



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

Khoa Công nghệ Thông tin

ĐỒ ÁN THỰC HÀNH

BUILDING AND MINING DATA WAREHOUSE

Student: 21127702 - Bùi Nguyễn Tin
19127407 - Nguyễn Huy Hoàng
20127189 - Nguyễn Quốc Huy
20127204 - Nguyễn Phụng Khanh

Course: CSC12107 - 21HTTT2

Teacher: Hồ Thị Hoàng Vy
Tiết Gia Hồng
Nguyễn Ngọc Minh Châu
Lê Nguyễn Hoài Nam

Thành phố Hồ Chí Minh – 2024

Mục lục

BẢNG PHÂN CÔNG CÔNG VIỆC.....	3
NỘI DUNG.....	4
1. Thiết kế NDS, DDS.....	4
1.1. Mô hình NDS.....	4
1.2. Mô hình DDS.....	5
2. Giải thích các Components.....	8
2.1. Stage to NDS.....	8
2.2. NDS to DDS.....	11

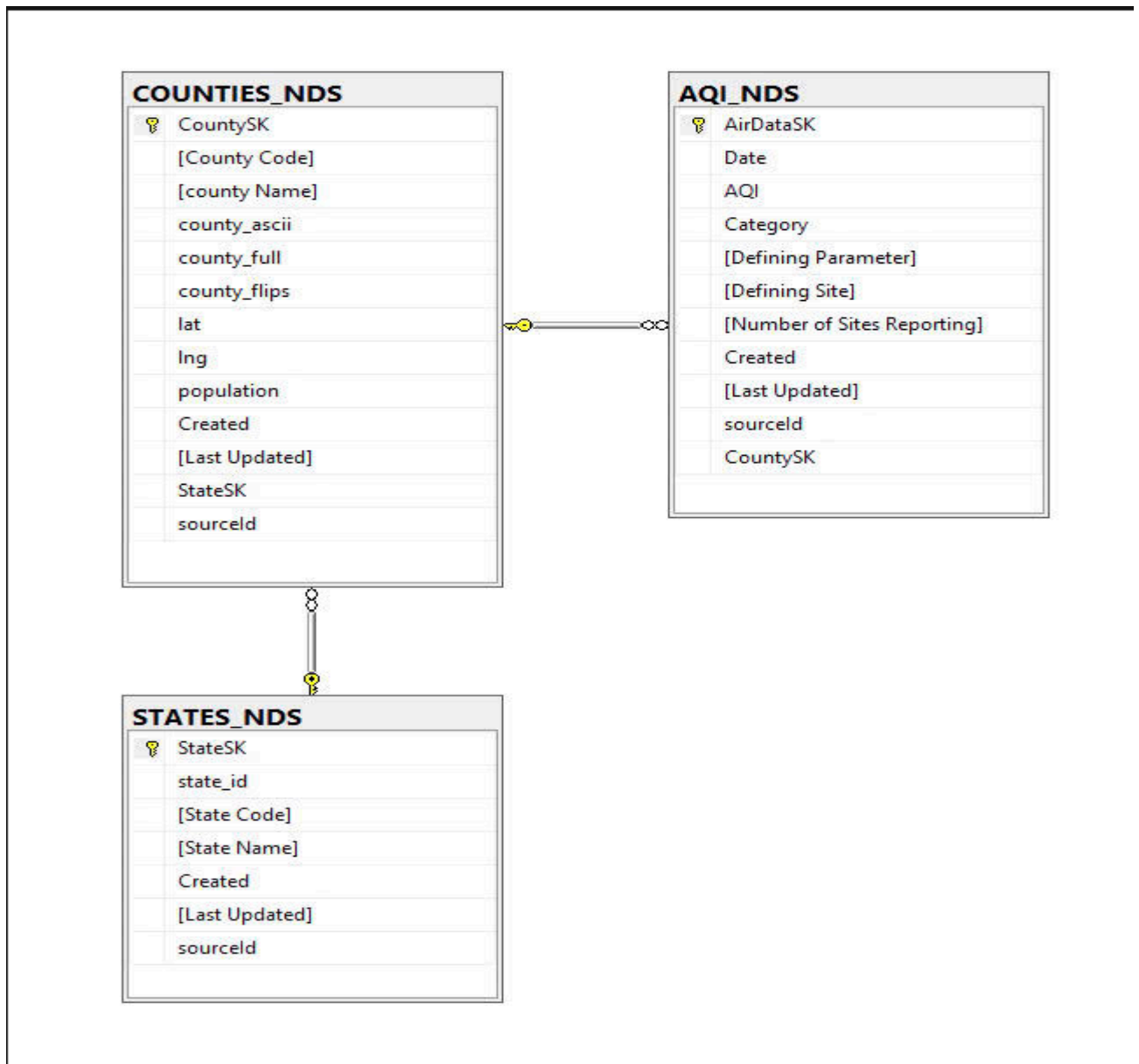
BẢNG PHÂN CÔNG CÔNG VIỆC

STT	Họ và Tên	MSSV	Công việc	Mức độ hoàn thành
1	Nguyễn Quốc Huy	20127189	DDS, báo cáo	100%
2	Nguyễn Phụng Khanh	20127204	Báo cáo	100%
3	Nguyễn Huy Hoàng	19127407	X	0%
4	Bùi Nguyễn Tin	21127702	DDS, báo cáo	100%

NỘI DUNG

1. Thiết kế NDS, DDS

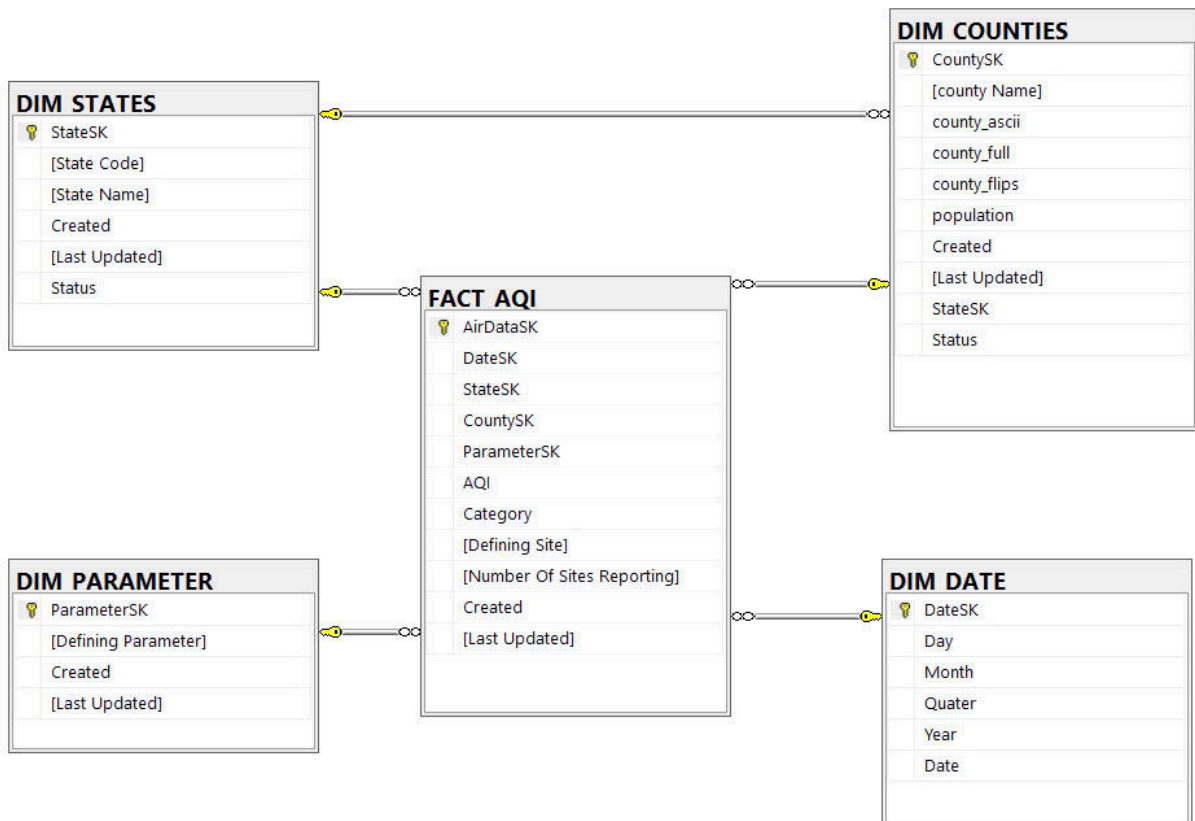
1.1. Mô hình NDS



Lấy dữ liệu từ giai đoạn 1 (source to stage) tiến hành thêm dữ liệu vào với các cột tương ứng bổ sung thêm các khóa chính, khóa ngoại, sourceId.

1.2. Mô hình DDS

- Thiết kế NDS:



● Fact table:

Bảng FACT_AQI:

Bảng FACT AQI là bảng thực tế (fact table) chính trong mô hình schema sao của hệ thống, chứa các giá trị đo lường và liên kết với các bảng chiều (dimension tables).

- AirDataSK: Khóa thay thế (surrogate key) cho bảng FACT AQI.
- DateSK: Khóa thay thế liên kết với bảng chiều DIM DATE, chứa thông tin về ngày.
- StateSK: Khóa thay thế liên kết với bảng chiều DIM STATES, chứa thông tin về bang.
- CountySK: Khóa thay thế liên kết với bảng chiều DIM COUNTIES.
- ParameterSK: Khóa thay thế liên kết với bảng chiều DIM PARAMETER, chứa thông tin về tham số đo lường.
- AQI: Giá trị đo lường chỉ số chất lượng không khí.
- Category: Danh mục của AQI.
- [Defining Site]: Địa điểm chính xác nơi đo lường AQI.
- [Number Of Sites Reporting]: Số lượng địa điểm báo cáo dữ liệu AQI.
- Created: Ngày và giờ tạo bản ghi.
- [Last Updated]: Ngày và giờ bản ghi được cập nhật lần cuối.

● Dimension table:

các bảng Dimension.

DIM STATES Table: Chứa thông tin về các bang.

- **StateSK:** Khóa thay thế cho bảng DIM STATES.
- **[State Code]:** Mã của bang.
- **[State Name]:** Tên của bang.
- **Created:** Ngày và giờ tạo bản ghi.
- **[Last Updated]:** Ngày và giờ bản ghi được cập nhật lần cuối.
- **Status:** Trạng thái của bản ghi.

DIM COUNTIES Table: Chứa thông tin về các hạt/quận.

- **CountySK:** Khóa thay thế cho bảng DIM COUNTIES.
- **[County Name]:** Tên của county.
- **county_ascii:** Tên ASCII của county.

- **county_full**: Tên đầy đủ của county.
- **county_fips**: Mã FIPS của county.
- **population**: Dân số của county.
- **Created**: Ngày và giờ tạo bản ghi.
- **[Last Updated]**: Ngày và giờ bản ghi được cập nhật lần cuối.
- **StateSK**: Khóa thay thế liên kết với bảng DIM STATES.
- **Status**: Trạng thái của bản ghi.

DIM PARAMETER Table: Chứa thông tin về các tham số đo lường.

- **ParameterSK**: Khóa thay thế cho bảng DIM PARAMETER.
 - **[Defining Parameter]**: Tham số đo lường.
 - **Created**: Ngày và giờ tạo bản ghi.
 - **[Last Updated]**: Ngày và giờ bản ghi được cập nhật lần cuối.
- Riêng bảng DIM_DATE: các cột ngày, tháng, quý, năm được tách ra riêng qua 1 PROCEDURE

```
CREATE PROCEDURE AddDateTo_DIM_DATE
AS
BEGIN
    DECLARE @reqStart_date datetime,
            @reqEnd_date datetime,
            @current_date datetime;

    SELECT @reqStart_date = MIN(Date), @reqEnd_date = MAX(Date)
    FROM [NDS_ISBI].[dbo].[AQI_NDS];

    SET @current_date = @reqStart_date;

    WHILE @current_date <= @reqEnd_date
    BEGIN
        INSERT INTO [dbo].[DIM_DATE] ([Day], [Month], [Quater], [Year], [Date])
        VALUES (
            DATEPART(DAY, @current_date),
            DATEPART(MONTH, @current_date),
            DATEPART(QUARTER, @current_date),
            DATEPART(YEAR, @current_date),
            CAST(@current_date AS DATE)
        );

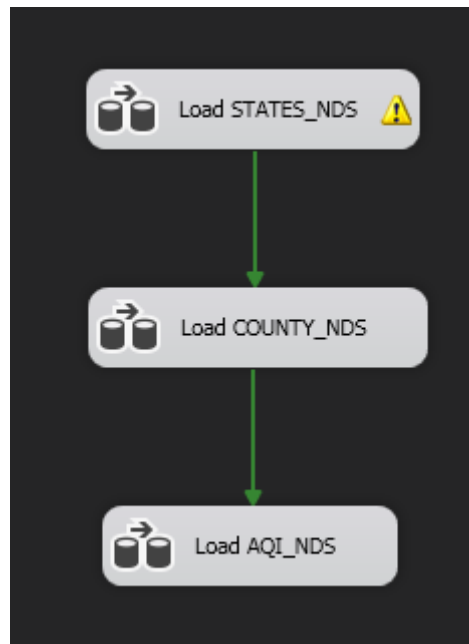
        SET @current_date = DATEADD(DAY, 1, @current_date);
    END;
END;
GO

EXEC AddDateTo_DIM_DATE
GO
```

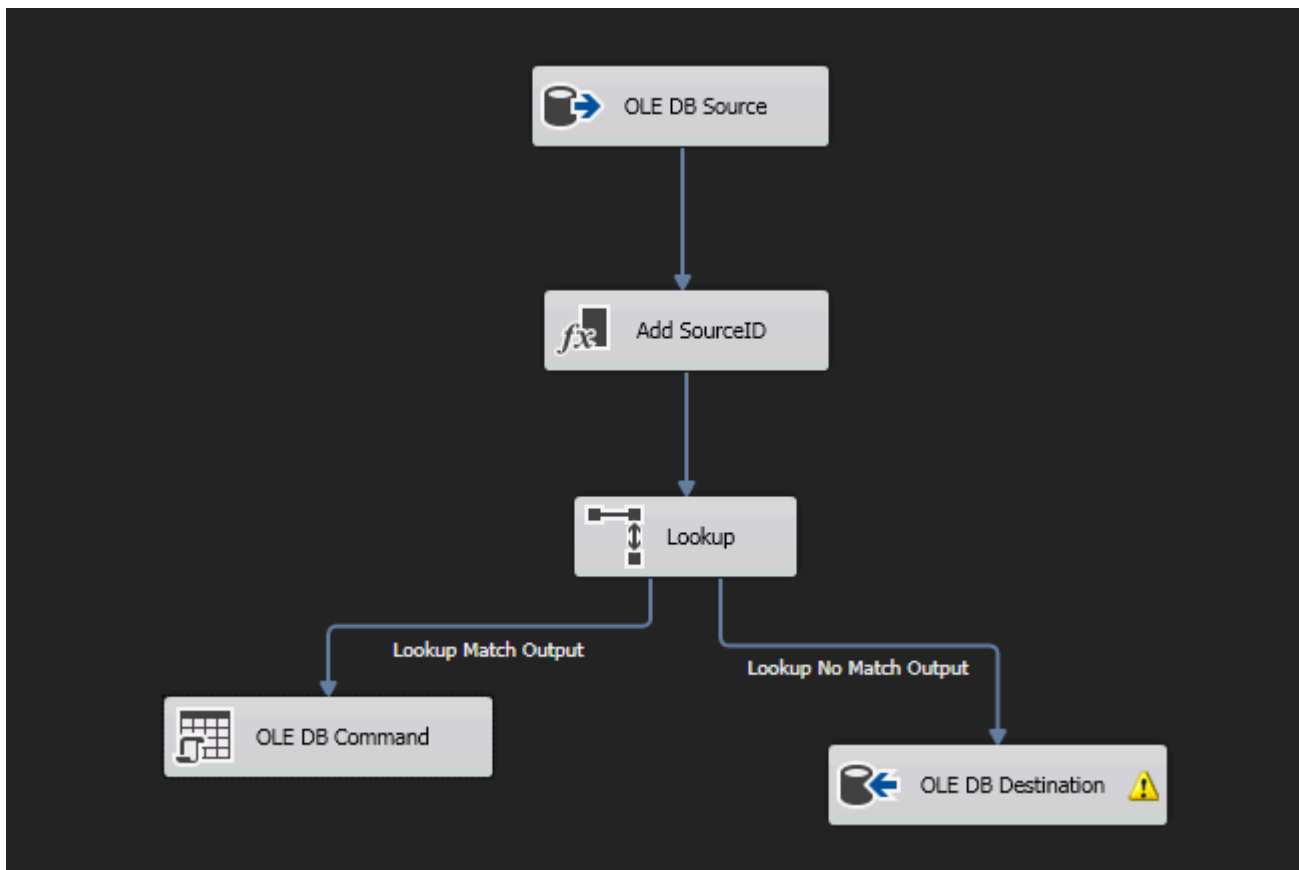
Thực hiện chạy vòng lặp bắt đầu với ngày nhỏ nhất trong dữ liệu NDS và kết thúc với ngày gần nhất, với mỗi lần lặp lại sẽ công thêm một current và thêm một ngày cho current đó.

2. Giải thích các Components

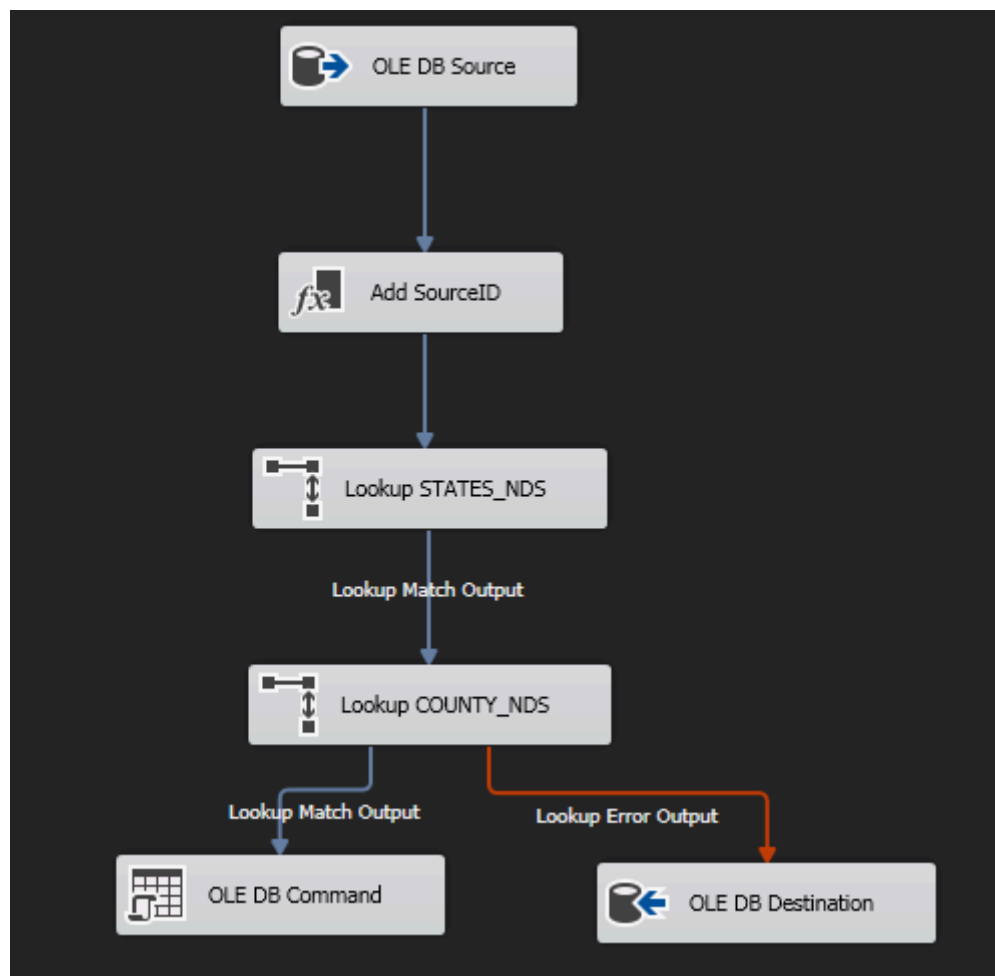
2.1. Stage to NDS



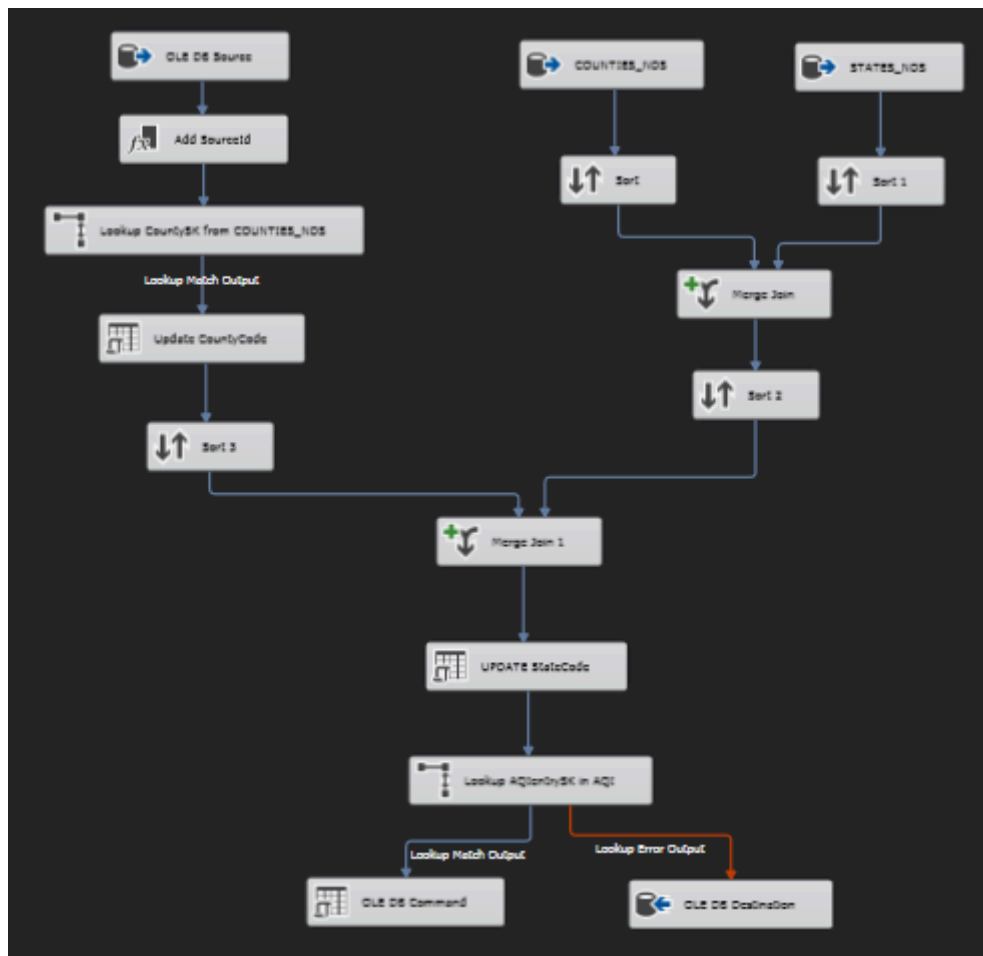
Các DataFlow thực hiện nạp dữ liệu từ các bảng Stage vào các bảng STATES_NDS, COUNTY_NDS, AQI_NDS.



Các component trong DataFlow 'Load STATES_NDS': Đầu tiên lấy dữ liệu trong bảng USCOUNTIES_STAGE, sau đó thêm cột SourceID và thực hiện Lookup lấy StateSK trong bảng STATES_NDS. Nếu có dữ liệu rồi thì thực hiện cập nhật lại, còn nếu chưa có dữ liệu trước đó thì tiến hành nạp vào bảng.



Tương tự với DataFlow 'Load COUNTY_NDS': Lấy dữ liệu từ bảng USCOUNTIES_STAGE, sau đó thêm cột SourceID và thực hiện Loopup lấy StateSK, [county_fips] từ các bảng STATES_NDS, COUNTY_NDS. Nếu có dữ liệu rồi thì thực hiện cập nhật lại, còn nếu chưa có dữ liệu trước đó thì tiến hành nạp vào bảng.

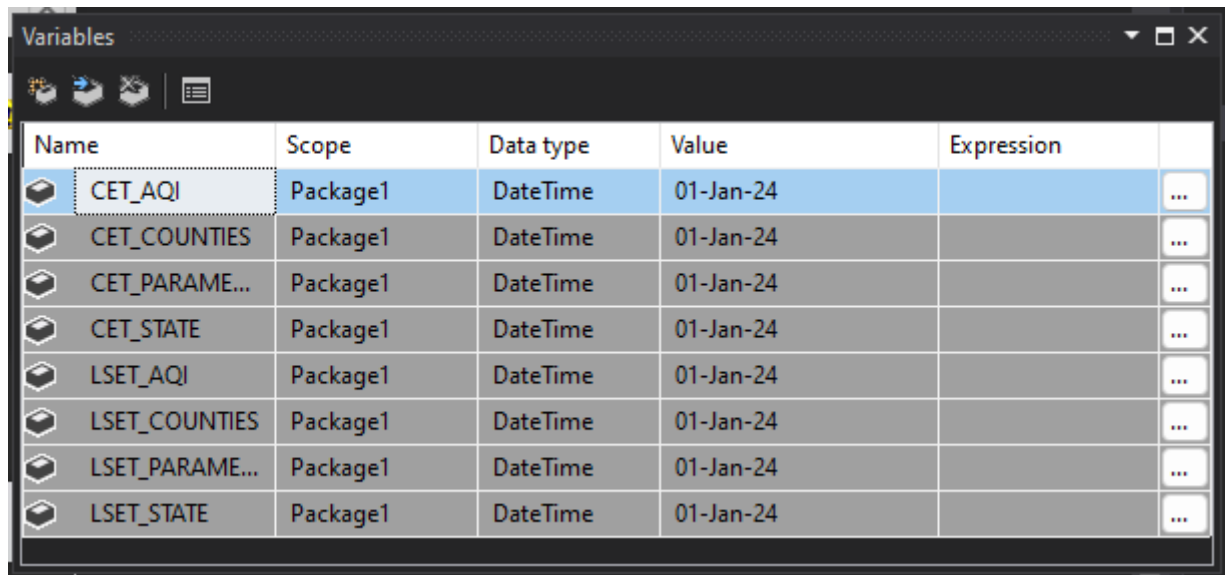


Tương tự với DataFlow ‘Load AQI_NDS’: Lấy dữ liệu từ bảng AIR_QUALITY_STAGE, sau đó thêm cột SourceID và thực hiện Lookup lấy CoutySK trong bảng COUNTY_NDS, cập nhật lại LastUpdated và CountyCode. Tiếp theo tiến hành Merge join hai bảng COUNTY_NDS và STATES_NDS và tổng Merge join lại với bảng AIR_QUALITY_STAGE. Cuối cùng thực hiện Lookup [AirDataSK] trong AQI_NDS. . Nếu có dữ liệu rồi thì thực hiện cập nhật lại, còn nếu chưa có dữ liệu trước đó thì tiến hành nạp vào bảng.

2.2. NDS to DDS

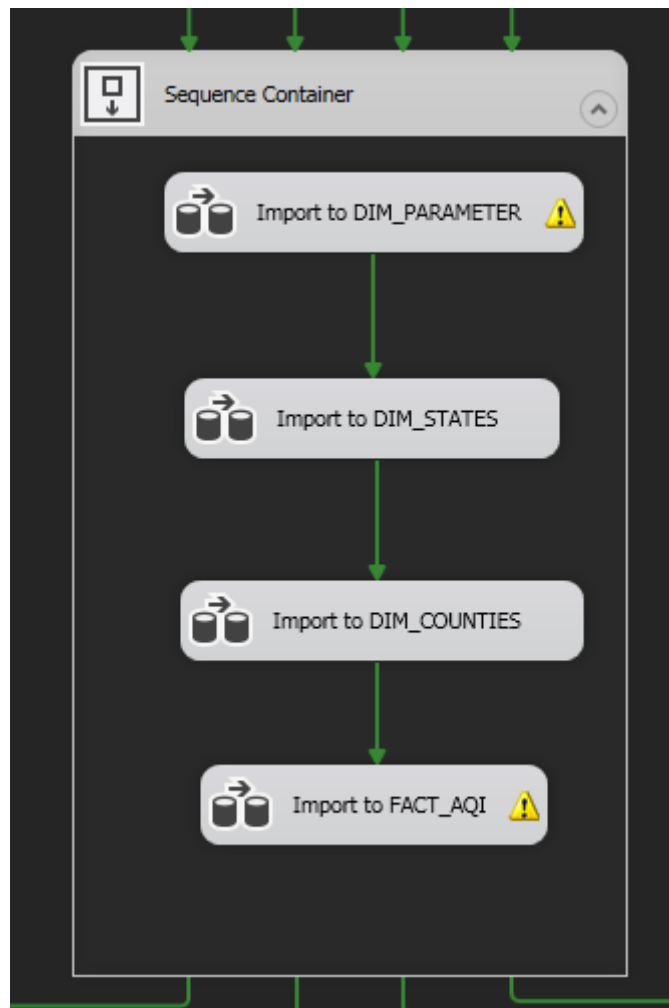


Các Component cập nhật CET và lấy LSET, CET của các bảng DIM_PARAMETER, DIM_STATES, DIM_COUNTIES, FACT_AQI trong Data_Flow.

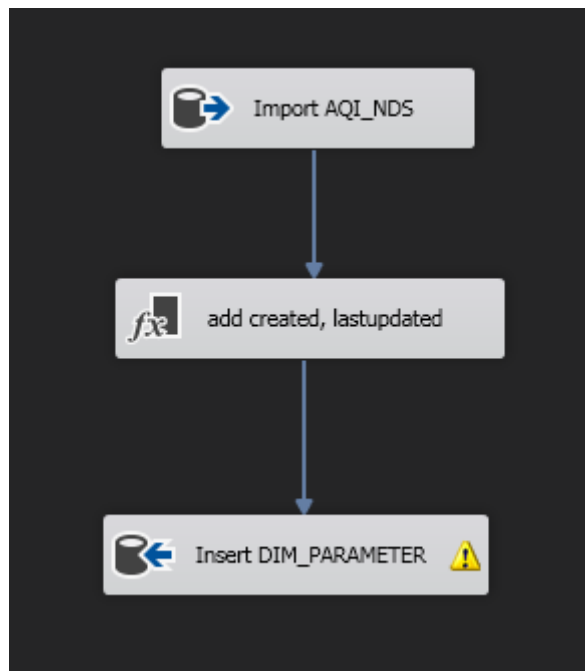


Name	Scope	Data type	Value	Expression	
CET_AQI	Package1	DateTime	01-Jan-24		...
CET_COUNTIES	Package1	DateTime	01-Jan-24		...
CET_PARAME...	Package1	DateTime	01-Jan-24		...
CET_STATE	Package1	DateTime	01-Jan-24		...
LSET_AQI	Package1	DateTime	01-Jan-24		...
LSET_COUNTIES	Package1	DateTime	01-Jan-24		...
LSET_PARAME...	Package1	DateTime	01-Jan-24		...
LSET_STATE	Package1	DateTime	01-Jan-24		...

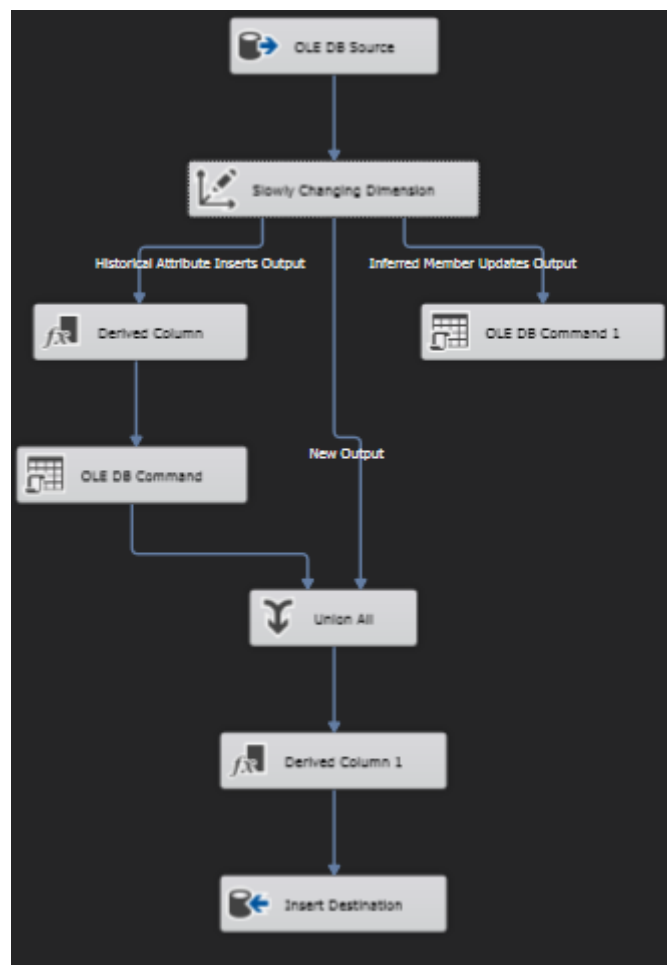
Thiết lập các giá trị LSET, CET tương ứng với từng bảng DIM_PARAMETER, DIM_STATES, DIM_COUNTIES, FACT_AQI.



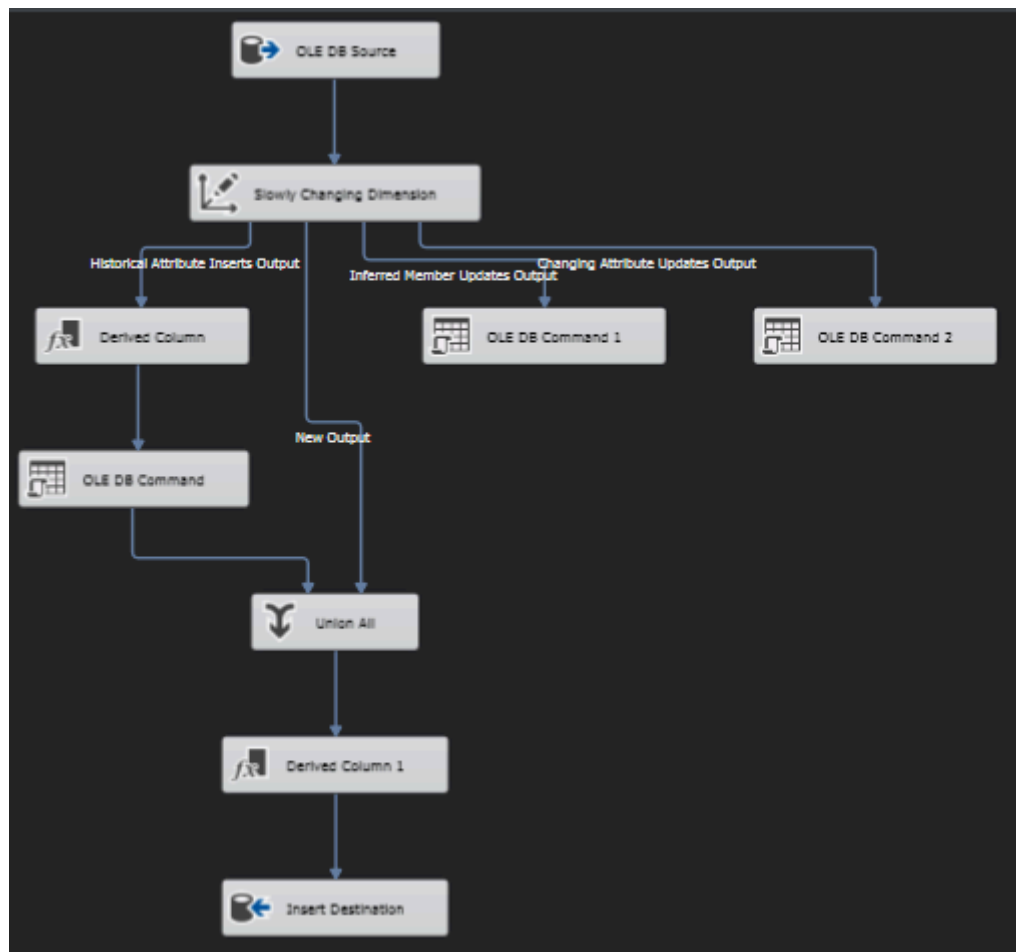
Container tổng của các DataFlow nạp dữ liệu từ các bảng trong NDS vào các bảng DIM_PARAMETER, DIM_STATES, DIM_COUNTIES, FACT_AQI.



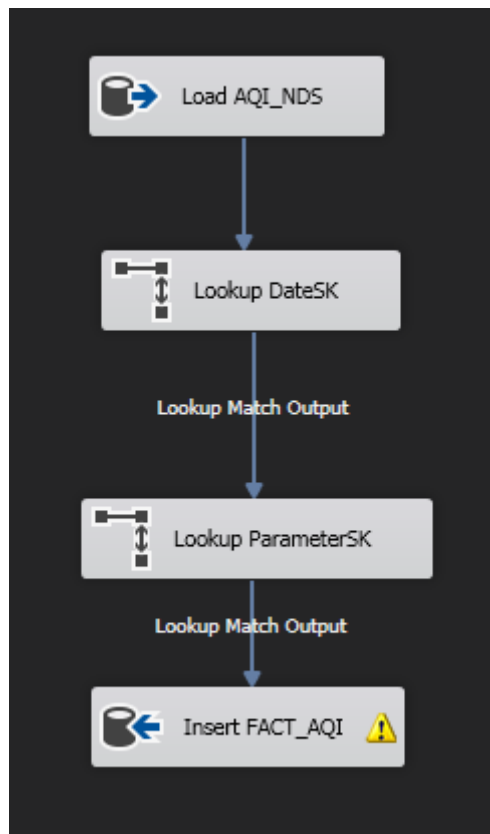
Trong phần nạp dữ liệu của bảng DIM_PARAMETER gồm: Lấy dữ liệu từ AQI_NDS, sau đó thêm cột ngày tạo [Created] và ngày cập nhật [LastUpdated] và cuối cùng nạp vào bảng DIM_PARAMETER.



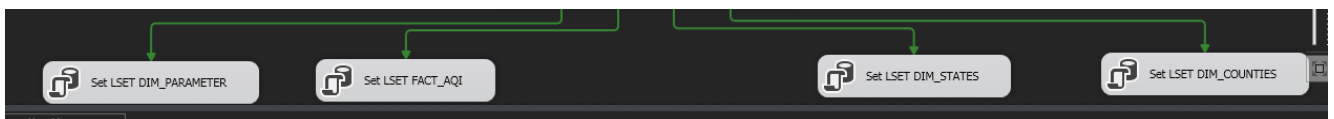
Trong phần nạp dữ liệu của bảng DIM_STATES gồm: Lấy dữ liệu từ bảng STATES_NDS, sau đó sử dụng component ‘Slowly Changing Dimension’ để tạo chiều và cuối cùng nạp vào bảng DIM_STATES.



Trong phần nạp dữ liệu của bảng DIM_COUNTIES gồm: Lấy dữ liệu từ bảng COUNTIES_NDS, sau đó sử dụng component ‘Slowly Changing Dimension’ để tạo chiều và cuối cùng nạp vào bảng DIM_COUNTIES.



Trong phần nạp dữ liệu của bảng FACT_AQI gồm: Lấy dữ liệu từ bảng AQI_NDS, sau đó thực hiện Lookup để lấy hai tham số DateSK và ParameterSK từ hai bảng DIM_DATE và DIM_PARAMETER, cuối cùng nạp dữ liệu vào bảng FACT_AQI.



Cuối cùng cập nhật LSET cho từng bảng DIM_PARAMETER, DIM_STATES, DIM_COUNTIES, FACT_AQI tương ứng.