TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐẠI HỌC QUỐC GIA TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO
ĐỒ ÁN MÔN HỌC

PROJECT ETL SOURCE – STAGE – NDS- DDS, OLAP

MÔN HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH

Giảng viên hướng dẫn : **Hồ Thị Hoàng Vy**
: Tiết Gia Hồng
: Nguyễn Ngọc Minh Châu

Nhóm sinh viên thực hiện : **Nhóm 17**

HỌC KỲ I – NĂM HỌC 2024-2025

THÔNG TIN NHÓM

Mã nhóm	***MSSV***	***Họ và tên***	***Ghi chú***
CQ2021/1 NHÓM 05	20120274	Nguyễn Linh Đăng Dương	Nhóm trưởng
	21120252	Võ Hoàng Nam Hưng	
	21120270	Huỳnh Lê Đăng Khoa	

BẢNG PHÂN CÔNG VÀ ĐÁNH GIÁ CÔNG VIỆC

Công việc thực hiện	***Người thực hiện***	***Mức độ hoàn thành***	***Đánh giá của nhóm***
Thực hiện ETL từ Source -> Stage -> NDS	20120274 Nguyễn Linh Đăng Dương	100%	10/10
Thực hiện ETL từ NDS-> DDS	21120252 Võ Hoàng Nam Hưng	100%	10/10
Thực hiện ETL từ DDS-> OLAP CUBE, viết báo cáo	21120270 Huỳnh Lê Đăng Khoa	100%	10/10

MỤC LỤC

I. Thiết kế databases (NDS, DDS)

1. NDS.

1.1. Cấu trúc Database

Database NDS bao gồm 3 bảng chính:

- STATE: Quản lý thông tin về các bang.
- COUNTY: Lưu trữ thông tin chi tiết về các hạt trong từng bang.
- AQI: Lưu trữ dữ liệu chất lượng không khí (Air Quality Index - AQI) tại từng hạt.

1.2. Bảng STATE

Cột chính:

- StateSK: Khóa chính, là số tự tăng để đảm bảo tính duy nhất.
- StateID: Tên viết tắt đặc trưng cho từng bang.
- StateCode: Mã đặc trưng cho từng bang.
- StateName: Tên đầy đủ của bang.
- CreatedTime: Thời gian tạo bản ghi.
- UpdatedTime: Thời gian cập nhật bản ghi.
- SourceID: ID của nguồn cung cấp dữ liệu.

Ý nghĩa:

- Chuẩn hóa dữ liệu thông tin các bang.
- Làm cha trong mối quan hệ cha-con với bảng COUNTY.

1.3. Bảng COUNTY

Cột chính:

- CountySK: Khóa chính, là số tự tăng duy nhất.
- CountyCode: Mã định danh cho từng hạt (dạng ký tự 5 chữ số).
- CountyName: Tên hạt.
- CountyNameAscii: Tên hạt dạng chuẩn hóa (ASCII).
- CountyFull: Tên đầy đủ của hạt.
- CountyFips: Mã FIPS chuẩn cho từng hạt.
- Latitude: Vĩ độ địa lý
- Longitude: Kinh độ địa lý.
- Population: Dân số tại hạt.
- StateSK: Khóa ngoại tham chiếu đến bảng STATE.

- CreatedTime: Thời gian tạo bản ghi.
- UpdatedTime: Thời gian cập nhật bản ghi.
- SourceID: ID nguồn cung cấp dữ liệu.

Ý nghĩa:

- Lưu trữ thông tin chi tiết của các hạt, liên kết với bang tương ứng.
- Dùng làm điểm nối với bảng AQI để xác định địa điểm cụ thể của dữ liệu không khí.

1.4. Bảng AQI

Cột chính:

- AQIEntrySK: Khóa chính, là số tự tăng duy nhất.
- Date: Ngày ghi nhận dữ liệu.
- AQI: Chỉ số chất lượng không khí.
- Category: Loại ô nhiễm không khí.
- DefiningParameter: Tham số quyết định AQI.
- DefiningSite: Vị trí đo lường cụ thể.
- NumberOfSitesReporting: Số lượng trạm báo cáo dữ liệu.
- CountySK: Khóa ngoại tham chiếu đến bảng COUNTY.
- CreatedTime: Thời gian tạo bản ghi.
- UpdatedTime: Thời gian cập nhật bản ghi.
- SourceID: ID nguồn cung cấp dữ liệu.

Ý nghĩa:

- Lưu trữ dữ liệu về chất lượng không khí theo thời gian.
- Liên kết với COUNTY để xác định vị trí địa lý cụ thể.

1.5. Ý nghĩa tổng thể

Chuẩn hóa dữ liệu:

- Loại bỏ trùng lặp thông tin bang và hạt.
- Mỗi bảng có vai trò rõ ràng, dễ dàng bảo trì và mở rộng.

Hỗ trợ phân tích:

- Dữ liệu có thể được phân tích theo các cấp địa lý khác nhau (bang, hạt) để thực hiện phân cấp chiều dữ liệu sau này dễ dàng hơn.
- Theo dõi xu hướng ô nhiễm không khí theo thời gian, không gian và loại ô nhiễm.

Khả năng tích hợp:

- Sử dụng các mã chuẩn như CountyFips và StateCode để tích hợp dữ liệu từ nhiều nguồn khác nhau.

2. DDS.

2.1. Cấu trúc các bảng

Cơ sở dữ liệu DDS được thiết kế theo mô hình Data Warehouse, bao gồm các bảng Dimension (chiều dữ liệu) và Fact (bảng sự kiện) để hỗ trợ phân tích dữ liệu chất lượng không khí (AQI). Hệ thống này chuẩn hóa dữ liệu từ cơ sở dữ liệu NDS (AirQualityData_NDS) thành các thành phần cấu trúc rõ ràng, dễ bảo trì và mở rộng.

2.2. Bảng DateDimension

Cột chính:

- DateSK: Khóa chính, số tự tăng.

Các cột bổ sung:

- Lưu thông tin thời gian chi tiết như ngày, tháng, quý, năm, và múi giờ mùa hè (DayLightSaving).

Ý nghĩa:

- Chuẩn hóa thông tin thời gian phục vụ phân tích xu hướng dữ liệu AQI theo ngày, tháng, quý, hoặc năm.

2.3. Bảng DefiningParamDim

Cột chính:

- DefParamSK: Khóa chính.

Các cột bổ sung:

- ParaName: Tên tham số xác định AQI.
- CreatedTime, UpdatedTime, SourceID: Thời gian tạo, cập nhật và nguồn dữ liệu.

Ý nghĩa:

- Chuẩn hóa các tham số định nghĩa AQI, như bụi mịn (PM2.5), PM10, CO, NO2, Ozone

2.4. Bảng StateDimension

Cột chính:

- StateSK: Khóa chính.

Các cột bổ sung:

- StateName, StateID: Tên và mã định danh của bang.
- CreatedTime, UpdatedTime, SourceID: Thông tin về thời gian tạo, cập nhật, và nguồn.

Ý nghĩa:

- Lưu trữ thông tin bang, làm cha trong mối quan hệ với bảng CountyDimension.

2.5. Bảng CountyDimension

Cột chính:

- CountySK: Khóa chính.

Các cột bổ sung:

- Các thông tin về hạt như mã, tên đầy đủ, mã FIPS, tọa độ địa lý (vĩ độ, kinh độ), và dân số.
- Liên kết với bảng StateDimension thông qua StateSK.

Ý nghĩa:

- Chuẩn hóa dữ liệu về các hạt, là nền tảng cho việc xác định vị trí địa lý trong dữ liệu AQI.

2.6. Bảng AQICategoryDim

Cột chính:

- CategorySK: Khóa chính.

Các cột bổ sung:

- Mức độ ô nhiễm không khí (LevelsOfConcern), giá trị chỉ số AQI tối thiểu và tối đa, và màu sắc đại diện (DailyAQIColor).

Ý nghĩa:

- Phân loại mức độ ô nhiễm không khí, hỗ trợ phân tích và trực quan hóa dữ liệu.

2.7. Bảng AQIFactTable

Cột chính:

- AQISK: Khóa chính, số tự tăng.

Các cột bổ sung:

- Khóa ngoại tham chiếu đến các bảng Dimension: DateDimension, CountyDimension, DefiningParamDim, và AQICategoryDim.
- Chỉ số AQI thực tế (AQI), thời gian tạo, cập nhật, và nguồn dữ liệu.

Ý nghĩa:

- Lưu trữ dữ liệu AQI chi tiết theo thời gian, địa điểm, loại ô nhiễm, và tham số xác định.

2.8. Ý nghĩa tổng thể

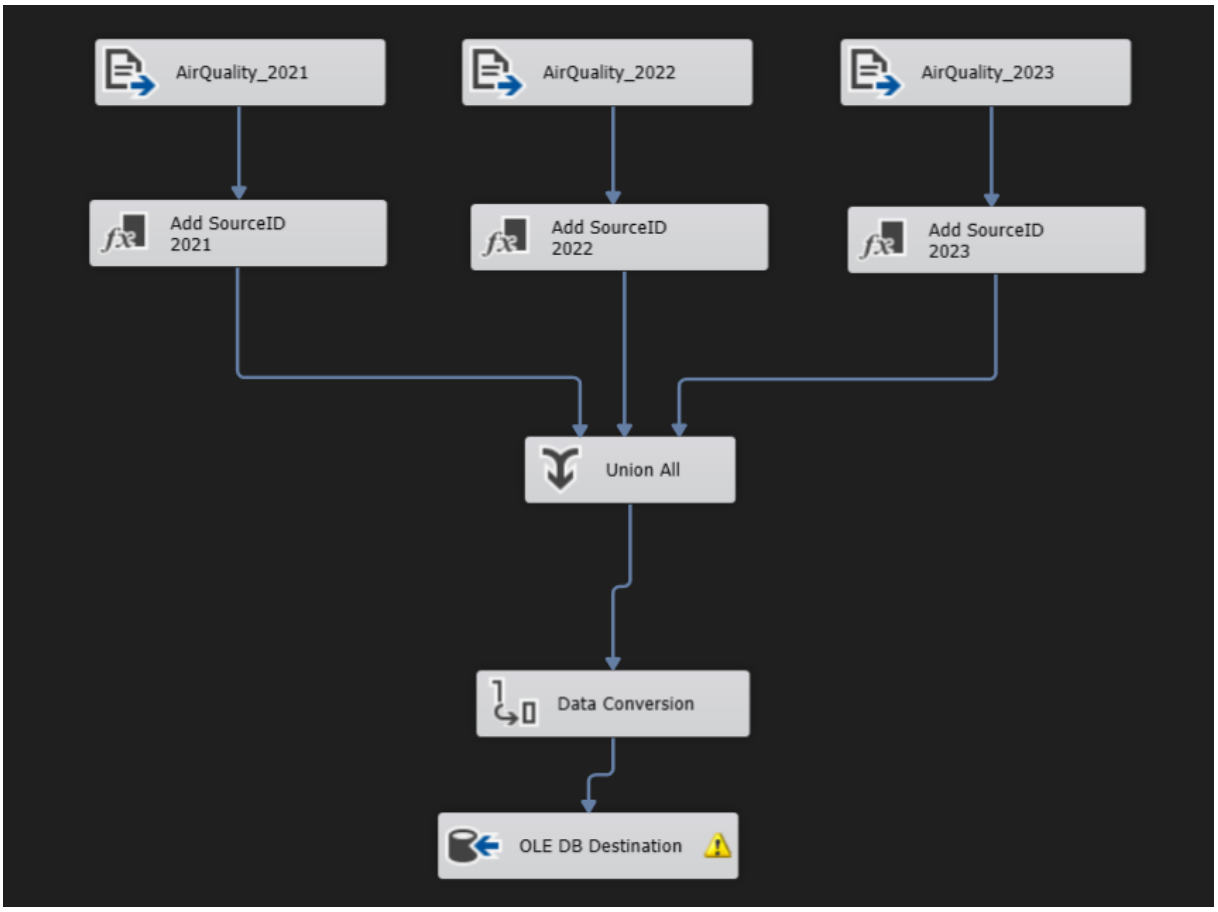
- Chuẩn hóa dữ liệu: Loại bỏ sự trùng lặp và đảm bảo dữ liệu nhất quán.

- Hỗ trợ phân tích: Dễ dàng phân tích dữ liệu theo thời gian, không gian, hoặc loại ô nhiễm không khí.
- Khả năng mở rộng: Cấu trúc dễ mở rộng, hỗ trợ thêm các chiều dữ liệu hoặc thông tin chi tiết.
- Tích hợp dữ liệu: Sử dụng các mã chuẩn như FIPS và StateID, dễ dàng tích hợp dữ liệu từ nhiều nguồn khác nhau.

II. Quá trình thực hiện ETL data (cleaning, transformation, data integration,…)

1. Qúa trình Source – Stage

1.1. AirQualityData



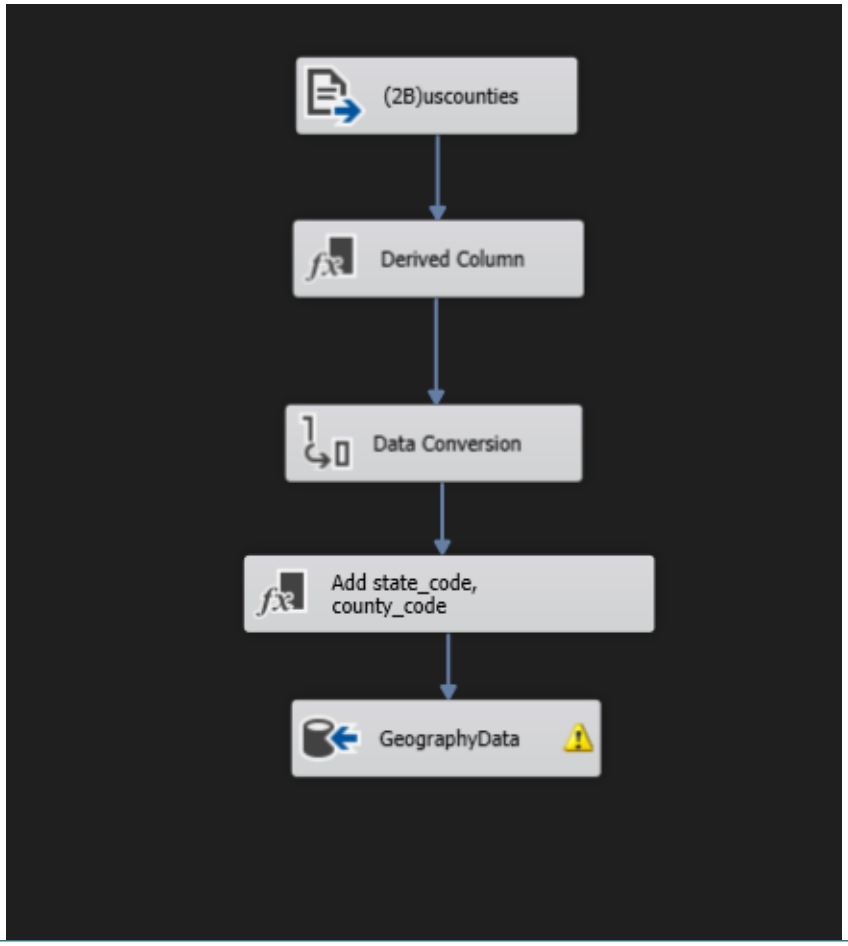
- Nguồn dữ liệu: Bộ dữ liệu chất lượng không khí (Air Quality) từ các năm 2021, 2022, và 2023, được lưu trong 3 file CSV:
  - 10_state_aqi_2021.csv

- 10_state_aqi_2022.csv
- 10_state_aqi_2023.csv

- Các bước thực hiện:

1. Đọc dữ liệu: Nạp dữ liệu từ ba tệp nguồn.
2. Thêm SourceID: Gán giá trị SourceID tương ứng để xác định nguồn dữ liệu:
   - SourceID = 1 cho dữ liệu năm 2021.
   - SourceID = 2 cho dữ liệu năm 2022.
   - SourceID = 3 cho dữ liệu năm 2023.
3. Hợp nhất dữ liệu: Sử dụng phép Union để gộp ba bảng lại với nhau, vì các tệp này có cùng cấu trúc cột.
4. Chuyển đổi kiểu dữ liệu: Chuẩn hóa kiểu dữ liệu của các cột để tương thích với bảng AirQualityData trong cơ sở dữ liệu Stage.
5. Lưu dữ liệu vào Stage: Nạp dữ liệu đã chuẩn hóa vào bảng AirQualityData trong Stage.

1.2. GeographyData



- Nguồn dữ liệu: Tệp dữ liệu địa lý uscounty.csv.
- Các bước thực hiện:

1. Đọc dữ liệu: Nạp dữ liệu từ tệp nguồn.
2. Chuẩn hóa dữ liệu:
   - Loại bỏ các ký tự không cần thiết, ví dụ: dấu ngoặc kép (""), bằng cách sử dụng Derived Column.
   - Chuyển đổi kiểu dữ liệu của các cột để phù hợp với định dạng trong bảng GeographyData của cơ sở dữ liệu Stage.
3. Tách dữ liệu cột:

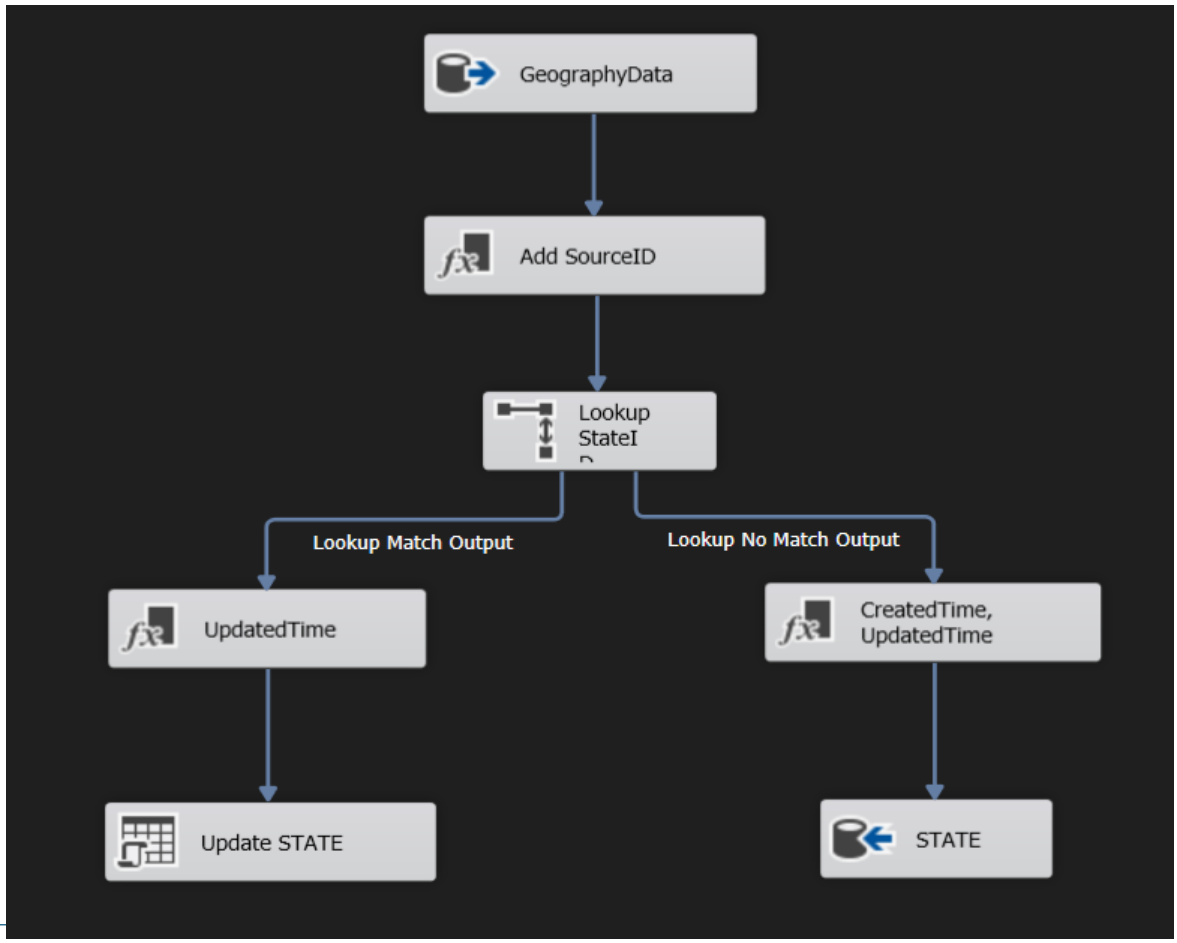
Tách cột county_fips thành hai cột mới:

- state_code = county_fips / 1000
- county_code = county_fips % 1000

4. Lưu dữ liệu vào Stage: Nạp dữ liệu đã chuẩn hóa vào bảng GeographyData.

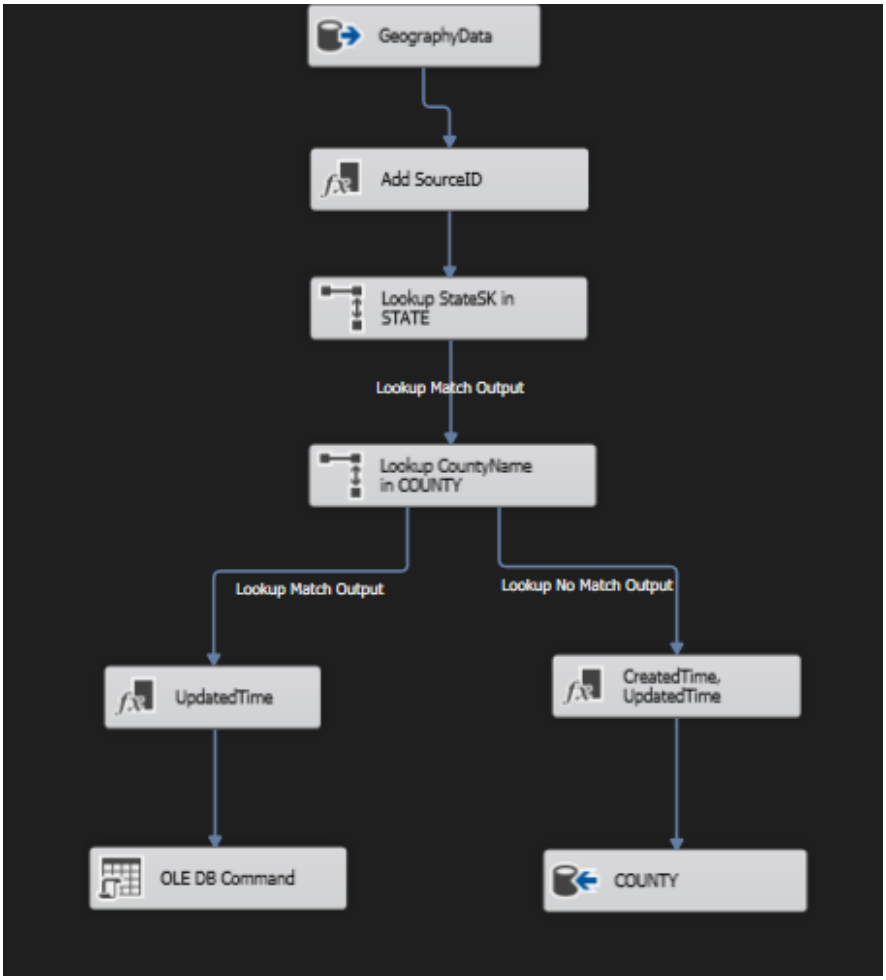
2. Qúa trình Stage – NDS

2.1. Load bảng STATE



- Chọn dữ liệu: Thực hiện SELECT DISTINCT từ bảng GeographyData để loại bỏ các bản ghi trùng lặp, đảm bảo chỉ lấy những dữ liệu duy nhất.
- Thêm SourceID: Gán giá trị SourceID = 4 cho tất cả các bản ghi lấy ra từ bảng GeographyData.
- Kiểm tra sự tồn tại của StateID: Sử dụng phép LookUp để kiểm tra xem StateID đã tồn tại trong bảng STATE chưa.
    - Nếu StateID chưa tồn tại, thực hiện thêm bản ghi mới vào bảng STATE.
    - Nếu StateID đã tồn tại, tiến hành cập nhật các thông tin tương ứng trong bảng STATE.
- Thêm cột thời gian: Với các bản ghi mới thêm vào bảng STATE, tạo thêm hai cột thời gian:
    - CreatedTime: Thời gian bản ghi được tạo mới.
    - UpdatedTime: Thời gian bản ghi được cập nhật lần cuối.
    - Với các bản ghi đã tồn tại (được cập nhật), chỉ cần thêm cột UpdatedTime để ghi lại thời gian cập nhật gần nhất.

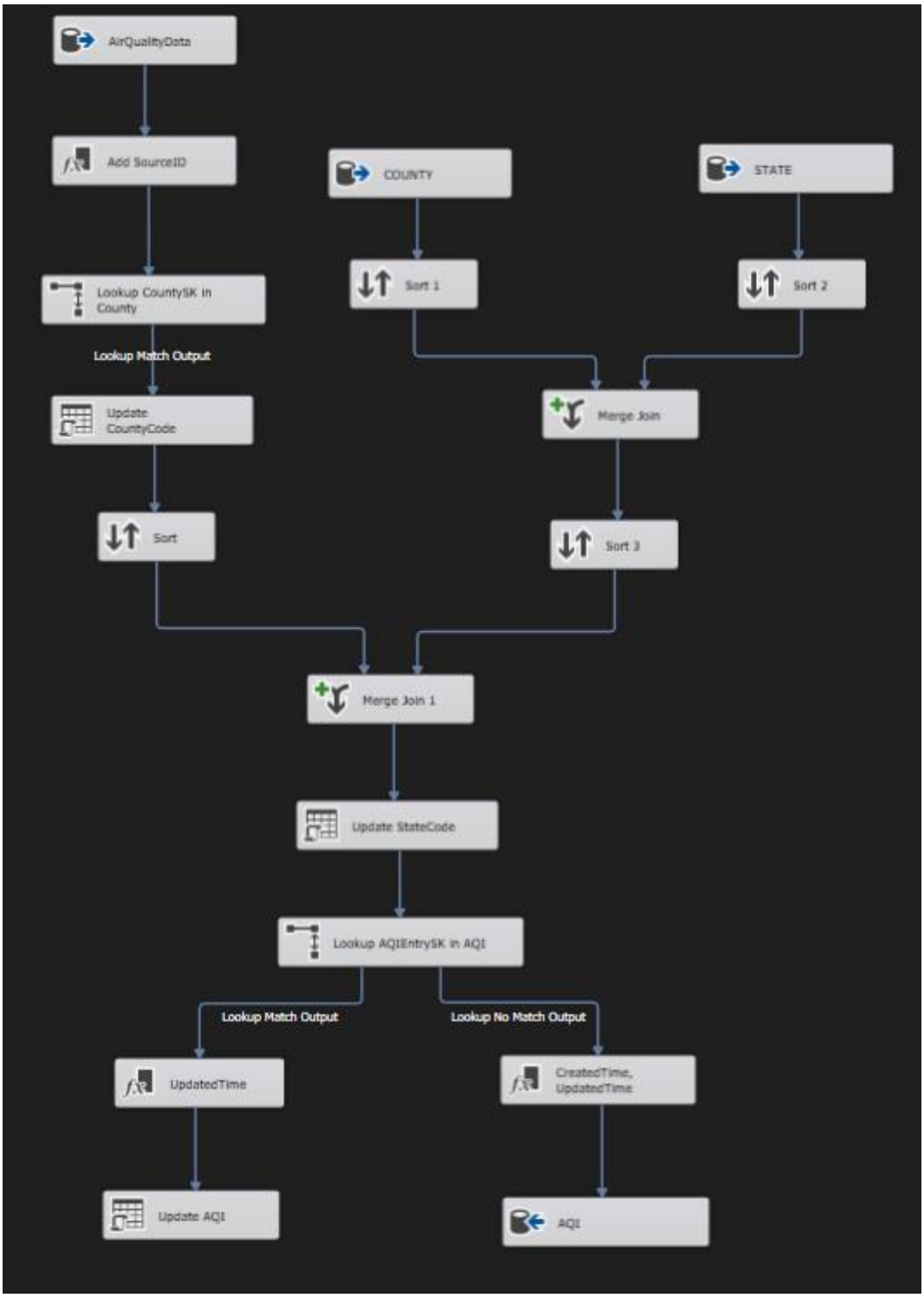
2.2. Load bảng COUNTY



- Chọn dữ liệu: Thực hiện SELECT từ bảng GeographyData.
- Thêm SourceID: Gán giá trị SourceID = 4 cho tất cả các bản ghi lấy ra từ bảng GeographyData.
- Kiểm tra sự tồn tại của StateSK: Sử dụng phép LookUp để kiểm tra xem StateSK đã tồn tại trong bảng STATE chưa. Nếu có rồi thì sẽ tiếp tục
- Kiểm tra sự tồn tại của CountyName trong COUNTY: Sử dụng phép LookUp để kiểm tra xem CountyName đã tồn tại trong bảng COUNTY chưa.
    - Nếu CountyName chưa tồn tại, thực hiện thêm bản ghi mới vào bảng COUNTY.
    - Nếu CountyName đã tồn tại, tiến hành cập nhật các thông tin tương ứng trong bảng COUNTY.
- Thêm cột thời gian: Với các bản ghi mới thêm vào bảng COUNTY, tạo thêm hai cột thời gian:
    - CreatedTime: Thời gian bản ghi được tạo mới.

- UpdatedTime: Thời gian bản ghi được cập nhật lần cuối.
- Với các bản ghi đã tồn tại (được cập nhật), chỉ cần thêm cột UpdatedTime để ghi lại thời gian cập nhật gần nhất.

2.3. Load bảng AQI.

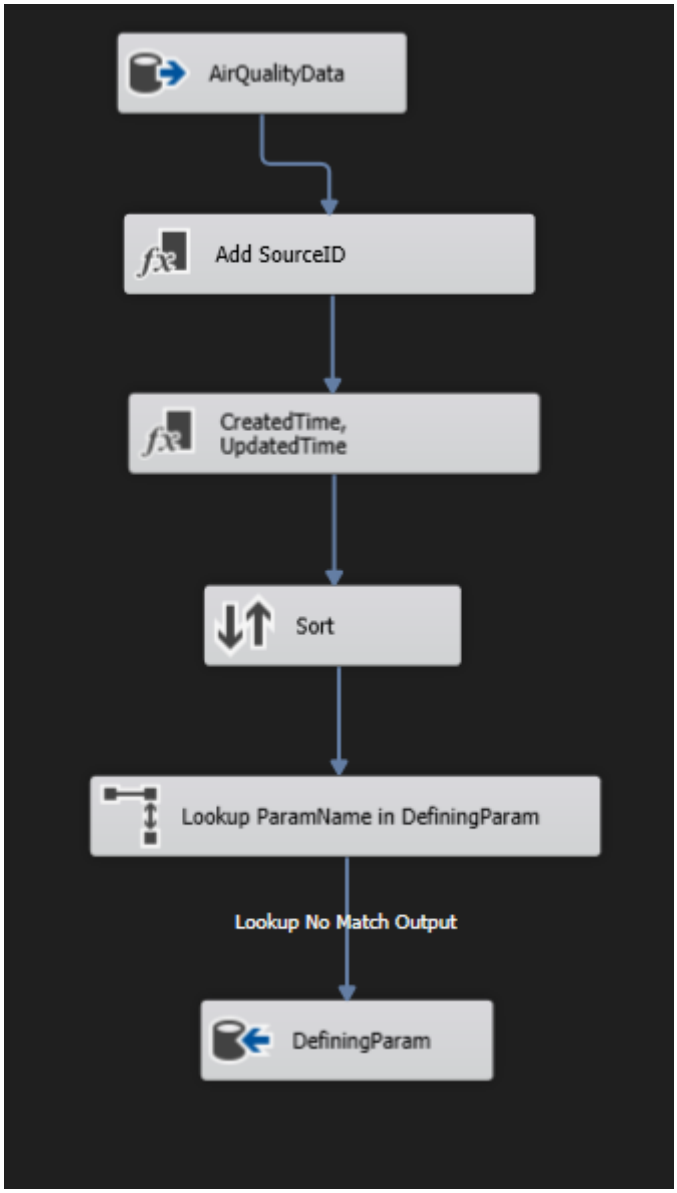


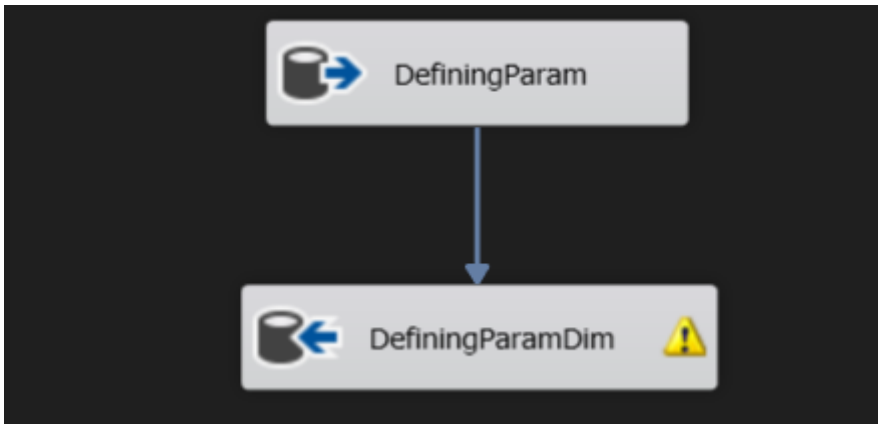
- Dữ liệu ban đầu: Dữ liệu từ bảng AirQualityData trong Stage chứa thông tin về chất lượng không khí.
- Thêm SourceID: Tại bước "Add SourceID", một giá trị SourceID = 4 được thêm vào dữ liệu để xác định nguồn gốc của từng bản ghi.
- Phân chia dữ liệu: Dữ liệu sẽ được phân chia theo các khu vực địa lý như COUNTY và STATE để đảm bảo dữ liệu được phân loại rõ ràng và dễ dàng xử lý.
- Sắp xếp dữ liệu:
  - Các bước "Sort 1" và "Sort 2" được thực hiện để sắp xếp dữ liệu theo các tiêu chí cụ thể, đảm bảo thứ tự hợp lý trước khi tiếp tục quá trình.
- Kết hợp và tiếp tục sắp xếp: Dữ liệu được kết hợp lại thông qua bước "Merge Join" để ghép nối các bản ghi từ các nguồn khác nhau.
- Sau đó, bước "Sort 3" sẽ tiếp tục sắp xếp dữ liệu lần cuối trước khi cập nhật các mã vùng.
- Cập nhật CountyCode và StateCode: Dữ liệu được cập nhật các thuộc tính CountyCode và StateCode từ các thông tin đã được phân chia và kết hợp.
- Hai nhánh xử lý dữ liệu:
  - Nhánh "Lookup Match Output": Cập nhật thuộc tính UpdatedTime cho các bản ghi đã tồn tại trong bảng AQI trong NDS.
  - Nhánh "Lookup No Match Output": Cập nhật cả hai thuộc tính CreatedTime và UpdatedTime cho các bản ghi mới, nhằm ghi lại thời gian tạo mới và cập nhật của bản ghi.
- Cập nhật vào bảng AQI: Cả hai nhánh đều cập nhật dữ liệu vào bảng AQI trong Non-Dimensional Storage (NDS), đảm bảo dữ liệu được chiết xuất, chuyển đổi và tải vào đúng bảng.

3. Qúa trình NDS – DDS

- Get và Set LSET và CET của các bảng trong Database DDS

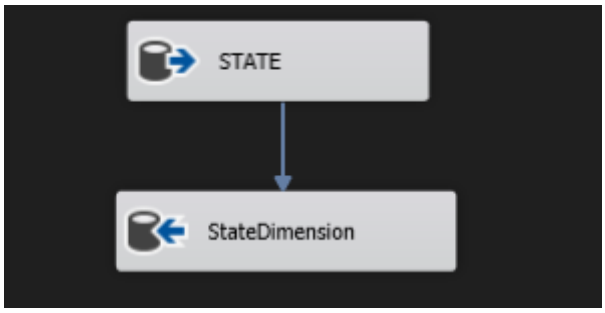
3.1 DefiningParamDim





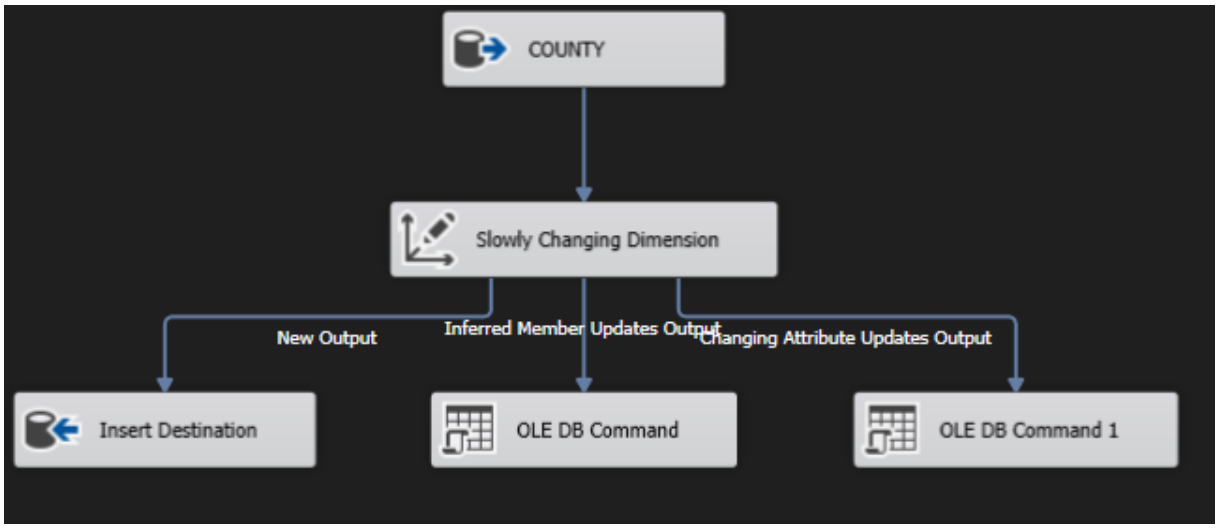
- Dữ liệu ban đầu: Dữ liệu từ bảng DefiningParam trong Database NDS chứa thông tin đã được lọc.
- Lưu dữ liệu vào DefiningParamDim trong Database DDS.

3.2 StateDimension



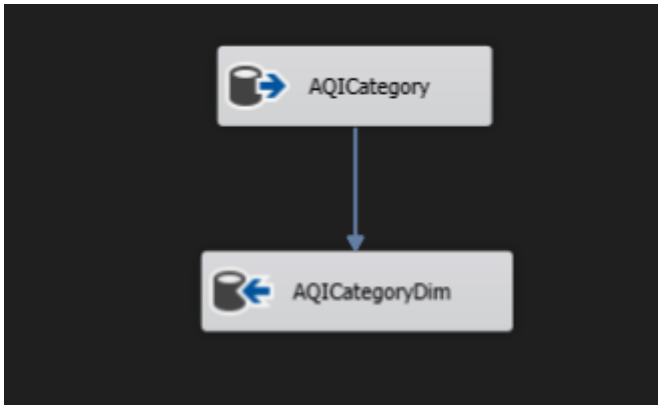
- Dữ liệu ban đầu: Dữ liệu từ bảng STATE trong Database NDS chứa thông tin đã được lọc.
- Lưu dữ liệu vào bảng StateDimension trong Database DDS.

3.3 CountyDimension



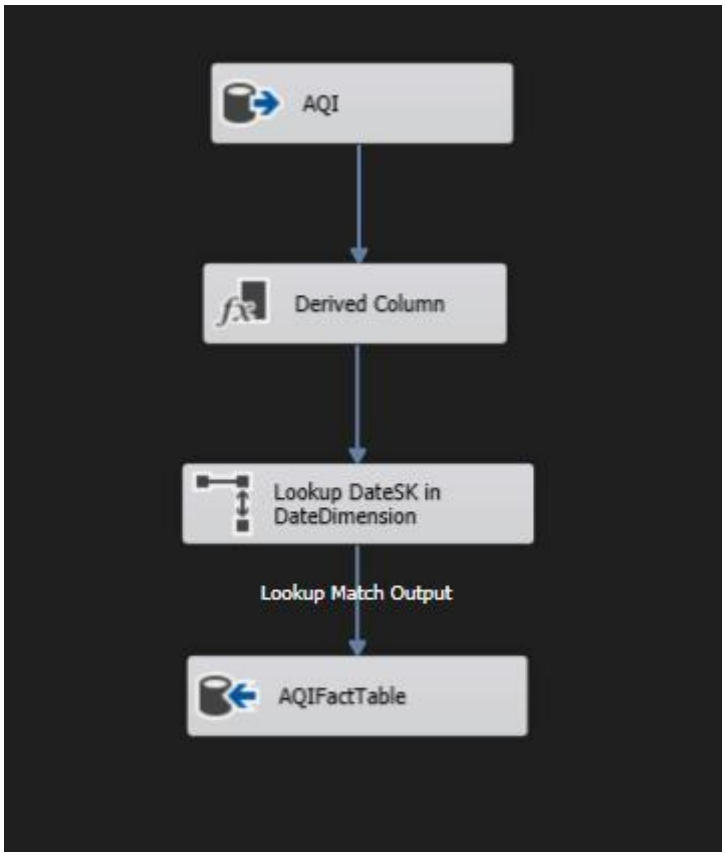
- Dữ liệu ban đầu: Dữ liệu từ bảng STATE trong Database NDS chứa thông tin đã được lọc.
- Thực hiện chiều thay đổi chậm với attribute Population sẽ có 3 nhánh:
    - New Output: với records mới chưa có tồn tại trong bảng dimension sẽ thực hiện insert
    - Inferred Member Updates Output: Record chưa đầy đủ cần được cập nhật thông tin
    - Changing Attribute Updates Output: Record đã có mà attribute đã thay đổi thì sẽ cần được cập nhật

3.4 AQICategoryDim



- Dữ liệu ban đầu: Dữ liệu từ bảng AQICategory trong Database NDS chứa thông tin đã được lọc.
- Lưu dữ liệu vào bảng AQI CategoryDim trong Database DDS.

3.5 AQIFactTable



- Dữ liệu ban đầu: Dữ liệu từ bảng AQI từ trong Database NDS chứa thông tin đã được lọc.
- Chuyển đổi dữ liệu: Chuyển đổi dữ liệu ngày tháng năm đúng định dạng.
- Lookup: Lookup thông tin ngày tháng năm từ bảng DateDimension được tạo bằng script trong SQL Server lấy ra dữ liệu ngày tháng năm.

- Đổ dữ liệu vào bảng AQIFactTable.

III. Quá trình Data mining

IV. Tài liệu tham khảo