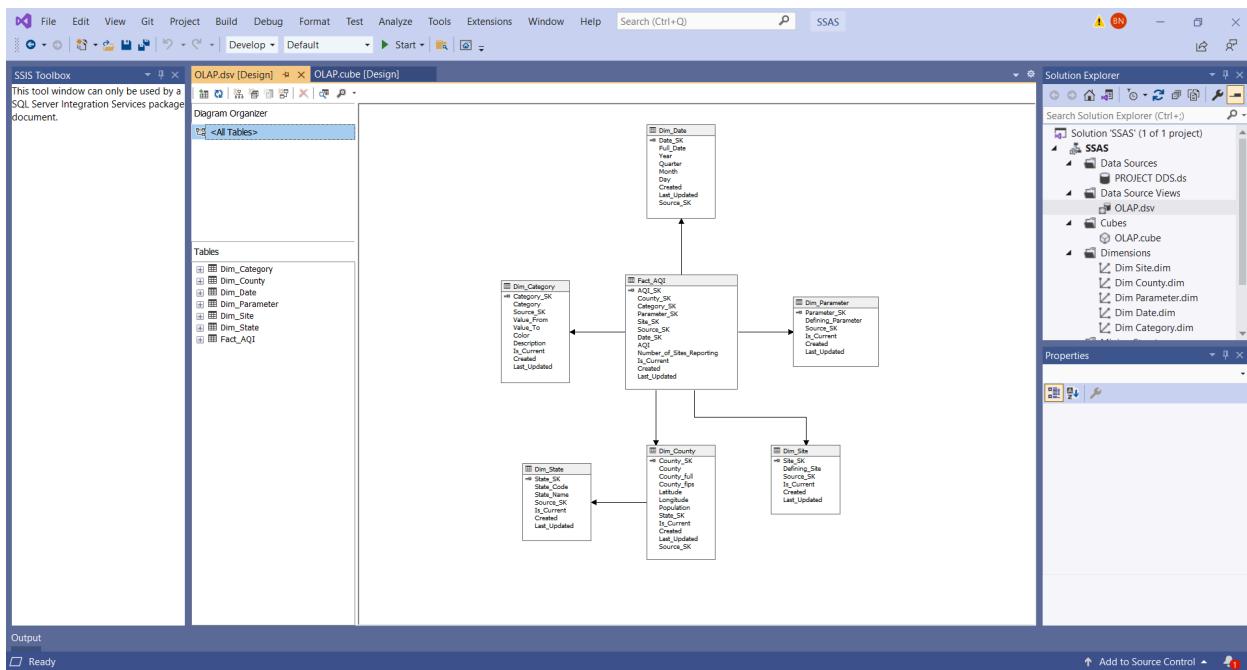
Do yêu cầu tiến độ chỉ tới phần tạo OLAP Cube, chưa thực hiện phân tích dữ liệu trong cube và viết MDX. Thế nên, phần OLAP cũng chỉ được nhóm tạo dựa trên DDS vừa được ETL vào.

Đầu tiên, nhóm tạo data source kết nối với DDS:



Kết nối tới bằng username/password của Window Account.

Sau đó tạo Data Source View và Cube cho DDS vừa thêm:



Ở đây, nhóm không lấy chiều Dim_Source bởi vì các phân tích không dùng tới Dim_Source.

Sau khi tạo Cube, ta sẽ có được các chiều:

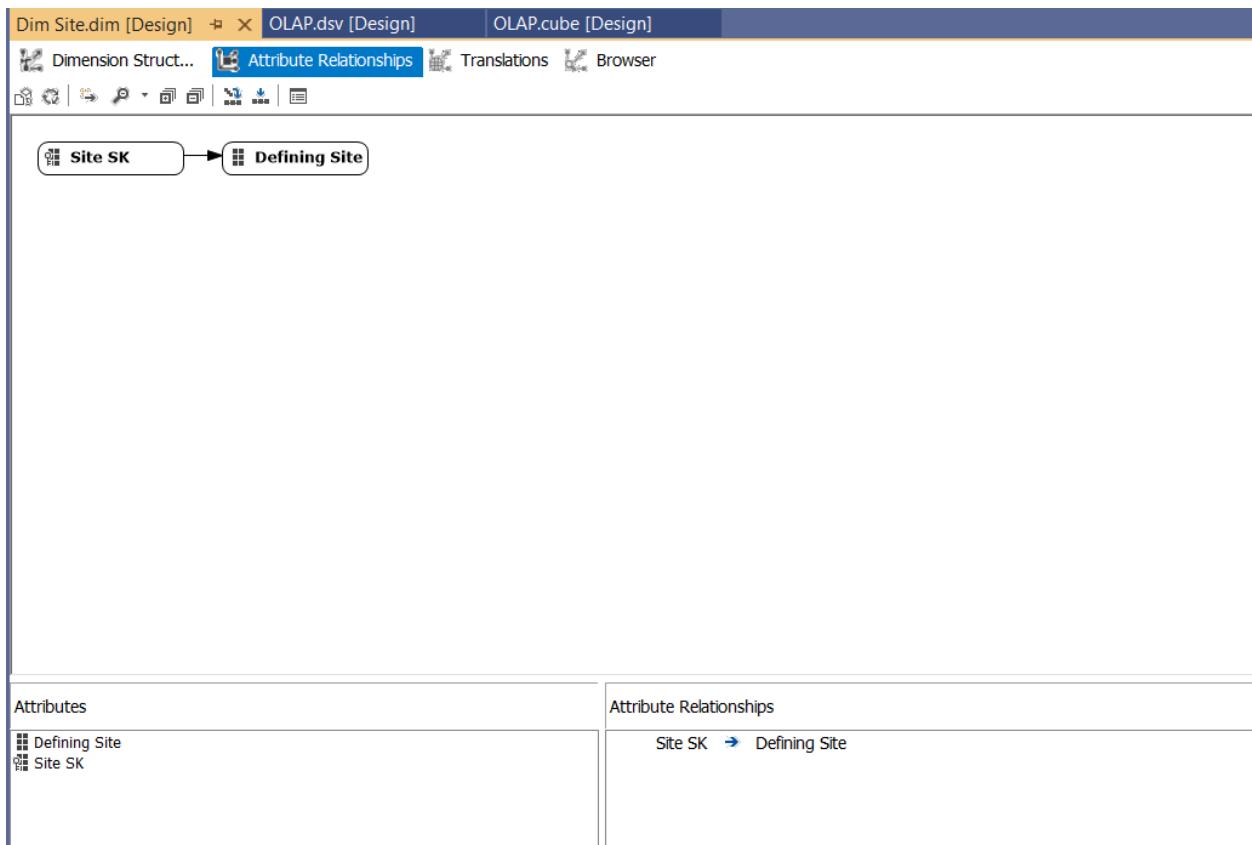


Nhấn đúp vào các chiều, nhóm tiến hành cấu trúc các chiều như sau:

Dim Site:

The screenshot shows the 'Dim Site.dim [Design]' tab in the Analysis Services Designer. The interface is divided into three main sections: Attributes, Hierarchies, and Data Source View.

- Attributes:** Shows the 'Dim Site' attribute selected, with sub-items 'Defining Site' and 'Site SK'.
- Hierarchies:** Shows a 'Hierarchy Site' node with 'Defining Site' and '<new level>' under it. A tooltip indicates: 'To create a new hierarchy, drag an attribute here.'
- Data Source View:** Shows the 'Dim_Site' table with columns: Site_SK, Defining_Site, Source_SK, Is_Current, Created, and Last_Updated.



Điều chỉnh Properties cho cột Defining Site như sau:

Source	
CustomRollupColumn	(none)
CustomRollupPropertiesColumn	(none)
KeyColumns	Dim_Site.Site_SK (Integer) ...
NameColumn	Dim_Site.Defining_Site (WChar)
ValueColumn	(none)

Ta thực hiện tương tự cho Dim Parameter, Dim Category.

Attributes

- Dim_Parameter
 - Defining Parameter
 - Parameter SK

Hierarchies

Dimension Attribute Properties

Defining Parameter DimensionAttribute

MembersWithData	NonLeafDataVisible
MembersWithDataCaption	
NamingTemplate	
RootMemberIf	ParentIsBlankSelfOrMissing
UnaryOperatorColumn	(none)
Source	
CustomRollupColumn	(none)
CustomRollupPropertiesColumn	(none)
KeyColumns	Dim_Parameter.Parameter_SK (Int...)
NameColumn	Dim_Parameter.Defining_Parameter
ValueColumn	(none)

KeyColumns
Specifies the details of the binding to the column(s) containing the member key(s).

Attributes

- Dim_Category
 - Category
 - Category SK

Hierarchies

Dimension Attribute Properties

Defining Parameter DimensionAttribute

MembersWithData	NonLeafDataVisible
MembersWithDataCaption	
NamingTemplate	
RootMemberIf	ParentIsBlankSelfOrMissing
UnaryOperatorColumn	(none)
Source	
CustomRollupColumn	(none)
CustomRollupPropertiesColumn	(none)
KeyColumns	Dim_Category.Category_SK (Int...)
NameColumn	Dim_Category.Defining_Parameter
ValueColumn	(none)

KeyColumns
Specifies the details of the binding to the column(s) containing the member key(s).

Dim Date:

The screenshot shows the 'Dimension Structure' tab selected in the top navigation bar. The interface is divided into three main sections: Attributes, Hierarchies, and Data Source View.

Attributes section:

- Dim Date
 - Date SK
 - Day
 - Month
 - Quarter
 - Year

Hierarchies section:

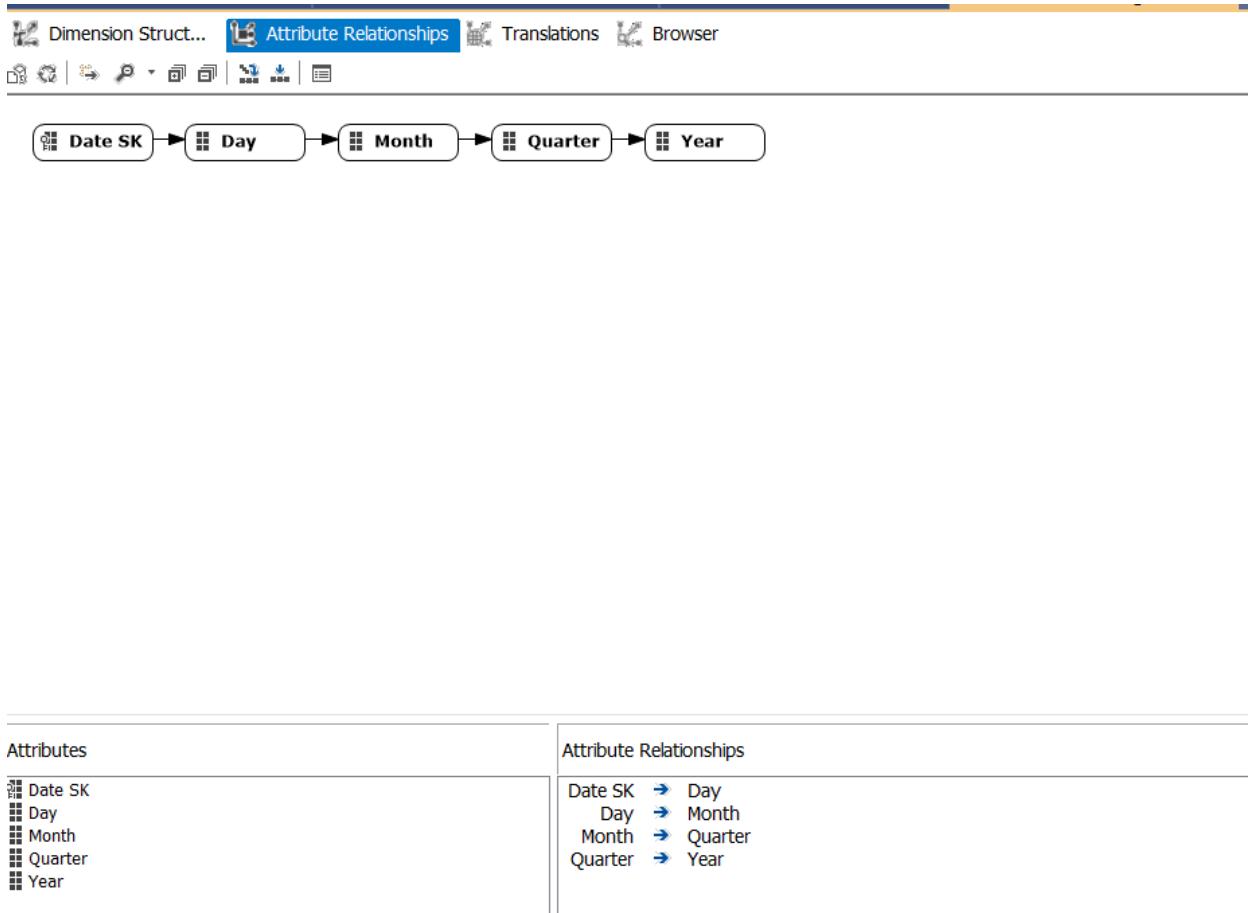
Hierarchy Date

- Year
- Quarter
- Month
- Day
- <new level>

To create a new hierarchy, drag an attribute here.

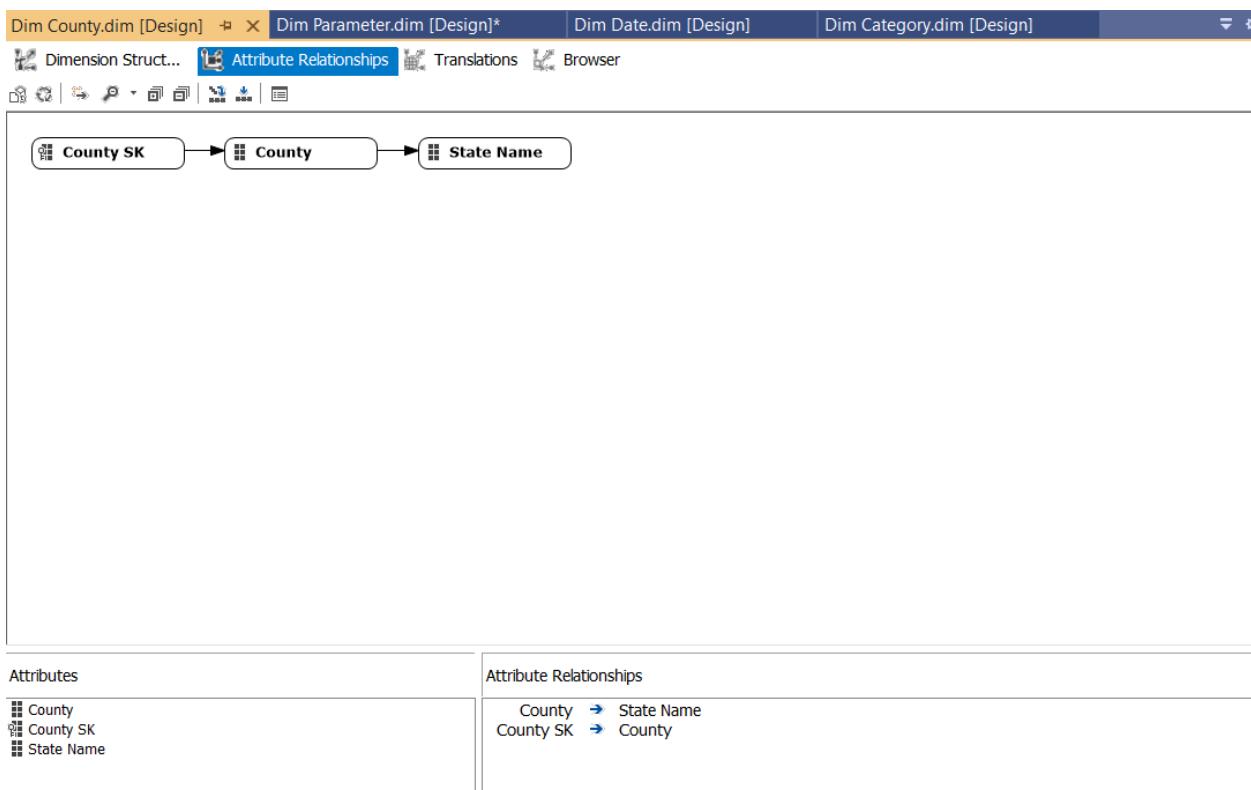
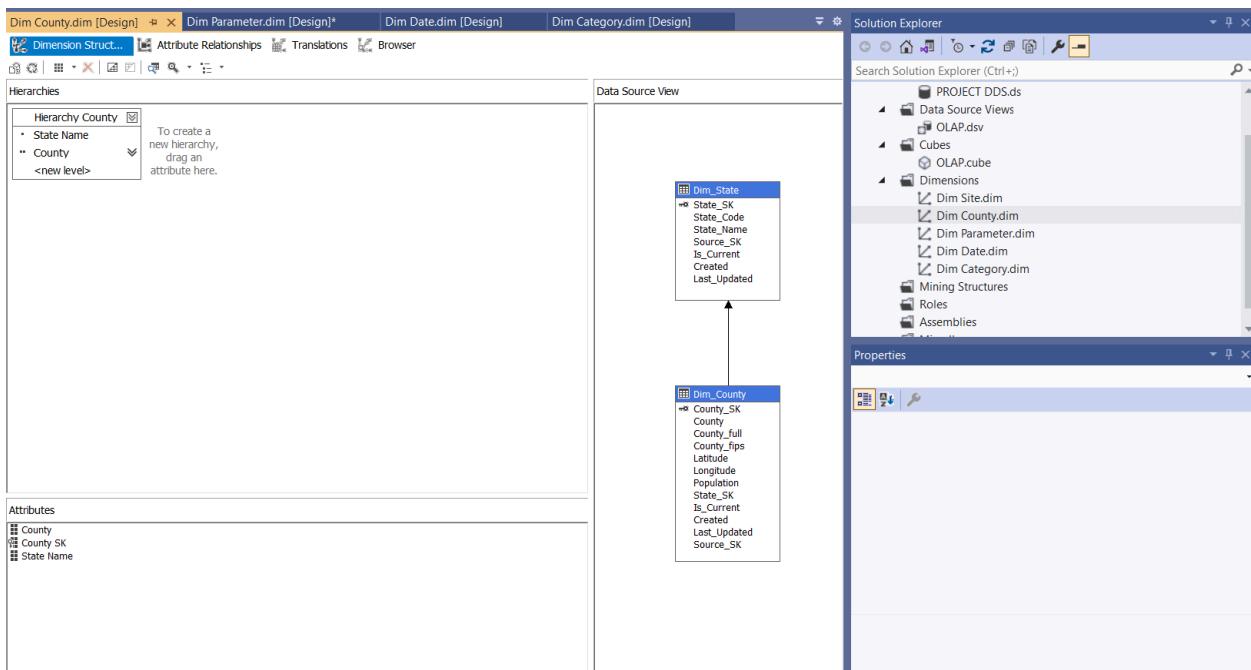
Data Source View section:

Dim Date
Date_SK
Full_Date
Year
Quarter
Month
Day
Created
Last_Updated
Source_SK



Cấu trúc key cho các thuộc tính Year -> Month -> Quarter -> Day là tổ hợp bao gồm các key của cấp lớn hơn và cột thuộc tính đó. Ví dụ Year có key là Year => Month có key là Month và Year. Và name column là tên của thuộc tính đó.

Dim County bao gồm cả Dim State được cấu trúc như sau:



Với Key Column và Name Column tương tự như Dim Date với các chiều là State => County.

Cuối cùng, vào lại cube và thực hiện process để chạy tạo cube:



Sau khi chạy xong, ta thực hiện thử Data Analyst:

A screenshot of the 'Data Analyst' window in SSMS, showing the results of a query. The window title is 'OLAP.cube [Design]'. The top menu bar includes 'Edit as Text', 'Import...', 'MDX', and various toolbars. The main area contains a table titled 'AvgAQI' with columns for 'Category' and 'AvgAQI'. The data shows the following values:

Category	AvgAQI
Good	41.5883084678904
Hazardous	207.962962962963
Moderate	55.1221488123329
Unhealthy	112.548951048951
Unhealthy for Sensitive Groups	80.4603524229075
Very Unhealthy	156.428571428571

Có thể thấy quá trình tạo Cube đã hoàn tất. Ta tiến hành sử dụng Cube trong giai đoạn tiếp theo của đồ án.

V. MDX

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

Sự kiện: Dữ liệu AQI của một hạt trong một bang được ghi nhận trong ngày

Bối cảnh sự kiện:

- Ai: Trạm AQI
- Ở đâu: hạt, tiểu bang
- Cái gì: dữ liệu AQI (Chỉ số AQI, phân loại, đơn vị)
- Khi nào: ngày ghi nhận

Đo lường: chỉ số AQI

- Các giá trị có sẵn từ nguồn: AQI
- Các giá trị phải tính toán: min(AQI), max(AQI)

Cấp chi tiết dữ liệu: mỗi dòng trong bảng fact tương ứng với một chỉ số chất lượng không khí của một thang đo cho một hạt của một bang trong một ngày tại một trạm đo lường.

MDX:

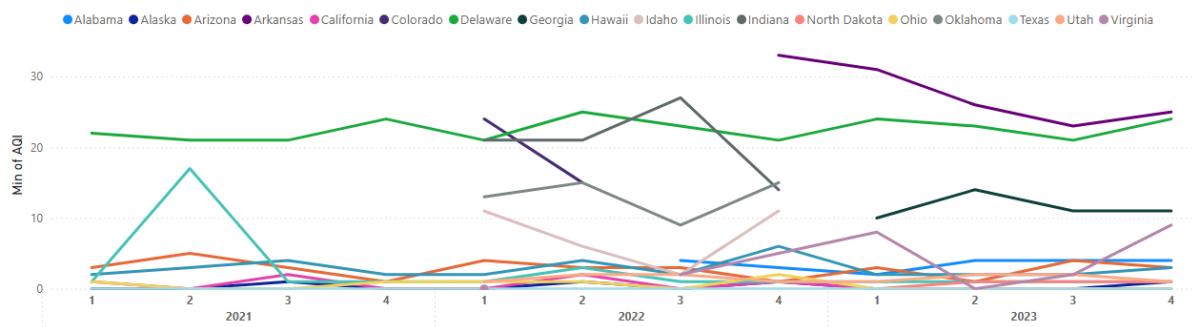
```
1 -- 1. Report the min and max of AQI value for each State during each quarter of years.
2 SELECT
3     { [Measures].[MinAQI],
4         [Measures].[MaxAQI]
5     } ON COLUMNS,
6     NON EMPTY
7     (
8         --[Dim State].[State Name].[State Name].MEMBERS *
9         [Dim County].[State Name].[State Name].MEMBERS *
10        [Dim Date].[Hierarchy Date].[Year].MEMBERS *
11        --[Dim Date].[Hierarchy Date].[Quarter].MEMBERS
12        [Dim Date].[Quarter].[Quarter].MEMBERS
13    ) ON ROWS
14 FROM [OLAP];
```

Kết quả:

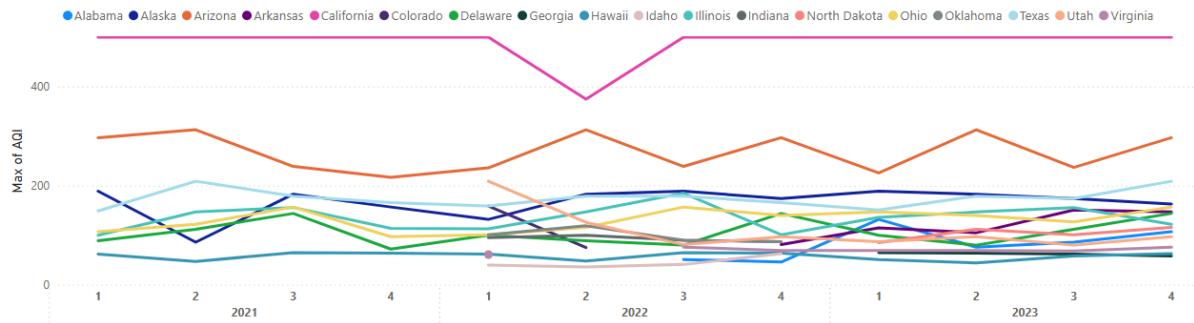
		Messages	Results	MinAQI	MaxAQI
Alabama	2022	3	4	51	
Alabama	2022	4	3	46	
Alabama	2023	1	2	132	
Alabama	2023	2	4	76	
Alabama	2023	3	4	86	
Alabama	2023	4	4	107	
Alaska	2021	1	0	189	
Alaska	2021	2	0	86	
Alaska	2021	3	1	183	
Alaska	2021	4	0	157	
Alaska	2022	1	0	132	
Alaska	2022	2	1	183	
Alaska	2022	3	0	189	
Alaska	2022	4	1	174	
Alaska	2023	1	0	189	
Alaska	2023	2	0	183	

Trực quan:

Giá trị AQI nhỏ nhất của từng tiểu bang theo quý



Giá trị AQI lớn nhất của từng tiểu bang theo quý



Nhận xét:

Trong biểu đồ MinAQI, ta có thể thấy hầu hết các bang đều có mức AQI Good (dưới 50). Có một số bang bị thiếu dữ liệu.

Trong biểu đồ MaxAQI, ta có thể thấy chỉ số MaxAQI có thể đạt được của hầu hết các bang nằm ở mức Unhealthy (151 đến 200). Tuy nhiên, có 2 bang đạt chỉ số rất cao (Very unhealthy and Hazardous) là California và Arizona

Đối với bang Arizona:

- Dựa vào đây, ta có thể thấy, bang Arizona là bang có dân số đông (khoảng 7 triệu người). Dân số đông sẽ dẫn đến việc sử dụng phương tiện giao thông như xe máy, xe hơi rất nhiều, đặc biệt là giờ cao điểm. Ngoài ra, xe tải trọng lớn (loại xe đóng vai trò quan trọng trong nền kinh tế) cũng có rất nhiều. Nhiên liệu mà họ đang sử dụng là nhiên liệu hóa thạch, cho nên khi sử dụng sẽ gây ra một lượng khí thải độc hại lớn.
- Ngoài nguyên nhân trên, khí thải của các nhà máy, xí nghiệp cũng đóng vai trò không hề nhỏ trong việc ô nhiễm không khí. Theo như hình, quý 2 và quý 4 mỗi năm thường có lượng khí thải lớn, nguyên nhân là do lúc này thời tiết lạnh nên cần nhiều nhiên liệu hơn để làm nóng.

Source: [Arizona Air Quality Index \(AQI\) and USA Air Pollution | IQAir](#)

Đối với bang California:

Mức độ ô nhiễm không khí ở bang này là rất cao. Có rất nhiều nguyên nhân dẫn đến việc này.

- Dân số đông (39 triệu người, gấp 5 lần bang Arizona), việc này dẫn đến mức độ sử dụng nhiên liệu cao, tương tự đối với các xí nghiệp.
- Cháy rừng tự nhiên và sau đó tạo ra khói gây ô nhiễm
- Vị trí địa lý không thuận lợi (Sức nóng gây ra ở bên trong không thể thoát ra ngoài do có núi bao quanh vùng biển)

Source:

[California Air Quality Index \(AQI\) and USA Air Pollution | IQAir](#)

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

Sự kiện: Dữ liệu AQI của một hạt trong một bang được ghi nhận trong ngày

Bối cảnh sự kiện:

- Ai: Trạm AQI
- Ở đâu: tiểu bang
- Cái gì: dữ liệu AQI (Chỉ số AQI, phân loại, đơn vị)
- Khi nào: ngày ghi nhận

Đo lường: chỉ số AQI

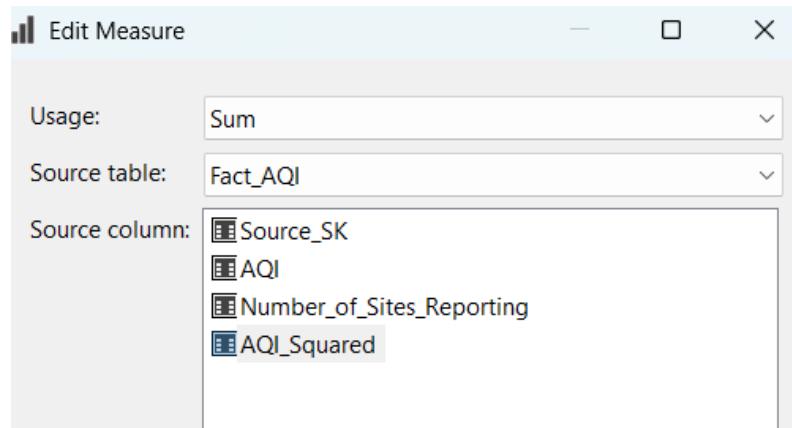
- Các giá trị có sẵn từ nguồn: AQI
- Các giá trị phải tính toán: mean(AQI), stddev(AQI)

Cấp chi tiết dữ liệu: mỗi dòng trong bảng fact tương ứng với một chỉ số chất lượng không khí của một thang đo cho một hạt của một bang trong một ngày tại một trạm đo lường.

Cài đặt bổ sung vào Fact_AQI trong DDS: Chèn cột AQI_Squared vào bảng Fact_AQI với công thức $AQI * AQI$ dùng để thực hiện tính toán độ lệch chuẩn.

```
510 | ALTER TABLE PROJECT_DDS.dbo.Fact_AQI ADD AQI_Squared AS (AQI * AQI);
```

Ở OLAP Cube, tạo MinAQI, MaxAQI và tạo measure Squared AQI với cài đặt là Sum của cột AQI_Squared:



Tạo các calculated measure Mean AQI, Standard Deviation:

The screenshot shows the Analysis Services Scripting Environment. The top menu bar has tabs for OLAP(cube [Design]), Dim Date.dim [Design], Dim County.dim [Design], and OLAP.dsv [Design]. The ribbon bar includes options like Cube Structure, Dimension Usage, Calculations, KPIs, Actions, Partitions, Aggregations, Perspectives, Translations, and Browser. Below the ribbon is a toolbar with various icons. On the left, there's a Script Organizer pane listing items: Command, CALCULATE, [Mean AQI] (selected), and [Standard Deviation]. A Calculation Tools pane on the right contains tabs for Metadata, Functions, and Templates, along with search and measure group selection fields. The main workspace displays the properties for the selected calculated measure [Mean AQI]. The 'Name' field is set to '[Mean AQI]'. Under 'Parent Properties', 'Parent hierarchy' is set to 'Measures'. Under 'Expression', the formula is '[Measures].[AQI]/[Measures].[CountFactAQI]'. A status bar at the bottom right shows 'Ln: 1 Ch: 43'.

This screenshot shows the continuation of creating calculated measures. The top menu bar and ribbon are identical to the previous screen. The Script Organizer pane lists the same items: Command, CALCULATE, [Mean AQI], and [Standard Deviation] (selected). The Calculation Tools pane is also similar. The main workspace shows the properties for the selected calculated measure [Standard Deviation]. The 'Name' field is set to '[Standard Deviation]'. Under 'Parent Properties', 'Parent hierarchy' is set to 'Measures'. Under 'Expression', the formula is an IIF statement:

```
IIF([Measures].[CountFactAQI] > 0,
    IIF(
        ((([Measures].[Squared AQI] / [Measures].[CountFactAQI]) -
        ((([Measures].[AQI])^2) / ([Measures].[CountFactAQI]))^2)) >= 0,
        ((([Measures].[Squared AQI] / [Measures].[CountFactAQI]) -
        ((([Measures].[AQI])^2) / ([Measures].[CountFactAQI])))^2)^0.5,
        0
    )
)
```

. A status bar at the bottom right shows 'Ln: 11 Ch: 6 SPC C1'. There is also a 'Change' button next to the 'Parent member' dropdown.

MDX:

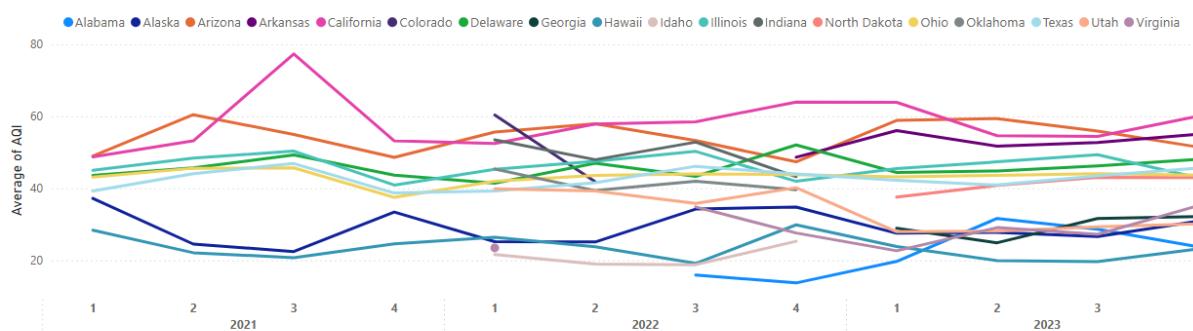
```
16 -- 2. Report the mean and the standard deviation of AQI value for each State during each quarter of years.
17 WITH
18 MEMBER [Measures].[Max_AQI] AS
19     CoalesceEmpty([Measures].[MaxAQI], 0)
20 MEMBER [Measures].[Min_AQI] AS
21     CoalesceEmpty([Measures].[MinAQI], 0)
22 MEMBER [Measures].[Mean_AQI] AS
23     CoalesceEmpty([Measures].[Mean AQI], 0)
24 MEMBER [Measures].[Standard_Deviation] AS
25     CoalesceEmpty([Measures].[Standard Deviation], 0)
26 SELECT
27 NON EMPTY { [Measures].[Max_AQI], [Measures].[Min_AQI], [Measures].[Mean_AQI], [Measures].[Standard_Deviation] } ON COLUMNS,
28 NON EMPTY {
29 FILTER([Dim County].[State Name].[State Name] * [Dim Date].[Year].[Year].ALLMEMBERS * [Dim Date].[Quarter].[Quarter],
30 [Dim County].[State Name].CURRENTMEMBER.NAME <> "Unknown" AND
31 [Dim Date].[Year].CURRENTMEMBER.NAME <> "Unknown" AND
32 [Dim Date].[Quarter].CURRENTMEMBER.NAME <> "Unknown" ) }
33 DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [OLAP];
```

Kết quả:

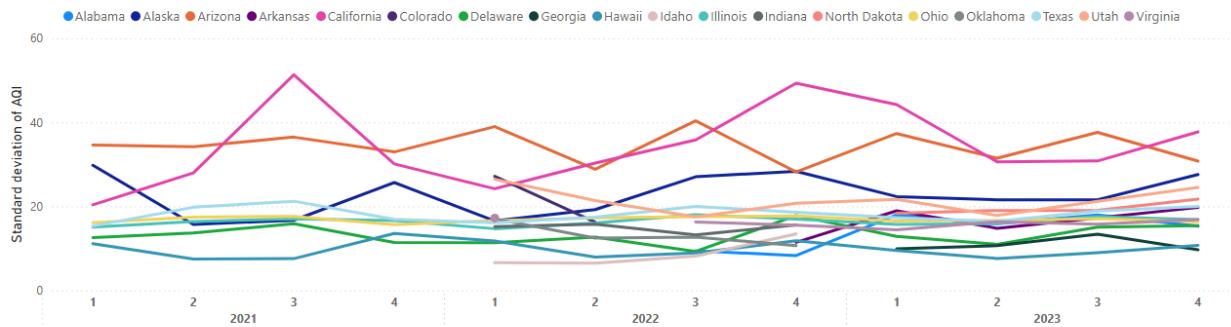
			Max_AQI	Min_AQI	Mean_AQI	Standard_Deviation
Alabama	2021	1	0	0	0	0
Alabama	2021	2	0	0	0	0
Alabama	2021	3	0	0	0	0
Alabama	2021	4	0	0	0	0
Alabama	2022	1	0	0	0	0
Alabama	2022	2	0	0	0	0
Alabama	2022	3	51	4	15.9272727272727	9.40377979219874
Alabama	2022	4	46	3	13.734693877551	8.2999274897351
Alabama	2023	1	132	2	19.7111111111111	17.9816229921578
Alabama	2023	2	76	4	31.6521739130435	15.8528798526302
Alabama	2023	3	86	4	28.6421052631579	17.9437224263873
Alabama	2023	4	107	4	23.780701754386	15.2934791584747
Alaska	2021	1	189	0	37.2346368715084	29.787029151755
Alaska	2021	2	86	0	24.5223880597015	15.713724592401
Alaska	2021	3	183	1	22.4338235294118	16.6869763437919
Alaska	2021	4	157	0	33.41666666666667	25.6661345543688

Trực quan:

Chất lượng không khí trung bình theo từng quý ở từng tiểu bang



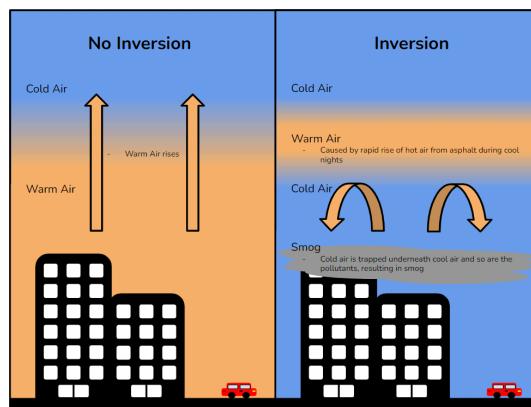
Chất lượng không khí trung bình theo từng quý ở từng tiểu bang



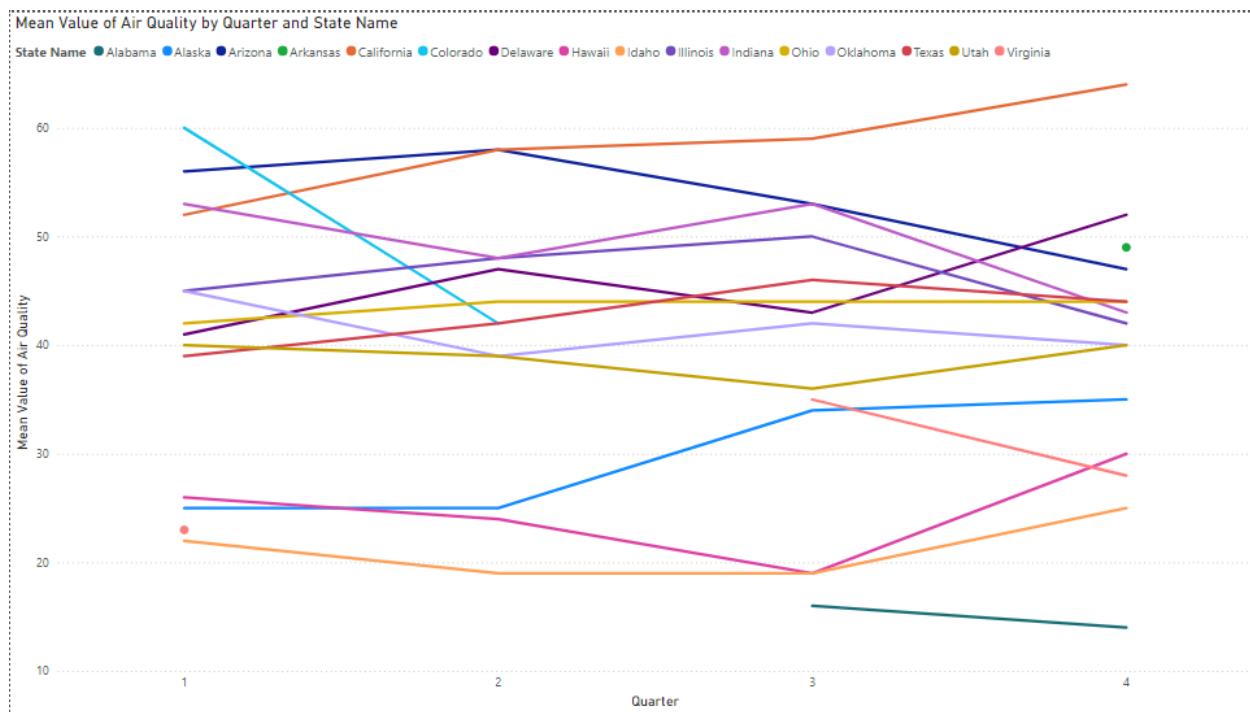
Nhận xét:

Biểu đồ thể hiện giá trị chất lượng không khí trung bình của các bang ở Mỹ trong năm 2021. Nhìn chung, chất lượng không khí ở các bang đều có xu hướng tăng nhưng giảm mạnh vào những tháng cuối năm.

Trong ba quý đầu năm, mức độ ô nhiễm không khí liên tục tăng, đỉnh điểm là vào tháng 3. Có nhiều nguyên nhân dẫn đến sự tăng trưởng này. Vào những tháng đầu năm, những tháng mùa đông ở lục địa Bắc Mỹ, hiện tượng nghịch nhiệt (temperature inversions) thường xuyên diễn ra, đặc biệt ở vùng thung lũng và đô thị. Hiện tượng này làm thay đổi thứ tự của các tầng khí có nhiệt độ khác nhau, lớp khí nóng bị kẹp giữa hai lớp khí lạnh hơn. Lớp khí nóng ngăn các chất ô nhiễm có thể được phân tán đi cho tới khi thời tiết thay đổi. Ngoài ra, vào những tháng mùa hè, đợt nắng nóng nghiêm trọng làm tăng nguy cơ cháy rừng ở các bang. Ví dụ, vào tháng 7, vụ cháy rừng Dixie ở California đã thả ra lượng lớn bụi mịn (PM2.5) và khí độc như CO, NOx vào không khí. Cũng vì lí do này, chất lượng không khí ở California bị đẩy lên mức nguy hiểm, ảnh hưởng sức khỏe ở mức độ nghiêm trọng.

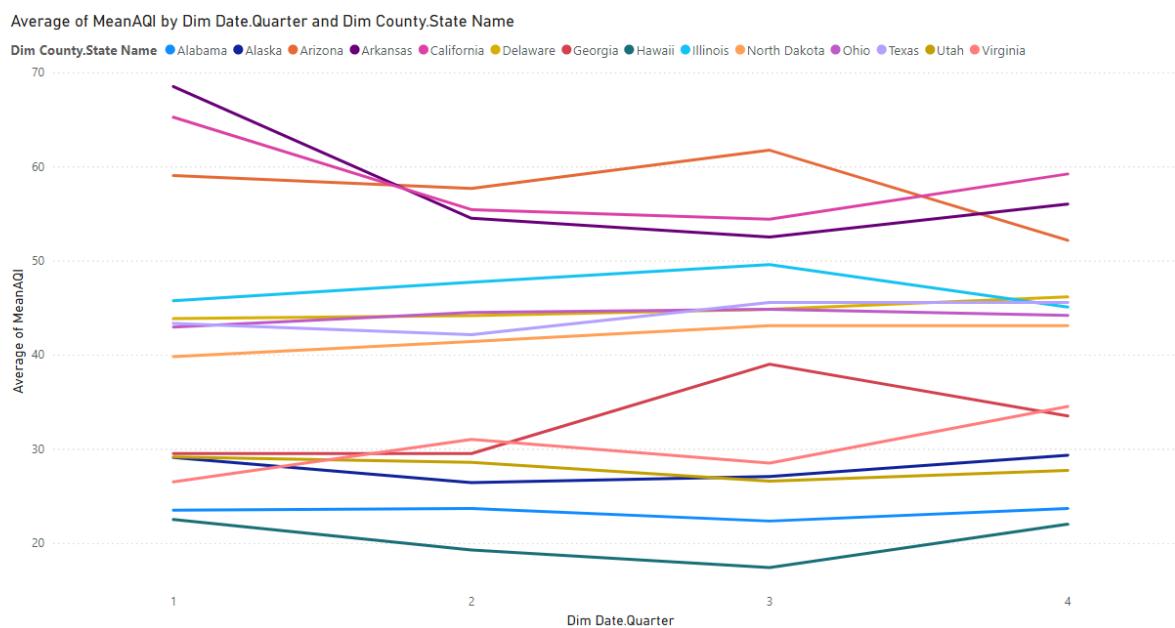


Mặc khác, nhờ nằm ở vị trí tách biệt so với phần lãnh thổ lục địa, Hawaii và Alaska không chịu ảnh hưởng nhiều từ các hoạt động công nghiệp và giao thông đông đúc – những nguyên nhân chính gây ô nhiễm không khí. Đặc biệt, điều kiện khí hậu của hai khu vực này cũng góp phần cải thiện chất lượng không khí. Hawaii với các luồng gió biển thường xuyên giúp phân tán các chất ô nhiễm, trong khi Alaska có mật độ dân cư thưa thớt và thiên nhiên rộng lớn giúp giảm thiểu các nguồn phát thải. Điều này được phản ánh rõ qua dữ liệu, khi giá trị chất lượng không khí thấp nhất được ghi nhận vào quý 3 lần lượt là 21 và 22, cho thấy mức độ ô nhiễm rất thấp so với các bang khác ở Mỹ.



Dữ liệu ở những năm 2022 ghi nhận thêm nhiều tiểu bang hơn. Mức độ ô nhiễm giảm dần từ quý 1 đến quý 4 ở hầu hết các bang, với các giá trị trong quý 4 thường thấp hơn đáng kể so với quý 1. Điều này có thể phản ánh tác động của thời tiết, khi lưu thông không khí và các điều kiện khí hậu mùa đông có thể giúp giảm nồng độ các hạt ô nhiễm. Cũng tương tự năm 2021, hiện tượng nghịch nhiệt tiếp tục làm chất lượng không khí ở các bang khá cao ở những tháng đầu năm. California có mức độ ô nhiễm ổn định ở mức cao (dao động từ 55.59 đến 64.10), nguyên nhân là do mật độ dân cư đông, lượng lớn phương tiện giao thông, các hoạt động công nghiệp, vấn đề ô nhiễm khí thải từ xe cộ và cháy rừng tại bang này. Trong khi đó, Hawaii, Virginia, Idaho, và Oklahoma có chỉ số ô nhiễm thấp hơn so với các bang khác trong hầu hết

các quý. Hawaii có chỉ số thấp nhất (dao động từ 18.80 (Q3) đến 28.83 (Q4)), nhờ vào khí hậu trong lành và nền kinh tế ít công nghiệp hóa.



Trong năm 2023, mức độ ô nhiễm ít biến động hơn so với năm 2022, với chỉ số giữa các quý không chênh lệch đáng kể trong nhiều bang. Hawaii vẫn là tiểu bang có chất lượng không khí tốt nhất và California vẫn ở mức tương đối cao (54.37 (Q3) đến 65.19 (Q1)). Cụ thể, các bang như Arkansas, Arizona, và California tiếp tục duy trì mức độ ô nhiễm cao. Arkansas có chỉ số tăng cao nhất trong quý 1 (68.50) nhưng giảm nhẹ ở các quý sau. California tiếp tục có vấn đề ô nhiễm nghiêm trọng, đặc biệt là trong quý 1 (65.19), do các nguyên nhân như cháy rừng, giao thông đông đúc, và công nghiệp. Ngược lại, Hawaii, Georgia, và Utah tiếp tục là những bang có chỉ số thấp nhất, phản ánh chất lượng không khí tốt hơn. Hawaii giữ mức thấp nhất trong cả năm (17.40 - 22.50) nhờ vào khí hậu biển và ít công nghiệp hóa. Georgia có chỉ số dao động từ 29.50 (Q1) đến 33.50 (Q4), cho thấy mức ô nhiễm ổn định và thấp hơn nhiều bang khác.

Trong giai đoạn từ năm 2021 đến 2023, chất lượng không khí tại các bang của Hoa Kỳ cho thấy sự khác biệt đáng kể về mức độ ổn định và biến động qua từng năm. Năm 2021, các bang như California (14.80 - 24.08) và Arizona (12.79 - 16.90) ghi nhận độ lệch chuẩn cao, phản ánh sự biến động lớn trong chất lượng không khí, có thể do tác động của cháy rừng, giao thông và hoạt động công nghiệp. Ngược lại, các

bang như Hawaii (3.02 - 4.17) và Georgia (3.18 - 4.11) duy trì sự ổn định đáng kể nhờ điều kiện tự nhiên và ít chịu ảnh hưởng từ hoạt động công nghiệp. Năm 2022, mức độ biến động tăng nhẹ ở các bang chịu ô nhiễm cao như California và Arizona, đặc biệt vào các quý mùa hè và mùa thu. Tuy nhiên, đến năm 2023, chất lượng không khí đã có sự cải thiện rõ rệt, với độ lệch chuẩn giảm mạnh tại nhiều bang. California giảm từ 24.12 (Q3, 2022) xuống 13.78 (Q3, 2023), và Alaska giảm từ 15.03 (Q2, 2022) xuống 6.71 (Q2, 2023), cho thấy hiệu quả từ các biện pháp kiểm soát môi trường. Nhìn chung, năm 2023 chứng kiến sự ổn định hơn trong chất lượng không khí trên khắp các bang, đánh dấu bước tiến tích cực trong nỗ lực bảo vệ môi trường.

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.

Sự kiện: Dữ liệu AQI của một hạt của một tiểu bang được ghi nhận trong một ngày

Bối cảnh sự kiện:

- Ai: Trạm ghi nhận AQI
- Ở đâu: hạt của tiểu bang
- Cái gì: chỉ số AQI
- Khi nào: mỗi ngày

Đo lường: số lượng ngày, chỉ số AQI trung bình của hạt của mỗi tiểu bang được đánh giá chất lượng không khí là “very unhealthy” hoặc hơn.

- Các giá trị có sẵn từ nguồn: AQI, Category, CategorySK, DateSK
- Các giá trị phải tính toán: Số lượng ngày = Count(Distinct DateSK)
- Điều kiện tính toán: AQI > 200 OR Category IN [“Very Unhealthy”, “Hazardous”]

Cấp chi tiết độ mịn: Một dòng trong fact tương ứng với chỉ số AQI của một Parameter ở một hạt trong một tiểu bang được đo truong một ngày.

MDX:

```
35  -- 3. Report the number of days, and the mean AQI value where the air quality is rated
36  --as "very unhealthy" or worse for each State and County.
37  SELECT NON EMPTY
38    { [Measures].[NumberOfDay], [Measures].[Mean AQI] } ON COLUMNS,
39    NON EMPTY
40    {
41      CROSSJOIN([Dim County].[Hierarchy County].[State Name],[Dim County].[County].[County])
42    } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS
43  FROM
44  (SELECT
45    {[Dim Category].[Hierarchy Category].[Category].&[Hazardous],
46    [Dim Category].[Hierarchy Category].[Category].&[Very Unhealthy]
47  } ON COLUMNS
48  FROM [OLAP]);
```

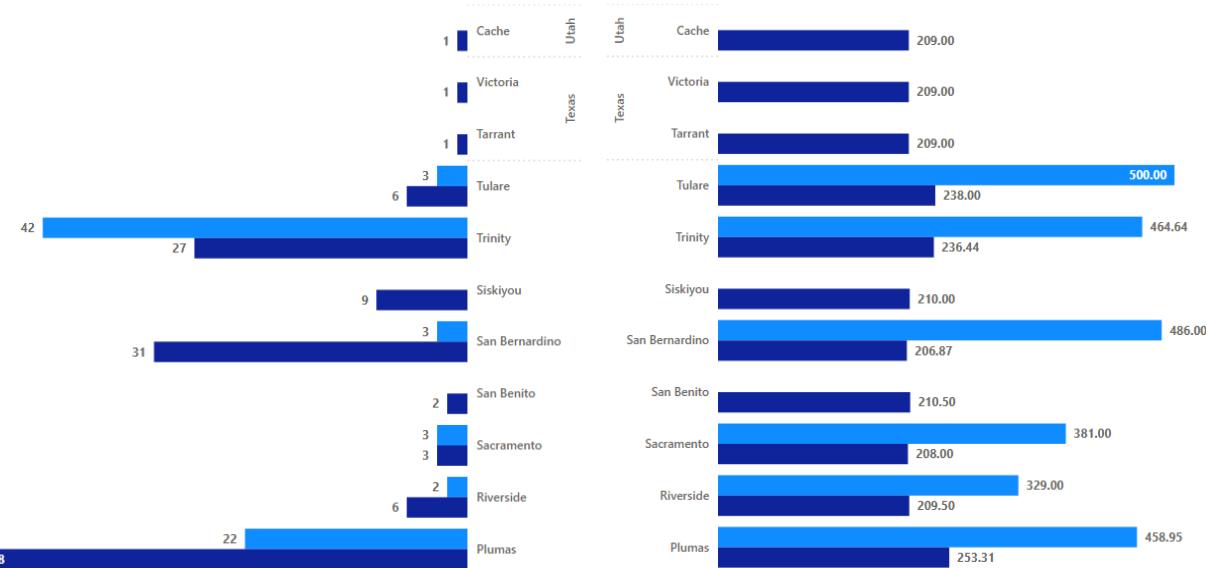
Kết quả:

		Messages	Results
		NumberOfDay	Mean AQI
Arizona	Gila	1	297
Arizona	La Paz	73	215.575342465753
Arizona	Maricopa	145	215.013793103448
Arizona	Navajo	1	313
Arizona	Pima	2	313
California	Amador	3	244.333333333333
California	Butte	6	244.333333333333
California	Calaveras	1	225
California	Colusa	2	225
California	Del Norte	2	324.5
California	El Dorado	7	296.714285714286
California	Fresno	3	241.333333333333
California	Humboldt	3	342
California	Imperial	13	330.538461538462
California	Inyo	13	314
California	Kern	5	326.8

Trực quan:

Số ngày có chất lượng không khí lớn hơn 200 theo từng hạt ở từ...

● Hazardous ● Very Unhealthy



Trung bình chất lượng không khí lớn hơn 200 theo từng hạt ở từ...

● Hazardous ● Very Unhealthy

Nhận xét:

Nhìn chung, tiểu bang Arizona và California có số ngày không khí nguy hiểm cao. Hạt Maricopa của tiểu bang Arizona có số ngày cao nhất với 145 ngày trong ba năm được ghi nhận, với chỉ số chất lượng không khí trung bình là 215. Tiểu bang chịu ảnh hưởng nghiêm trọng bởi Ozone và bụi mịn 2.5, đặc biệt nghiêm trọng vào các tháng mùa hè. Ở cùng tiểu bang, hạt La Paz cũng ghi nhận số lượng ngày ô nhiễm lên đến 73 ngày.

Ở tiểu bang California, các bang như Trinity và Plumas chịu ảnh hưởng nghiêm trọng từ các vụ cháy rừng vào mùa hè cũng làm tăng số lượng ngày ô nhiễm lên lần lượt là 69 và 70 ngày. Đặc biệt là đám cháy rừng Dixie đã làm cho nhiều ngày liên tiếp đạt chỉ số bụi mịn 2.5 lên đến mức cao nhất 500, với con số trung bình là 464. Plumas cũng chịu ảnh hưởng tương tự của đám cháy trong thời gian dài với con số trung bình là 458 trong những ngày cực kỳ nguy hiểm.

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.

Sự kiện: Khi trạm ghi nhận được AQI của một trong 4 bang (Hawaii, Alaska, Illinois và Delaware).

Bối cảnh:

- Ai: Trạm.
- Ở đâu:
 - Bang (State): Hawaii, Alaska, Illinois, Delaware.
 - Quận (County): Các quận của các State trên
- Cái gì: Chỉ số chất lượng không khí (AQI)
- Khi nào: Mỗi ngày quan sát.

Đo lường:

- Giá trị có sẵn: Từng ngày được ghi nhận AQI
- Giá trị cần tính toán: Số ngày được ghi nhận trong từng loại chất lượng không khí.

Số ngày = COUNT(DISTINCT Date_SK) GROUPBY (Category, State)

Độ mịn:

Mỗi dòng dữ liệu trong fact tương ứng với một quan sát về chất lượng không khí của một ngày tại một quận trong một bang

MDX:

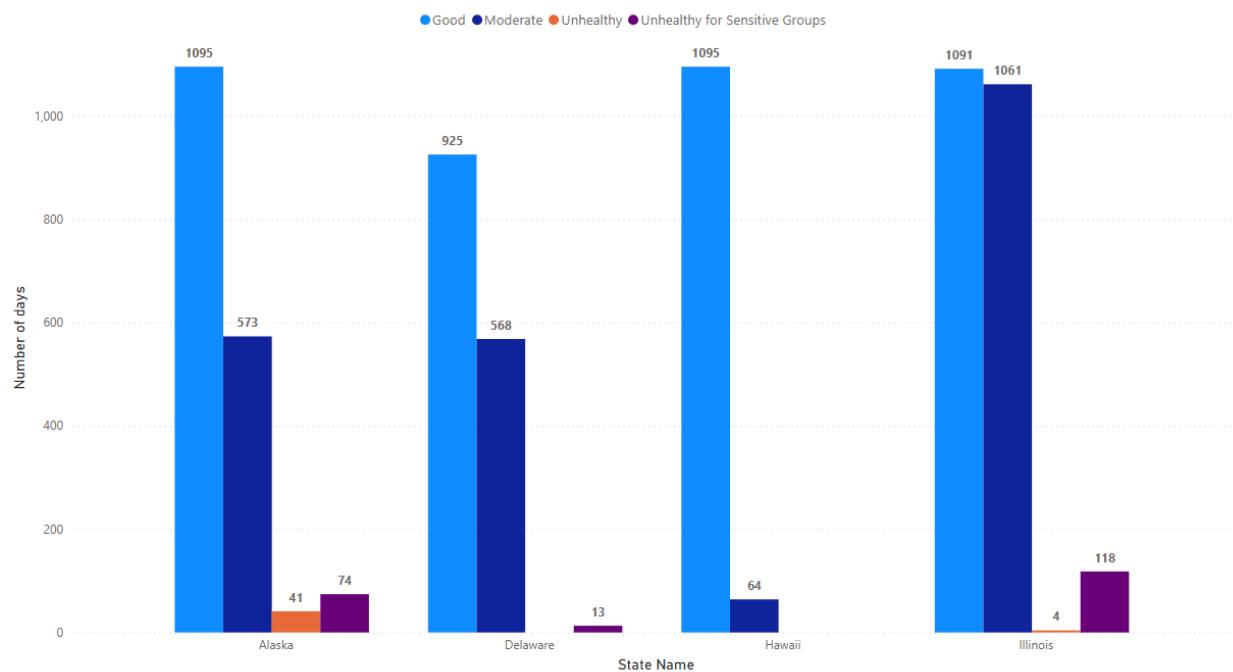
```
51  -- 4. For the four following states: Hawaii, Alaska, Illinois and Delaware,
52  -- count the number of days in each air quality Category (Good, Moderate,etc.) by County
53 WITH
54 MEMBER [Measures].[CoalescedMeasure] AS
55     CoalesceEmpty([Measures].[NumberOfDay], 0) -- Null to 0
56 SELECT
57     FILTER(
58         [Dim Category].[Category].MEMBERS,
59         NOT(
60             [Dim Category].[Category].CURRENTMEMBER.NAME = "UNKNOWN")
61         ) ON COLUMNS,
62         NON EMPTY
63     { [Dim County].[State Name].[Hawaii],
64     [Dim County].[State Name].[Alaska],
65     [Dim County].[State Name].[Illinois],
66     [Dim County].[State Name].[Delaware] }
67     * FILTER(
68         {[Dim County].[County].MEMBERS},
69         [Dim County].[County].CURRENTMEMBER.NAME <> "All"
70     ) ON ROWS
71 FROM [OLAP]
72 WHERE [Measures].[CoalescedMeasure];
```

Kết quả:

		All	Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Hawaii	Hawaii	1088	1029	0	59	0	0	0
Hawaii	Honolulu	1095	1092	0	3	0	0	0
Hawaii	Kauai	454	454	0	0	0	0	0
Hawaii	Maui	1027	1023	0	4	0	0	0
Alaska	Aleutians East	360	299	0	60	0	1	0
Alaska	Anchorage	1095	939	0	138	7	11	0
Alaska	Denali	1051	854	0	158	13	26	0
Alaska	Fairbanks North Star	1095	800	0	247	19	29	0
Alaska	Juneau	1063	966	0	94	1	2	0
Alaska	Kenai Peninsula	360	319	0	38	1	2	0
Alaska	Matanuska-Susitna	1053	973	0	77	1	2	0
Alaska	North Slope	335	324	0	10	0	1	0
Illinois	Adams	487	440	0	47	0	0	0
Illinois	Champaign	1049	738	0	311	0	0	0
Illinois	Clark	1033	897	0	136	0	0	0
Illinois	Cook	1095	499	0	554	2	40	0

Trực quan:

Tổng số ngày của từng loại chất lượng không khí của 4 tiểu bang Alaska, Delaware, Hawaii và Illinois



Nhận xét:

- Biểu đồ thể hiện tổng số ngày của từng loại chất lượng không khí tại 4 tiểu bang: Alaska, Delaware, Hawaii và Illinois. Từ biểu đồ, ta nhận ra:
- Cả 4 bang thì chất lượng không khí chủ yếu được đo là mức "Good". Trong đó thì Hawaii và Illinois có số ngày chất lượng không khí "Good" cao nhất - 1095 ngày. Còn Alaska và Delaware có số ngày "Good" thấp hơn, lần lượt là 1095 và 925 ngày.
- Alaska và Delaware có số ngày "Moderate" tương đương nhau, khoảng 573 và 568 ngày. Trong khi Illinois là nơi có nhiều ngày với chất lượng không khí "Moderate" nhất. Ở chiều hướng ngược lại thì Hawaii lại có ít ngày nhất.
- So với mức "Good" và "Moderate" thì 4 mức còn lại rất ít ỏi, cho thấy chất lượng không khí tại cả 4 bang đều ở mức tốt.
- "Unhealthy" và "Unhealthy for sensitive group" là rất nhỏ (<100).
- Ở 2 mức nguy hiểm là "Very Unhealthy" và "Hazardous" thì không hề tồn tại trên biểu đồ.

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

Sự kiện: Khi trạm ghi nhận được AQI của một trong 4 bang (Hawaii, Alaska, Illinois và Delaware).

Bối cảnh:

- Ai: Trạm.
- Ở đâu: Bang (State): Hawaii, Alaska, Illinois, Delaware.
- Cái gì: Chỉ số chất lượng không khí (AQI)
- Khi nào: Mỗi ngày quan sát.

Đo lường:

- Giá trị có sẵn: Từng ngày được ghi nhận AQI
- Giá trị cần tính toán: Giá trị trung bình AQI theo quý

Độ mịn:

Mỗi dòng dữ liệu trong fact tương ứng với một quan sát về chất lượng không khí của một ngày tại một quận trong một bang

MDX:

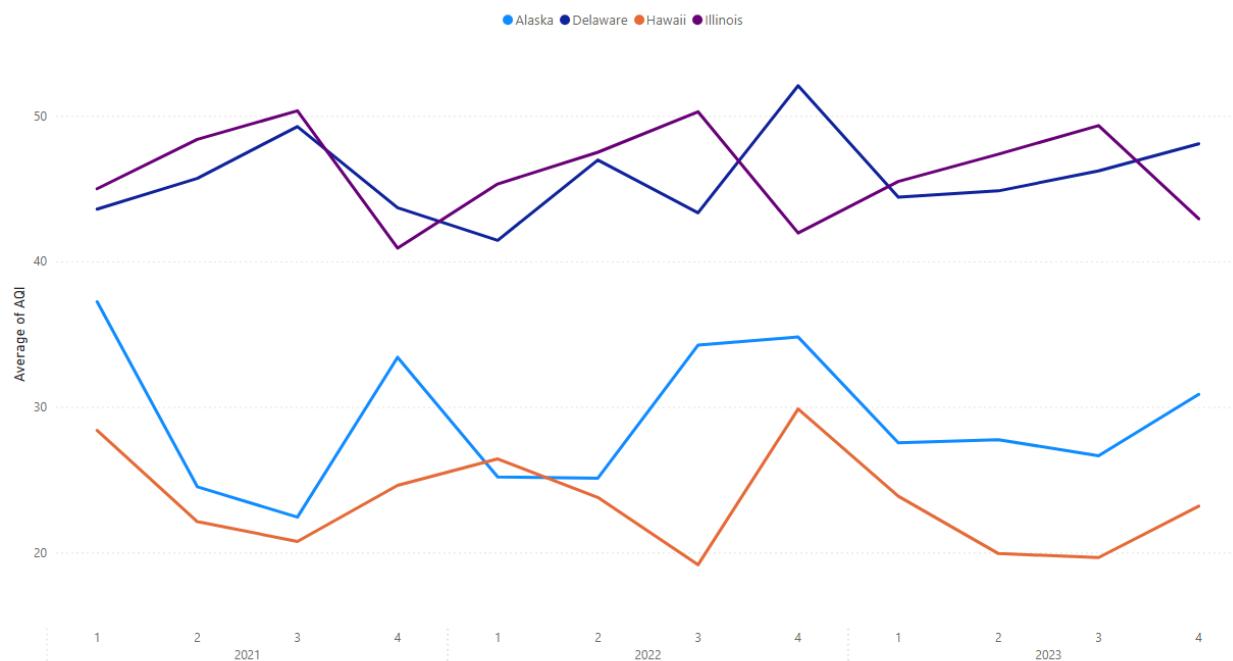
```
76  -- 5. For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the
77  --mean AQI value by quarters.
78 WITH
79 Member [Measures].[Trung bình AQI] AS
80 CoalesceEmpty([Measures].[Mean AQI],0) --Null to 0
81 SELECT
82   FILTER ({*
83     [Dim Date].[Hierarchy Date].[Year].Members *
84     [Dim Date].[Quarter].Members
85   }, ([Dim Date].[Year].CURRENTMEMBER.NAME <> "Unknown" AND [Dim Date].[Quarter].CURRENTMEMBER.NAME <> "Unknown"))
86   ) ON COLUMNS,
87 NON EMPTY
88 { [Dim County].[State Name].[Hawaii],
89   [Dim County].[State Name].[Alaska],
90   [Dim County].[State Name].[Illinois],
91   [Dim County].[State Name].[Delaware] }
92 * FILTER(
93   {[Dim County].[County].MEMBERS},
94   [Dim County].[County].CURRENTMEMBER.NAME <> "All"
95 ) ON ROWS
96 FROM [OLAP]
97 WHERE [Measures].[Trung bình AQI];
```

Kết quả:

		Messages	Results								
		2021	2021	2021	2021	2021	2021	2022	2022	2022	2022
		All	1	2	3	4	All	All	1	2	3
Hawaii	Hawaii	29.2602739726027	34.8888888888889	24.032967032967	20.7391304347826	37.445652173913	29.2988826815642	35.22891			
Hawaii	Honolulu	28.1671232876712	32.9111111111111	27.2967032967033	23.8369565217391	28.7173913043478	28.027397260274	32.35555			
Hawaii	Kauai	17.0906593406593	19.7	17.1538461538462	18.9021739130435	12.6153846153846	20.7555555555556	20.75555			
Hawaii	Maui	21.2617079899807	26.0777777777778	20	19.5760869565217	19.4444444444444	18.4459833795014	18.07777			
Alaska	Aleutians East	15.0169491525424	17.5	13.7096774193548	13.6	15.3333333333333	34.9173553719008	35			
Alaska	Anchorage	32.3424657534247	38.7777777777778	29.8021978021978	23.304347826087	37.5978260869565	32.6054794520548	28.26666			
Alaska	Denali	32.4876033057851	38	35.989010989011	25.4945054945055	30.5274725274725	41.8461538461538	33.66666			
Alaska	Fairbanks North Star	52.4794520547945	74.7888888888889	38.0659340659341	37.5434782608696	59.8478260869565	41.3424657534247	37.41111			
Alaska	Juneau	24.5434173669468	28.1555555555556	17.6470588235294	22.1111111111111	29.7608695652174	20.9473684210526	17.68493			
Alaska	Kenai Peninsula	11.0082644628099	8.93103448275862	13.4838709677419	12.6	8.93548387096774	23.5416666666667	11.13333			
Alaska	Matanuska-Susitna	22.5480225988701	31.2444444444444	13.0697674418605	15.25	30.7325581395349	17.9442815249267	15.65168			
Alaska	North Slope	6.90350877192982	8	7.53333333333333	3.96296296296296	7.93103448275862	8.3	4.53333			
Illinois	Adams	40.3264462809917	39.741935483871	44.1590909090909	40.2282608695652	30.3225806451613	0	0	0		
Illinois	Champaign	45.3808219178082	45.5777777777778	46.989010989011	50.1413043478261	38.8369565217391	41.717868338558	40.47727			

Trực quan:

Chất lượng không khí trung bình theo từng quý của 4 tiểu bang Alaska, Delaware, Hawaii và Illinois



Nhận xét:

Nhìn chung, chất lượng không khí ở các bang có sự thay đổi liên tục trong suốt năm và có xu hướng tăng cao ở vào quý 4. Cụ thể, các tiểu bang không nằm trên phần lãnh thổ lục địa như Hawaii và Alaska có chất lượng không khí tốt hơn với con số trung bình cao nhất được ghi nhận lần lượt là 29 và 37. Xu hướng thay đổi của hai tiểu bang cũng có điểm chung là sẽ giảm ở quý 2, 3 và tăng trở lại vào quý 4 đến quý 1

năm sau. Vào quý 1, Alaska phải đối mặt với tình trạng nghịch nhiệt, giữ chất ô nhiễm ở gần mặt đất. Do ở xa đất liền, khí hậu ôn hòa và nền kinh tế đặc thù mà Hawaii giữ được chất lượng không khí ở mức tốt trong suốt năm, dù ở thời điểm cao nhất vẫn nằm ở mức tốt.

Ngược lại, các tiểu bang thuộc Bắc Mỹ, Delaware và Illinois, xu hướng thay đổi có sự khác biệt khi mức độ ô nhiễm ở mức cao và tiếp tục tăng cho đến quý 3 và bắt đầu suy giảm vào quý 4 và tiếp tục tăng lại khi bắt đầu vào mùa đông năm sau. Vào mùa đông (Q1), hiện tượng nghịch nhiệt làm ảnh hưởng phần lớn các tiểu bang đất liền, kết hợp với lượng khí thải công nghiệp, sưởi ấm và phương tiện giao thông làm tăng mức độ ô nhiễm. Mùa hè (Q3) là thời điểm ô nhiễm nhất đối với không khí. Các tiểu bang này chịu ảnh hưởng bởi ô nhiễm ozone trong thời tiết nắng nóng. Con số trung bình cao nhất được ghi nhận ở Delaware và Illinois lần lượt là 52 và 50, đều thuộc những tháng mùa hè. Quý 4 là nửa cuối mùa thu, thời tiết ôn hòa góp phần làm giảm lượng khí thải ozone trong không khí, không khí vào quý này được cải thiện.

6. Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California.

Sự kiện: Dữ liệu AQI của một hạt trong một bang được ghi nhận trong ngày

Bối cảnh sự kiện:

- Ai: Trạm AQI
- Ở đâu: tiểu bang
- Cái gì: dữ liệu AQI (Chỉ số AQI, phân loại, đơn vị)
- Khi nào: ngày ghi nhận

Đo lường: chỉ số AQI

- Các giá trị có sẵn từ nguồn: AQI
- Các giá trị phải tính toán: min(AQI), max(AQI), mean(AQI), stddev(AQI)

Cấp chi tiết dữ liệu: mỗi dòng trong bảng fact tương ứng với một chỉ số chất lượng không khí của một thang đo cho một hạt của một bang trong một ngày tại một trạm đo lường.

MDX:

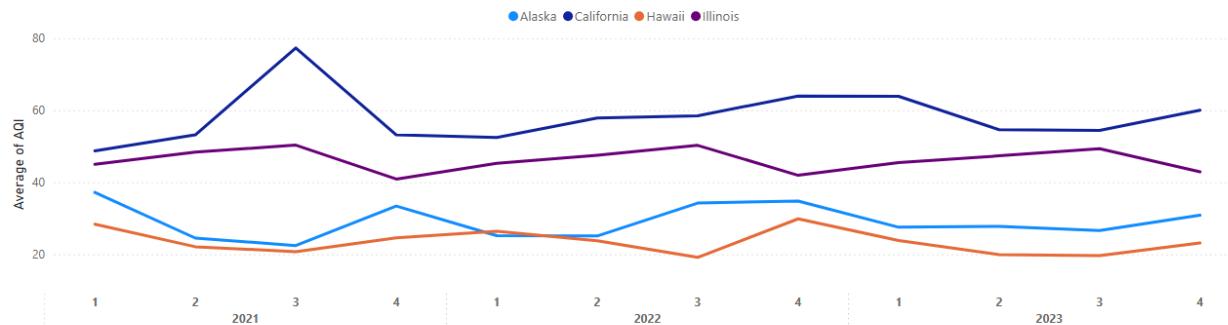
```
99  -- 6. Design a report to demonstrate the AQI fluctuation trends over the year for the four
100 --following states: Hawaii, Alaska, Illinois and California.
101 WITH
102   MEMBER [Measures].[Max_AQI] AS
103     CoalesceEmpty([Measures].[MaxAQI], 0)
104   MEMBER [Measures].[Min_AQI] AS
105     CoalesceEmpty([Measures].[MinAQI], 0)
106   MEMBER [Measures].[Mean_AQI] AS
107     CoalesceEmpty([Measures].[Mean AQI], 0)
108   MEMBER [Measures].[Standard_Deviation] AS
109     CoalesceEmpty([Measures].[Standard Deviation], 0)
110 SELECT
111 NON EMPTY { [Measures].[Max_AQI], [Measures].[Min_AQI], [Measures].[Mean_AQI], [Measures].[Standard_Deviation] } ON COLUMNS,
112 NON EMPTY
113   {
114     [Dim County].[State Name].[Hawaii],
115     [Dim County].[State Name].[Alaska],
116     [Dim County].[State Name].[Illinois],
117     [Dim County].[State Name].[California]
118   }
119 *
120 FILTER(
121   [Dim Date].[Year].[Year].ALLMEMBERS *
122   [Dim Date].[Quarter].[Quarter].ALLMEMBERS,
123   [Dim Date].[Year].CURRENTMEMBER.NAME <> "Unknown" AND
124   [Dim Date].[Quarter].CURRENTMEMBER.NAME <> "Unknown"
125 )
126 DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME
127 ON ROWS
128 FROM [OLAP];
```

Kết quả:

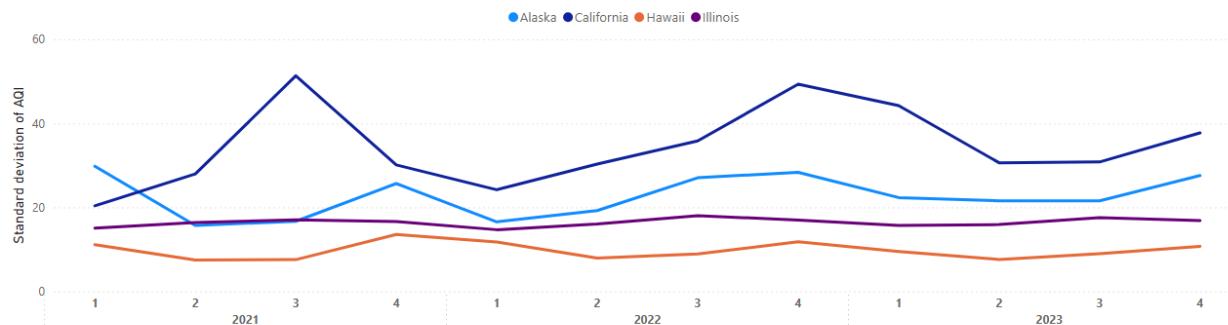
			Max_AQI	Min_AQI	Mean_AQI	Standard_Deviation
Hawaii	2021	1	62	2	28.3944444444444	11.1193011482148
Hawaii	2021	2	47	3	22.1208791208791	7.47903320548083
Hawaii	2021	3	65	4	20.7635869565217	7.55854542044125
Hawaii	2021	4	64	2	24.6164383561644	13.5369915690064
Hawaii	2022	1	62	2	26.4334277620397	11.77680602732
Hawaii	2022	2	48	4	23.7838827838828	7.93246426525791
Hawaii	2022	3	65	2	19.16	8.89554740509901
Hawaii	2022	4	64	6	29.8644688644689	11.8268932609149
Hawaii	2023	1	51	2	23.8674242424242	9.50764442206281
Hawaii	2023	2	44	2	19.9274809160305	7.60012608809744
Hawaii	2023	3	58	2	19.6531365313653	8.9733913407803
Hawaii	2023	4	63	3	23.1822033898305	10.7240355052543
Alaska	2021	1	189	0	37.2346368715084	29.787029151755
Alaska	2021	2	86	0	24.5223880597015	15.713724592401
Alaska	2021	3	183	1	22.4338235294118	16.6869763437919
Alaska	2021	4	157	0	33.41666666666667	25.6661345543688

Trực quan:

Chất lượng không khí trung bình theo từng quý ở từng tiểu bang



Chất lượng không khí trung bình theo từng quý ở từng tiểu bang



Nhận xét:

Nhìn vào biểu đồ thể hiện giá trị chất lượng không khí trung bình, hai tiểu bang nằm trong lục địa Bắc Mỹ là California và Illinois có xu hướng tăng từ đầu năm đến cuối quý ba và bắt đầu giảm về cuối năm và tăng lại khi bắt đầu vào mùa đông (cuối quý 4) trong khi hai tiểu bang nằm xa vùng lãnh thổ này lại không khí tốt hơn vào quý 2 và 3, sau đó tăng nhẹ ở quý 4 và quý 1 năm sau. Ngoài ra ta có thể thấy tiểu bang có chất lượng không khí tệ nhất là California, trong khi đó Hawaii và Alaska lại có không khí tốt hơn. Để giải thích cho xu hướng này, nguyên nhân có thể là do sự khác biệt về địa lý và nền công nghiệp ở các tiểu bang. Ở Hawaii và Alaska có khí hậu khác biệt so với vùng lục địa và nền kinh tế có mức độ công nghiệp hóa thấp. Ngược lại, ở vùng lục địa, các bang bị ảnh hưởng bởi hiệu ứng thời tiết nghịch nhiệt, khiến cho các chất gây ô nhiễm bị giữ lại ở gần mặt đất, làm tăng mức độ ô nhiễm. Ngoài ra, vào mùa hè (quý 2-3), các bang này còn đối mặt với khả năng cháy rừng cao do những đợt nắng nóng. Đỉnh điểm là ở tiểu bang California vào tháng 7-8 năm 2021, vụ cháy

rừng Dixie đã đẩy mức độ ô nhiễm ở đây lên mức cao nhất do lượng lớn bụi mịn và khói từ đám cháy. Tuy đã kiểm soát được chất lượng không khí ở mức ổn định, Illinois cũng chịu những ảnh hưởng từ những nguy cơ tương tự các tiểu bang khác ở đất liền. Tiểu bang cũng đã nhận một lượng lớn chất gây ô nhiễm từ đám cháy rừng ở Canada vào tháng 6/2023.

Về sự biến động, tiểu bang Illinois lại làm tốt trong việc duy trì chất lượng không khí. Tiểu bang đã thực hiện nhiều hành động nhằm bảo vệ môi trường ở đây, như giảm lượng khí thải công nghiệp và phương tiện cá nhân, đưa xe điện vào vận hành. Ngược lại, California lại rất nhạy cảm đối với những tác động bên ngoài, khiến cho tốc độ ô nhiễm tăng rất nhanh ở nửa cuối năm. Diện tích rừng ở California cũng rất lớn, khiến những tháng mùa hè là thời điểm nhạy cảm với cháy rừng, làm cho chất lượng không khí có xu hướng tăng vọt.

7. Build graphs/charts for the above reports.

Biểu đồ, đồ thị đã được chèn chung vào các câu trên.

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

Sự kiện: Dữ liệu AQI của một hạt trong một bang được ghi nhận trong ngày

Bối cảnh sự kiện:

- Ai: Trạm AQI
- Ở đâu: hạt, tiểu bang
- Cái gì: dữ liệu AQI (Chỉ số AQI, phân loại, đơn vị)
- Khi nào: ngày ghi nhận

Đo lường: chỉ số AQI

- Các giá trị có sẵn từ nguồn: AQI
- Các giá trị phải tính toán: Mean(AQI)

Cấp chi tiết dữ liệu: mỗi dòng trong bảng fact tương ứng với một chỉ số chất lượng không khí của một thang đo cho một hạt của một bang trong một ngày tại một trạm đo lường.

MDX:

```

132 --8. Use a regional map to visually represent (by color) the mean AQI value in regions during a year.
133 --Version 1
134 SELECT
135     {[Measures].[Mean AQI]} ON COLUMNS,
136     NON EMPTY
137     CROSSJOIN(
138         [Dim Date].[Hierarchy Date].[Year].&[2023],           -- Year 2023
139         [Dim Date].[Month].[Month].Members,                  -- All Months
140         [Dim County].[State Name].[State Name].Members      -- All States
141     ) ON ROWS
142 FROM [OLAP];
143
144 --Version 2
145 WITH
146     Member [Measures].[Trung binh AQI] AS
147     CoalesceEmpty([Measures].[Mean AQI],0) --Doi ten
148     SELECT
149         FILTER ({
150             [Dim Date].[Hierarchy Date].[Year].&[2023] *
151             [Dim Date].[Quarter].Members
152         }, ([Dim Date].[Year].CURRENTMEMBER.NAME <> "Unknown" AND [Dim Date].[Quarter].CURRENTMEMBER.NAME <> "Unknown"))
153     ) ON COLUMNS,
154     NON EMPTY [Dim County].[state Name].[State Name].Members ON ROWS
155     FROM [OLAP]
156 WHERE [Measures].[Trung binh AQI];

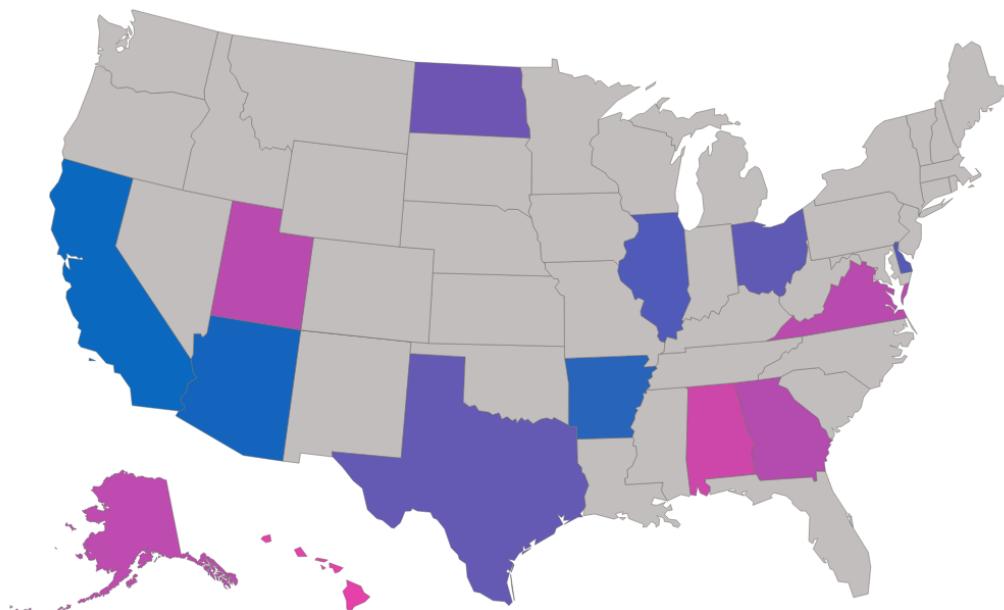
```

Kết quả:

	2023	2023	2023	2023	2023
	All	1	2	3	4
Alabama	25.8772378516624	19.7111111111111	31.6521739130435	28.6421052631579	23.780701754386
Alaska	28.2024965325936	27.54166666666667	27.7537037037037	26.6442831215971	30.8676470588235
Arizona	56.4197301070265	58.8664799253035	59.3780260707635	55.9056776556777	51.4844486333647
Arkansas	53.625	56.063829787234	51.7294117647059	52.7558139534884	55.0348837209302
California	58.2009722741599	63.8869281045752	54.6273675250053	54.4055649241147	60.0574050362783
Colorado	0	0	0	0	0
Connecticut	0	0	0	0	0
Delaware	45.7735618115055	44.4156378600823	44.8461538461538	46.2244897959184	48.0819672131148
Georgia	29.6721698113208	28.9340659340659	24.8865979381443	31.6224489795918	32.1376811594203
Hawaii	21.6060019361084	23.8674242424242	19.9274809160305	19.6531365313653	23.1822033898305
Idaho	0	0	0	0	0
Illinois	46.3605207226355	45.4904601571268	47.3662044170519	49.3326572008114	42.9293924466338
Indiana	0	0	0	0	0
North Dakota	41.2381516587678	37.5748031496063	40.810551558753	43.0314606741573	42.9820224719101
Ohio	43.6503264789007	43.1896625373772	43.6233956729006	44.0460057061341	43.6680619506069

Trực quan:

Trung bình chất lượng không khí năm 2023 theo tiểu bang



Trung bình chất lượng không khí tăng dần theo thang màu từ hồng => tím => xanh.

Nhận xét:

Nhìn chung, trong năm 2023, các bang được khảo sát có mức độ AQI trung bình khá tốt. Tuy nhiên, vẫn có các bang có mức độ AQI trung bình cao hơn các bang còn lại, như California, Arizona, Arkansas. Các bang có mức độ AQI trung bình thấp như Alaska, Hawaii, Alabama.

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

Sự kiện: Dữ liệu AQI của một hạt trong một bang được ghi nhận trong ngày

Bối cảnh sự kiện:

- Ai: Trạm AQI
- Ở đâu: hạt, tiểu bang
- Cái gì: dữ liệu AQI (Chỉ số AQI, phân loại, đơn vị)
- Khi nào: ngày ghi nhận

Đo lường: chỉ số AQI

- Các giá trị có sẵn từ nguồn: AQI
- Các giá trị phải tính toán: min(AQI), max(AQI), mean(AQI), stddev(AQI)

Cấp chi tiết dữ liệu: mỗi dòng trong bảng fact tương ứng với một chỉ số chất lượng không khí của một thang đo cho một hạt của một bang trong một ngày tại một trạm đo lường.

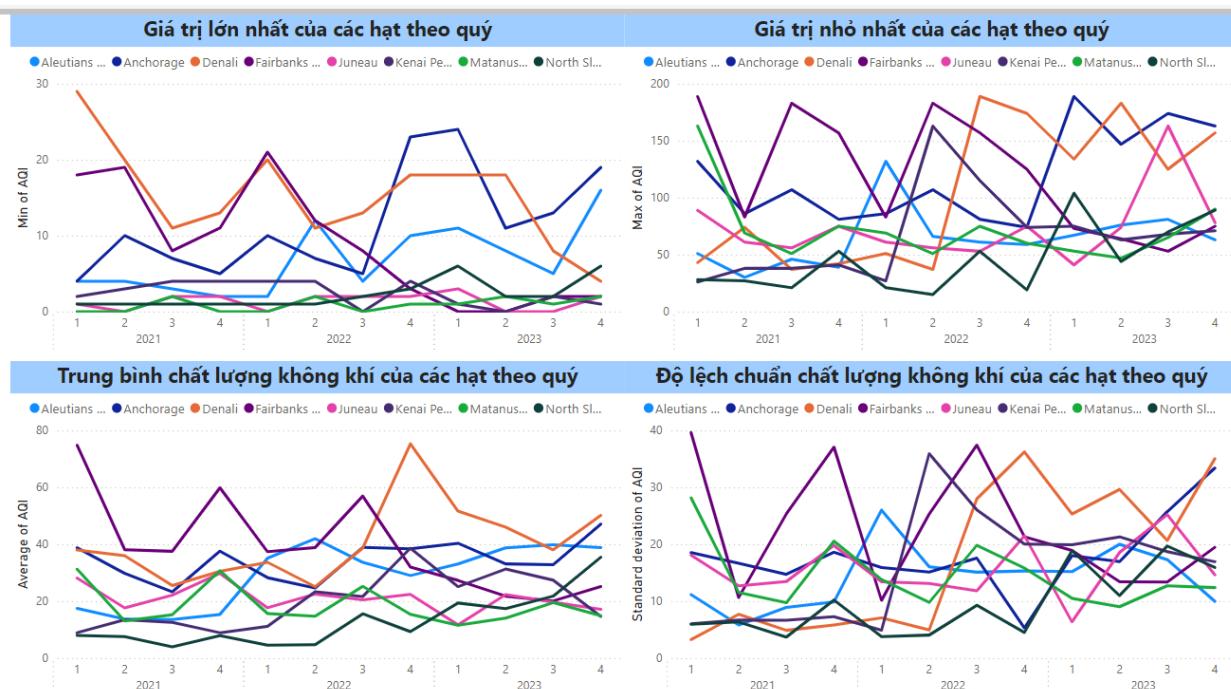
MDX:

```
158 -- 9. Report the mean, the standard deviation, min and max of AQI value group by
159 -- State and County during each quarter of the year.
160 WITH
161 MEMBER [Measures].[Max_AQI] AS
162     CoalesceEmpty([Measures].[MaxAQI], 0)
163 MEMBER [Measures].[Min_AQI] AS
164     CoalesceEmpty([Measures].[MinAQI], 0)
165 MEMBER [Measures].[Mean_AQI] AS
166     CoalesceEmpty([Measures].[Mean AQI], 0)
167 MEMBER [Measures].[Standard_Deviation] AS
168     CoalesceEmpty([Measures].[Standard Deviation], 0)
169 SELECT
170 NON EMPTY { [Measures].[Max_AQI], [Measures].[Min_AQI], [Measures].[Mean_AQI], [Measures].[Standard_Deviation] } ON COLUMNS,
171 NON EMPTY {
172     FILTER([Dim County].[State Name].[State Name] * [Dim County].[County].[County] * [Dim Date].[Year].[Year].ALLMEMBERS * [Dim Date].[Quarter].[Quarter],
173     [Dim County].[State Name].CURRENTMEMBER.NAME <> "Unknown" AND
174     [Dim Date].[Year].CURRENTMEMBER.NAME <> "Unknown" AND
175     [Dim Date].[Quarter].CURRENTMEMBER.NAME <> "Unknown" )
176 DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [OLAP];
```

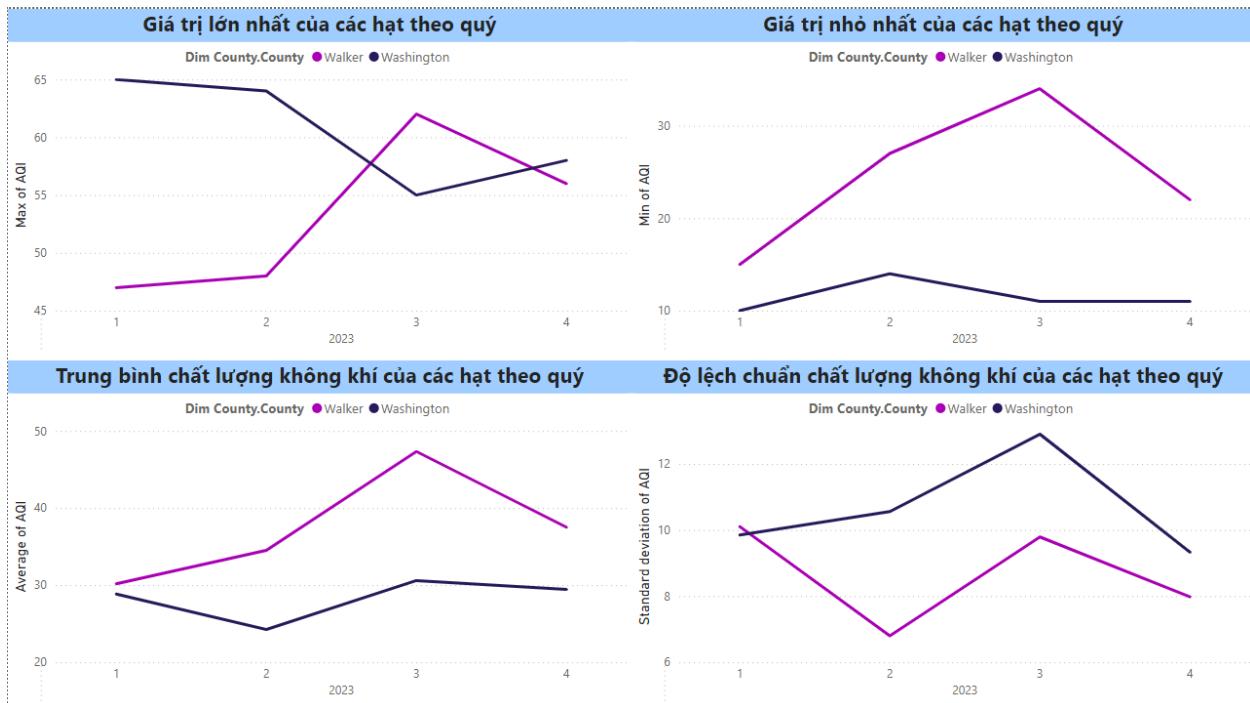
Kết quả:

			Max_AQI	Min_AQI	Mean_AQI	Standard_Deviation
Alabama	Sumter	2021	1	0	0	0
Alabama	Sumter	2021	2	0	0	0
Alabama	Sumter	2021	3	0	0	0
Alabama	Sumter	2021	4	0	0	0
Alabama	Sumter	2022	1	0	0	0
Alabama	Sumter	2022	2	0	0	0
Alabama	Sumter	2022	3	0	0	0
Alabama	Sumter	2022	4	0	0	0
Alabama	Sumter	2023	1	0	0	0
Alabama	Sumter	2023	2	7	7	0
Alabama	Sumter	2023	3	11	4	7.666666666666667
Alabama	Sumter	2023	4	51	4	18.7272727272727
Alabama	Tuscaloosa	2021	1	0	0	0
Alabama	Tuscaloosa	2021	2	0	0	0
Alabama	Tuscaloosa	2021	3	0	0	0
Alabama	Tuscaloosa	2021	4	0	0	0

Trực quan:



Tiểu bang Georgia:



Nhận xét cho tiểu bang Georgia:

Nhìn chung, Walker có chất lượng không khí tệ hơn Washington trong suốt năm. Hai hạt thuộc tiểu bang đều có sự thay đổi chất lượng không khí đáng kể trong quý 3. Cả hai tiểu bang đều ghi nhận được mức độ ô nhiễm trung bình cao nhất trong năm vào quý này ở Walker và Washington lần lượt là 47 và 30. Mức ô nhiễm cao nhất ở Walker là vào quý 3 với chỉ số 62 trong khi cao nhất ở Washington là 65 vào quý 1. Cả hai đều ghi nhận mức thấp nhất của mình là vào quý 1 với Walker là 15 và Washington là 10. Tốc độ thay đổi ở Walker chậm lại là vào quý 2 sau đó nhanh trở lại vào quý 3. Ngược lại, Washington lại cho thấy tốc độ gia tăng nồng độ chất gây ô nhiễm tăng dần trong năm và chậm lại vào cuối năm.

10. Report the mean AQI value by State, Category, DayLightSaving over years.

Cài đặt thêm cột DayLightSaving trong Dim_Date ở DDS:

```
-- Added for Data Analysis requirements
ALTER TABLE PROJECT_DDS.dbo.Dim_Date
ADD DayLightSaving AS
CASE
    WHEN Full_Date BETWEEN '2023-03-12' AND '2023-11-05' THEN CAST(1 AS BIT)
    ELSE CAST(0 AS BIT)
END;
```

Thêm DayLightSaving vào một thuộc tính của Dim_Date ở SSAS OLAP Cube:

The screenshot shows the SSAS Dimension Designer interface with the following details:

- OLAP cube [Design]***: The current cube being edited.
- Dim Date.dim [Design]**: The dimension being modified.
- Dim County.dim [Design]**: Another dimension listed in the tabs.
- OLAP.dsv [Design]**: A data source view tab.
- Dimension Structure**: The active tab in the ribbon.
- Attribute Relationships**: A button in the ribbon.
- Translations**: A button in the ribbon.
- Browser**: A button in the ribbon.
- Attributes** pane: Shows the attributes of the Dim Date dimension, including Date_SK, Day, Day Light Saving, Month, Quarter, and Year.
- Hierarchies** pane: Displays the "Hierarchy Date" structure:
 - Year
 - Quarter
 - Month
 - Day
 - <new level>A tooltip states: "To create a new hierarchy, drag an attribute here."
- Data Source View** pane: Shows the structure of the Dim Date table, listing columns: Date_SK, Full_Date, Year, Quarter, Month, Day, Created, Last_Updated, Source_SK, and DayLightSaving.

Sự kiện: Khi trạm ghi nhận được AQI trong một ngày Daylight Saving

Bối cảnh:

- Ai: Trạm.
- Ở đâu: Các bang (State)
- Cái gì: Chỉ số chất lượng không khí (AQI)
- Khi nào: Mỗi ngày quan sát.

Đo lường:

- Giá trị có sẵn: Giá trị AQI quan sát được tại các bang, Loại chất lượng không khí, trạng thái DayLight Saving (True hoặc False).
- Giá trị cần tính toán: Chỉ số AQI trung bình của từng loại chất lượng không khí trong ngày DayLight Saving.

$$\text{MeanAQI} = \text{AVG(AQI)} \text{ GROUPBY (Category, DayLightSaving = 1)}$$

Độ mịn:

Một dòng trong Fact tương ứng với một quan sát AQI trong một ngày tại một bang, với thông tin về loại chất lượng không khí và trạng thái DayLightSaving.

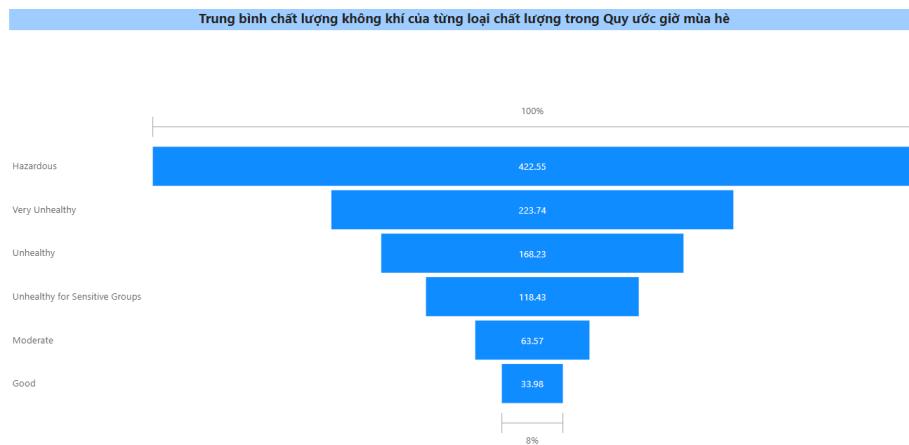
MDX:

```
199 --Version 2
200 WITH
201 MEMBER [Measures].[Mean_AQI] AS
202     CoalesceEmpty([Measures].[Mean AQI], 0) -- Null to 0
203 SELECT
204     NON EMPTY
205         FILTER(
206             [Dim Category].[Category].[Category].MEMBERS,
207             [Dim Category].[Category].CURRENTMEMBER.NAME <> "Unknown")
208     ON COLUMNS, -- Categories as columns
209     NON EMPTY
210         FILTER(
211             [Dim County].[State Name].[State Name].MEMBERS,
212             [Dim County].[State Name].CURRENTMEMBER.NAME <> "Unknown"
213         ) ON ROWS -- States as rows
214     FROM [OLAP]
215     WHERE (
216         [Measures].[Mean_AQI],
217         [Dim Date].[Day Light Saving].[TRUE] -- Filter for Daylight Saving Time
218     );
```

Kết quả:

	Messages	Results					
		Good	Hazardous	Moderate	Unhealthy	Unhealthy for Sensitive Groups	Very Unhealthy
Alabama	22.7378640776699	0	57.6842105263158	0	0	0	0
Alaska	21.6256983240223	0	61.6447368421053	165.714285714286	119.1666666666667	0	0
Arizona	38.1296296296296	313	66.6716981132075	177.315789473684	117.944	214.9111111111111	0
Arkansas	39.4782608695652	0	61.0894308943089	151	111.285714285714	0	0
California	36.6694148076301	427.761904761905	66.215667311412	167.271604938272	119.949843260188	235.852941176471	0
Colorado	0	0	0	0	0	0	0
Connecticut	0	0	0	0	0	0	0
Delaware	38.5672514619883	0	59.7431693989071	0	115.75	0	0
Georgia	25.9633865546218	0	54.3809523809524	0	0	0	0
Hawaii	20.4776334776335	0	58	0	0	0	0
Idaho	0	0	0	0	0	0	0
Illinois	37.0466061933062	0	61.9019189765458	156	113.326086956522	0	0
Indiana	0	0	0	0	0	0	0
North Dakota	30.5	0	62.2853470437018	0	103.75	0	0
Ohio	34.939151813153	0	61.5634167385677	0	110.285714285714	0	0
Oklahoma	0	0	0	0	0	0	0

Trực quan:



Nhận xét:

Biểu đồ thể hiện trung bình chất lượng không khí (AQI) theo từng loại chất lượng (Category) trong mùa hè, có các nhận xét sau:

- Giá trị trung bình AQI của nhóm "Hazardous" là **422.55**, vượt xa các mức khác. Cho thấy trong những ngày DayLight Saving thì mức độ ô nhiễm vô cùng cao, chất lượng không khí vô cùng thấp, cực kỳ nguy hiểm cho người dân.
- Biểu đồ có xu hướng giảm đều khi từ các nhóm nguy hiểm về nhóm tốt hơn. Cho thấy chất lượng không khí an toàn **cực kì thấp** so với sự ô nhiễm.
- Nhóm "Good" có giá trị trung bình AQI chỉ **33.98**, tức là chỉ có một phần nhỏ khu vực hoặc thời gian đạt tiêu chuẩn không khí tốt.
- Sự chênh lệch lớn giữa các mức, đặc biệt giữa "Moderate" (**63.57**) và "Unhealthy for Sensitive Groups" (**118.43**), cho thấy sự chênh lệch về chất lượng không khí là rất đáng kể.

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

Sự kiện: Khi trạm ghi nhận được AQI của các Bang.

Bối cảnh:

- Ai: Trạm.
- Ở đâu: Bang (State)
- Cái gì: Chỉ số chất lượng không khí (AQI)
- Khi nào: Mỗi ngày quan sát. Month > Day

Đo lường:

- Giá trị có sẵn: Từng ngày được ghi nhận AQI, Loại chất lượng không khí, các Bang, Tháng
- Giá trị cần tính toán: Số ngày được ghi nhận trong từng loại chất lượng không khí.
Số ngày = COUNT(DISTINCT Date_SK) GROUPBY(Month, Category, State)

Độ mịn:

Mỗi dòng dữ liệu trong fact tương ứng với một quan sát AQI của một ngày trong một tháng tại một bang.

MDX:

```
220  -- 11. Count the number of days by State, Category in each month
221  WITH
222    MEMBER [Measures].[CoalescedMeasure] AS
223      CoalesceEmpty([Measures].[NumberOfDay], 0) -- Null to 0
224  SELECT
225    NON EMPTY FILTER({[Dim Date].[Year].[Year].MEMBERS * [Dim Date].[Quarter].[Quarter] * [Dim Date].[Month].[Month].MEMBERS},
226      [Dim Date].[Year].CURRENTMEMBER.Name <> "Unknown" AND
227      [Dim Date].[Quarter].CURRENTMEMBER.Name <> "Unknown" AND
228      [Dim Date].[Month].CURRENTMEMBER.Name <> "Unknown"
229    ) ON COLUMNS,
230    NON EMPTY
231    CROSSJOIN(
232      FILTER(
233        [dim County].[State Name].MEMBERS,
234        [Dim County].[State Name].CURRENTMEMBER.Name <> "All" AND [Dim County].[State Name].CURRENTMEMBER.Name <> "Unknown"
235      ),
236      FILTER(
237        [dim Category].[Category].MEMBERS,
238        [Dim Category].[Category].CURRENTMEMBER.Name <> "All" AND [Dim Category].[Category].CURRENTMEMBER.Name <> "Unknown"
239      )
240    ) ON ROWS
241  FROM [OLAP]
242  WHERE [Measures].[CoalescedMeasure];
```

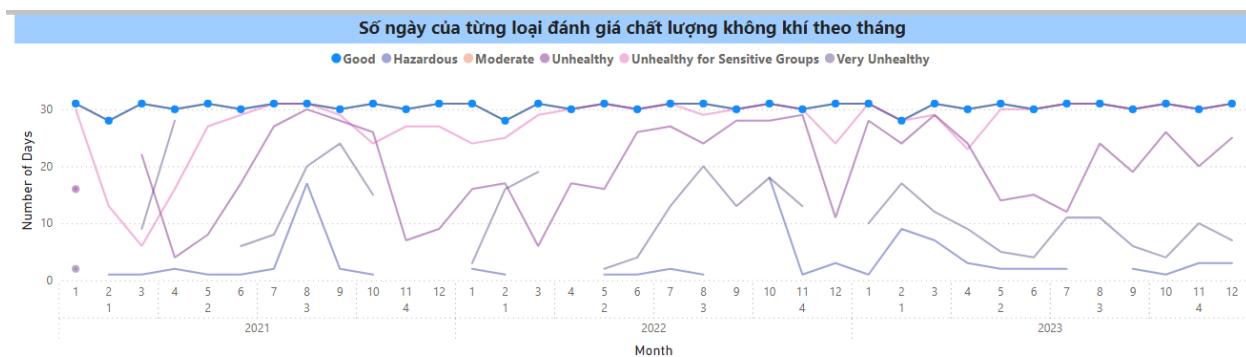
Kết quả:

		2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2021	2022	2022	2022	2022	
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5
Alabama	Good	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alabama	Hazardous	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alabama	Moderate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alabama	Unhealthy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alabama	Unhealthy for Sensitive Groups	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alabama	Very Unhealthy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alaska	Good	31	28	31	30	31	30	31	31	30	31	31	31	28	31	30	31	31
Alaska	Hazardous	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alaska	Moderate	25	21	20	17	5	7	15	5	3	15	25	27	16	7	14	18	11
Alaska	Unhealthy	7	0	0	0	0	0	1	0	0	0	2	3	0	0	0	1	0
Alaska	Unhealthy for Sensitive Groups	11	5	0	0	0	0	1	0	1	1	2	4	0	1	0	1	0
Alaska	Very Unhealthy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Arizona	Good	31	28	31	30	31	30	31	31	30	31	30	31	28	31	30	31	31

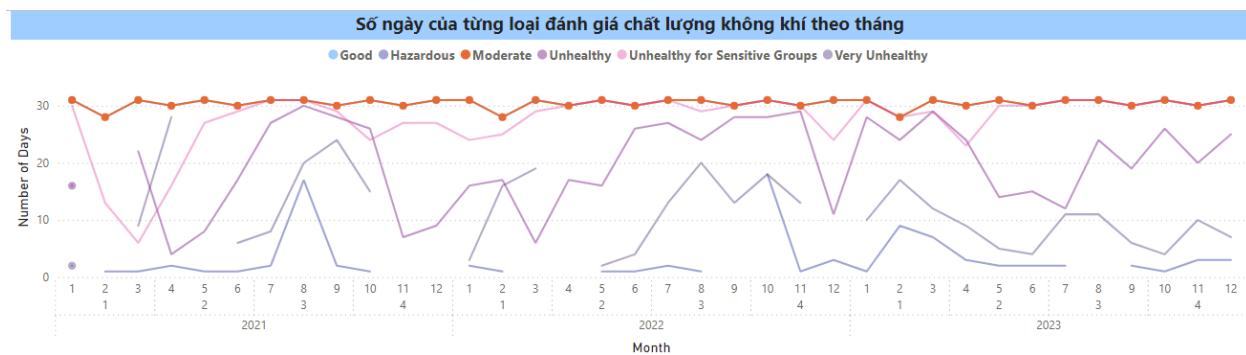
Trực quan:



Good:



Moderate:



Nhận xét:

Ta nhận thấy rằng:

- Số ngày đo được chất lượng không khí “Good” qua các năm là có xu hướng khá ổn định. Nhưng có sự giảm nhẹ vào giữa mùa xuân (tháng 2) và nhanh chóng quay trở lại, có thể do đây là thời điểm tốt để mọi người đi ra ngoài, góp phần giảm chất lượng không khí so với ngày thường. Tuy nhiên hầu hết thì số ngày “Good” lại luôn trên cao, cho thấy không khí ở nhiều nơi vẫn được duy trì ở mức rất tốt.
- Chất lượng không khí “Moderate” thì không có sự khác biệt nhiều với “Good” (Cả 2 đè lên lẫn nhau trên biểu đồ). Tương tự “Good” thì “Moderate” không có sự thay đổi nhiều qua các tháng và có sự giảm nhẹ vào giữa mùa Xuân (Tháng 2) và quay trở lại như cũ.
- Ở “Unhealthy” thì có sự biến động vô cùng lớn. Trong 2 năm 2021 và 2022 thì có xu hướng giảm vào cuối đông (tháng 11-12) sau đó duy trì nhẹ cho tới đầu mùa hè (tháng 4) thì tăng lên mạnh mẽ. Tuy nhiên vào 2023 thì biểu đồ cho thấy việc tăng lại đỉnh gần như là lập tức vào đầu xuân (tháng 1) sau khi giảm mạnh vào cuối đông theo chu kỳ cũ. Điều này cho thấy là số ngày ô nhiễm đang càng ngang càng tăng cao một cách vượt bậc, đặc biệt là khi nền kinh tế đã quay lại hoạt động ổn định sau khi kết thúc mùa dịch Covid.
- Tuy nhiên, nguy hiểm hơn là “Unhealthy for sensitive group” khi chỉ giảm mạnh vào mùa xuân 2021 thi tăng lên mạnh mẽ và liên tục duy trì ở trên cao. Có thể covid đã giúp việc giảm ô nhiễm trong nhất thời xong khi covid đi qua, không khí đã ngay lập quay trở lại ô nhiễm đáng kể để ảnh hưởng tới những nhóm người nhạy cảm như người già, trẻ em,....
- Với “Very Unhealthy” thì nó có xu hướng cực kì khó đoán vì có những thời điểm hoàn toàn biến mất xong xuất hiện trở lại một cách mạnh mẽ. Và sự thay đổi bất thường này là không hề rõ ràng. Ta chỉ nhận ra được một ít như nó sẽ gia tăng mạnh vào tháng 7. Và thời gian mà “Very unhealthy” tồn tại gia tăng dần qua các năm khi trong 2021 thì chỉ tồn tại 7/12 tháng thì tới 2022 là 10/12 tháng, còn 2023 thì đã tồn tại 6/6 tháng. Cho thấy rằng không khí đang trở nên dần trở nên ô nhiễm nặng hơn, và có thể trong tương lai thì sẽ trở nên

- Cuối cùng là về “Hazardous”, ta thấy đa số thời gian thì nó đều nằm ở cuối biểu đồ, cho thấy sự xuất hiện vô cùng ít ỏi của ngày đó được “Hazardous”. Chỉ có 2 khoảng thời gian tăng mạnh vào 8/2021,10/2022 và gia tăng nhẹ vào 2/2023.

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

Sự kiện: Khi dữ liệu AQI được ghi nhận

Bối cảnh sự kiện:

- Ai: Trạm AQI
- Ở đâu: hạt, tiểu bang
- Cái gì: dữ liệu AQI (Chỉ số AQI, phân loại, đơn vị)
- Khi nào: ngày ghi nhận

Đo lường: Số ngày ứng với Category và Defining Parameter

- Các giá trị có sẵn từ nguồn: AQI
- Các giá trị phải tính toán: Count Distinct (Date)

Cấp chi tiết dữ liệu: mỗi dòng trong bảng fact ứng với số ngày mà Category và Defining Parameter được ghi nhận

MDX:

```
-- 244 -- 12. Report the number of days by Category and Defining Parameter.
245 SELECT
246     [Measures].[NumberOfDay] ON COLUMNS,
247     NON EMPTY
248     (
249         [Dim Category].[Category].[Category].MEMBERS *
250         [Dim Parameter].[Defining Parameter].[Defining Parameter].MEMBERS
251     ) ON ROWS
252     FROM [OLAP];
253
```

Kết quả:

Messages Results		
		NumberOfDay
Good	CO	207
Good	NO2	1070
Good	Ozone	1095
Good	PM10	1091
Good	PM2.5	1095
Hazardous	Ozone	19
Hazardous	PM10	13
Hazardous	PM2.5	66
Moderate	CO	10
Moderate	NO2	502
Moderate	Ozone	986
Moderate	PM10	857
Moderate	PM2.5	1095
Unhealthy	NO2	7
Unhealthy	Ozone	507
Unhealthy	PM10	72

Trực quan:

Category	CO	NO2	Ozone	PM10	PM2.5
Good	207	1070	1095	1091	1095
Hazardous			19	13	66
Moderate	10	502	986	857	1095
Unhealthy		7	507	72	377
Unhealthy for Sensitive Groups		15	792	195	677
Very Unhealthy			225	34	114

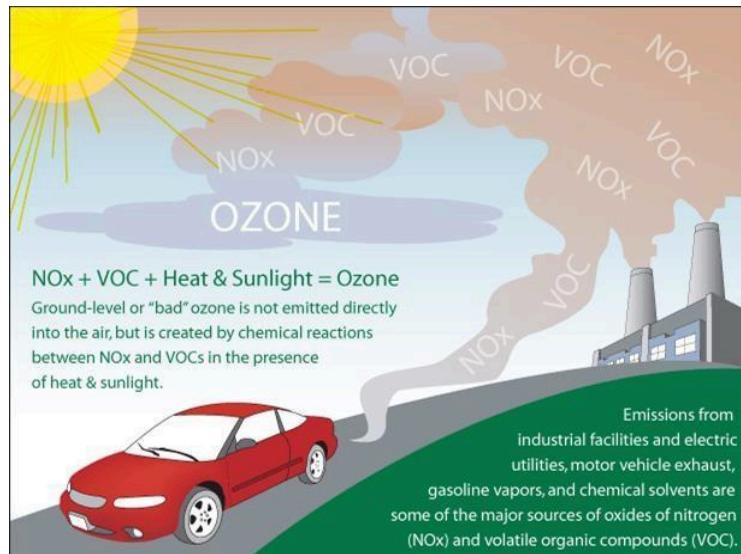
Nhận xét:

Dựa vào heatmap sau, ta có thể thấy tình hình ô nhiễm không khí trong toàn nước Mỹ. Trong không khí thường có các chất gây ô nhiễm như: CO, NO2, Ozone (dù bảo vệ Trái Đất ở tầng khí quyển, đây là một khí độc có hại cho sức khỏe con người nếu tiếp xúc trực tiếp), PM10, PM2.5.

Đối với AQI loại trung bình (moderate), không tốt cho sức khỏe (Unhealthy, very unhealthy), không khí bị chịu ảnh hưởng nhiều bởi Ozone, PM10, PM2.5.

Tuy nhiên, đối với độc hại (Hazardous), thì phần lớn chịu tác động bởi PM10 và PM2.5

=> Như vậy, 3 chất ô nhiễm phổ biến: Ozone, PM10, PM2.5



PM10, PM2.5, Ozone thường được sản sinh ra do hoạt động các nhà máy, sử dụng xe cá nhân với nhiên liệu hóa thạch.

VI. Mining

1. Mô hình ARIMA

a. Tổng quan về ARIMA

ARIMA là một mô hình thống kê được sử dụng để phân tích và dự đoán dữ liệu chuỗi thời gian không dừng. Mô hình này kết hợp ba thành phần chính:

- AR - Autoregressive: Liên quan đến việc dự đoán giá trị hiện tại của chuỗi dựa trên các giá trị quá khứ.
- I - Integrated: Sử dụng phương pháp lấy sai phân để biến chuỗi không dừng thành chuỗi dừng, giúp phân tích dễ dàng hơn.
- MA - Moving Average: Mô hình hóa mối quan hệ giữa chuỗi thời gian và các sai số dự báo trong quá khứ.

Các tham số trong ARIMA:

- p (số bậc của AR): Xác định số lượng giá trị quá khứ cần sử dụng trong mô hình tự hồi quy.

- d (số lần lấp sai phân): Xác định số lần cần lấp sai phân để chuỗi trở thành chuỗi dừng.
- q (số bậc của MA): Xác định số lượng sai số dự báo quá khứ cần sử dụng trong mô hình Moving Average.
- Ví dụ: ARIMA (1, 1, 1): sử dụng một bậc tự hồi quy (AR), áp dụng lấp sai phân bậc một (I) để làm chuỗi dừng và sử dụng một bậc trung bình trượt (MA).
 - b. Tính dừng:

Tính dừng là một đặc tính quan trọng khi phân tích chuỗi thời gian. Một chuỗi thời gian được gọi là dừng nếu các thuộc tính thống kê của nó, như trung bình, phương sai và tự tương quan, không thay đổi theo thời gian.

Đặc điểm của chuỗi dừng:

- Trung bình không đổi: Giá trị trung bình của chuỗi không thay đổi theo thời gian.
- Phương sai không đổi: Độ biến thiên của chuỗi không thay đổi theo thời gian.
- Tự tương quan không đổi: Sự tương quan giữa các giá trị tại các thời điểm khác nhau chỉ phụ thuộc vào độ trễ (lag) chứ không phụ thuộc vào thời gian.
- c. Lựa chọn tham số ARIMA (p, d, q)

Tự tương quan (Auto Correlation Function) là một khái niệm quan trọng trong phân tích chuỗi thời gian. Nó biểu thị mối quan hệ giữa các giá trị trong chuỗi tại các thời điểm khác nhau. Thông thường, các giá trị gần nhau trong chuỗi có xu hướng tương quan mạnh hơn, hoặc các giá trị thuộc cùng một chu kỳ của chuỗi (ví dụ: cùng tháng trong năm hoặc cùng quý) sẽ có mức độ tương quan cao. Đây chính là lý do thuật ngữ "tự tương quan" được sử dụng.

Hệ số tự tương quan, viết tắt là ACF, thường được dùng để xác định độ trễ trong các mô hình như trung bình trượt MV(q) và hỗ trợ xây dựng các mô hình dự báo phổ biến như ARIMA, GARCH, ARIMAX,... Bên cạnh đó, ACF còn được sử dụng để kiểm tra tính mùa vụ trong chuỗi thời gian.

Hệ số tự tương quan bậc kkk được tính toán theo công thức sau:

$$\rho(s, t) = \frac{cov(x_s, x_t)}{\sqrt{\sigma_s \sigma_t}}$$

Giá trị $p(s, t)$ đo lường khả năng dự báo của biến x_t nếu chỉ sử dụng biến x_s . Trong trường hợp 2 đại lượng có tương quan hoàn hảo tức $p(s, t) = 1$ hoặc -1 ta có thể biểu diễn $x_t = \beta_0 + \beta_1 x_s$. Hệ số của β_1 sẽ ảnh hưởng lên chiều của hệ số tương quan. Theo đó, $p(s, t)$ bằng 1 nếu $\beta_1 > 0$ và bằng -1 nếu $\beta_1 < 0$.

Nhìn chung bậc q không nên quá lớn, thông thường q tối đa là 5.

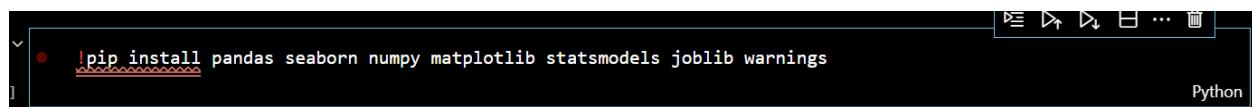
Tương quan riêng phần (Partial AutoCorrelation Function), tương tự như ACF, là một chỉ số dùng để đo lường mức độ tương quan trong chuỗi thời gian. Điểm khác biệt chính là PACF loại bỏ tác động của các chuỗi độ trễ trung gian, giúp đánh giá mối tương quan trực tiếp giữa một giá trị hiện tại và một giá trị tại một độ trễ cụ thể.

Quá trình tính PACF bao gồm việc xây dựng một phương trình hồi quy tuyến tính, trong đó giá trị hiện tại được dự đoán dựa trên các giá trị tại các độ trễ trước đó. Từ đó, PACF được tính bằng cách loại bỏ ảnh hưởng của các độ trễ trung gian thông qua việc trừ đi giá trị ước lượng từ phương trình hồi quy.

PACF thường được sử dụng để xác định bậc tự do p của thành phần tự hồi quy (AR) trong mô hình AR(p). Thông thường, giá trị của p nhỏ hơn hoặc bằng 5. Do đó, người dùng có thể chọn bậc tự do của PACF là một giá trị từ 1 đến 5 dựa trên phân tích cụ thể.

2. Mining data

Tải các thư viện cần thiết:



```
• pip install pandas seaborn numpy matplotlib statsmodels joblib warnings
```

Tiếp theo là khai báo dữ liệu và đọc dữ liệu từ file excel, in ra các bảng, kiểm tra xem dữ liệu đã đầy đủ hay chưa:

```

import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.stattools import adfuller
from statsmodels.tools.eval_measures import aic, bic
import joblib
import warnings

warnings.filterwarnings('ignore')

```

Python

```

df = pd.read_excel('../datasets/data.xlsx', sheet_name=None)
print(df.keys())

```

Python

```

dict_keys(['Dim_Category', 'Dim_County', 'Dim_Date', 'Dim_Parameter', 'Dim_Site', 'Dim_State', 'Fact_AQI'])

```

Nhóm quyết định lựa chọn việc dự đoán dựa trên từng state khác nhau. Nên cần phải gộp các bảng dữ liệu lại và lựa ra các cột dữ liệu cần thiết.

```

# get sheet
df_aqi = df['Fact_AQI']
df_county = df['Dim_County']
df_date = df['Dim_Date']
df_state = df['Dim_State']
# merge sheet

df_merged = pd.merge(df_aqi, df_county[['County_SK', 'County', 'State_SK']], on='County_SK', how='left')
df_merged = pd.merge(df_merged, df_date[['Date_SK', 'Full_Date']], on='Date_SK', how='left')
df_merged = pd.merge(df_merged, df_state[['State_SK', 'State_Name']], on='State_SK', how='left')

# format
df_merged['Full_Date'] = pd.to_datetime(df_merged['Full_Date'])
df_merged = df_merged[['State_Name', 'Full_Date', 'AQI']]

print(df_merged.head(5))

```

Python

	State_Name	Full_Date	AQI
0	Alaska	2021-01-01	7
1	Alaska	2021-01-04	8
2	Alaska	2021-01-07	11
3	Alaska	2021-01-10	4
4	Alaska	2021-01-13	4

Về chọn 3 chỉ số p,d ,q của ARIMA thì nhóm ban đầu lựa chọn sử dụng auto_arima của pmdarima. Tuy nhiên thư viện pmdarima bị lỗi đồng bộ với phiên bản mới nhất của numpy, nên nhóm đã quyết định sử dụng việc tự động tìm chỉ số thông qua AIC/BIC.

Đầu tiên, sẽ kiểm tra tính dừng và thực hiện việc dừng lại cho series dữ liệu, từ đó có thể chọn $d = 1$ trong quá trình train ARIMA.

```
# Kiểm tra tính dừng bằng ADF test
def adf_test(series):
    result = adfuller(series)
    if result[1] > 0.05:
        return False # Không dừng
    else:
        return True # Dừng

# Kiểm tra tính dừng và thực hiện sai phân cho đến khi chuỗi dừng
def make_stationary(series, max_diff=3):
    diff_series = series
    for i in range(max_diff):
        if adf_test(diff_series):
            return diff_series
        diff_series = diff_series.diff().dropna()
    return diff_series # Trả về chuỗi đã sai phân tối đa max_diff lần
```

Tiếp theo, tìm ra chỉ số q , p tốt nhất thông qua quá trình thử và check. Như ở trên ta đã biết rằng q và p sẽ không thể đi quá 5 ($\max = 5$). Nên ta sẽ thử từ 0 đến 5 cho cả 2 chỉ số rồi thử nghiệm liên tục, và sau đó tìm ra chỉ số thích hợp nhất khi so sánh chỉ số AIC và BIC của model:

```
# Kiểm tra AIC và BIC
if model_fit.aic < best_aic:
    best_aic = model_fit.aic
    best_bic = model_fit.bic
    best_order = (p, 1, q)
```

Từ đó chọn được chỉ số p , d , q tốt nhất cho model.

Và cuối cùng ta sẽ train model ARIMA dựa trên kết quả các chỉ số tốt nhất. Vì mỗi state gần như là riêng biệt và không ảnh hưởng nhau trong quá trình train model, nên em quyết định tạo ra 18 model cho 18 state rồi lưu lại trong folder models:

```

# Hàm huấn luyện mô hình ARIMA và lưu mô hình
def train_and_save_arima_model(county_name, aqi_series):
    # Tìm tham số tốt nhất (p, d, q) tự động
    p, d, q = find_best_arima_params(aqi_series)
    print(p, d, q)
    # Huấn luyện mô hình ARIMA với tham số tốt nhất
    model = ARIMA(aqi_series, order=(p, d, q))
    model_fit = model.fit()

    # Lưu mô hình
    model_path = f'../models/{county_name}_arima.pkl'
    joblib.dump(model_fit, model_path)

```

Ta tiến hành train model dựa trên dữ liệu:

Ta cần set dữ liệu dựa trên state và tính toán theo tuần. Vì nếu tính theo ngày thì dữ liệu nó không thể đảm bảo được tính liên tục xuyên suốt (ngày nào cũng có dữ liệu) và dữ liệu cần thêm vào (dữ liệu bị thiếu) để xuất hiện sự liên tục là cực kì nhiều nên em quyết định gom dữ liệu thành từng tuần thay vì ngày.

Lý do chọn “Tuần” thay vì “Tháng” vì:

- Tuần sẽ đảm bảo được số lượng dữ liệu sẽ được liên tục, dù thiếu nhưng không nhiều và có thể thay thế bằng giá trị mean mà không sai lệch quá nhiều.
- Chọn “Tuần” sẽ đảm bảo số lượng dòng dữ liệu của các State khi đưa vào ARIMA được nhiều nhất, tăng độ tin cậy cho model được train.

```

# Huấn luyện và lưu mô hình ARIMA cho từng county
unique_states = df_merged['State_Name'].unique()
for state in unique_states:
    print(f'Training ARIMA model for {state}...')
    df_filtered = df_merged[df_merged['State_Name'] == state].set_index('Full_Date').sort_index()

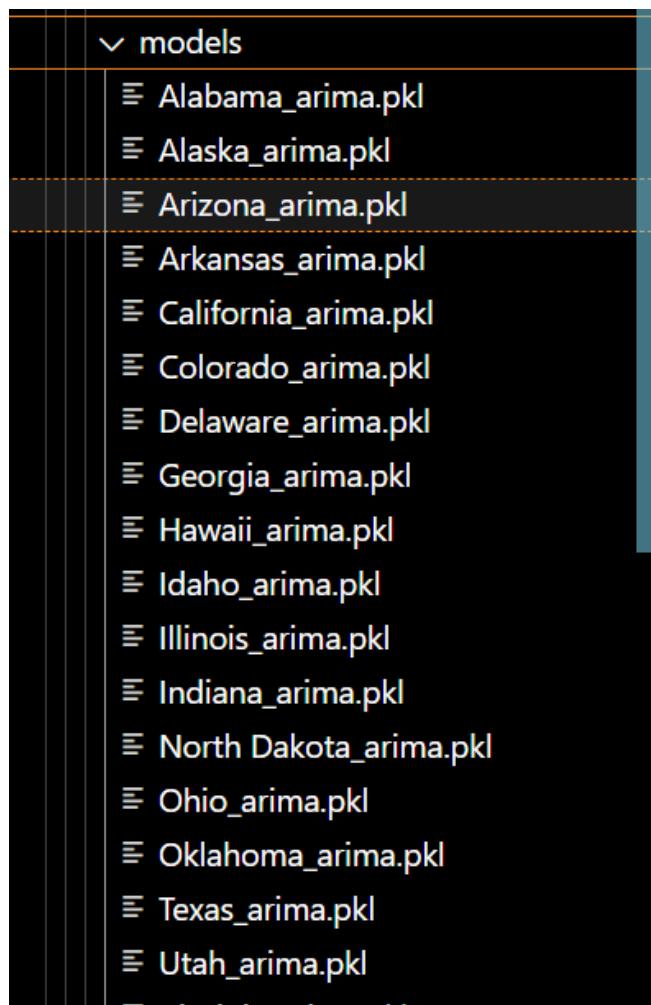
    # Resample dữ liệu theo tuần và tính giá trị trung bình của AQI mỗi tuần
    df_resampled = df_filtered.resample('W').mean(numeric_only=True) # Thêm numeric_only=True

    # Lấy chuỗi AQI để huấn luyện ARIMA
    aqi_series = df_resampled['AQI']
    aqi_series = aqi_series.fillna(aqi_series.mean())
    # Huấn luyện mô hình ARIMA
    train_and_save_arima_model(state, aqi_series)

```

Python

Ta thu được các model:



Sau khi train model thành công, ta cần sử dụng model để dự đoán:

```
# Hàm để tải mô hình ARIMA dựa trên tên location
def load_arima_model(location):
    model_path = f'../models/{location}_arima.pkl' # Xác định đường dẫn tự động theo tên location
    try:
        model = joblib.load(model_path) # Tải mô hình từ file
        print(f'Model for {location} loaded successfully.')
        return model
    except FileNotFoundError:
        print(f'Model for {location} not found at {model_path}.')
        return None

```

✓ 0.0s Python

```
location = input("Nhập tên location: ")

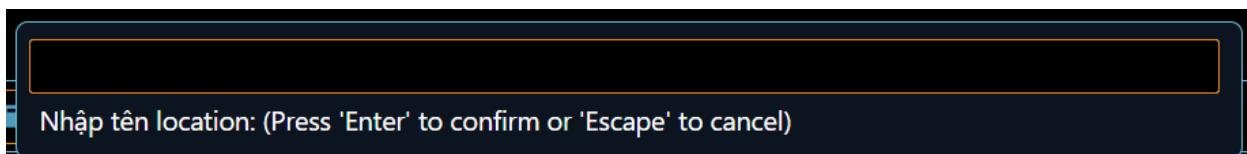
# Tải mô hình ARIMA cho location
model = load_arima_model(location)

if model:
    # Dự báo 1 tháng (4 tuần)
    forecast_1_month = model.forecast(steps=4)
    print(f'1 Month Forecast for {location}:')
    print(forecast_1_month)

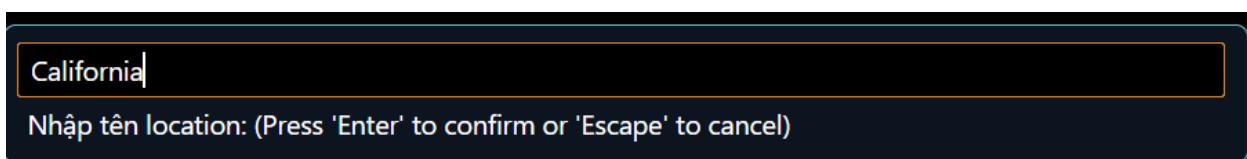
    # Dự báo 1 quý (12 tuần)
    forecast_1_quarter = model.forecast(steps=12)
    print(f'1 Quarter Forecast for {location}:\n')
    print(forecast_1_quarter)
```

✓ 8.2s Python

Khi chạy dòng code ở dưới, sẽ xuất hiện nơi nhập input, bạn chỉ cần nhập chính xác tên State thì sẽ chạy thành công:



Ví dụ chọn 'California'.



Nhấn enter.

Ta được:

```
Model for California loaded successfully.  
1 Month Forecast for California:  
2024-01-07      55.509224  
2024-01-14      58.326487  
2024-01-21      56.491312  
2024-01-28      57.367915  
Freq: W-SUN, Name: predicted_mean, dtype: float64  
1 Quarter Forecast for California:  
  
2024-01-07      55.509224  
2024-01-14      58.326487  
2024-01-21      56.491312  
2024-01-28      57.367915  
2024-02-04      58.806767  
2024-02-11      56.426901  
2024-02-18      58.229568  
2024-02-25      58.698291  
2024-03-03      56.931517  
2024-03-10      58.485949  
2024-03-17      58.296585  
2024-03-24      57.631189  
Freq: W-SUN, Name: predicted_mean, dtype: float64
```

Vẽ lại dữ liệu dự đoán:

```
# Tải dữ liệu thực tế (ví dụ là df_merged với dữ liệu AQI cho từng state và full date)
df_filtered = df_merged[df_merged['State_Name'] == location].set_index('Full_Date').sort_index().resample('W').mean()

# Dự báo 1 quý (12 tuần)
forecast_1_quarter = model.forecast(steps=12)

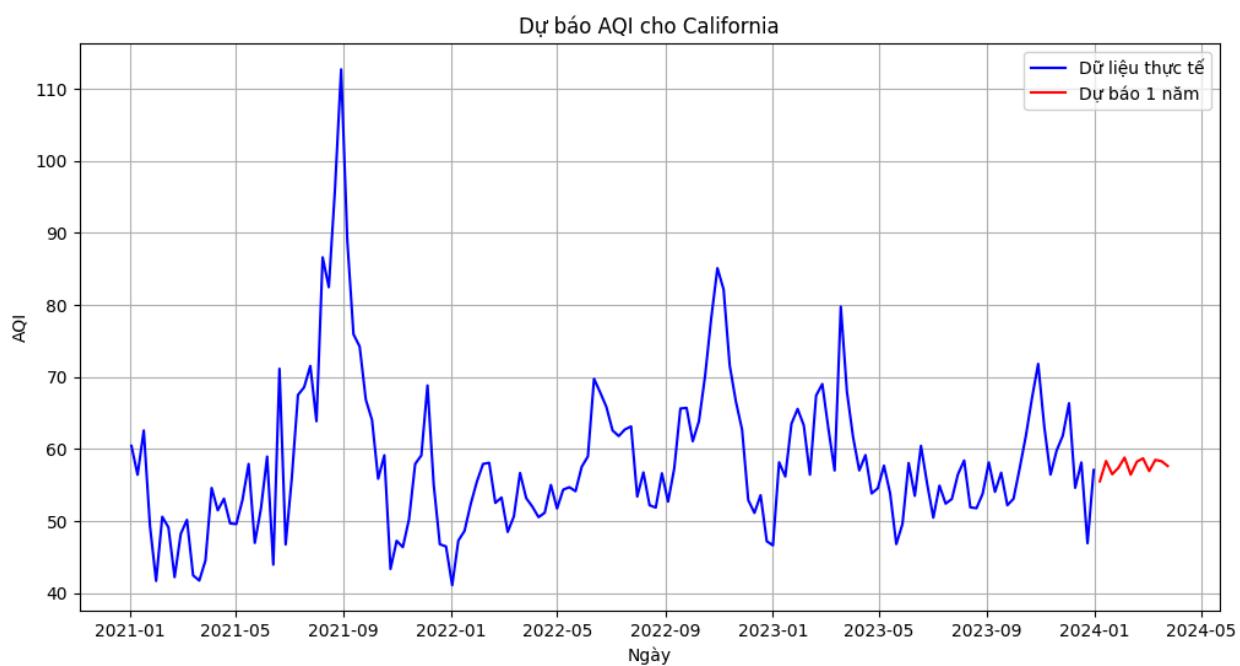
# Vẽ dữ liệu thực tế
plt.figure(figsize=(12, 6))
plt.plot(df_filtered.index, df_filtered['AQI'], label='Dữ liệu thực tế', color='blue')

# Tạo chuỗi thời gian cho dự báo 1 quý (12 tuần)
forecast_index_1_quarter = pd.date_range(start=df_filtered.index[-1] + pd.Timedelta(weeks=1), periods=12, freq='W')

# Vẽ dự báo 1 quý
plt.plot(forecast_index_1_quarter, forecast_1_quarter, label='Dự báo 1 năm', color='red')

# Thêm nhãn và tiêu đề
plt.title(f'Dự báo AQI cho {location}')
plt.xlabel('Ngày')
plt.ylabel('AQI')
plt.legend()
plt.grid(True)

# Hiển thị biểu đồ
plt.show()
```



Tham Khảo

<https://phamdinhkhanh.github.io/2019/12/12/ARIMAmode.html>