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 Phượng 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

STT	Họ và Tên	MSSV	Công việc	Mức độ hoàn thành
1	Nguyễn Quốc Huy	20127189	ETL, OLAP, MDX, Báo cáo	100%
2	Nguyễn Phượng Khanh	20127204	Báo cáo	100%
3	Nguyễn Huy Hoàng	19127407	Không có đóng góp	100%
4	Bùi Nguyễn Tin	21127702	ETL, OLAP, Báo cáo	100%

NỘI DUNG

1. Đổ dữ liệu từ file source vào STAGE

1.1. Thiết kế DATA_FLOW

Tạo bảng DATA_FLOW trong Database

```
CREATE TABLE DATA_FLOW
(
    ID INT NOT NULL IDENTITY(1,1),
    NAME VARCHAR(50) NOT NULL,
    STATUS INT,
    LSET DATETIME,
    CET DATETIME,
    CONSTRAINT PK_DATA_FLOW
    PRIMARY KEY CLUSTERED (ID)
)
```

Các thuộc tính bao gồm ID của từng Database, tên các bảng cụ thể của từng giai đoạn, ngày tạo và ngày thay đổi.

Dữ liệu nguồn gồm các file AIR_QUALITY và USCOUNTIES được lưu vào dataflow và thay đổi ngày tạo và ngày cập nhật. Tương tự như vậy cho các bảng trong database DDS:

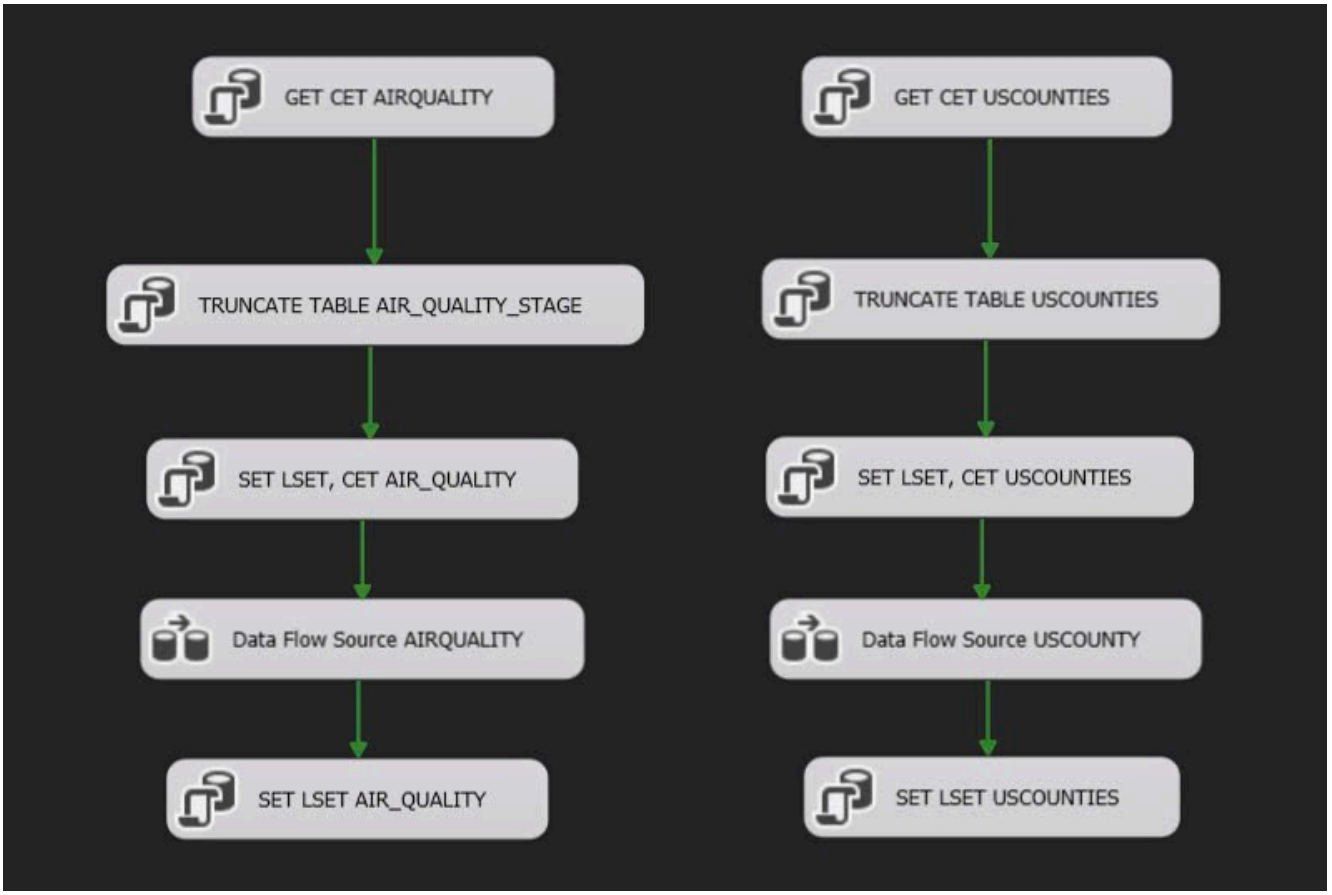
```
--Nap du lieu vao DataFlow
INSERT INTO DATA_FLOW (NAME, STATUS, LSET, CET)
VALUES ('AIR_QUALITY', 1, '2023-01-01 00:00:00','2024-01-01 00:00:00')
INSERT INTO DATA_FLOW (NAME, STATUS, LSET, CET)
VALUES('USCOUNTIES', 1, '2023-01-01 00:00:00','2024-01-01 00:00:00')
--DDS
INSERT INTO DATA_FLOW (NAME, STATUS, LSET, CET)
VALUES('DIM_STATES', 1, '2023-01-01 00:00:00','2024-01-01 00:00:00')
INSERT INTO DATA_FLOW (NAME, STATUS, LSET, CET)
VALUES('DIM_COUNTIES', 1, '2023-01-01 00:00:00','2024-01-01 00:00:00')
INSERT INTO DATA_FLOW (NAME, STATUS, LSET, CET)
VALUES('DIM_DATE', 1, '2023-01-01 00:00:00','2024-01-01 00:00:00')
--INSERT INTO DATA_FLOW (NAME, STATUS, LSET, CET)
--VALUES('DIM_PARAMETER', 1, '2023-01-01 00:00:00','2024-01-01 00:00:00')
INSERT INTO DATA_FLOW (NAME, STATUS, LSET, CET)
VALUES('FACT_AQI', 1, '2023-01-01 00:00:00','2024-01-01 00:00:00')
```

1.2. Quá trình nạp dữ liệu vào source sang Stage

Thiết kế database cho Stage: tạo hai bảng AIR_QUALITY_STAGE và USCOUNTIES_STAGE với các cột thuộc tính dựa trên file nguồn.

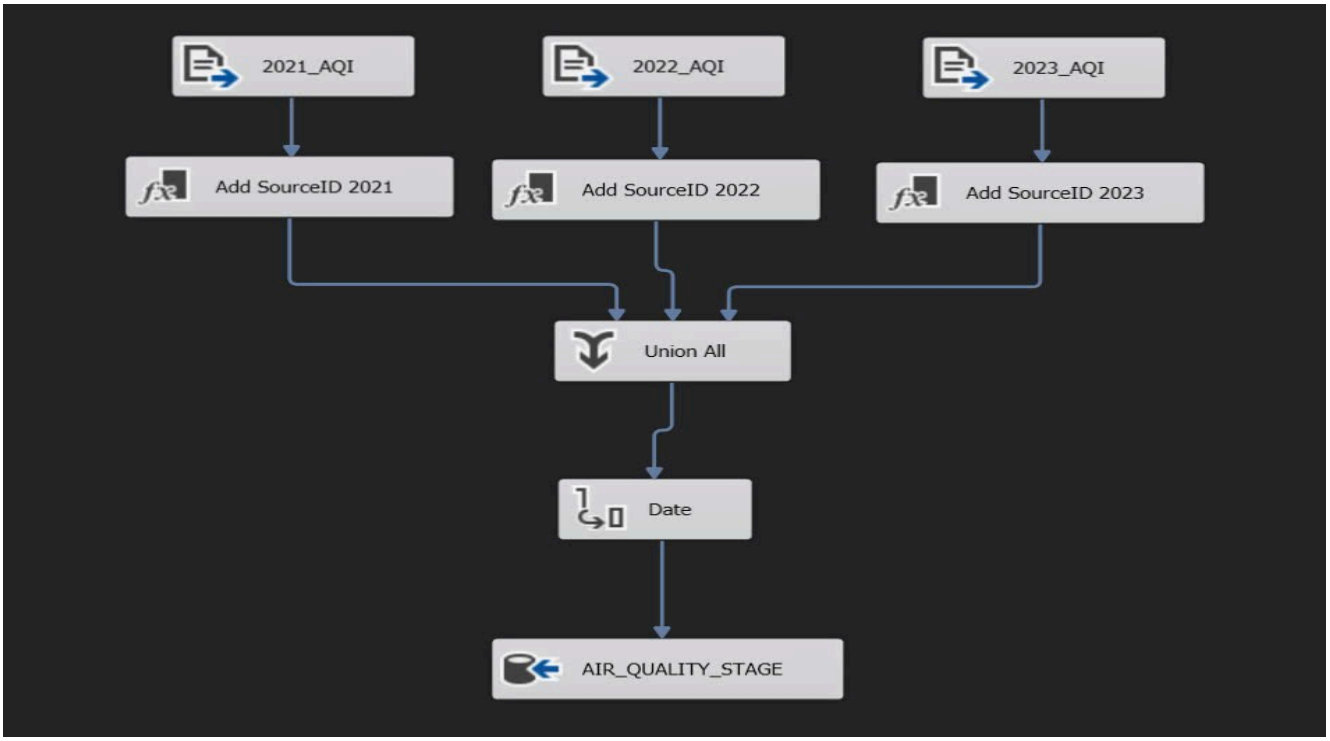
```
--Tao cac bang trong STAGE
CREATE TABLE AIR_QUALITY_STAGE (
    [State Name] varchar(50),
    [county Name] varchar(255),
    [State Code] int,
    [County Code] int,
    [Date] datetime,
    [AQI] int,
    [Category] varchar(255),
    [Defining Parameter] varchar(255),
    [Defining Site] varchar(255),
    [Number of Sites Reporting] int,
    [Created] datetime,
    [Last Updated] datetime,
    [SourceID] int
)
GO

CREATE TABLE USCOUNTIES_STAGE (
    [county] varchar(50),
    [county_ascii] varchar(225),
    [county_full] varchar(225),
    [county_fips] int,
    [state_id] varchar(50),
    [state_name] varchar(225),
    [lat] float,
    [lng] float,
    [population] int,
    [Created] datetime,
    [Last Updated] datetime
)
```

Quy trình xử lý dữ liệu:



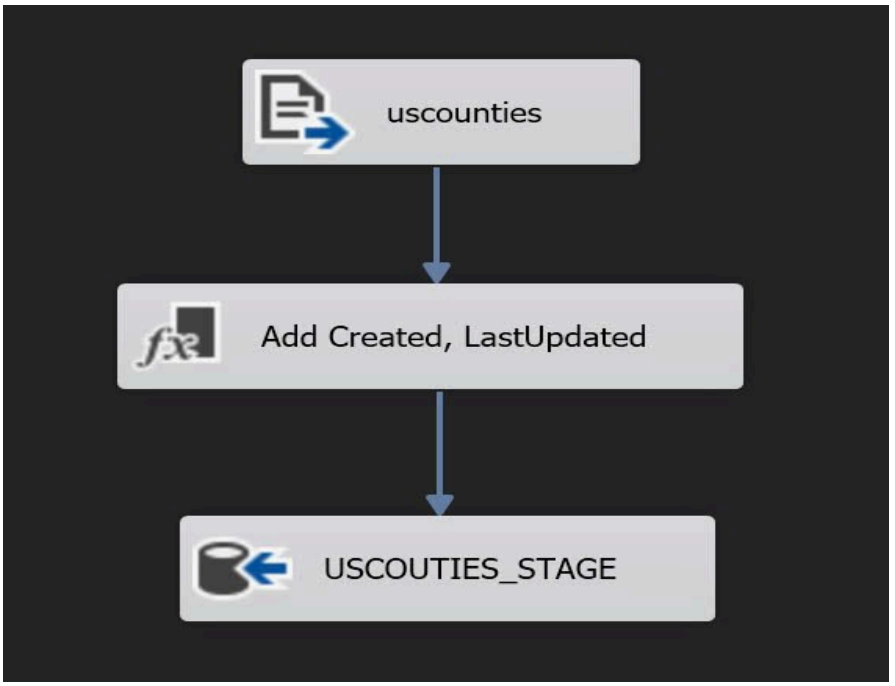
- Lấy ra ngày tạo và ngày cập nhật lại vào trong DATAFLOW của nguồn.
- Tiến hành truncate từng bảng nguồn.
- Cập nhật ngày tạo và ngày cập nhật lại vào trong DATAFLOW của nguồn.
- Tiến hành tải dữ liệu vào từ file dữ liệu nguồn.

Ở bước cập tải dữ liệu từ source trong component như sau:



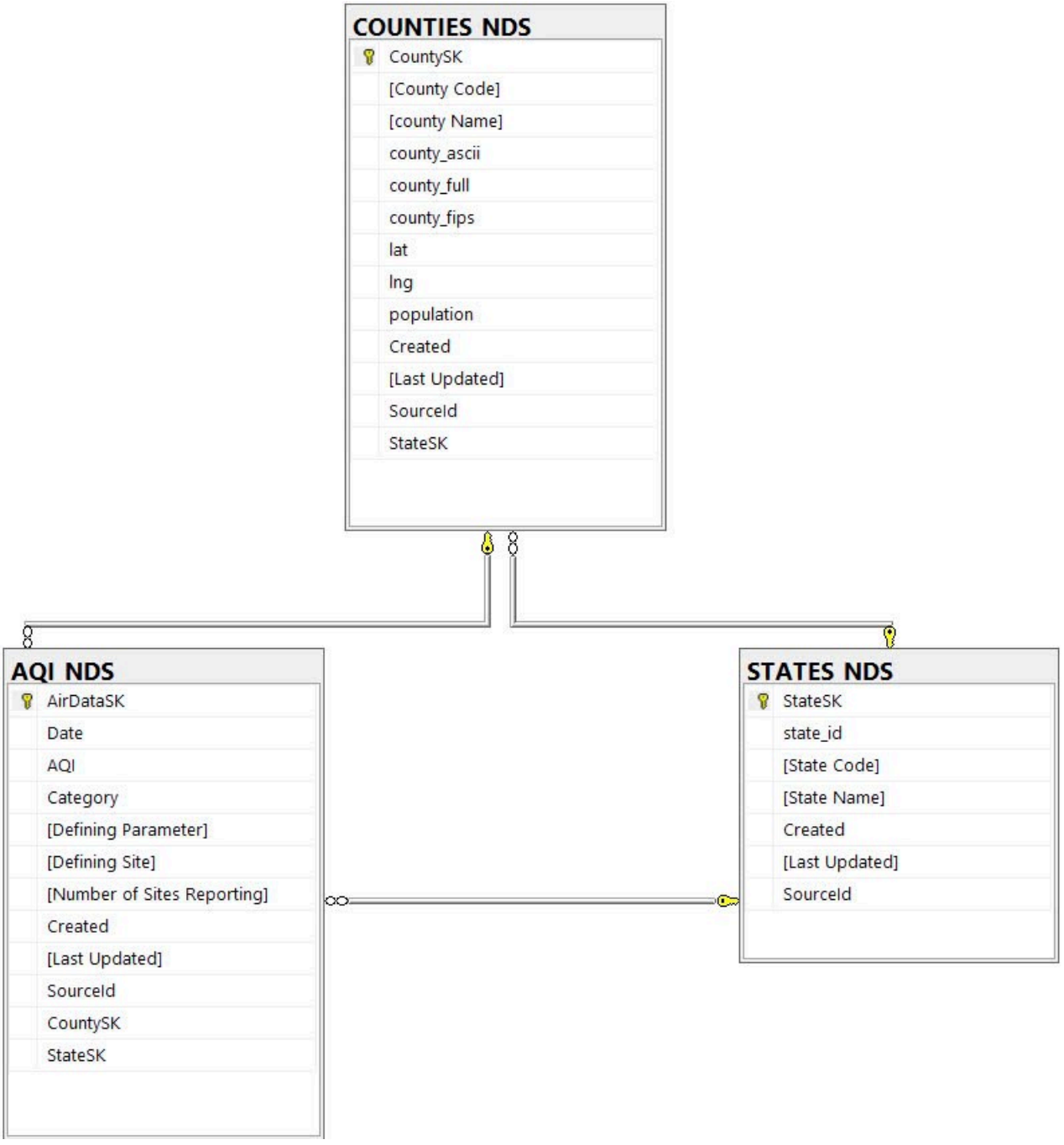
Tiến hành tải dữ liệu từ file nguồn lên và add thêm cột sourceID, vì có cuộc thuộc tính nên dữ liệu sẽ được thêm vào cùng 1 bảng ở cả 3 file (bước UNION ALL), cuối cùng là tải dữ liệu từ file nguồn vào bảng STAGE.

Tương tự cho file USCOUNTIES còn lại:


uscounties
Add Created, LastUpdated
USCOUTIES_STAGE

2. Stage sang NDS:

2.1. Sơ đồ NDS:



2.2. Thiết kế database cho các bảng tương ứng trong NDS:

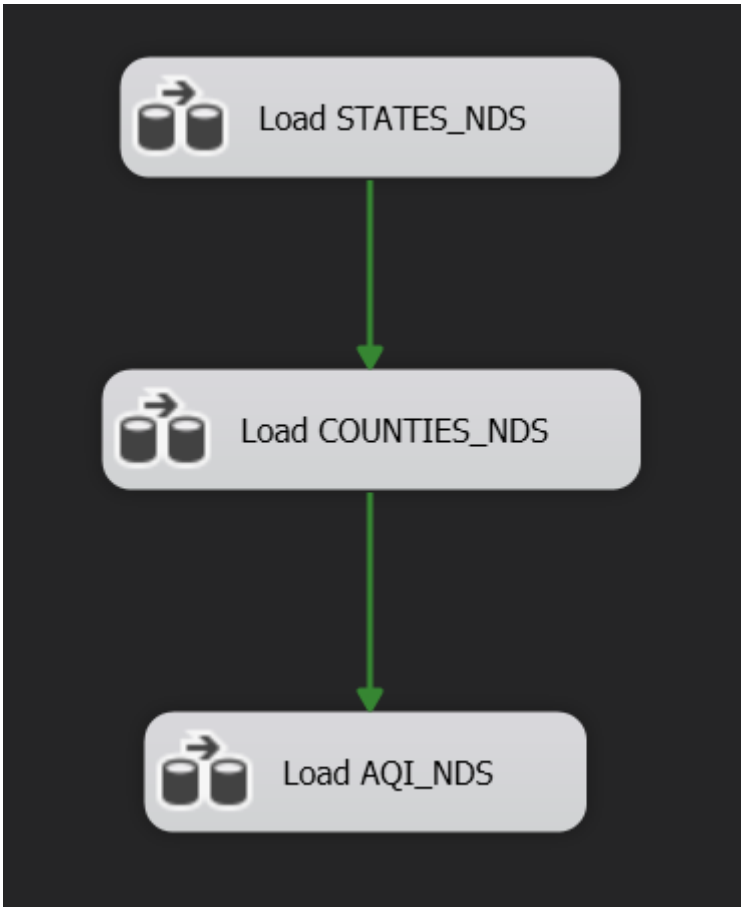
Dựa theo sơ đồ database NDS gồm 3 bảng STATES_NDS, COUNTIES_NDS, AQI_NDS với các thuộc tính lấy từ source tương ứng.

```
CREATE TABLE AQI_NDS (
    [AirDataSK] INT IDENTITY (1, 1),
    [Date] datetime,
    [AQI] int,
    [Category] varchar(255),
    [Defining Parameter] varchar(255),
    [Defining Site] varchar(255),
    [Number of Sites Reporting] int,
    [Created] datetime,
    [Last Updated] datetime,
    [SourceId] int,

    FOREIGN KEY (CountySK) REFERENCES COUNTIES_NDS(CountySK),
    [CountySK] INT NOT NULL,
    FOREIGN KEY (StateSK) REFERENCES STATES_NDS(StateSK),
    [StateSK] INT NOT NULL,

    CONSTRAINT PK_AQI_NDS
    PRIMARY KEY CLUSTERED (AirDataSK)
)
GO
```

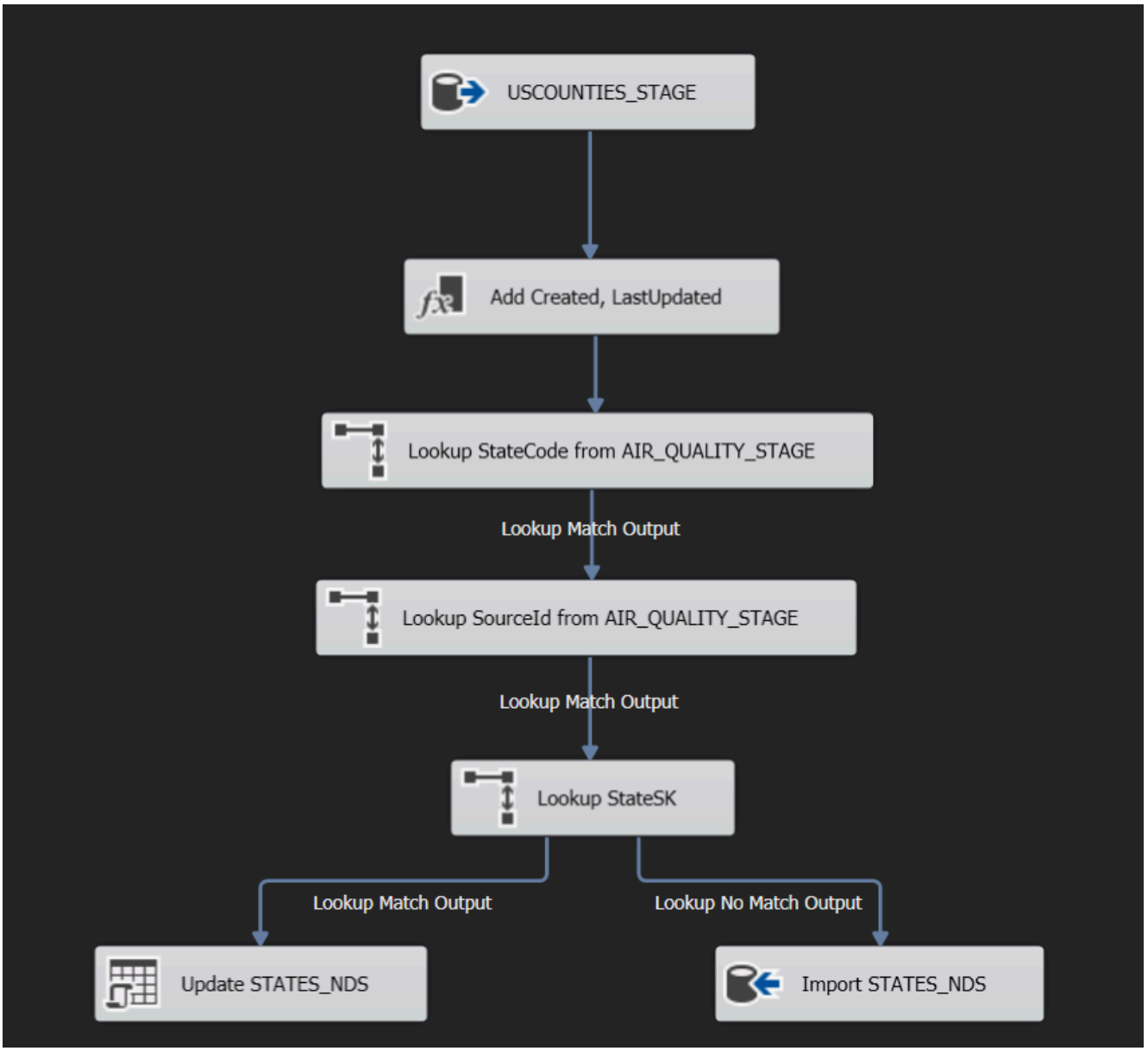
```
CREATE TABLE [STATES_NDS](
    [StateSK] INT IDENTITY (1, 1),
    [state_id] varchar(50),
    [State Code] int,
    [State Name] varchar(255),
    [Created] datetime,
    [Last Updated] datetime,
    [SourceId] int,

    CONSTRAINT PK_STATES_NDS
    PRIMARY KEY CLUSTERED (StateSK)
)
GO
```

```
CREATE TABLE COUNTIES_NDS(
    [CountySK] INT IDENTITY (1, 1),
    [County Code] int,
    [county Name] varchar(255),
    [county_ascii] varchar(225),
    [county_full] varchar(225),
    [county_fips] int,
    [lat] float,
    [lng] float,
    [population] int,
    [Created] datetime,
    [Last Updated] datetime,
    [SourceId] int,

    FOREIGN KEY (StateSK) REFERENCES STATES_NDS(StateSK),
    [StateSK] INT NOT NULL,

    CONSTRAINT PK_COUNTIES_NDS
    PRIMARY KEY CLUSTERED (CountySK)
)
GO
```

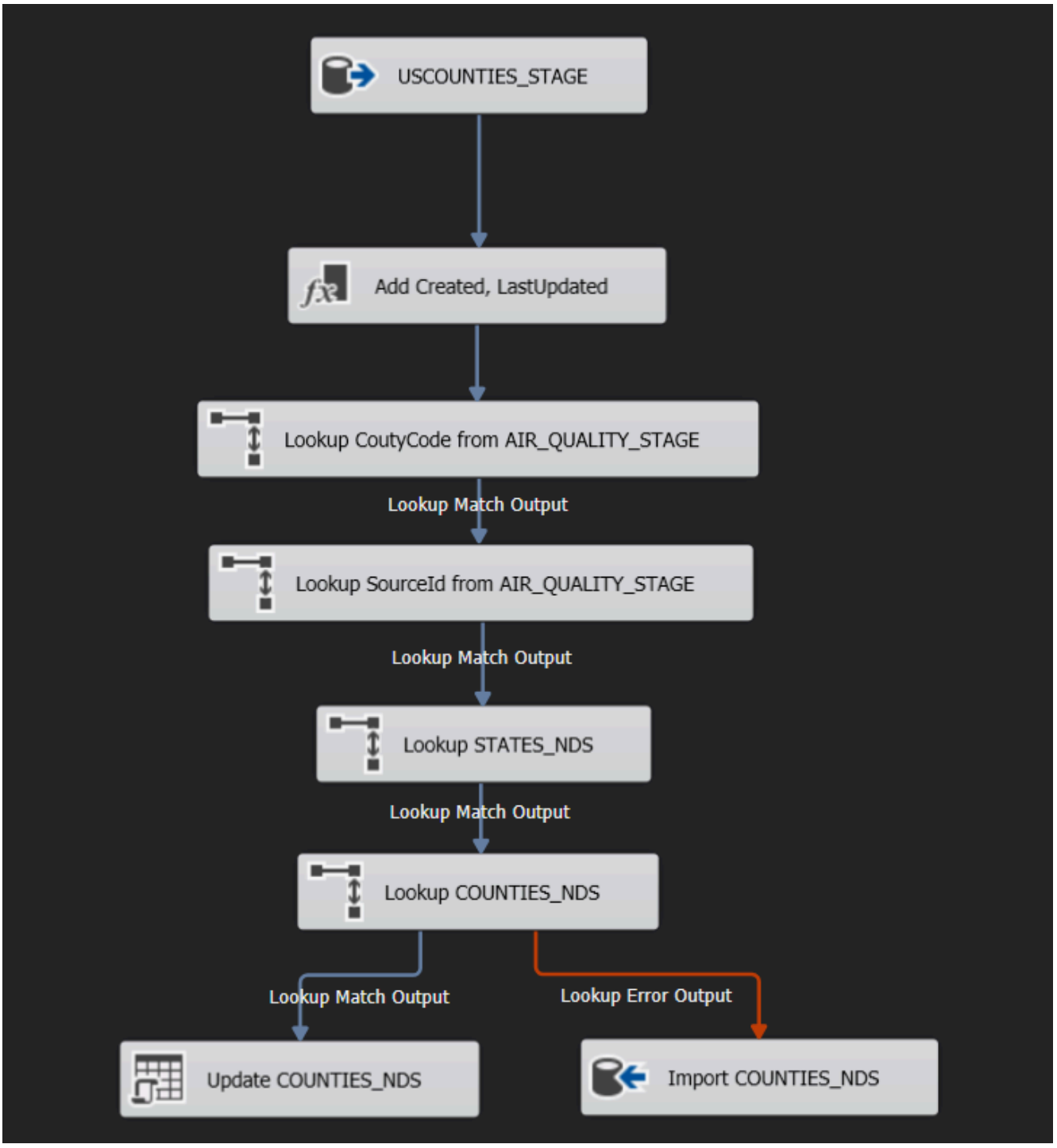
Quy trình xử lý dữ liệu:



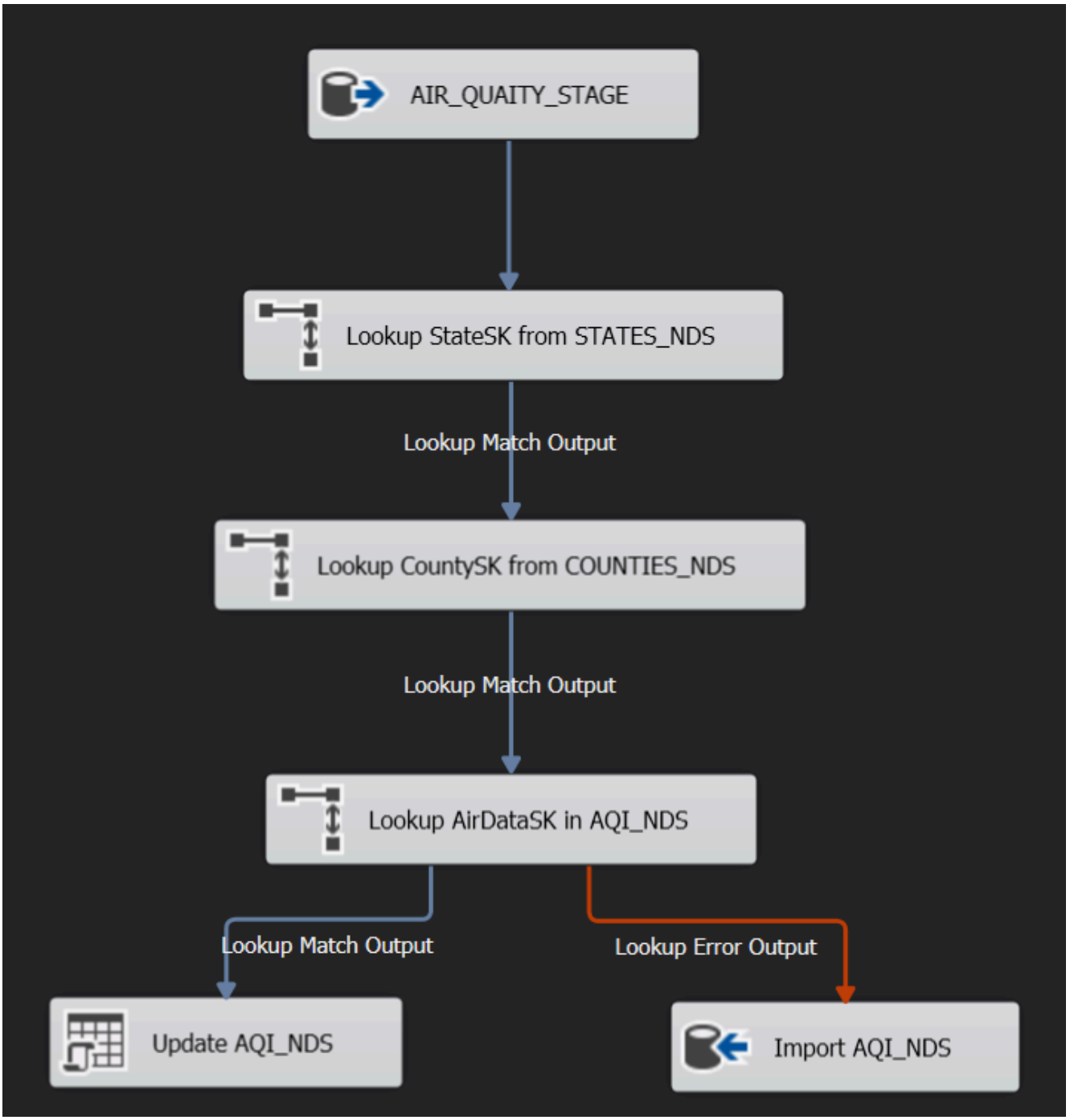
Đây là các Data Flow tổng quát của quá trình, thứ tự nạp dữ liệu cho các bảng lần lượt là: STATES_NDS -> COUNTIES_NDS -> AQI_NDS.



Các component trong DataFlow 'Load STATES_NDS': Đầu tiên lấy dữ liệu trong bảng USCOUNTIES_STAGE, sau đó thêm cột Created, LastUpdated và thực hiện Lookup lấy StateCode và SourceID trong bảng AIR_QUALITY_STAGE. 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.



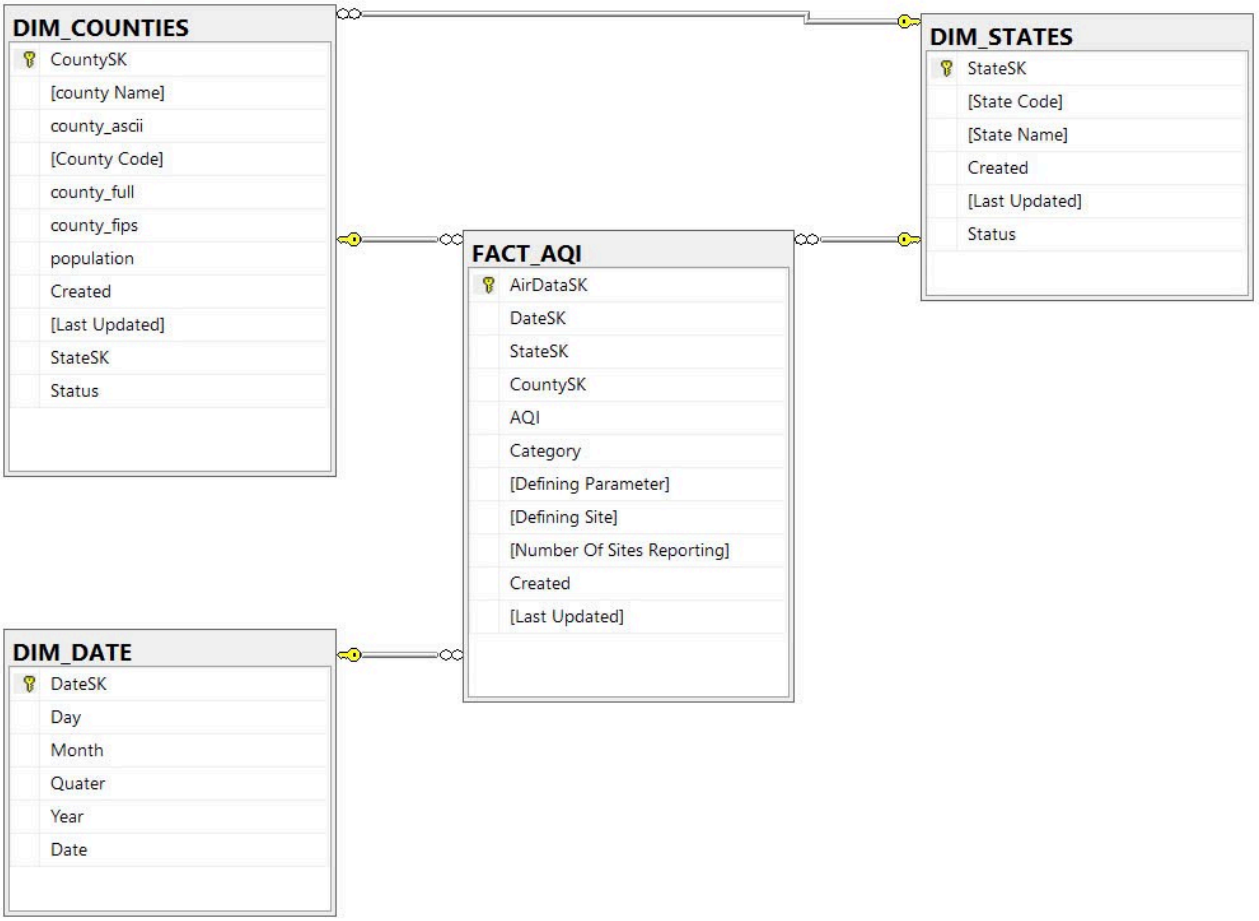
Tương tự với DataFlow 'Load COUNTIES_NDS': Lấy dữ liệu từ bảng USCOUNTIES_STAGE, sau đó thêm cột Created, LastUpdated và thực hiện Lookup lấy CountyCode, SourceID từ bảng AIR_QUALITY_STAGE và StateSK, CountySK từ các bảng STATES_NDS, COUNTIES_NDS. Nếu có dữ liệu rồi thì thực hiện cập nhật lại, còn nếu 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 và thực hiện Lookup lấy StateSK, CountySK từ các bảng STATES_NDS, COUNTIES_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.

3. NDS sang DDS

3.1. Sơ đồ DDS



3.2. Thiết kế database tương ứng

Dựa theo sơ đồ DDS gồm 3 bảng DIM và 1 bảng FACT, script tương ứng cho từng table DIM và FACT:

```
CREATE TABLE [DIM_STATES](
    [StateSK] INT,

    [State Code] int,
    [State Name] VARCHAR(255) NULL,
    [Created] DATETIME NULL,
    [Last Updated] DATETIME NULL,
    [Status] bit NULL,

    CONSTRAINT [PK_DIM_States] primary key clustered([StateSK] ASC)
    WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
```

```
CREATE TABLE [DIM_COUNTIES] (
    [CountySK] int,

    [county Name] varchar(255) NULL,
    [county_ascii] varchar(225) NULL,
    [County Code] int,
    [county_full] varchar(225) NULL,
    [county_fips] int NULL,
    [population] int NULL,
    [Created] datetime NULL,
    [Last Updated] datetime NULL,
    [StateSK] INT,
    [Status] bit NULL,

    CONSTRAINT PK_DIM_COUNTIES PRIMARY KEY CLUSTERED (CountySK ASC)
    WITH(PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dbo].[DIM_DATE](
    [DateSK] int IDENTITY(1,1),

    [Day] int NULL,
    [Month] int NULL,
    [Quater] int NULL,
    [Year] int NULL,
    [Date] datetime NULL,
    CONSTRAINT [PK_DIM_DATE] PRIMARY KEY CLUSTERED ([DateSK] ASC)
    WITH(PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
```

Để thêm dữ liệu vào DIM DATE, thực câu lệnh Procedure để lấy các giá trị năm, quý, tháng, ngày

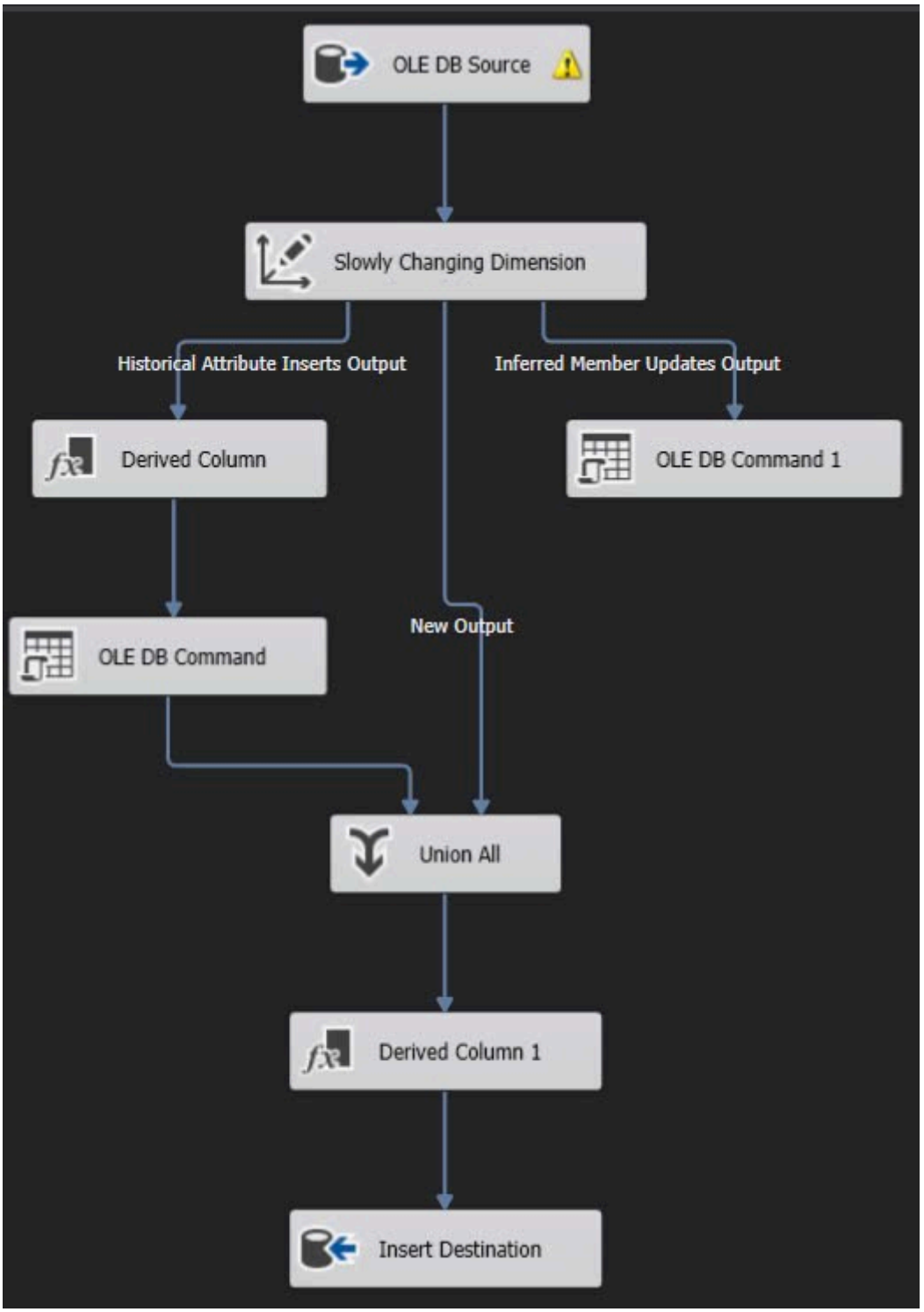
```
CREATE PROCEDURE AddDateTo_DIM_DATE
 AS
BEGIN
    DECLARE @reqStart_date datetime,
            @reqEnd_date datetime,
            @current_date datetime;

    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;
```

Quy trình nạp dữ liệu:

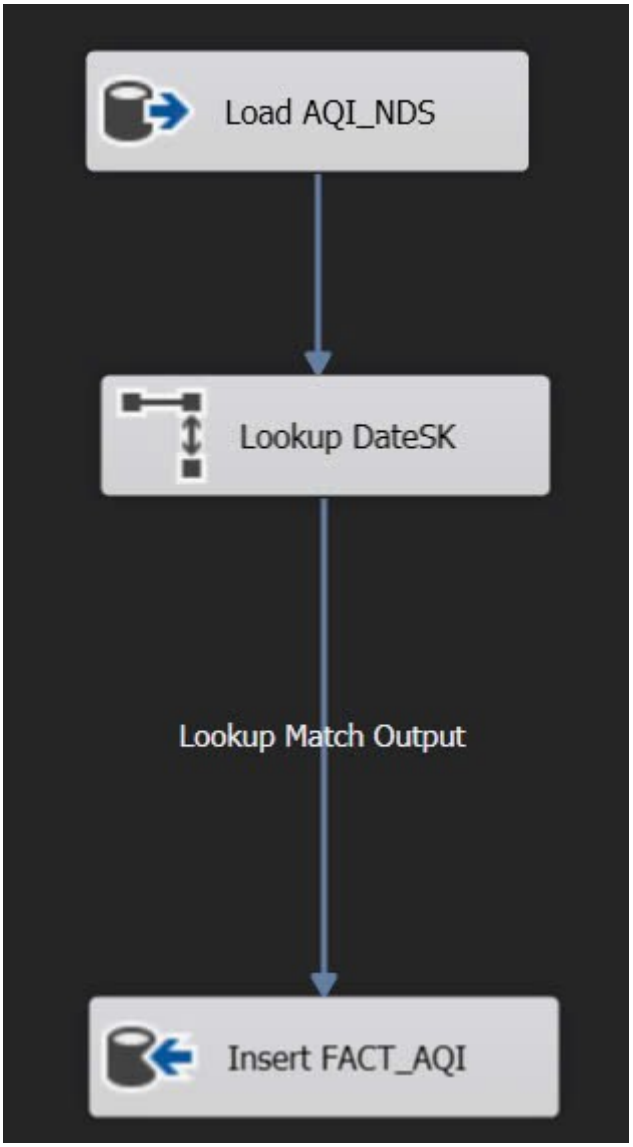
Thực hiện việc truy vấn đến CET và LSET trong dataflow của các bảng DIM COUNTIES, DIM STATES và FACT.

Tiến hành nạp dữ liệu vào các bảng trong DDS:

Ở các bảng dimension, trong các component import thực hiện tuần tự theo các bước sau:



Ở bảng FACT:



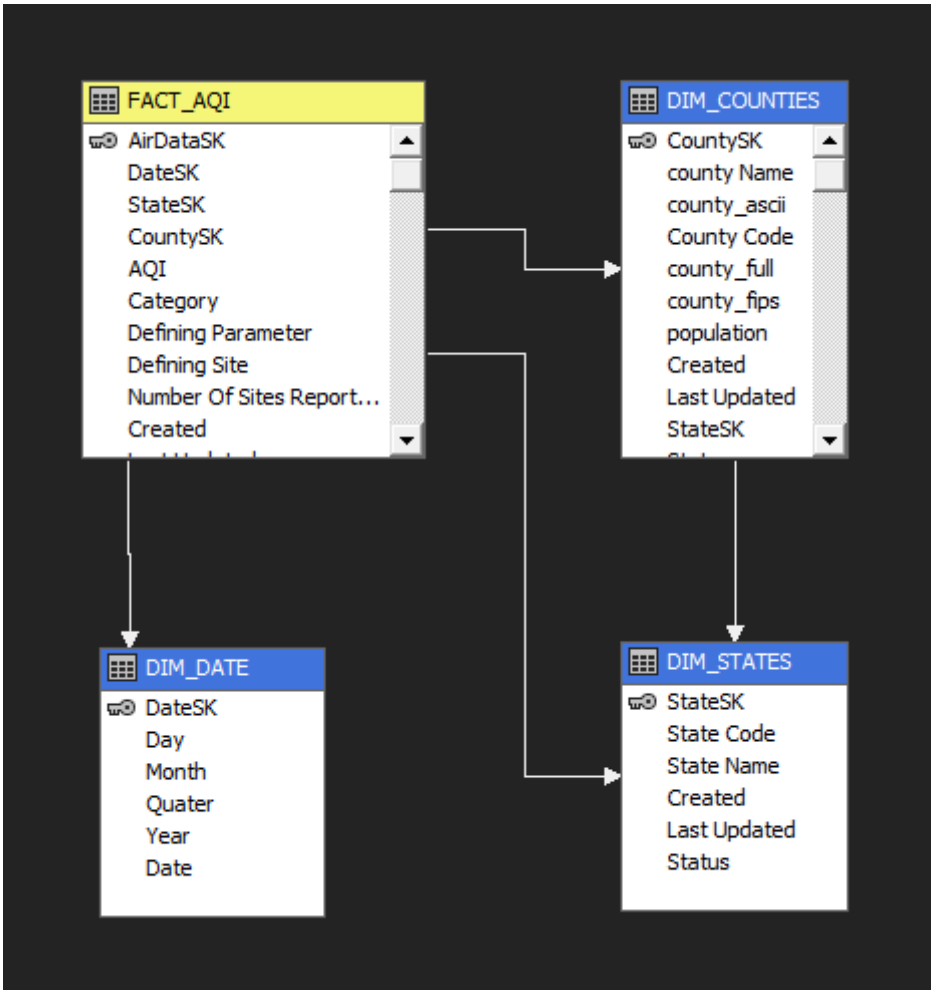
Thực hiện tải dữ liệu từ AQI_NDS, lookup DateSK từ Dim_DATE là khóa ngoại trỏ đến bảng FACT_AQI

Sau đó nhập toàn bộ dữ liệu đã mapping vào trong bảng.

4. OLAP & MDX

4.1. OLAP

4.1.1. Data Source Views



4.2. MDX

4.2.1. Min và Max của giá trị AQI cho mỗi bang trong từng quý của các năm

```
WITH
  MEMBER [Measures].[Min AQI] AS MIN([Date].[Quarter].MEMBERS, [Measures].[AQI])
  MEMBER [Measures].[Max AQI] AS MAX([Date].[Quarter].MEMBERS, [Measures].[AQI])
SELECT
  { [Measures].[Min AQI], [Measures].[Max AQI] } ON COLUMNS,
  { [State].[State Code].MEMBERS } ON ROWS
FROM [AQI Measurements]
WHERE [Date].[Quarter].MEMBERS
```

4.2.2. Mean và độ lệch chuẩn của giá trị AQI cho mỗi bang trong từng quý của các năm

```
WITH
  MEMBER [Measures].[Mean AQI] AS AVG([Date].[Quarter].MEMBERS, [Measures].[AQI])
  MEMBER [Measures].[Std Dev AQI] AS STDDEV([Date].[Quarter].MEMBERS, [Measures].[AQI])
SELECT
  { [Measures].[Mean AQI], [Measures].[Std Dev AQI], [Measures].[Min AQI], [Measures].[Max AQI] } ON COLUMNS,
  { [State].[State Code].MEMBERS } ON ROWS
FROM [AQI Measurements]
WHERE [Date].[Quarter].MEMBERS
```

4.2.3. Số ngày và giá trị AQI trung bình với chất lượng không khí được đánh giá là "rất không lành mạnh" hoặc tệ hơn

```
WITH
  MEMBER [Measures].[Very Unhealthy Days] AS
    COUNT(
      FILTER(
        [Date].[Date].MEMBERS,
        [Measures].[AQI] > 200
      )
    )
  MEMBER [Measures].[Mean Very Unhealthy AQI] AS
    AVG(
      FILTER(
        [Date].[Date].MEMBERS,
        [Measures].[AQI] > 200
      ),
      [Measures].[AQI]
    )
SELECT
  { [Measures].[Very Unhealthy Days], [Measures].[Mean Very Unhealthy AQI] } ON COLUMNS,
  { [State].[State Code].MEMBERS * [County].[County Code].MEMBERS } ON ROWS
FROM [AQI Measurements]
```

4.2.4. Số ngày trong mỗi hạng mục chất lượng không khí (Good, Moderate, etc.) cho các bang Hawaii, Alaska, Illinois và Delaware theo quận

```
SELECT
  NON EMPTY { [Date].[Date].MEMBERS } ON COLUMNS,
  {
    [State].[State Name].&[Hawaii],
    [State].[State Name].&[Alaska],
    [State].[State Name].&[Illinois],
    [State].[State Name].&[Delaware]
  } * [County].[County Name].MEMBERS ON ROWS
FROM [AQI Measurements]
WHERE (
  [AQI Category].[Category].MEMBERS
)
```

4.2.5. Giá trị trung bình AQI theo quý cho các bang Hawaii, Alaska, Illinois và Delaware

```
WITH
  MEMBER [Measures].[Mean AQI] AS AVG([Date].[Quarter].MEMBERS, [Measures].[AQI])
SELECT
  { [Measures].[Mean AQI] } ON COLUMNS,
  {
    [State].[State Name].&[Hawaii],
    [State].[State Name].&[Alaska],
    [State].[State Name].&[Illinois],
    [State].[State Name].&[Delaware]
  } ON ROWS
FROM [AQI Measurements]
WHERE [Date].[Quarter].MEMBERS
```

4.2.6. Báo cáo về xu hướng biến động AQI trong năm cho các bang Hawaii, Alaska, Illinois và California

```
WITH
  MEMBER [Measures].[AQI Trend] AS AVG([Date].[Quarter].MEMBERS, [Measures].[AQI])
SELECT
  { [Measures].[AQI Trend] } ON COLUMNS,
  {
    [State].[State Name].&[Hawaii],
    [State].[State Name].&[Alaska],
    [State].[State Name].&[Illinois],
    [State].[State Name].&[California]
  } ON ROWS
FROM [AQI Measurements]
WHERE [Date].[Quarter].MEMBERS
```

NGUỒN THAM KHẢO

[1] Quy trình ETL: Source - Stage - NDS - DDS. - Quí Nguyễn Phú

https://www.youtube.com/watch?v=UTMpdVfNYV4

[2] Demo Source to Stage to NDS to DDS - Quốc Trần

https://www.youtube.com/watch?v=2B9XiN9yFkA

[3] Change Data Capture Source to Staging

https://docs.varigence.com/bimlflex/delivering-solutions/delivering-scenarios/source-cdc