**BỘ MÔN HỆ THỐNG THÔNG TIN – KHOA CÔNG NGHỆ THÔNG TIN**

Giảng viên: Hồ Thị Hoàng Vy

Tiết Gia Hồng

Nguyễn Ngọc Minh Châu

Nhóm: Nhóm 14

bài tập môn học  - HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH

HỌC KỲ I – NĂM HỌC 2024 -2025

**HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM**

**BẢNG THÔNG TIN CHI TIẾT NHÓM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mã nhóm:** | **CQ.BI.2425.14** | | | |
| **Tên nhóm:** | **Nhóm 14** | | | |
| **Số lượng:** | **4 thành viên** | | | |
| **MSSV** | **Họ tên** | **Email** | **Điện thoại** | **Hình ảnh** |
| 20120424 | Dương Khánh An | [20120424@student.hcmus.edu.vn](mailto:20120093@student.hcmus.edu.vn) |  |  |
| 20120524 | Võ Đức Lợi | 20120524@student.hcmus.edu.vn |  |  |
| 20120632 | Trần Thái Vỹ | 20120632@student.hcmus.edu.vn |  |  |
| 19120048 | Hồ Nguyễn Trâm Anh | 19120048@student.hcmus.edu.vn |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Bảng phân công & đánh giá hoàn thành công việc** | | | |
| **Công việc thực hiện** | **Người thực hiện** | **Mức độ hoàn thành** | **Đánh giá của nhóm** |
| Xác định các transformations cần thực hiện để chuyển dữ liệu từ Stage sang NDS  Tạo ETL NDS -> DDS  Thao tác với Cube, truy vấn MDX | Dương Khánh An | 100% | 100% |
| Tạo ETL NDS -> DDS  Thao tác với Cube, truy vấn MDX | Trần Thái Vỹ | 100% | 100% |
| Tạo ETL Source -> Stage, Stage -> NDS  Thao tác với Cube, truy vấn MDX | Võ Đức Lợi | 100% | 100% |
| Xác định và phân tích cấu trúc CSDL NDS  Phân tích thiết kế metadata.  Kiểm thử, tổng hợp và viết báo cáo các giai đoạn | Hồ Nguyễn Trâm Anh | 100% | 100% |

**MỤC LỤC**

[**A.** **Yêu cầu của Đồ án/Bài tập** 4](#_Toc185382552)

[**B.** **Kết quả** 5](#_Toc185382553)

[**1.** **Source:** 5](#_Toc185382554)

[**2.** **Stage:** 7](#_Toc185382555)

[**3.** **NDS:** 7](#_Toc185382556)

[**4.** **DDS:** 9](#_Toc185382557)

[**4.1.** **Các bảng dimension:** 9](#_Toc185382558)

[**4.2.** **Bảng Fact:** 10](#_Toc185382559)

[**4.3.** **Phân tích thiết kế các Measure:** 11](#_Toc185382560)

[**5.** **ETL:** 13](#_Toc185382561)

[**5.1.** **ETL từ Soure vào Stage:** 13](#_Toc185382562)

[a. ETL County từ Source vào StageDB 13](#_Toc185382563)

[b. ETL AQI Measurements từ Source vào StageDB 18](#_Toc185382564)

[c. Kết quả chạy: 23](#_Toc185382565)

[**5.2.** **ETL từ Stage vào NDS:** 25](#_Toc185382566)

[a. ETL dữ liệu vào bảng State trong NDSDB: 25](#_Toc185382567)

[b. ETL dữ liệu vào bảng County trong NDSDB: 31](#_Toc185382568)

[c. ETL dữ liệu vào bảng AQI NDS: 37](#_Toc185382569)

[d. Kết quả chạy: 45](#_Toc185382570)

[**5.3.** **ETL từ NDS vào DDS:** 48](#_Toc185382571)

[a. ETL dữ liệu vào bảng Dim\_State: 48](#_Toc185382572)

[b. ETL dữ liệu vào bảng Dim\_County: 54](#_Toc185382573)

[c. ETL dữ liệu vào bảng Dim\_ Category: 63](#_Toc185382574)

[d. ETL dữ liệu vào bảng Dim\_Parameter: 70](#_Toc185382575)

[e. ETL dữ liệu vào bảng Dim\_Date: 77](#_Toc185382576)

[f. ETL dữ liệu vào bảng Dim\_FactAQI: 84](#_Toc185382577)

[g. Kết quả chạy: 100](#_Toc185382578)

[**6.** **OLAP:** 105](#_Toc185382579)

[**7.** **Link Github và video demo:** 110](#_Toc185382580)

**YÊU CẦU ĐỒ ÁN- BÀI TẬP**

|  |  |
| --- | --- |
| **Loại bài tập** | **□ Lý thuyết 🗹 Thực hành 🗹 Đồ án □ Bài tập** |
| **Ngày bắt đầu** | **10/12/2024** |
| **Ngày kết thúc** | **20/12/2024** |

# **Yêu cầu của Đồ án/Bài tập**

Giai đoạn 1: Thực hiện ETL Soure -> Stage -> NDS

Yêu cầu bài nộp:

* Đặt tên: MaNhom\_ETL
* Báo cáo phân tích CSDL: NDS, metadata
* Xác định các transformation
* Source ETL project
* Script tạo csdl
* Video giải thích từng component trong ETL process

Giai đoạn 2: Thực hiện NDS -> DDS, OLAP

Yêu cầu bài nộp:

* Đặt tên: MaNhom\_DDS
* Báo cáo phân tích CSDL: NDS, metadata, DDS
* Xác định các transformation: phân cấp chiều, mô hình hóa DDS
* Source ETL project, SSAS project (build OLAP cube)
* Script tạo csdl: stage, NDS, DDS, metadata
* Video giải thích từng component trong ETL process (NDS -> DDS, OLAP Cube)

Giai đoạn 3: MDX, reporting

Yêu cầu bài nộp:

* Đặt tên: MaNhom\_DDS
* Script MDX (có comment nội dung câu hỏi trước MDX script của từng câu)
* Sử dụng powerBI visualize các yêu cầu trong đồ án
* SSAS project
* Video giải thích.

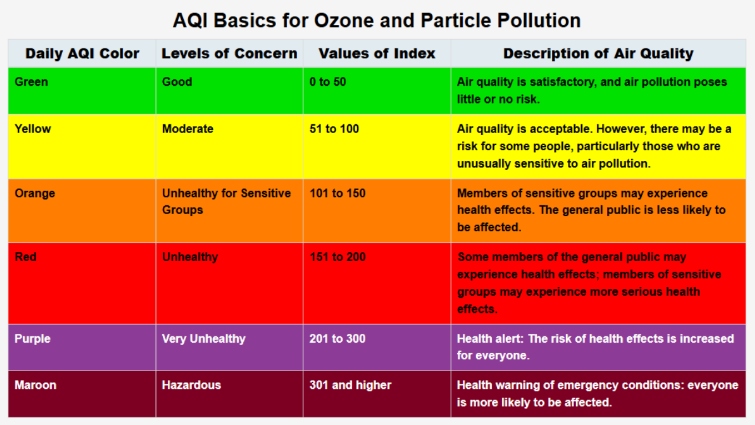
# **Kết quả**

1. **Source:**

Mô tả soure data:

Dữ liệu nguồn được lấy từ Báo cáo tóm tắt hàng ngày về chất lượng không khí (AQI) theo các Hạt của Cơ quan Bảo vệ Môi trường (EPA) trên 10 Tiểu bang Hoa Kỳ trong 3 năm (2021-2023)

* Gồm 3 file .cvs (10\_state\_aqi\_2021, 10\_state\_aqi\_2022, 10\_state\_aqi\_2023) với các cột thuộc tính như sau:
* State Name (Tên Bang)
* County Name (Tên Hạt)
* State Code (Mã code Bang)
* County Code (Mã code Hạt)
* Date (Ngày ghi nhận)
* AQI (Chỉ số AQI)
* Category (Loại chất lượng không khí)
* Defining Parameter (Tham số xác định AQI)
* Defining Site (Địa điểm xác định)
* Number of Sites Reporting (Số lượng địa điểm báo cáo)
* Created (Ngày tạo dữ liệu)
* Last Updated (Lần cuối cập nhật dữ liệu)
* Một bảng dữ liệu phân loại AQI:



* Ngoài ra còn có 1 file .csv ((2B)uscounties.csv) chứa dữ liệu về địa lý với các cột thuộc tính sau:
* county (Tên Hạt)
* county\_ascii (Tên ASCII của Hạt)
* county\_full (Tên đầy đủ của Hạt)
* county\_fips (Mã fips của Hạt)
* state\_id (Tên viết tắt của Bang)
* state\_name (Tên Bang)
* lat (Vĩ độ)
* lng (Kinh độ)
* population (Dân số)

1. **Stage:**

Cấu trúc của các bảng trong Stage tương tự như các bảng trong các file .csv ở Source

|  |  |  |  |
| --- | --- | --- | --- |
| **County** | **Lưu trữ thông tin của Hạt** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **county\_fips** | **Mã fips của Hạt** | **INT** | **PRIMARY KEY** |
| county | Tên Hạt | VARCHAR(50) |  |
| county\_asii | Tên ASCII của Hạt | VARCHAR(5) |  |
| county\_full | Tên đầy đủ của Hạt | VARCHAR(100) |  |
| state\_id | Tên Bang | VARCHAR(5) |  |
| state\_name | Tên Bang | VARCHAR(50) |  |
| Latitude | Vĩ độ | FLOAT |  |
| Longitude | Kinh độ | FLOAT |  |
| Populations | Dân số | INT |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **AQI\_Measurement** | **Lưu trữ thông tin của các lần đo AQI** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| state\_name | Tên Bang | VARCHAR(50) |  |
| county\_name | Tên Hạt | VARCHAR(50) |  |
| **state\_code** | **Mã code Bang** | **VARCHAR(50)** | **PRIMARY KEY** |
| **county\_code** | **Mã code Hạt** | **VARCHAR(50)** | **PRIMARY KEY** |
| date | Ngày ghi nhận AQI | DATE |  |
| AQI | Chỉ số AQI | INT |  |
| category | Loại AQI | VARCHAR(50) |  |
| defining\_parameter | Tham số xác định AQI | VARCHAR(50) |  |
| defining\_site | Địa điểm đo | VARCHAR(50) |  |
| num\_sites\_reporting | Số lượng địa điểm báo cáo | INT |  |
| **created** | **Ngày tạo** | **DATETIME** | **PRIMARY KEY** |
| **last\_updated** | **Lần cập nhật gần nhất** | **DATETIME** | **PRIMARY KEY** |

1. **NDS:**

Sau khi chuẩn hóa các bảng từ Stage ta được cơ sở dữ liệu như bên dưới:

|  |  |  |  |
| --- | --- | --- | --- |
| **State** | **Lưu trữ thông tin bang** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **state\_id** | **id Bang. Mỗi Bang có 1 id duy nhất dùng để định danh Bang** | **INT IDENTITY(1,1)** | **PRIMARY KEY** |
| state\_code | Mã code Bang. Mỗi Bang có 1 mã code duy nhất | VARCHAR(2) | NOT NULL UNIQUE |
| state\_name | Tên Bang | VARCHAR(255) | NOT NULL |
| state\_abbr | Tên viết tắt của Bang | VARCHAR(2) | NOT NULL |

|  |  |  |  |
| --- | --- | --- | --- |
| **County** | **Lưu trữ thông tin Hạt** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **county\_id** | **id Hạt. Mỗi Hạt có 1 id duy nhất dùng để định danh Hạt** | **INT IDENTITY(1,1)** | **PRIMARY KEY** |
| county\_fips | Mã fips của Hạt. Mỗi Hạt có 1 mã fips duy nhất | VARCHAR(5) | NOT NULL UNIQUE |
| county\_name | Tên Hạt | VARCHAR(255) |  |
| county\_ascii | Tên ASCII của Hạt | VARCHAR(255) |  |
| county\_full | Tên đầy đủ của Hạt | VARCHAR(255) |  |
| county\_code | Mã code Hạt | VARCHAR(3) | NOT NULL |
| lat | Vĩ độ | DECIMAL(9, 6) |  |
| lng | Kinh độ | DECIMAL(9, 6) |  |
| population | Dân số | INT |  |
| *state\_code* | *Mã code Bang. Khóa ngoại, tham chiếu đến state\_code của bảng State* | *VARCHAR(2)* | *FOREIGN KEY NOT NULL* |
| created\_date | Ngày tạo dữ liệu | DATETIME DEFAULT GETDATE() |  |
| updated\_date | Ngày cập nhật dữ liệu | DATETIME DEFAULT GETDATE() |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **AirQuality\_Measurement** | **Lưu trữ các lần đo lường chỉ số chất lượng không khí** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **measurement\_id** | **id mỗi lần đo lường AQI. Mỗi lần đo sẽ có 1 id duy nhất định danh lần đo đó** | **INT IDENTITY(1,1)** |  |
| county\_code | Mã code Hạt | VARCHAR(3) |  |
| date | Ngày ghi nhận AQI | DATE |  |
| aqi | Giá trị AQI đo được | INT |  |
| defining\_parameter | Chỉ số xác định AQI | VARCHAR(255) |  |
| defining\_site | Địa điểm xác định | VARCHAR(255) |  |
| num\_sites\_reporting | Số lượng địa điểm báo cáo | INT |  |
| *category\_id* | *Mã loại chất lượng không khí. Khóa ngoại tham chiếu đến category\_id của bảng Category* | *INT* | *FOREIGN KEY NOT NULL* |
| *county\_fips* | *Mã fips Hạt. Khóa ngoại tham chiếu đến county\_fips của bảng County* | *VARCHAR(50)* | *FOREIGN KEY NOT NULL* |
| created\_date | Ngày tạo dữ liệu | DATETIME DEFAULT GETDATE() |  |
| updated\_date | Ngày cập nhật dữ liệu | DATETIME DEFAULT GETDATE() |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **AirCategory** | **Lưu trữ các loại chất lượng không khí** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **category\_id** | **Khóa chính, mỗi loại chất lượng không khí có 1 category\_id duy nhất** | **INT IDENTITY(1,1)** | **PRIMARY KEY** |
| catagory\_name | Tên loại chất lượng không khí | VARCHAR(255) | NOT NULL |
| lower\_bound | Mức AQI tối thiểu cho loại chất lượng không khí | INT | NULL |
| upper\_bound | Mức AQI tối đa cho loại chất lượng không khí | INT | NULL |
| aqi\_color | Màu sắc AQI | VARCHAR(255) | NULL |
| description | Mô tả AQI | VARCHAR(255) | NULL |

1. **DDS:**

### **Các bảng dimension:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dim\_State** | **Chiều bang** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **state\_id** | **Khóa chính. Mỗi bang có 1 id duy nhất định danh bang** | **INT** | **PRIMARY KEY** |
| state\_code | Mã code Bang. Mỗi bang có 1 mã code duy nhất | VARCHAR(2) | NOT NULL UNIQUE |
| state\_name | Tên bang | VARCHAR(255) | NOT NULL |
| state\_abbr | Tên viết tắt của bang | INT | NOT NULL |

|  |  |  |  |
| --- | --- | --- | --- |
| **Dim\_County** | **Chiều hạt** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **county\_id** | **Khóa chính. Mỗi hạt có 1 id duy nhất định danh hạt** | **INT** | **PRIMARY KEY** |
| county\_name | Tên hạt | VARCHAR(255) | NOT NULL |
| county\_fips | Mã fip hạt. Mỗi hạt có 1 mã fip duy nhất | VARCHAR(5) | NOT NULL UNIQUE |
| *state\_id* | *Tên viết tắt của bang. Khóa ngoại tham chiếu đến state\_id của bảng DimState* | *INT* | *NOT NULL* |

|  |  |  |  |
| --- | --- | --- | --- |
| **Dim\_Date** | **Chiều ngày** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **date\_id** | **Khóa chính. Mỗi ngày có 1 id duy nhất định danh hạt** | **INT** | **PRIMARY KEY** |
| date | Ngày tháng năm | DATE | NOT NULL |
| quater | Quý | INT | NULL |
| month | Tháng | INT | NULL |
| year | Năm | INT | NULL |
| day | Ngày | INT | NULL |
| daylightsaving | Số ngày chiếu sáng.  0: false, 1: true | BIT | NOT NULL |

|  |  |  |  |
| --- | --- | --- | --- |
| **Dim\_DefiningParameter** | **Chiều tham số** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **parameter\_id** | **Khóa chính. Mỗi tham số có 1 id duy nhất định danh** | **INT IDENTITY(1,1)** | **PRIMARY KEY** |
| parameter\_name | Tên tham số | VARCHAR(255) |  |

### **Bảng Fact:**

|  |  |  |  |
| --- | --- | --- | --- |
| **AQI\_Fact** | **Bảng Fact AQI** | | |
| **Thuộc tính** | **Mô tả** | **Kiểu dữ liệu** | **Ràng buộc** |
| **fact\_id** | **Khóa chính. Mỗi AQI sẽ có 1 id duy nhất định danh** | **INT IDENTITY(1,1)** | **PRIMARY KEY** |
| *county\_id* | *Id Hạt. Khóa ngoại tham chiếu đến county\_id bảng Dim\_County* | *INT* | *NOT NULL* |
| *date\_id* | *Id ngày tháng năm. Khóa ngoại tham chiếu đến date\_id bảng Dim\_Date* | *INT* | *NOT NULL* |
| *category\_id* | *Id loại AQI. Khóa ngoại tham chiếu đến category\_id bảng Dim\_Caterogy* | *INT* | *NOT NULL* |
| *parameter\_id* | *Id tham số. Khóa ngoại tham chiếu đến parameter\_id bảng Dim\_Parameter* | *INT* | *NOT NULL* |
| mean\_aqi | Giá trị trung bình | FLOAT | NOT NULL |
| std\_aqi | Độ lệch chuẩn | FLOAT | NOT NULL |
| min\_aqi | Giá trị nhỏ nhất của AQI | INT | NOT NULL |
| max\_aqi | Giá trị lớn nhất của AQI | INT | NOT NULL |
| count\_day | Số ngày có loại AQI là “Very Unhealthy” | INT | NOT NULL |
| sum\_aqi | Tổng AQI | INT | NOT NULL |
| sum\_squares\_aqi | Căn bậc 2 tổng AQI | FLOAT | NOT NULL |
| count\_aqi | Số ngày trong từng loại AQI | INT | NOT NULL |

### **Phân tích thiết kế các Measure:**

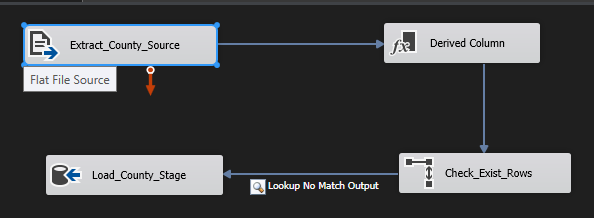
* min\_aqi/max\_aqi: Min/Max AQI theo bang và quý
* What: Min và Max AQI
* Where: Từng bang (State), phân cấp từ quận/huyện (County)
* When: Theo từng quý trong năm
* Why: Để hiểu rõ sự dao động AQI giữa các bang
* Dimension cần:
* Dim\_Date: Chứa hierarchy Year > Quarter > Month > Day
* Dim\_County: Phân cấp đến State
* Measure: Min(AQI\_Value), Max(AQI\_Value)
* Granularity (độ mịn): Một bản ghi tương ứng với AQI của từng quận trong từng ngày
* mean\_aqi và std\_aqi: Mean và StdDev AQI theo bang và quý
* What: Giá trị trung bình và độ lệch chuẩn AQI
* Where: Từng bang
* When: Theo từng quý
* Why: Để đánh giá sự biến động AQI giữa các bang
* Dimension cần: Dim\_Date, Dim\_County.
* Measure: AVG(AQI\_Value), STDDEV(AQI\_Value)
* Granularity: Một bản ghi tương ứng với AQI của từng quận trong từng ngày
* count\_day: Số ngày và Mean AQI khi "Very Unhealthy"
* What: Số ngày (Days\_Count) và Mean AQI
* Where: Từng bang và quận
* When: Không cần phân chia theo thời gian cụ thể
* Why: Để hiểu mức độ nghiêm trọng của tình trạng ô nhiễm
* Dimension cần: Dim\_County, Dim\_Category (nếu cần phân loại chi tiết "Very Unhealthy")
* Measure: COUNT(Days\_Count), Mean(AQI\_Value)
* Granularity: Một bản ghi tương ứng với quận/ngày, kèm loại "Very Unhealthy"
* count\_aqi: Số ngày theo từng loại AQI của các bang cụ thể
* What: Số ngày trong từng loại AQI (Good, Moderate,...)
* Where: 4 bang: Hawaii, Alaska, Illinois, Delaware
* When: Không phân chia thời gian
* Why: So sánh sự phân phối chất lượng không khí giữa các bang
* Dimension cần: Dim\_Category, Dim\_County
* Measure: COUNT(Days\_Count)
* Granularity: Một bản ghi tương ứng với loại AQI trong từng quận/ngày

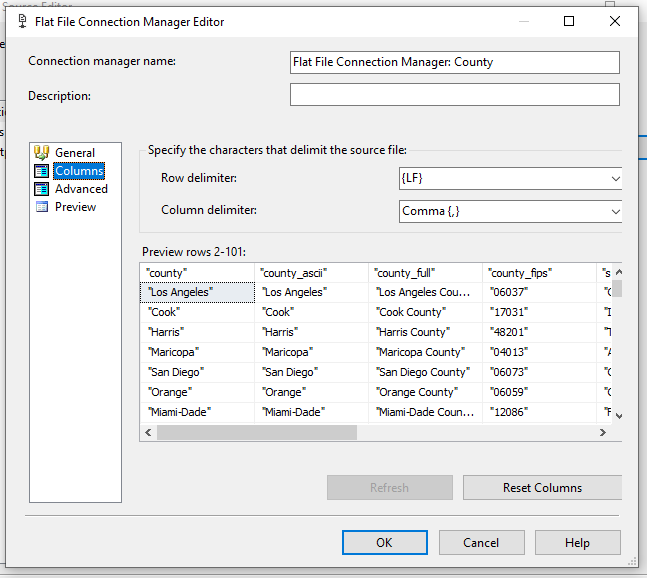
1. **ETL:**

### **ETL từ Soure vào Stage:**

#### ETL County từ Source vào StageDB

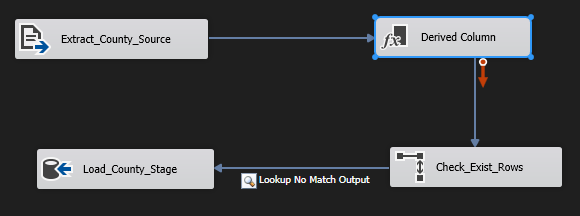
**Bước 1: Lấy dữ liệu County từ file (2B)uscounties.csv**

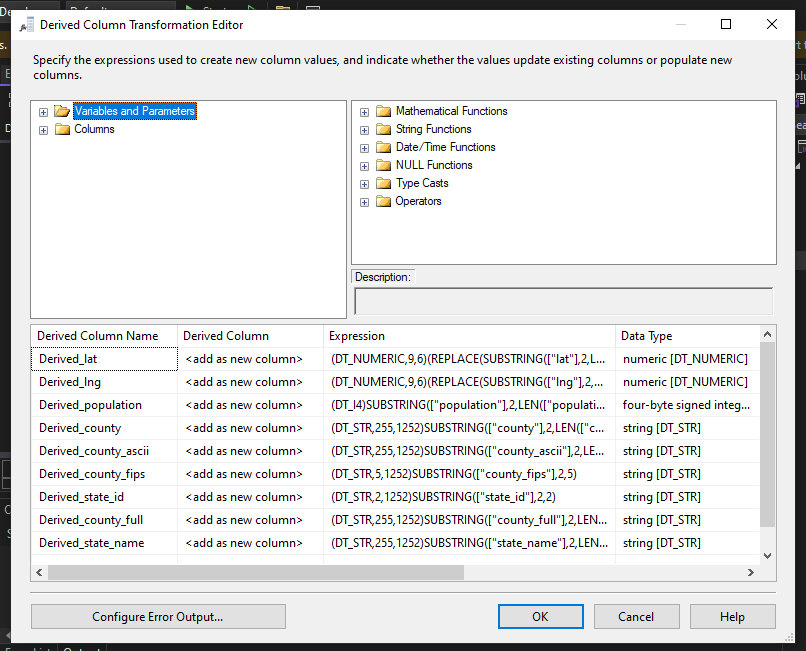




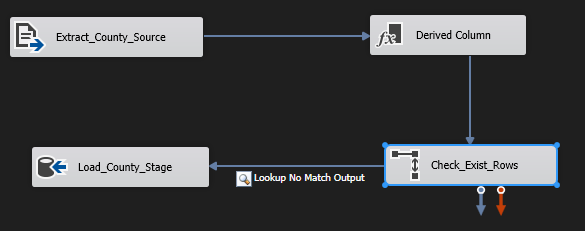
*Connect đến file (2B)uscounties.csv và lấy dữ liệu từ các cột*

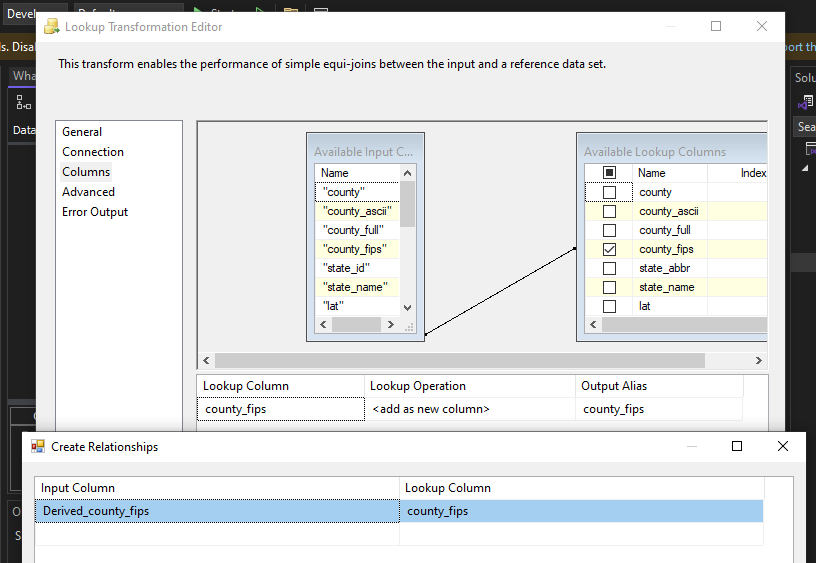
**Bước 2: Dẫn xuất, chuyển đổi và định dạng dữ liệu các cột**



**

**Bước 3: Kiểm tra để lọc ra các dòng đã tồn tại trong bảng**





*Mapping thuộc tính và Lookup thuộc tính county\_fips vì đây là khóa định danh duy nhất của mỗi Hạt*

Sau khi lọc ra các dòng đã tồn tại sẽ không được load vào Stage

**Bước 4: Load vào bảng Counties trong StageDB**

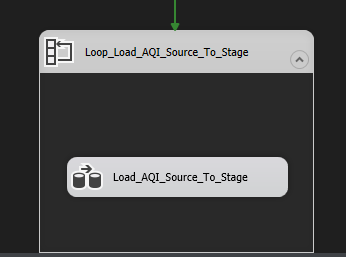




*Load dữ liệu vào bảng Counties trong StageDB*

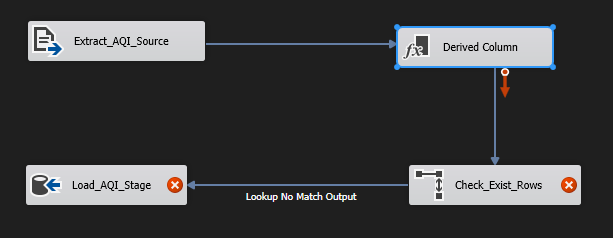
#### ETL AQI Measurements từ Source vào StageDB

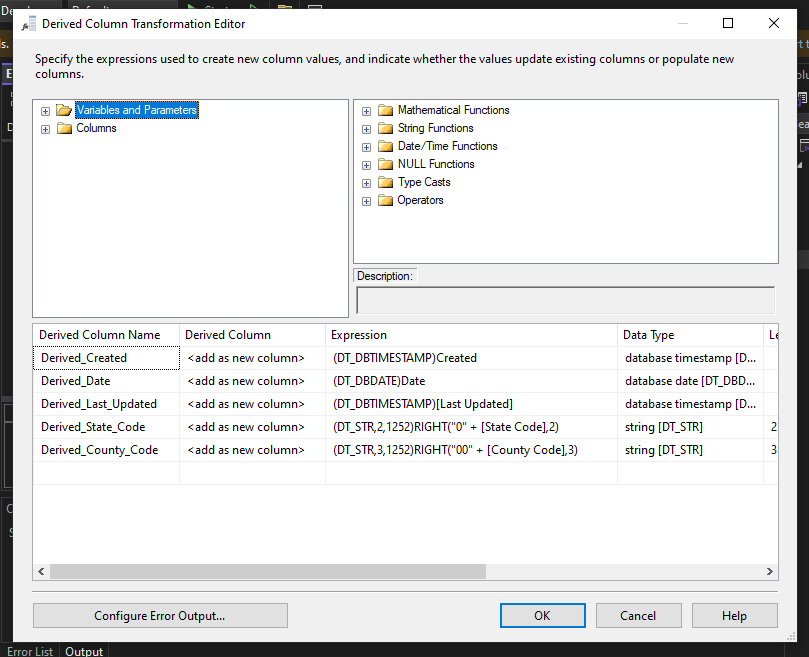
**Bước 1: Lấy dữ liệu từ các file AQI của từng năm**



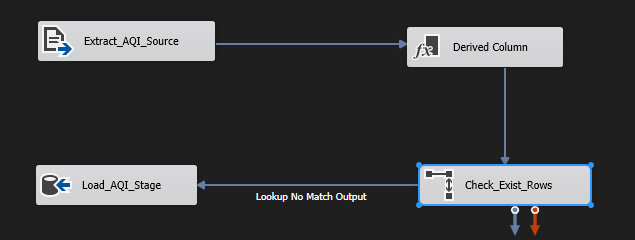
*Lấy dữ liệu từ 3 file* *10\_state\_aqi\_2021.csv, 10\_state\_aqi\_2022.csv, 10\_state\_aqi\_2023.csv*

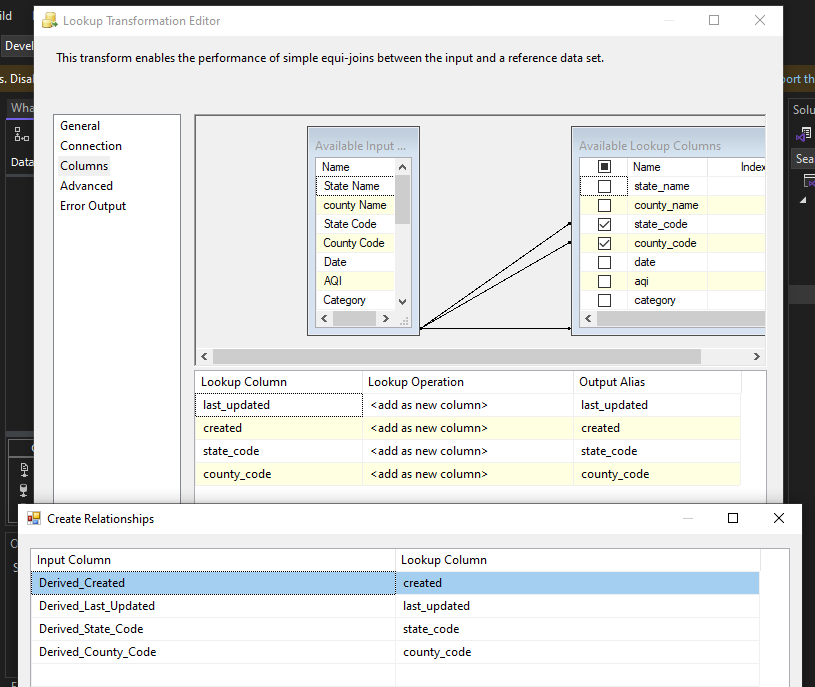
**Bước 2: Dẫn xuất, chuyển đổi và định dạng dữ liệu của các cột**

****

****

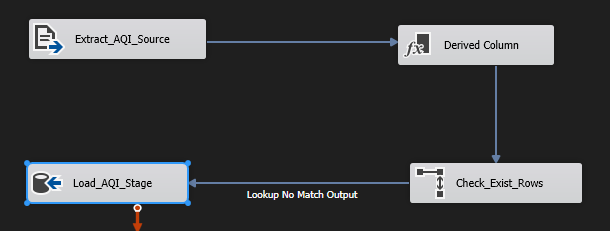
**Bước 3: Kiểm tra để lọc ra các dòng dữ liệu đã tồn tại**

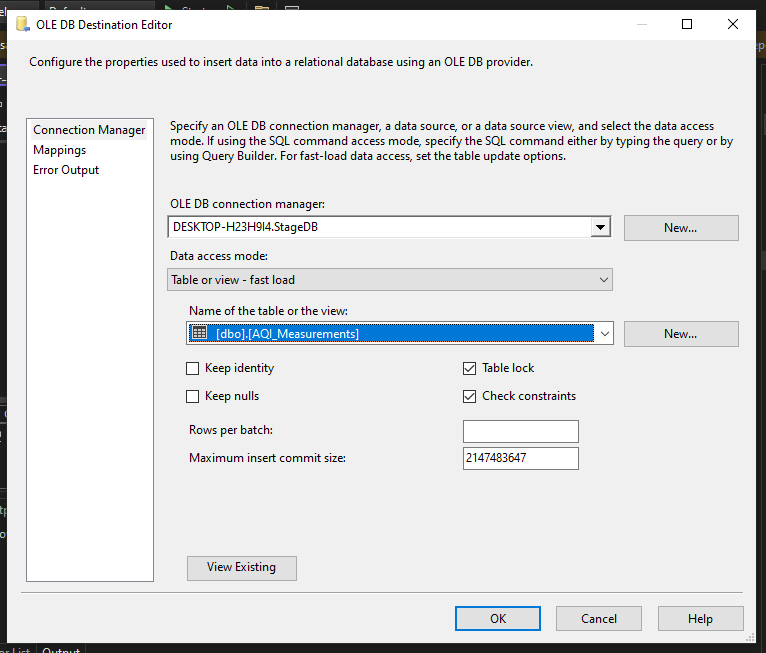
****

****

*Mapping các thuộc tính và Lookup trên thuộc tính last\_updated, created, state\_code, county\_code vì đây là bộ khóa định danh AQI*

**Bước 4: Load vào bảng AQI\_Measurements trong StageDB**

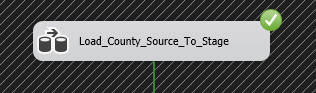
****

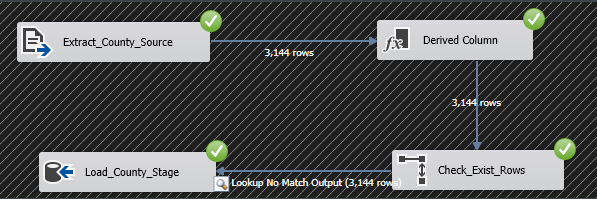
****

*Load dữ liệu vào bảng AQI\_Measurements trong StageDB*

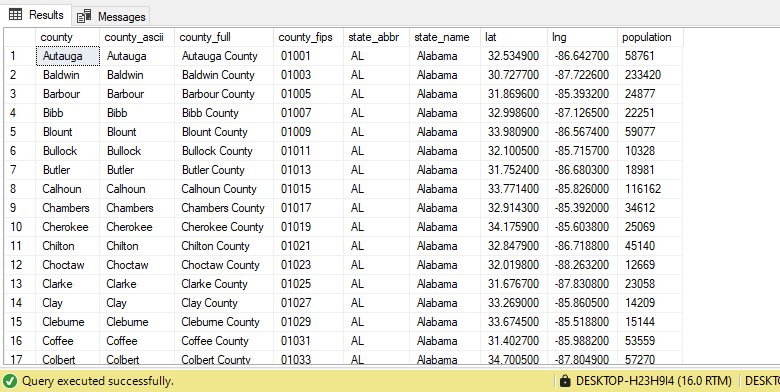
#### Kết quả chạy:

* **ETL County từ Soure vào StageDB**

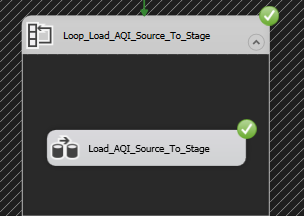
****

****

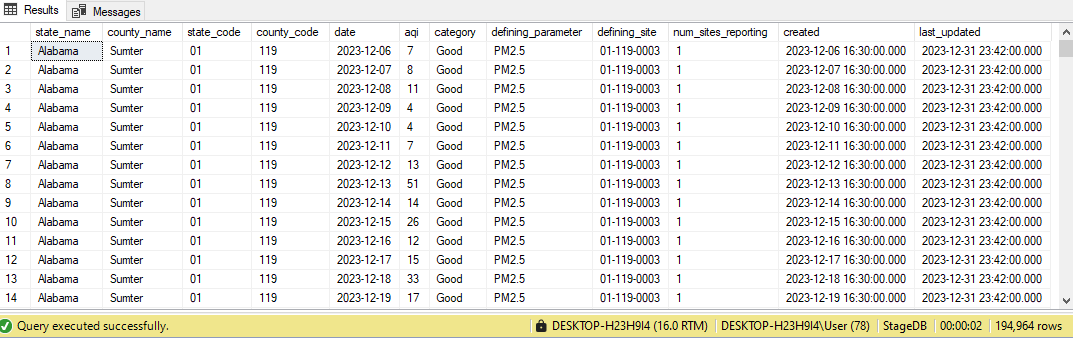
Vào StageDB refresh và query bảng Counties để kiểm tra kết quả

****

* **ETL AQI measurements từ Soure vào StageDB**

****

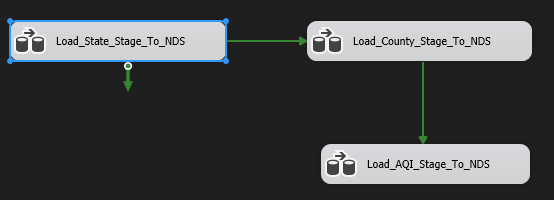
Vào StageDB refresh và query bảng AQI\_measurements để kiểm tra kết quả

****

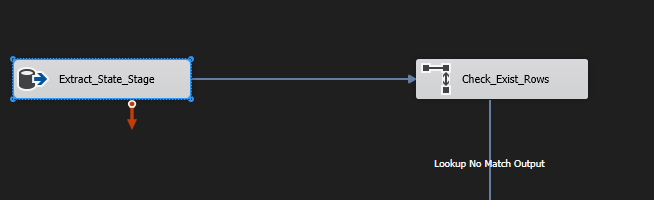
### **ETL từ Stage vào NDS:**

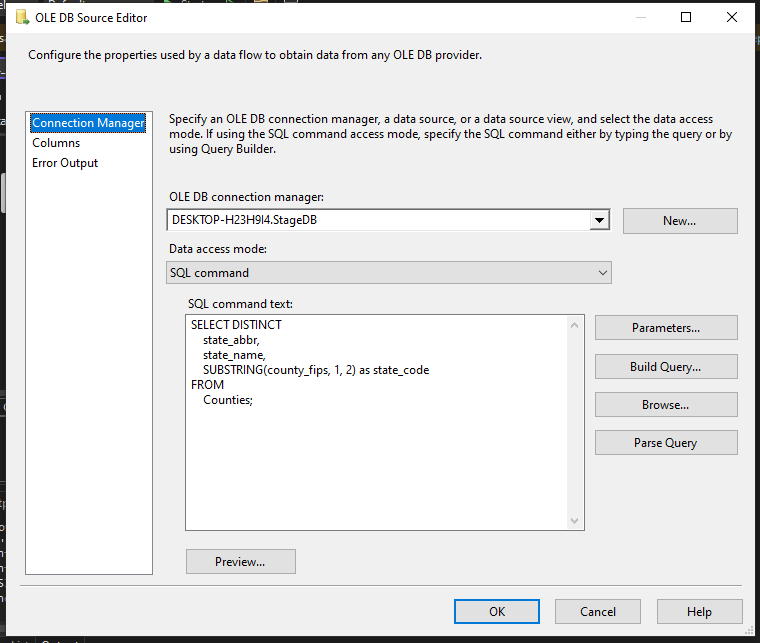
Trước khi nạp dữ liệu vào NDS thì cần thiết kế cơ sở dữ liệu cho NDS trước (các bảng dữ liệu thiết kế đã trình bày chi tiết ở mục 3.)

#### ETL dữ liệu vào bảng State trong NDSDB:

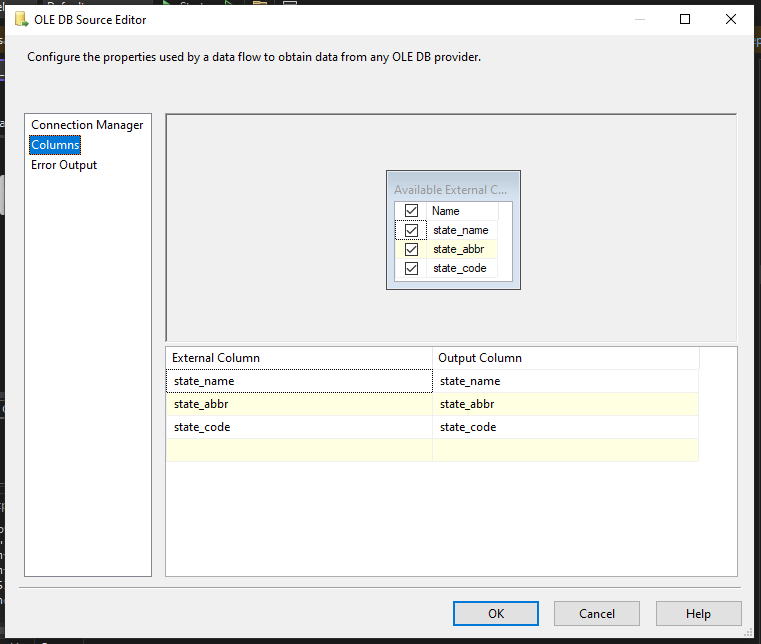


**Bước 1: Trích xuất các cột liên quan đến State trong bảng County trong Stage database**

****

****

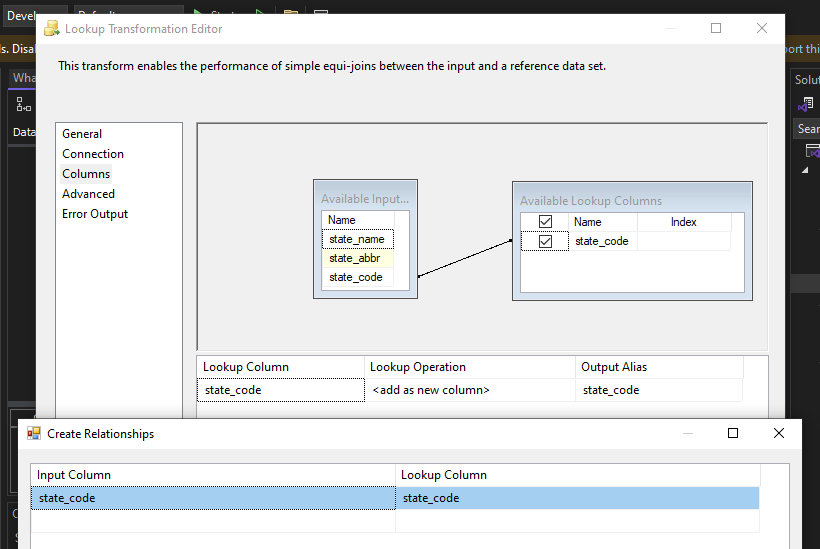
*Sử dụng code SQL để trích xuất thuộc tính liên quan đến Bang như tên viết tắt của Bang, tên Bang và mã code Bang trích từ 2 ký tự đầu tiên của thuộc tính county\_fips trong bảng Counties*



*Các column thu được*

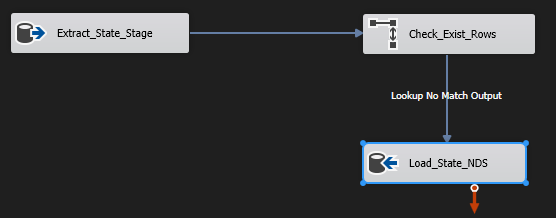
**Bước 2: Kiểm tra dòng tồn tại**

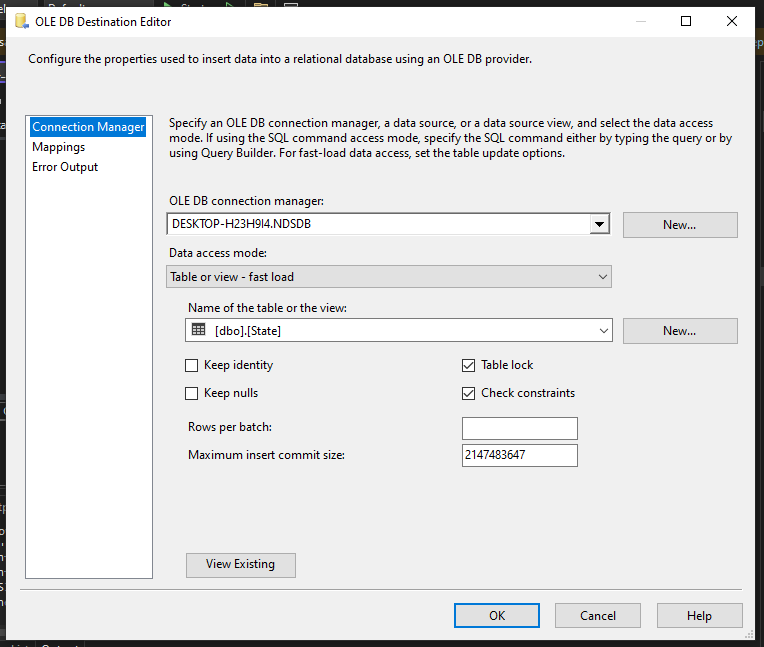
****



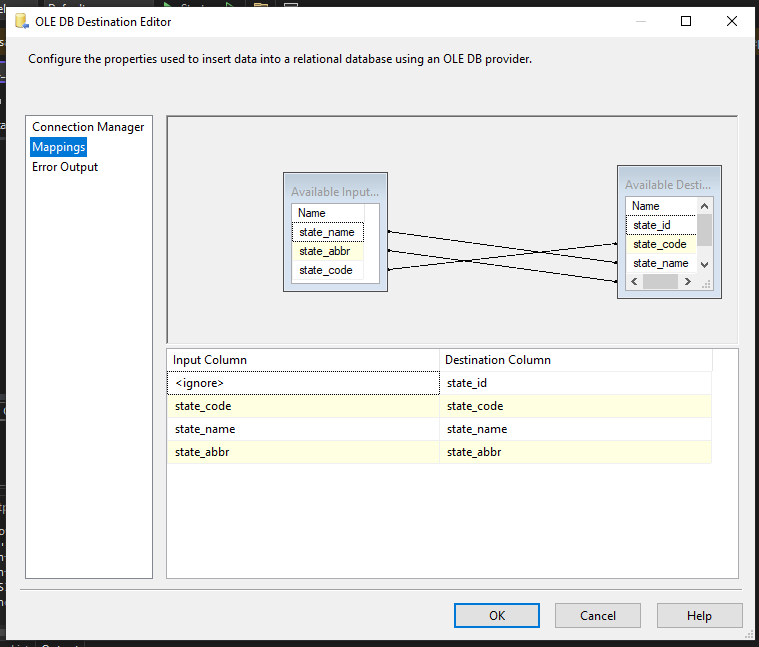
*Mapping thuộc tính và Lookup trên thuộc tính state\_code bởi vì trong bảng State thì stage\_code là unique*

**Bước 3: Load dữ liệu vào bảng State trong StageNDSDB:**



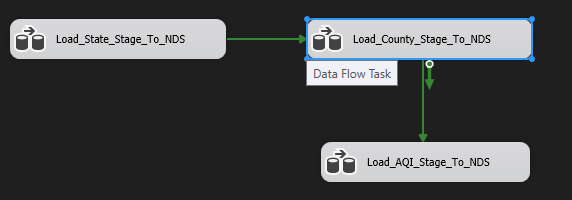


*Load dữ liệu vào bảng State trong NDSDB*

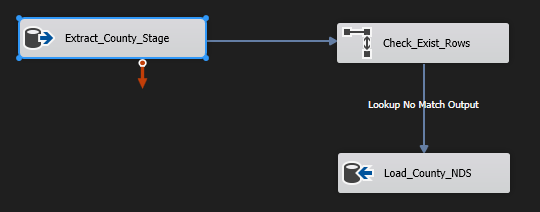


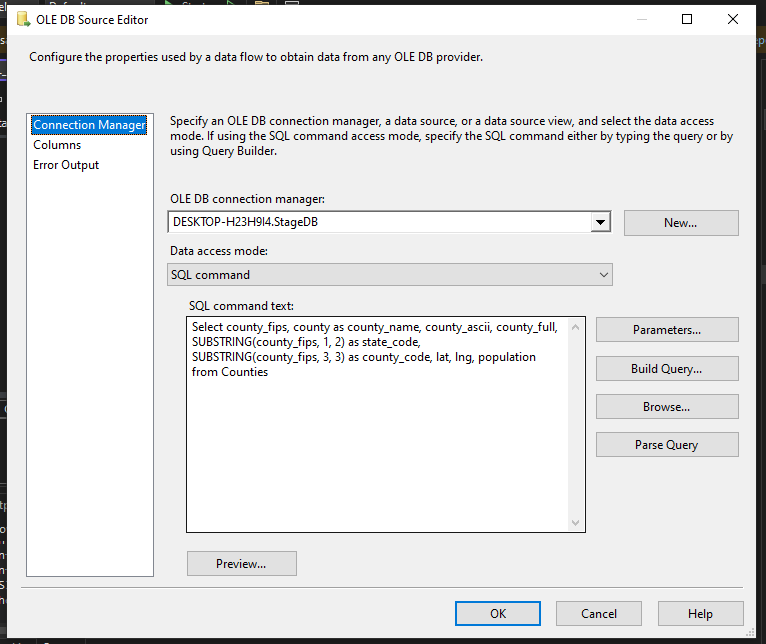
*Mapping (do state\_id là khóa tự tăng nên không cần map)*

#### ETL dữ liệu vào bảng County trong NDSDB:

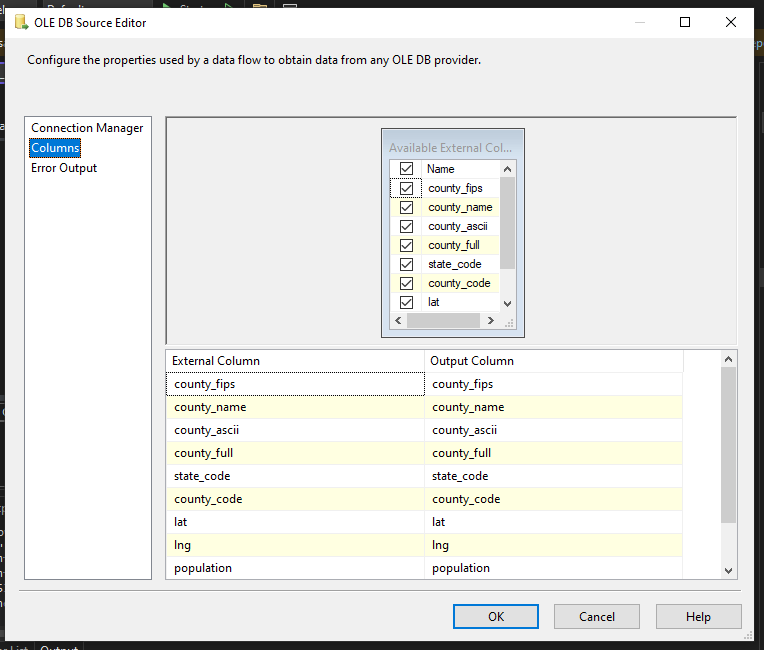


**Bước 1: Trích xuất các cột liên quan đến County trong bảng Counties trong StageDB**

**

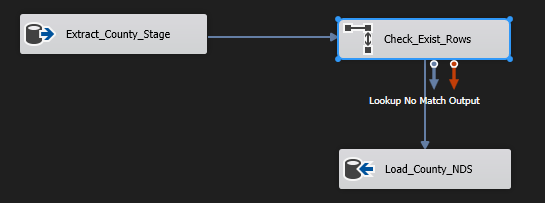


*Dùng code SQL trích xuất các thuộc tính liên quan đến County trong bảng Counties ở StageDB*



*Các column thu được*

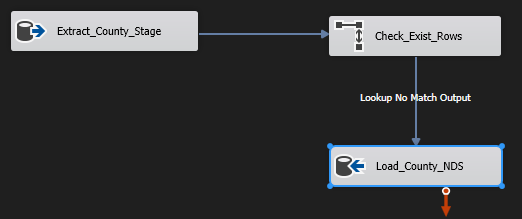
**Bước 2: Kiểm tra dòng tồn tại**

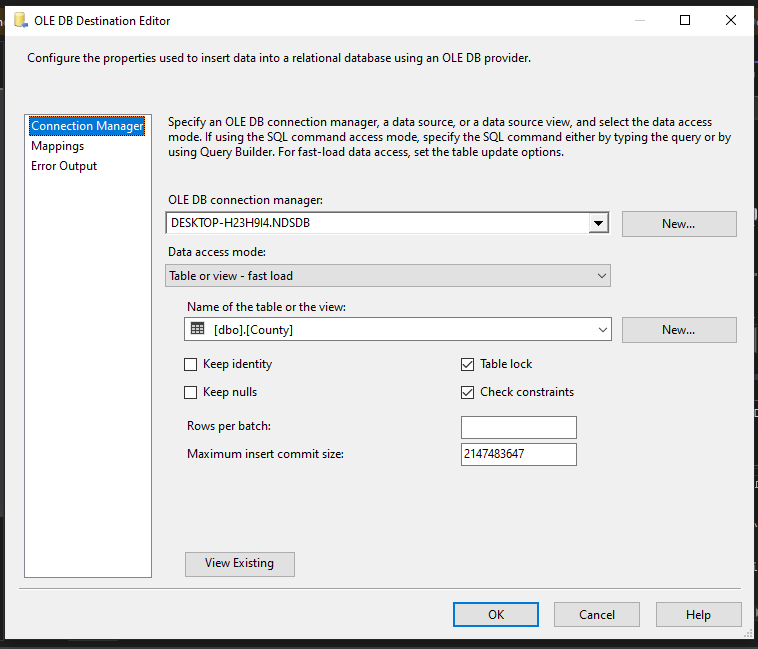
****

**

*Mapping các thuộc tính và Lookup trên 3 thuộc tính county\_fips, created\_date, updates\_date*

**Bước 3: Nạp dữ liệu vào bảng Counties của NDS**

**

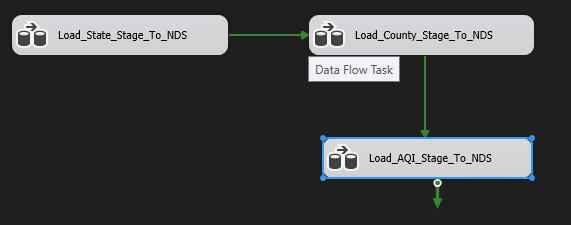
**

*Load vào bảng County trong NDSDB*

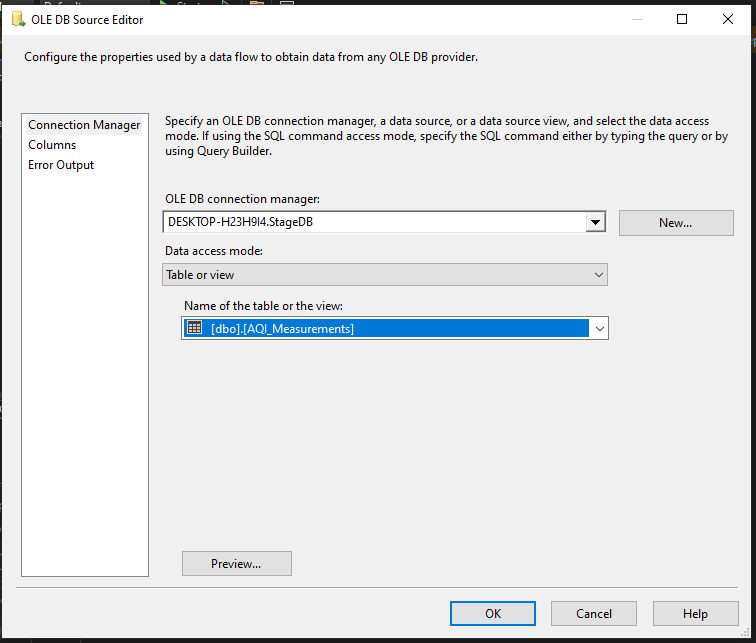
**

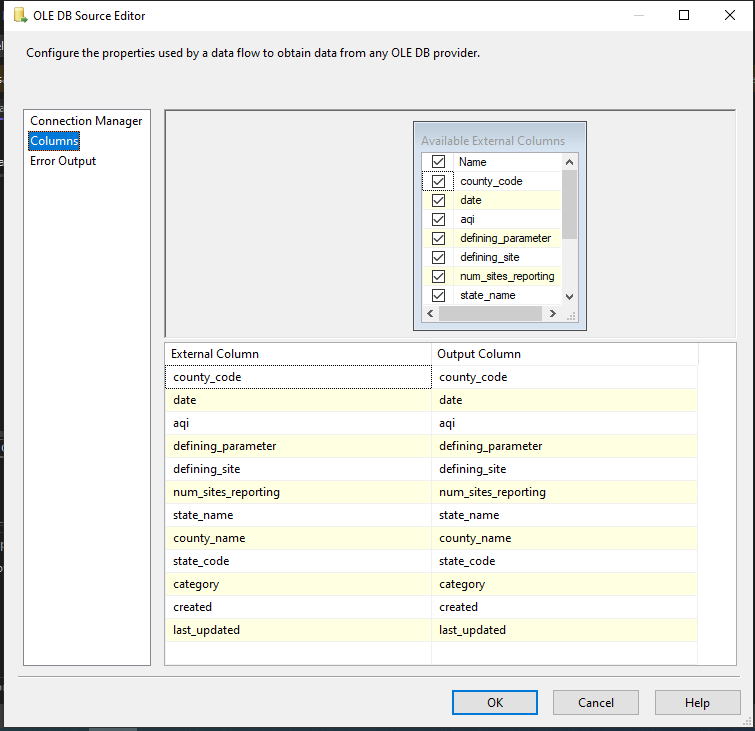
*Mapping (do county\_id là khóa tự tăng nên không cần map)*

#### ETL dữ liệu vào bảng AQI NDS:



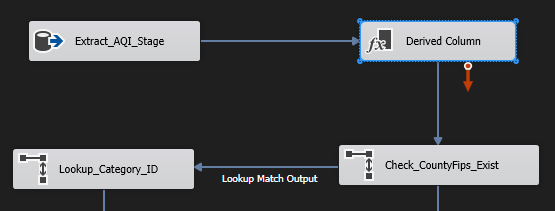
**Bước 1: Trích xuất dữ liệu trong bảng AQI\_Measurements trong StageDB**

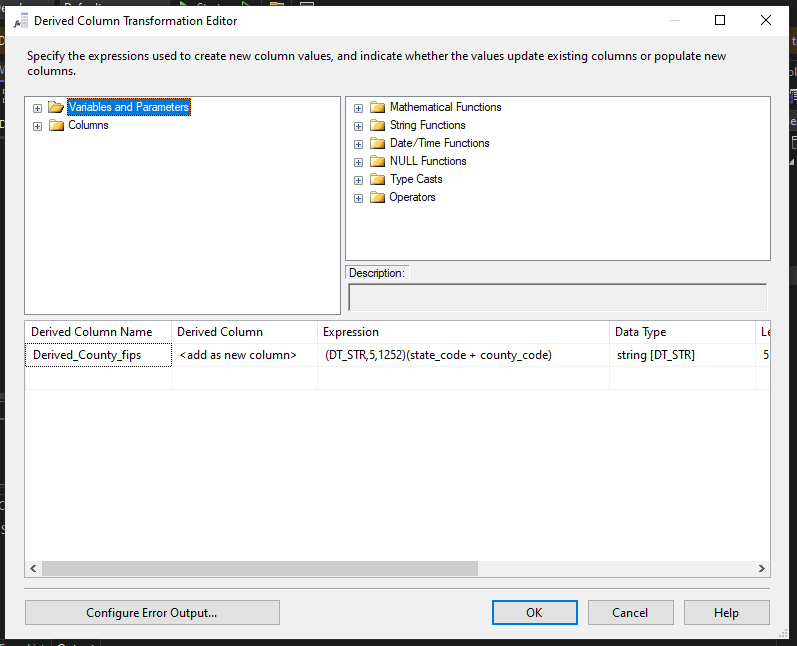




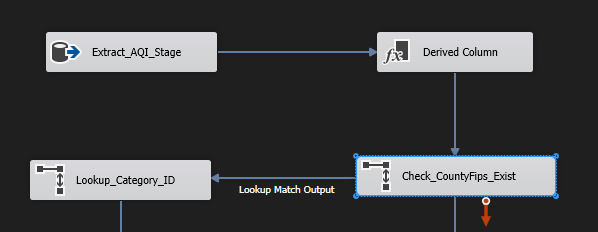
*Các column thu được*

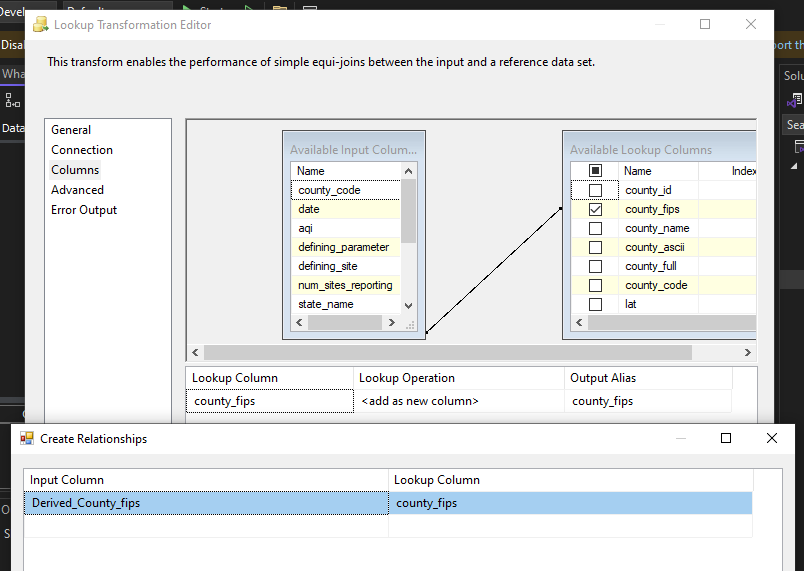
**Bước 2: Dẫn xuất, chuyển đổi và định dạng dữ liệu của các cột**





**Bước 3: Kiểm tra dòng tồn tại (county\_fips) trong bảng County**

****

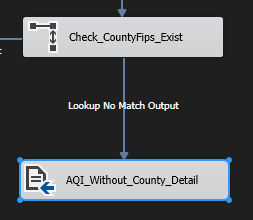
**

*Mapping thuộc tính và Lookup trên county\_fips vì đây là thuộc tính unique*

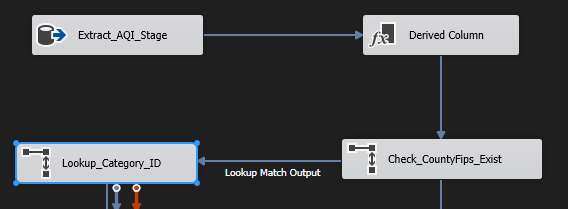
Ở bước này, sẽ có 2 luồng như sau:

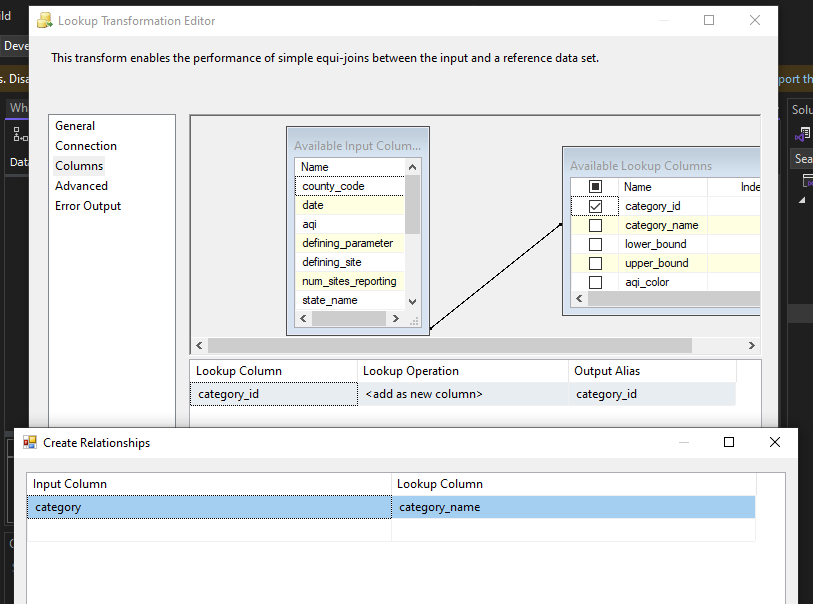
* Những data lookup no match sẽ được chuyển xuông bước 5
* Những data look up match sẽ được chuyển xuống bước 6

**Bước 5: Xử lý error data**

****

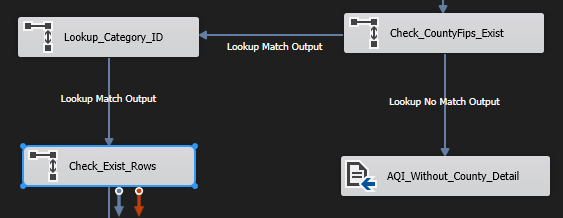
**Bước 6: Kiểm tra dòng tồn tại (category\_id) trong bảng Category**

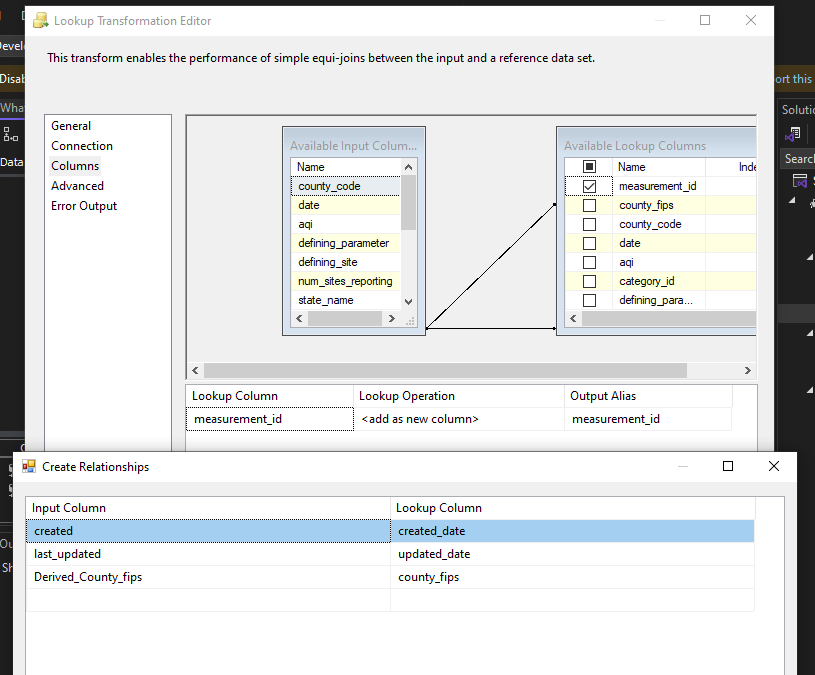
****

****

*Mapping thuộc tính và Lookup trên category\_id vì đây là thuộc tính unique*

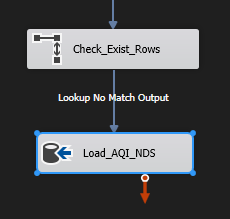
**Bước 7: Kiểm tra dòng tồn tại (measurement\_id) trong bảng AQI\_Measurement**

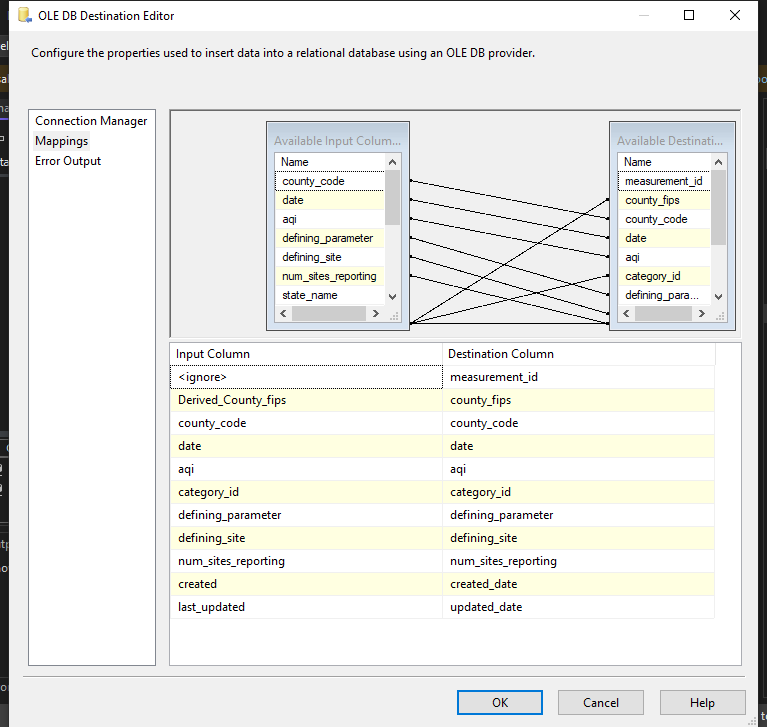
****

****

*Mapping các thuộc tính và Lookup trên measurement\_id vì đây là thuộc tính unique*

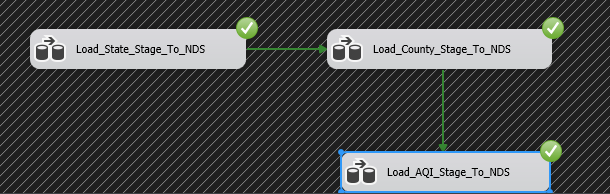
**Bước 8: Nạp dữ liệu vào bảng AQI NDS**

****

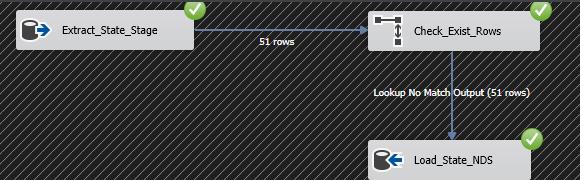
****

*Mapping*

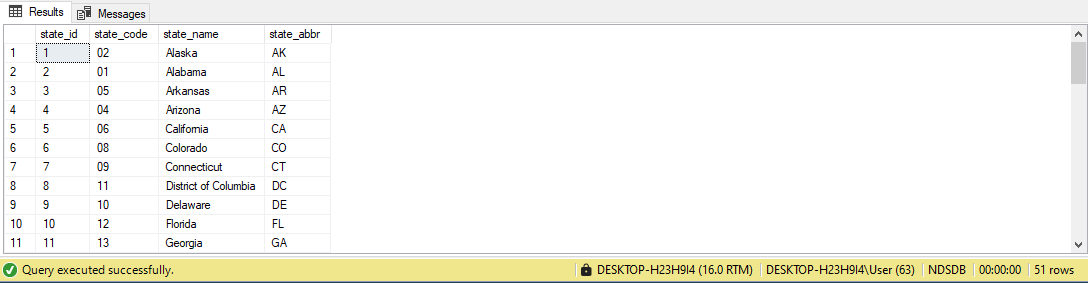
#### Kết quả chạy:



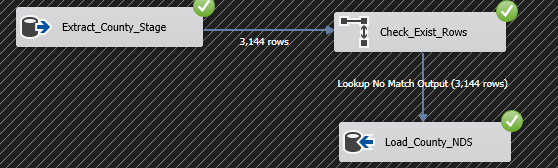
* **ETL dữ liệu vào bảng State:**

****

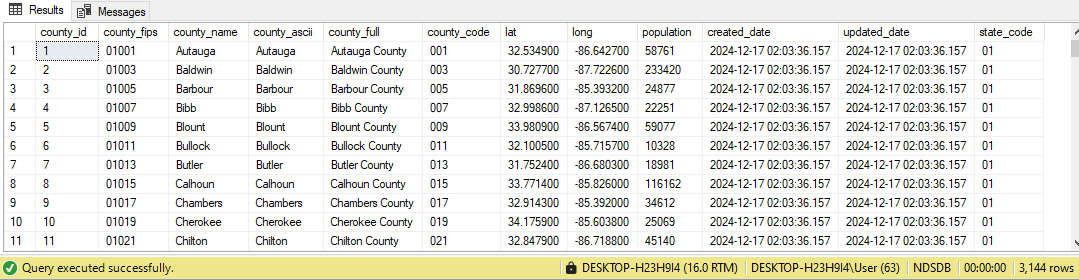
Vào NDSDB refresh và query bảng State để kiểm tra kết quả:

****

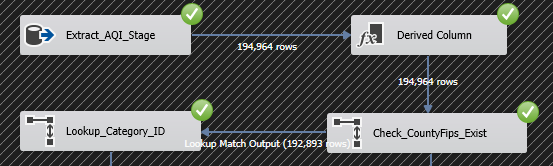
* **ETL dữ liệu vào bảng County:**

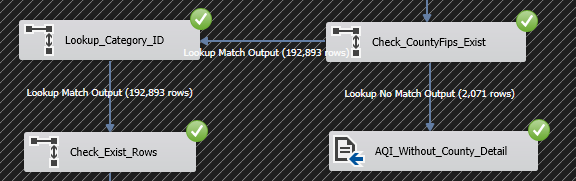
****

Vào NDSDB refresh và query bảng County để kiểm tra kết quả

****

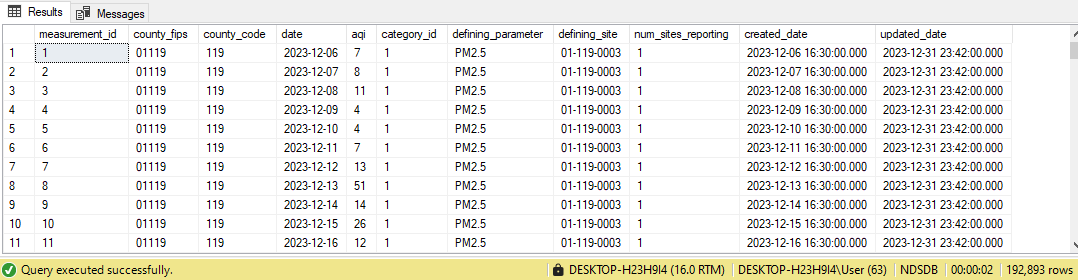
* **ETL dữ liệu vào bảng AQI\_measurement:**

****

****

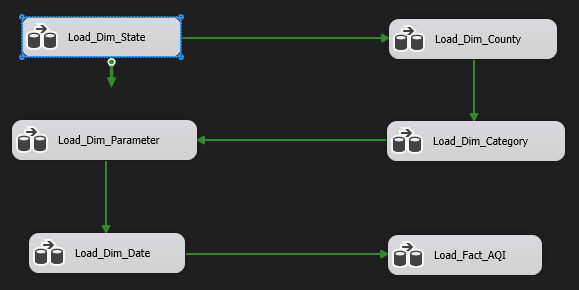
****

Vào NDSDB refresh và query bảng AQI\_measurement để kiểm tra kết quả

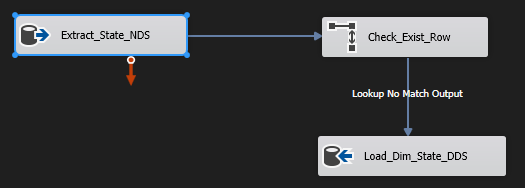


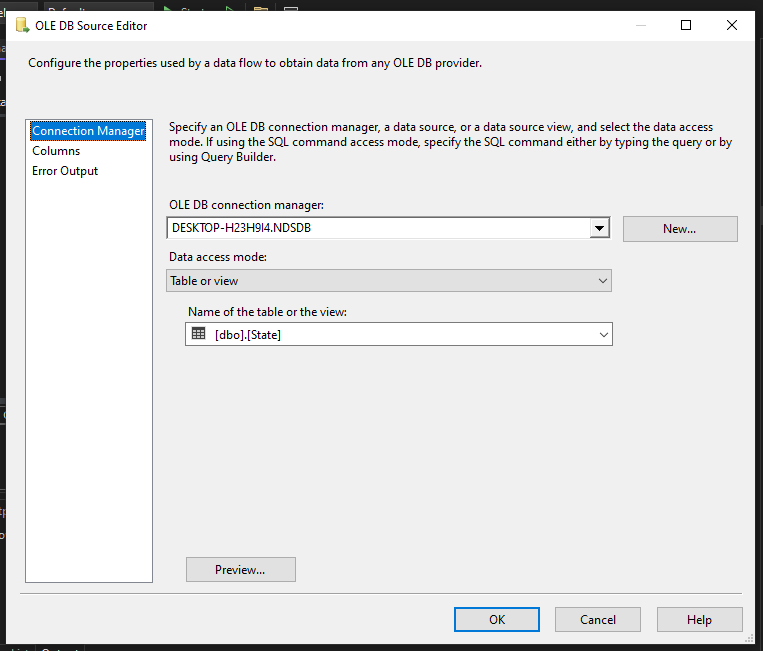
### **ETL từ NDS vào DDS:**

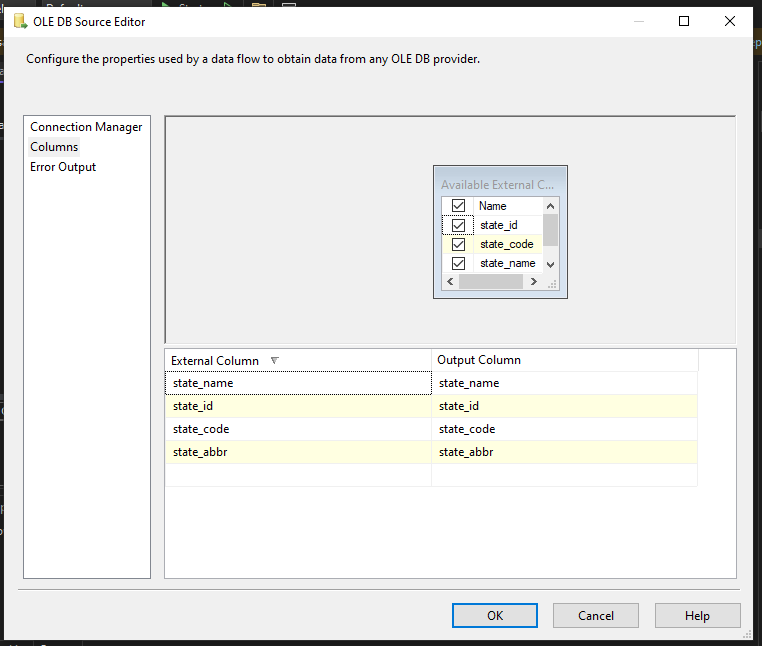
#### ETL dữ liệu vào bảng Dim\_State:

****

**Bước 1: Trích xuất dữ liệu từ bảng State trong NDS**

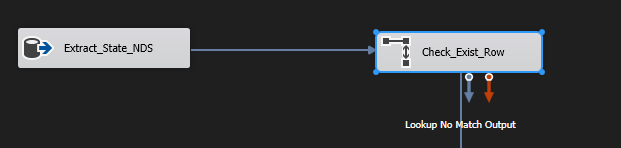
****

****

****

*Các column thu được*

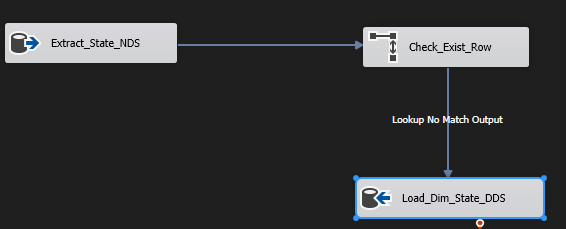
**Bước 2: Kiểm tra dòng tồn tại**

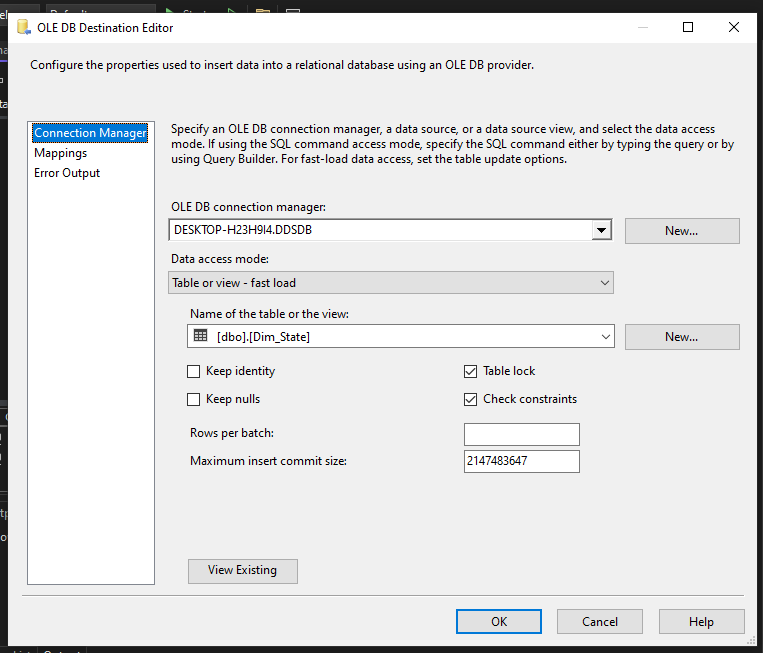
****

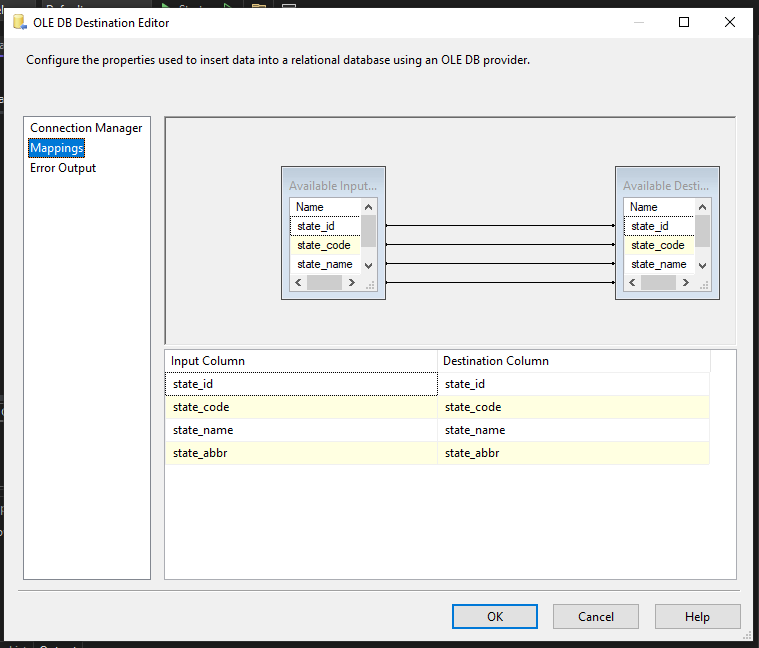


*Mapping thuộc tính và Lookup trên state\_id vì đây là khóa chính của bảng State*

**Bước 3: Load vào bảng Dim\_State DDS**

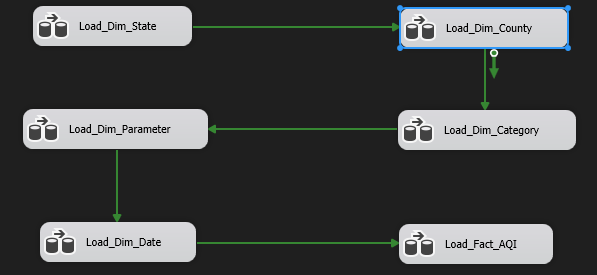
****

****

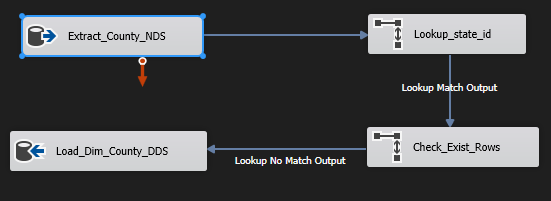
****

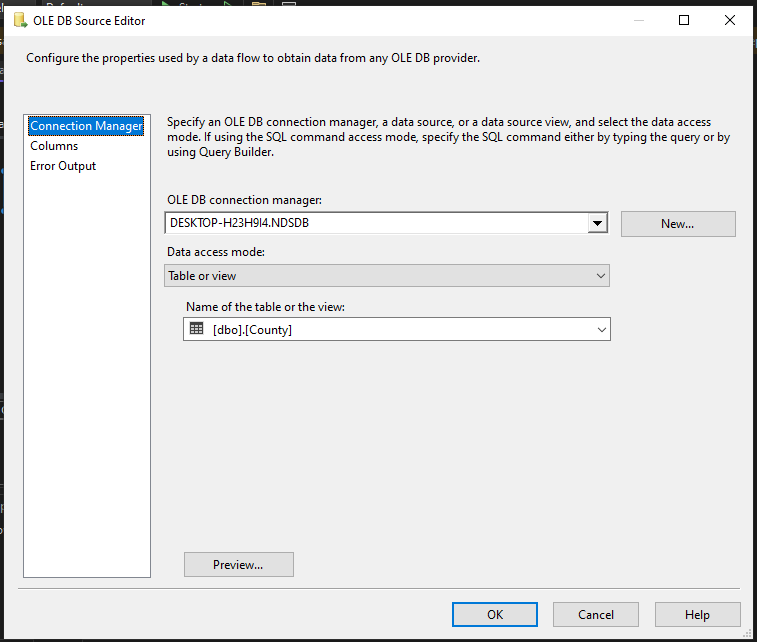
*Mapping*

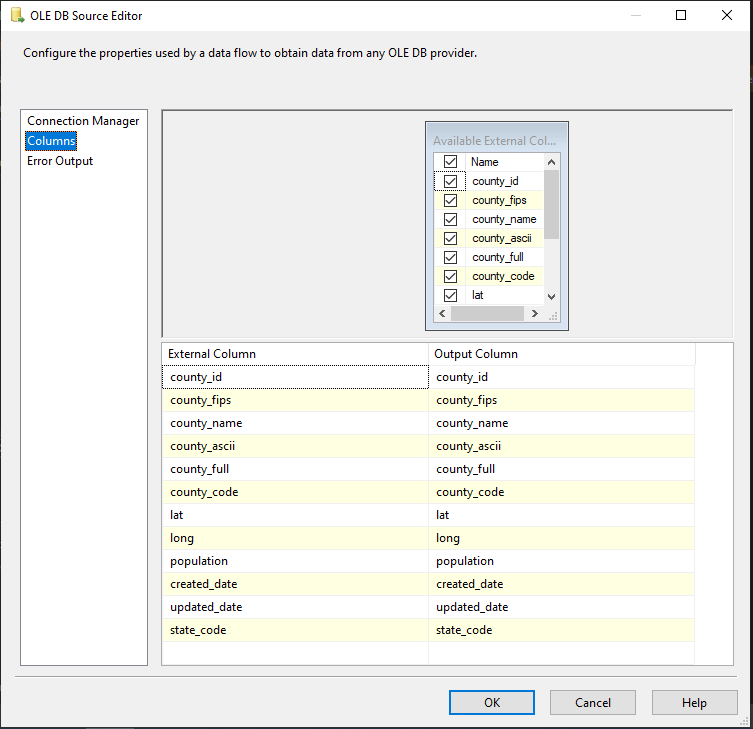
#### ETL dữ liệu vào bảng Dim\_County:



**Bước 1: Trích xuất dữ liệu từ bảng County trong NDS**

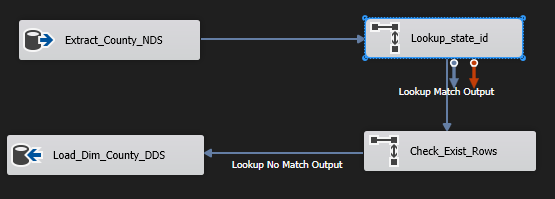
****

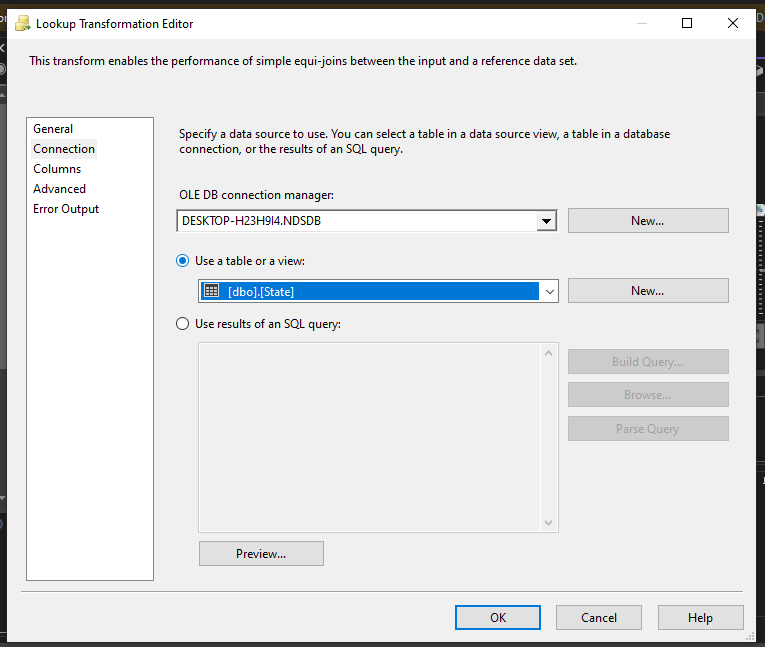
****

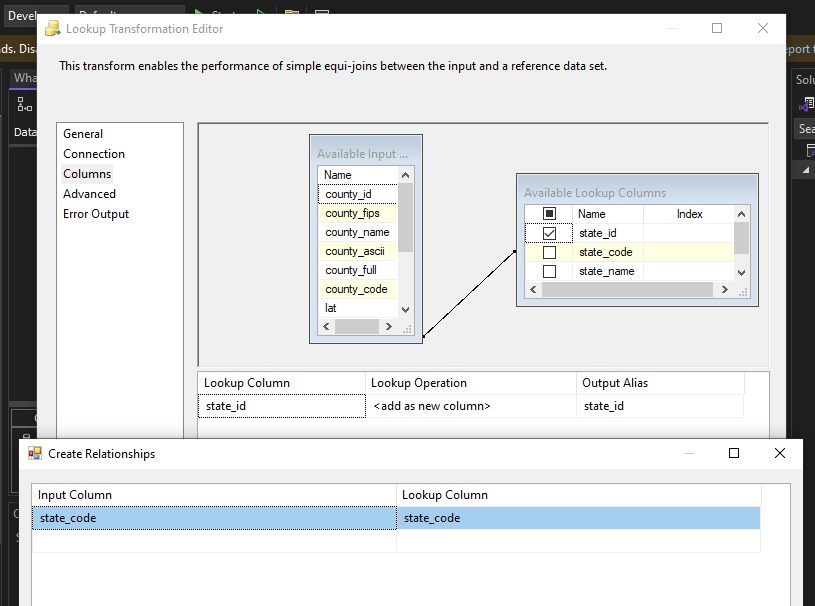
****

*Các column thu được*

**Bước 2: Kiểm tra dòng tồn tại (state\_id) trong bảng State NDSDB**

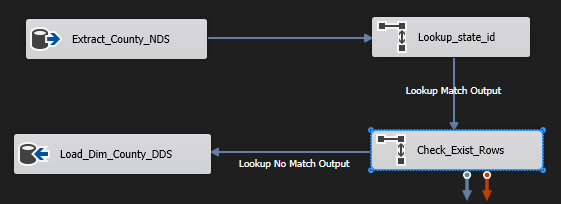
****

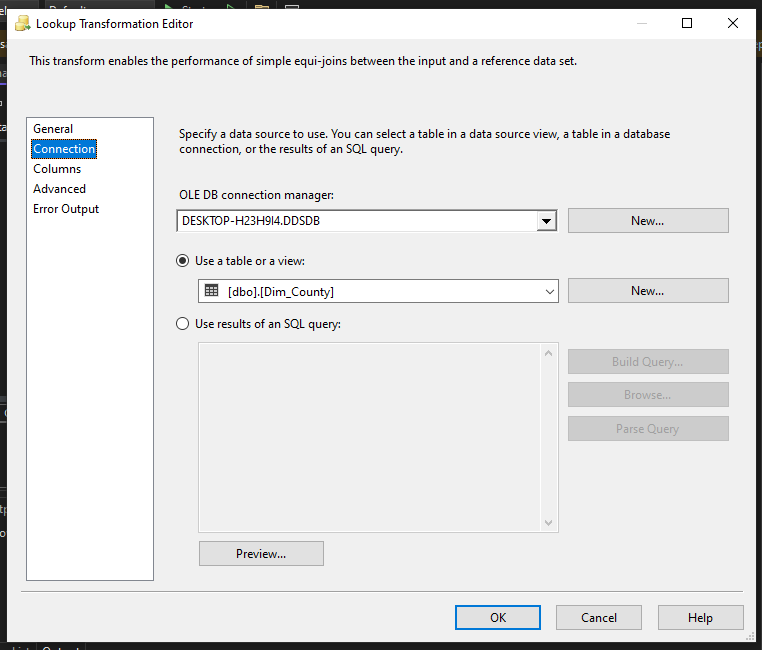
****

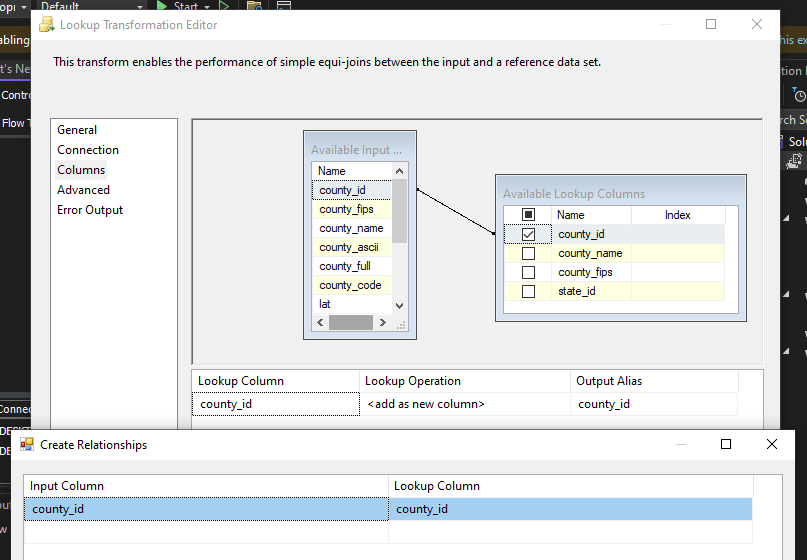


*Mapping thuộc tính và Lookup trên state\_id vì đây là khóa chính của bảng State*

**Bước 3: Kiểm tra dòng tồn tại (county\_id) trong bảng Dim\_County DDSDB**

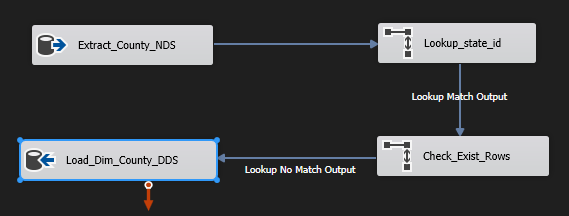


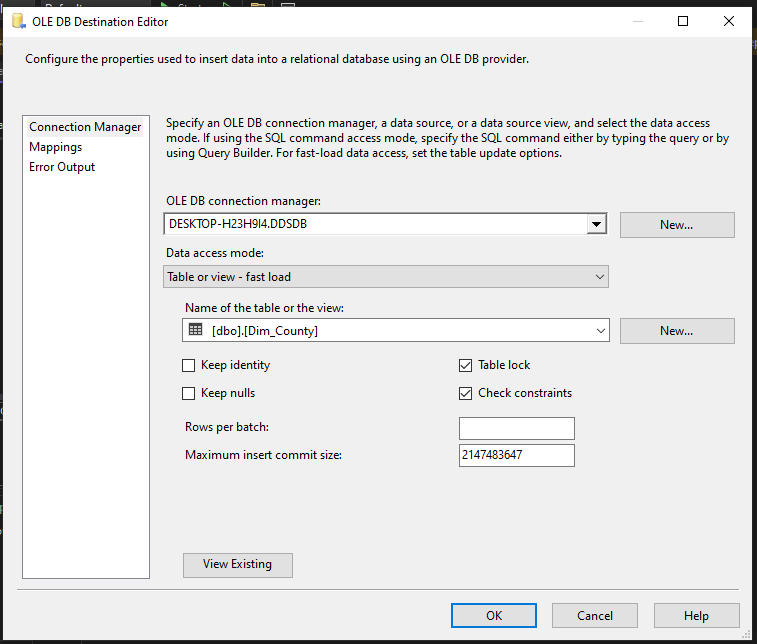


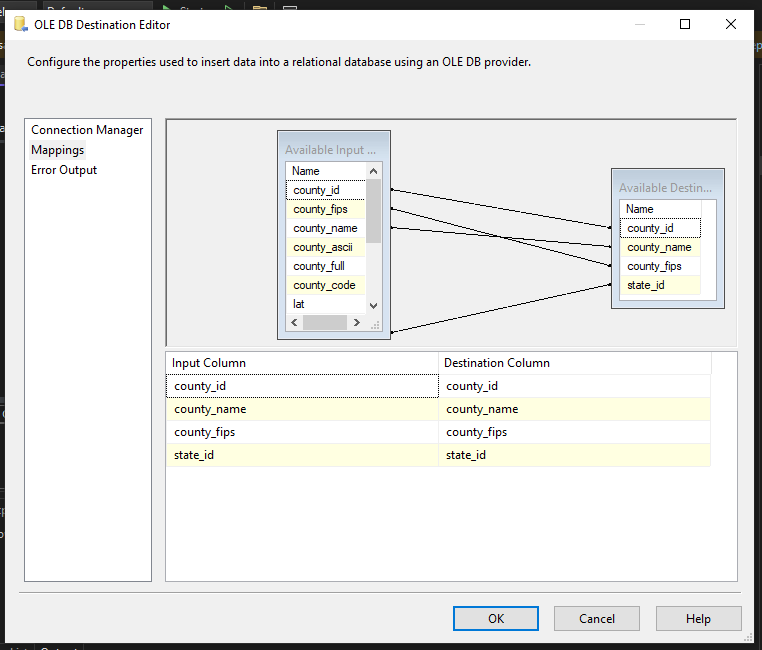
****

*Mapping thuộc tính và Lookup trên county\_id vì đây là khóa chính của bảng State*

**Bước 4: Load vào bảng Dim\_County DDS**

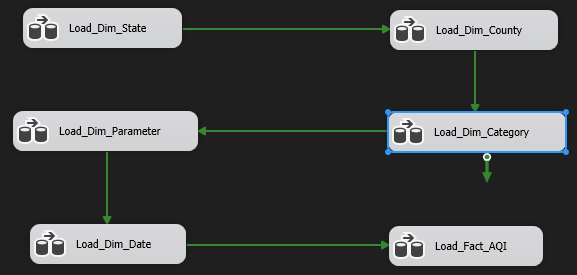
****

****

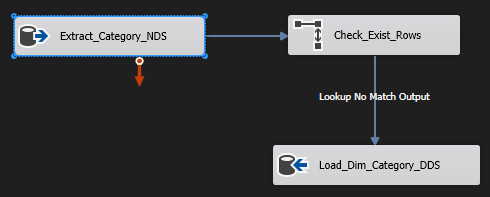
****

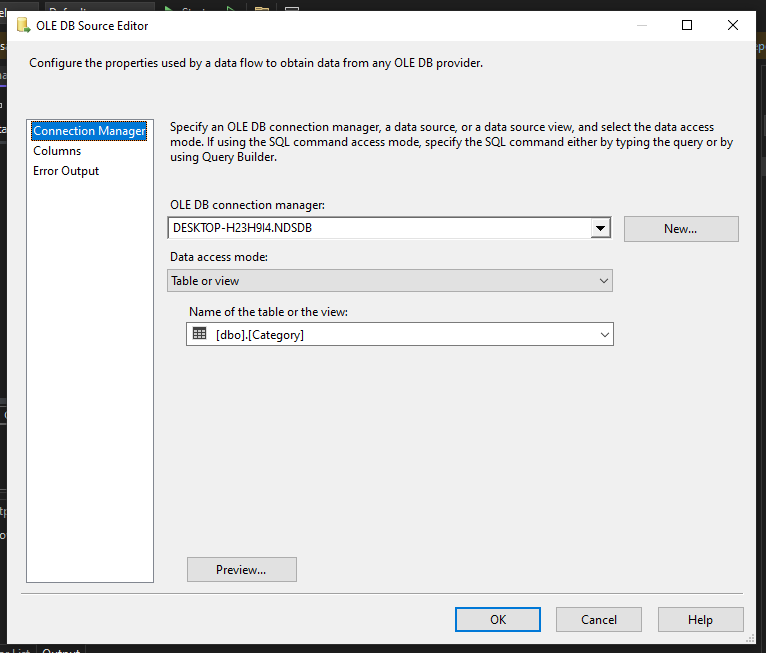
*Mapping*

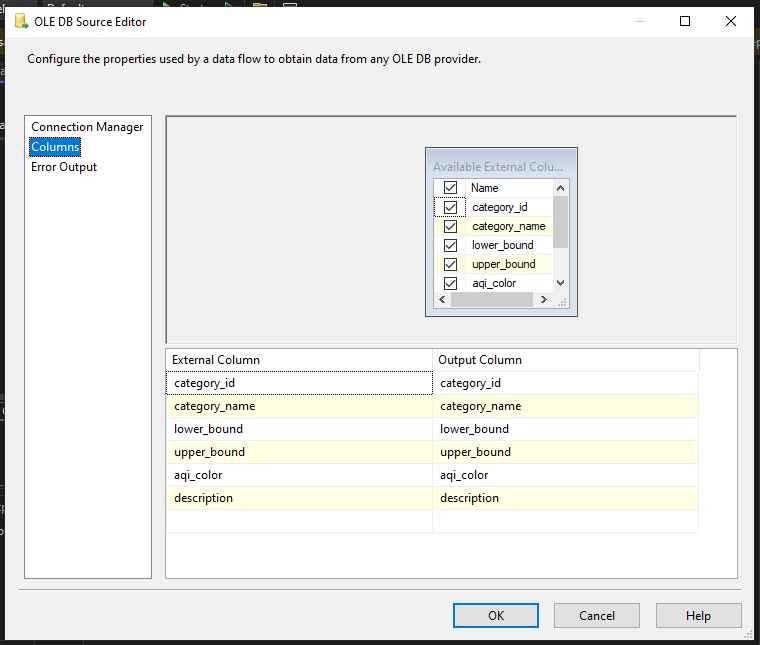
#### ETL dữ liệu vào bảng Dim\_ Category:

****

**Bước 1: Trích xuất dữ liệu từ bảng Category trong NDS**

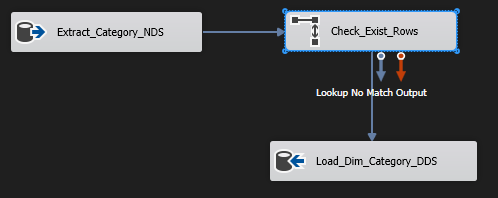
****

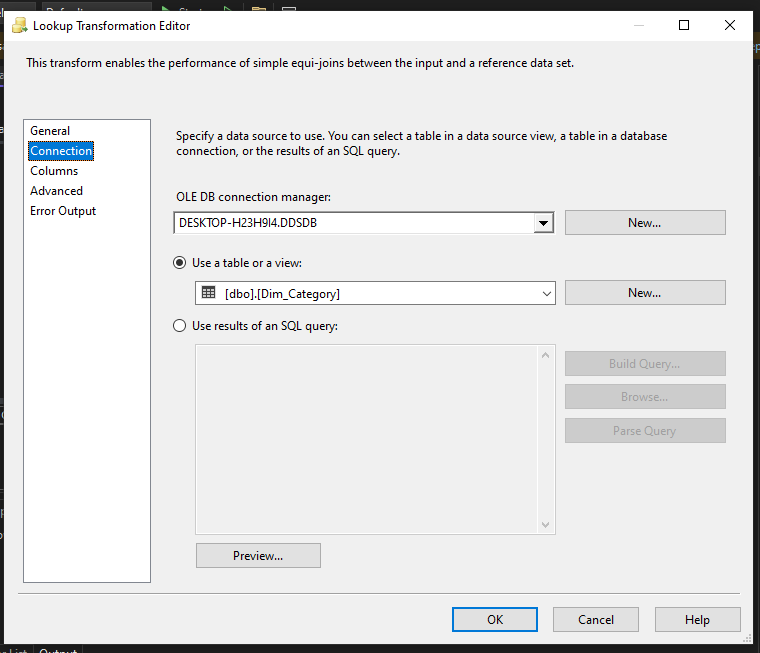
****

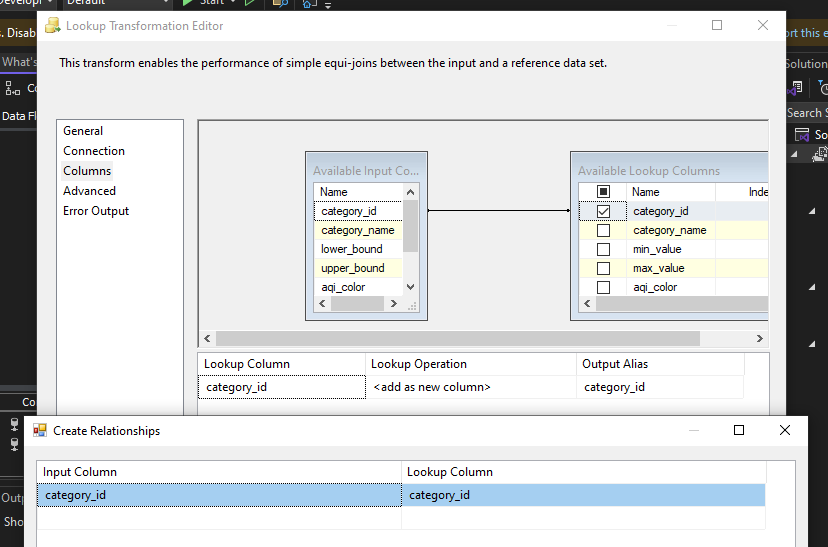
****

*Các column thu được*

**Bước 2: Kiểm tra dòng tồn tại**

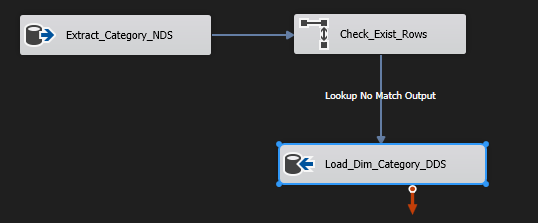
****

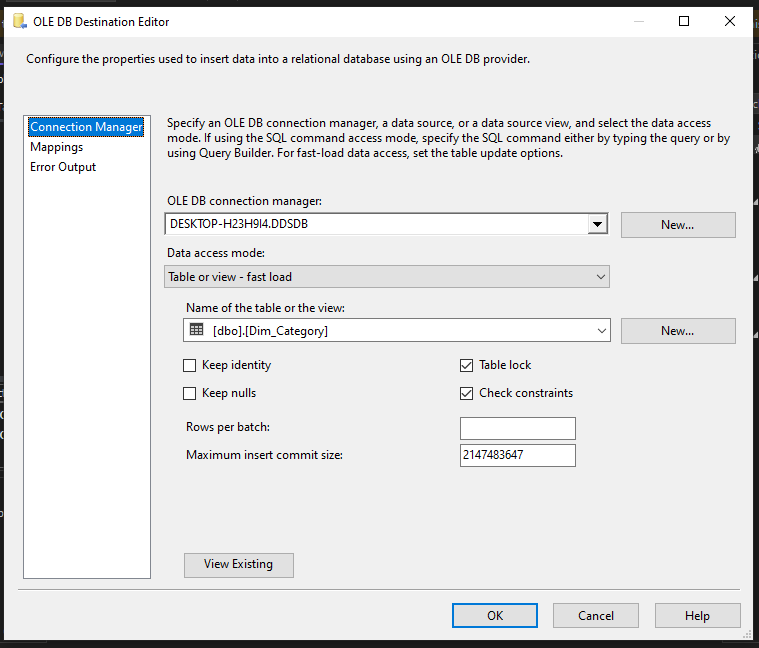
****

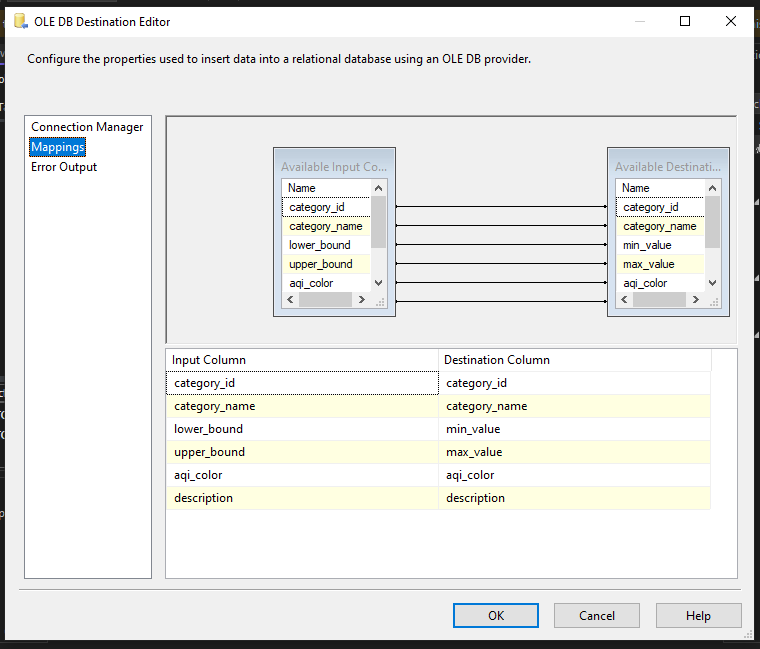


*Mapping thuộc tính và Lookup trên category\_id vì đây là khóa chính của bảng Dim\_Category*

**Bước 4: Load vào bảng Dim\_Category DDS**

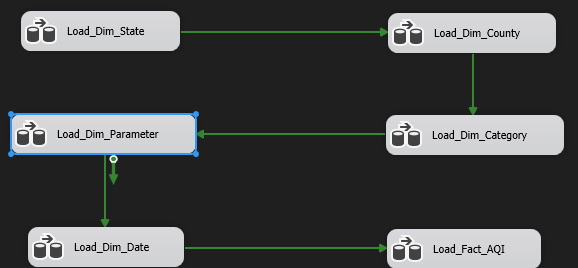
****

****

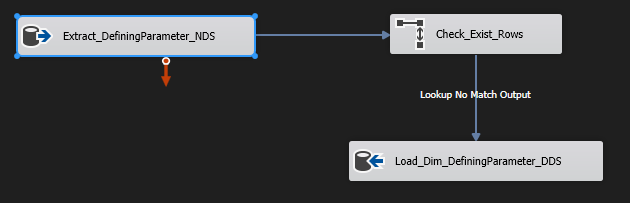
****

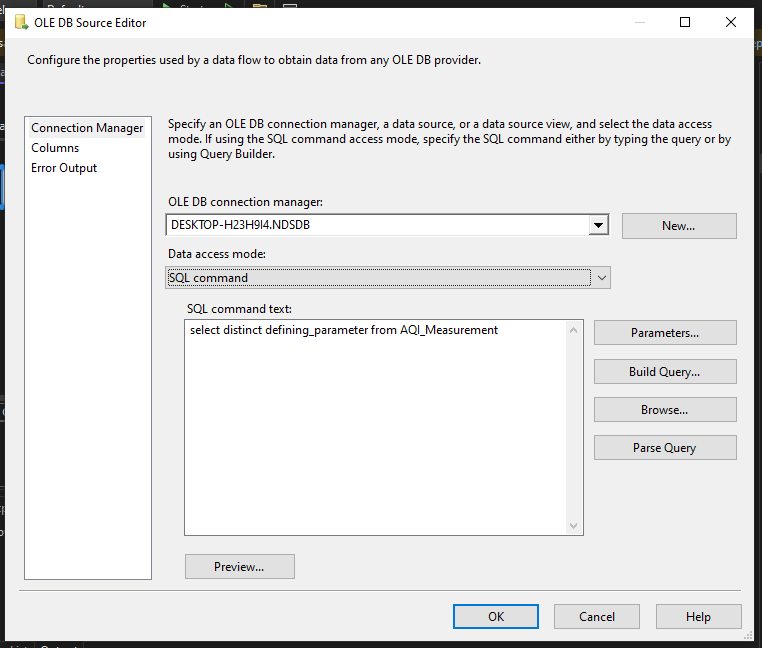
*Mapping*

#### ETL dữ liệu vào bảng Dim\_Parameter:

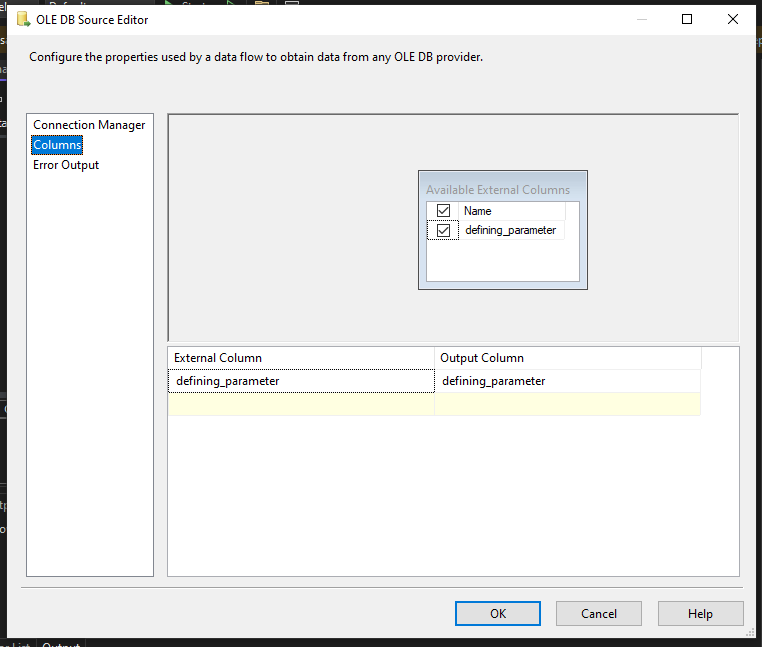
****

**Bước 1: Trích xuất dữ liệu liên quan đến paramenter từ bảng AQI\_Measurement trong NDS**

****

****

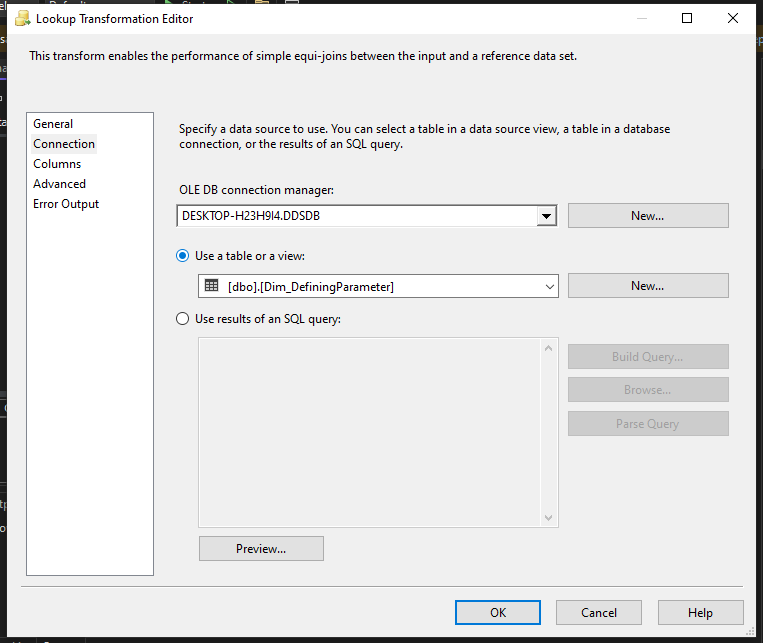
*Dùng code SQL trích xuất thuộc tính liên quan đến parameter trong bảng AQI\_Measurement ở NDSDB*

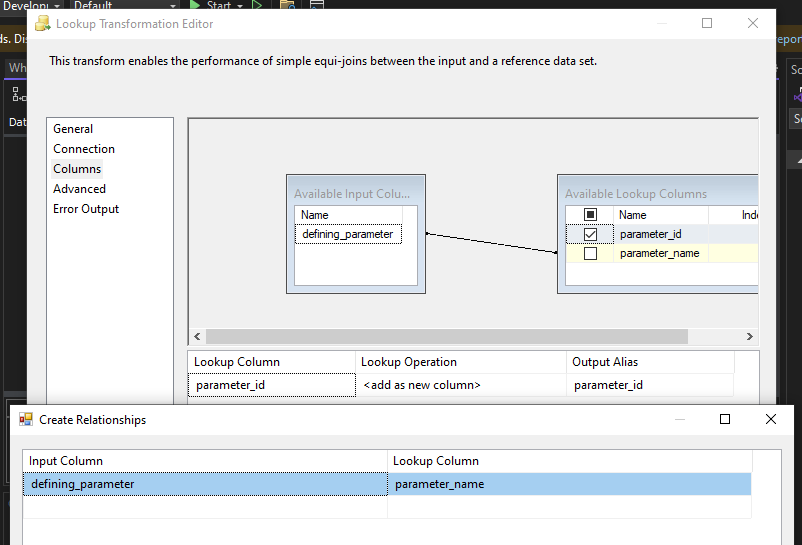


*Các column thu được*

**Bước 2: Kiểm tra dòng tồn tại**

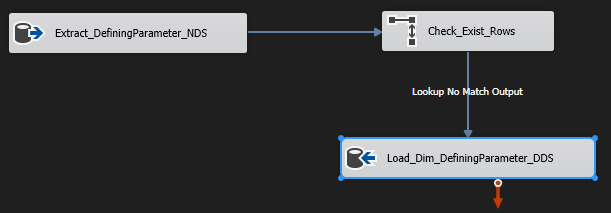
****

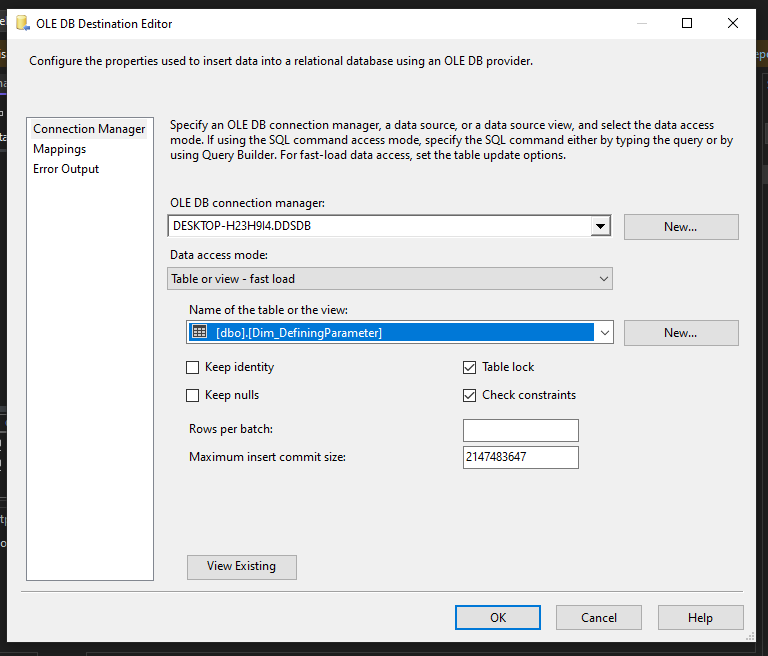
****

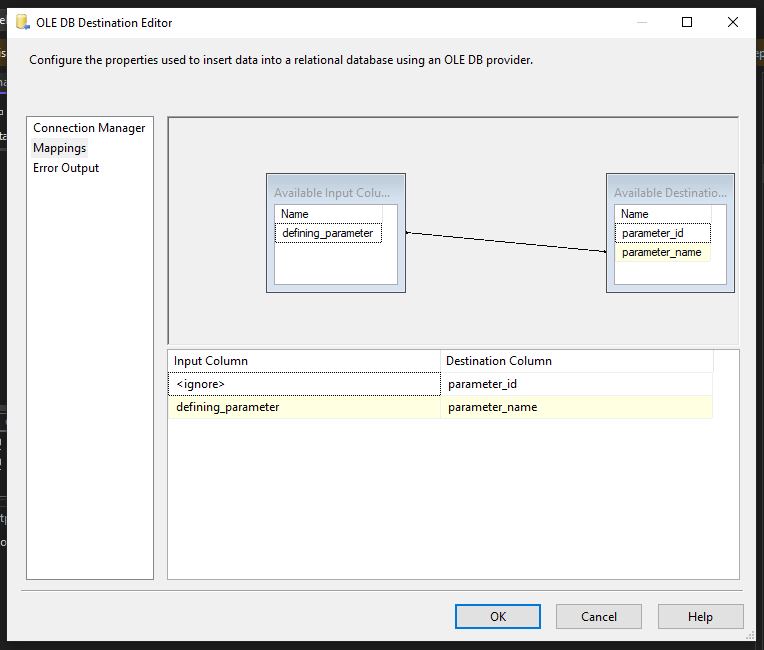
****

*Mapping thuộc tính và Lookup trên defining\_parameter vì đây là khóa chính của bảng Dim\_Definingparameter*

**Bước 4: Load vào bảng Dim\_Definingparameter DDS**

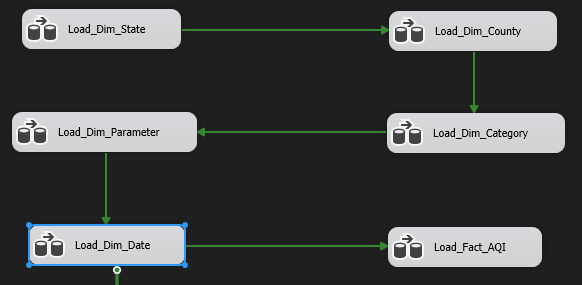
****

****

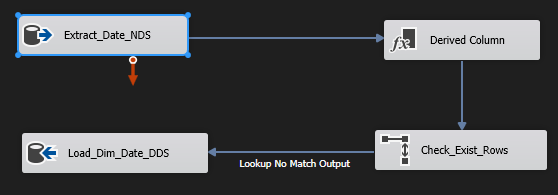
****

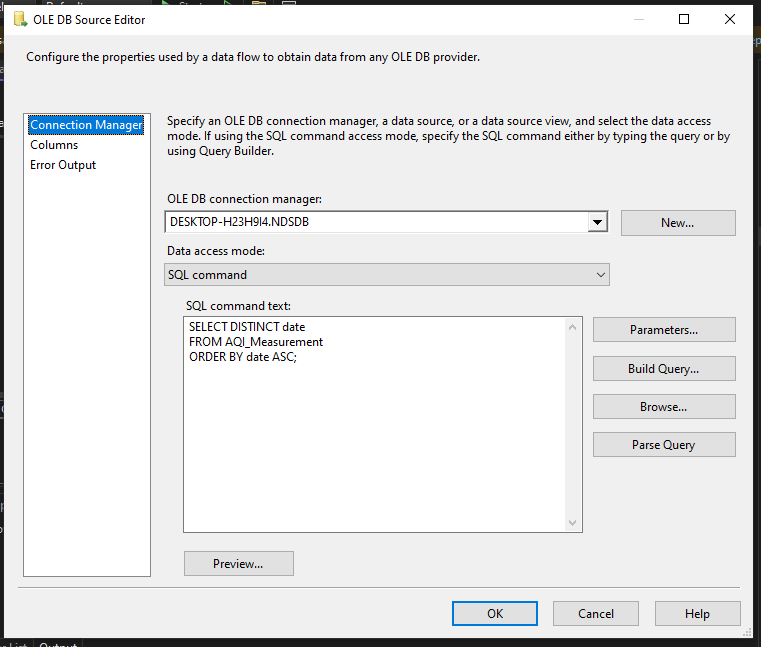
*Mapping (do parameter\_id là khóa tự tăng nên không cần map)*

#### ETL dữ liệu vào bảng Dim\_Date:

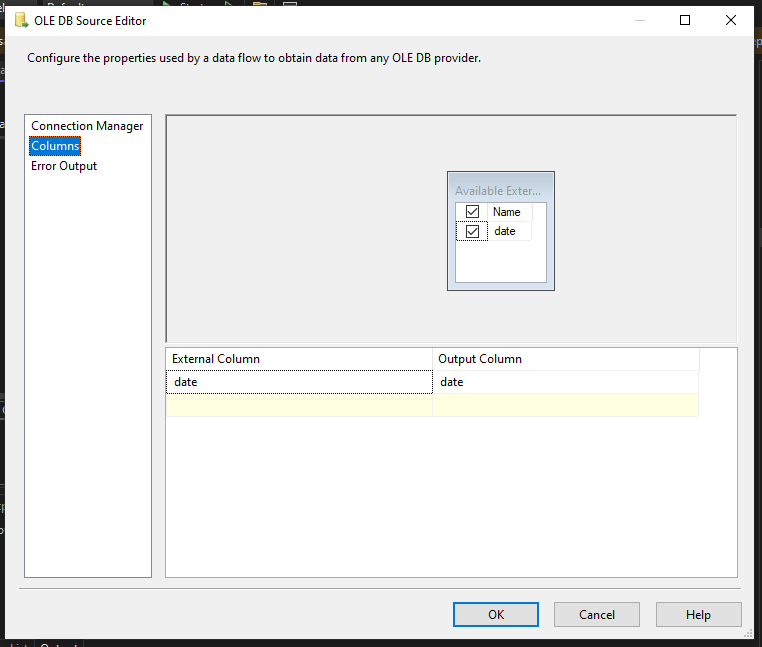
****

**Bước 1: Trích xuất dữ liệu liên quan đến date từ bảng AQI\_Measurement trong NDS**

****

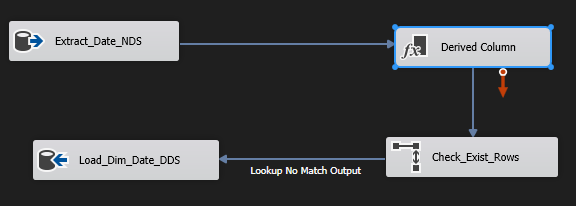


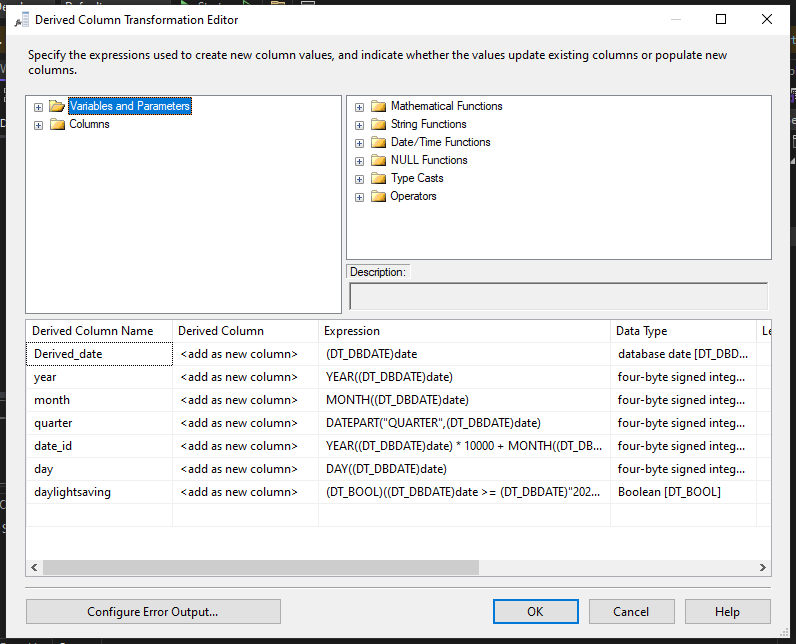
*Dùng code SQL trích xuất thuộc tính liên quan đến date trong bảng AQI\_Measurement ở NDSDB*

****

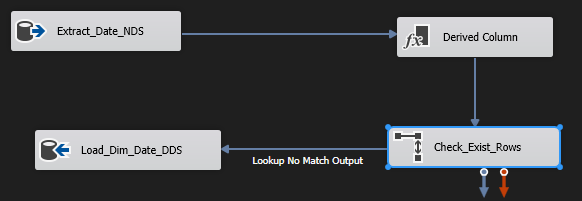
*Các column thu được*

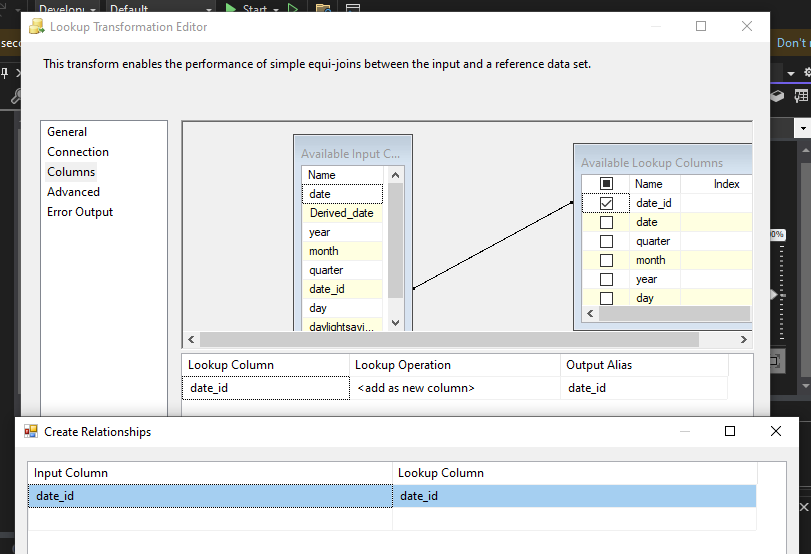
**Bước 2: Dẫn xuất, chuyển đổi và định dạng dữ liệu các cột**



**

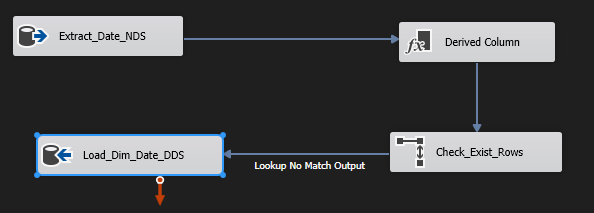
**Bước 3: Kiểm tra dòng tồn tại**

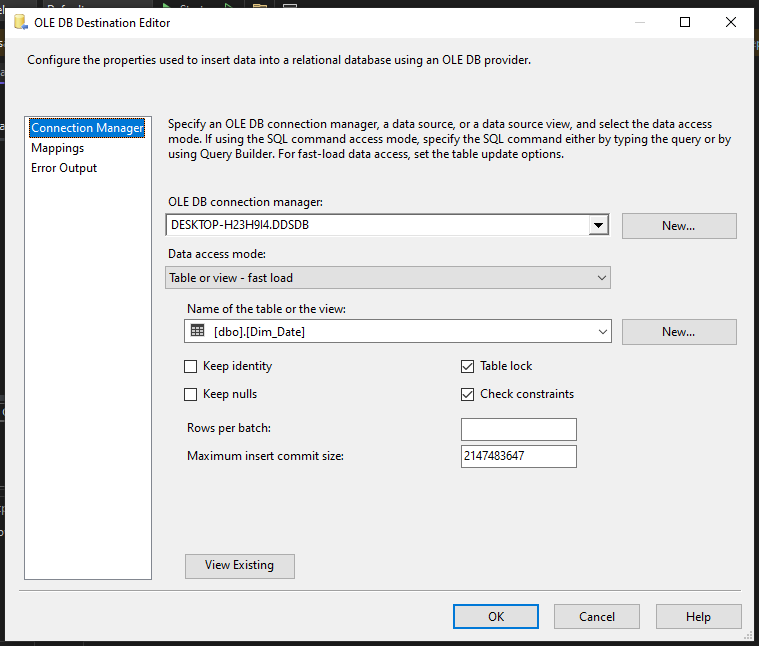




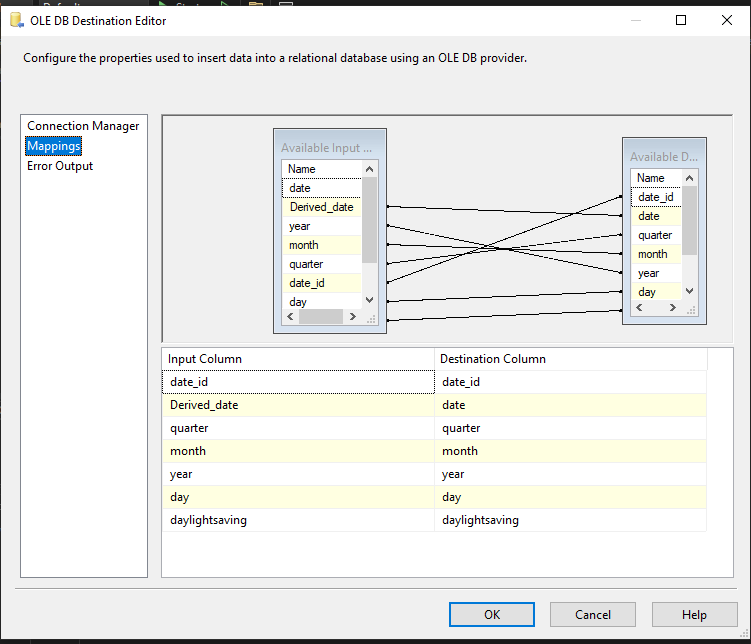
*Mapping thuộc tính và Lookup thuộc tính date\_id vì đây là khóa chính của bảng Dim\_Date*

**Bước 4: Load vào bảng Dim\_Date DDS**

****

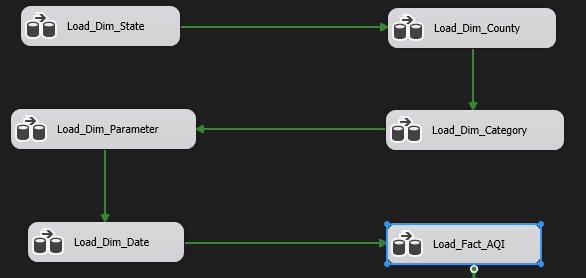


*Load dữ liệu vào bảng Dim\_Date trong DDSDB*

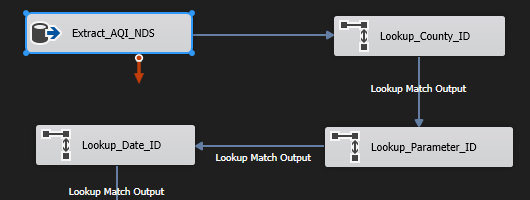
****

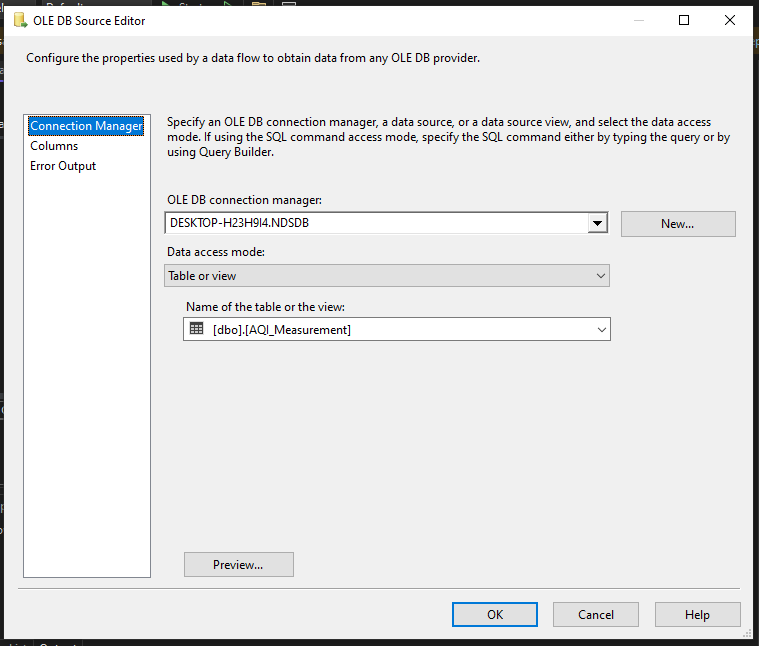
*Mapping*

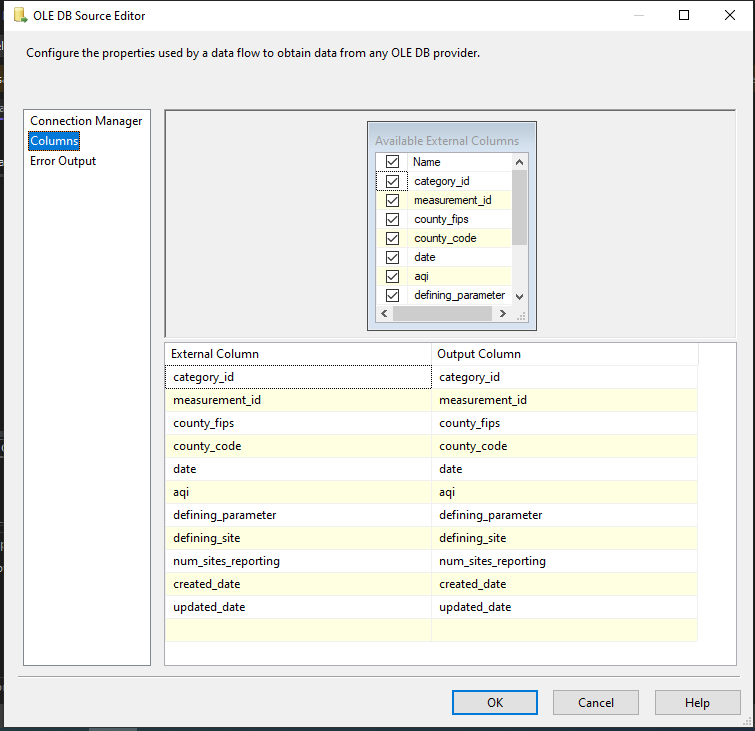
#### ETL dữ liệu vào bảng Dim\_FactAQI:

****

**Bước 1: Trích xuất dữ liệu từ bảng AQI\_Measurement trong NDS**

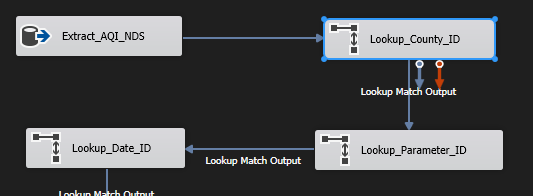
****

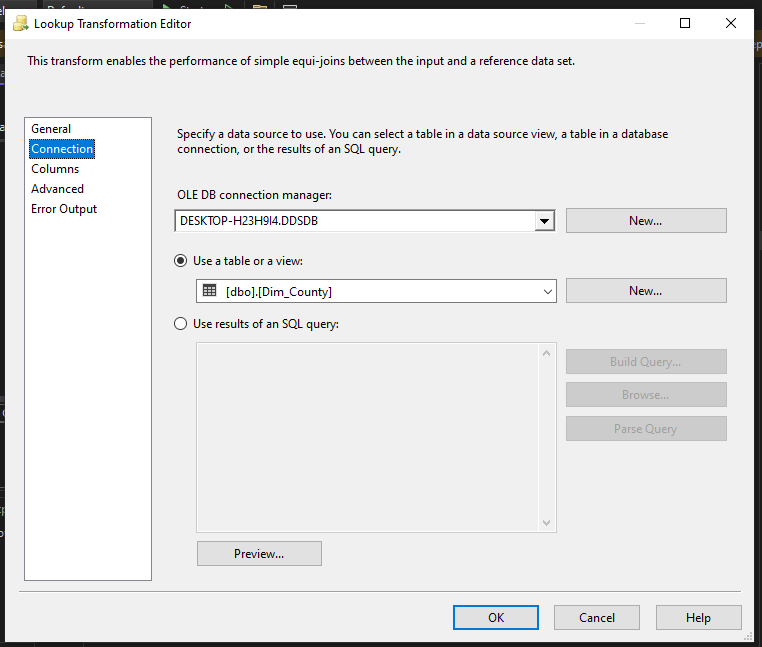


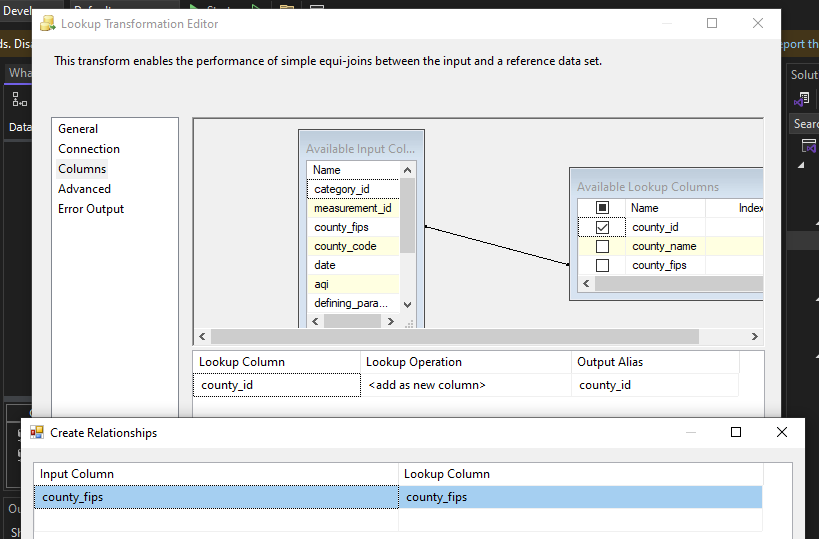
****

*Các column thu được*

**Bước 2: Kiểm tra dòng tồn tại (county\_id) trong bảng Dim\_County DDS**

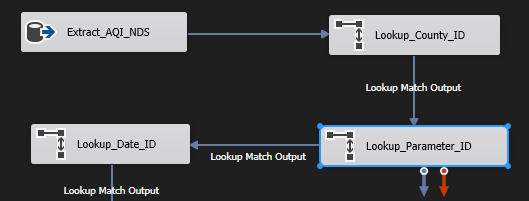


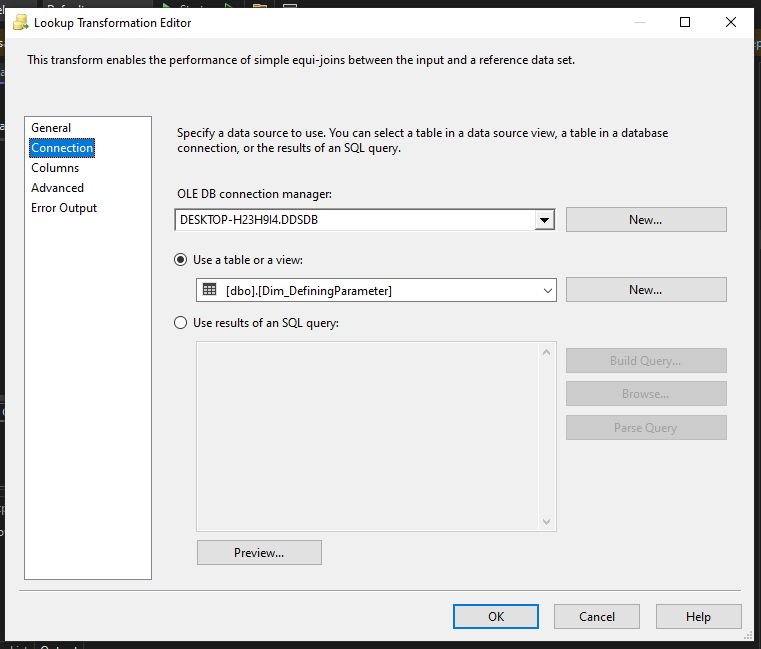


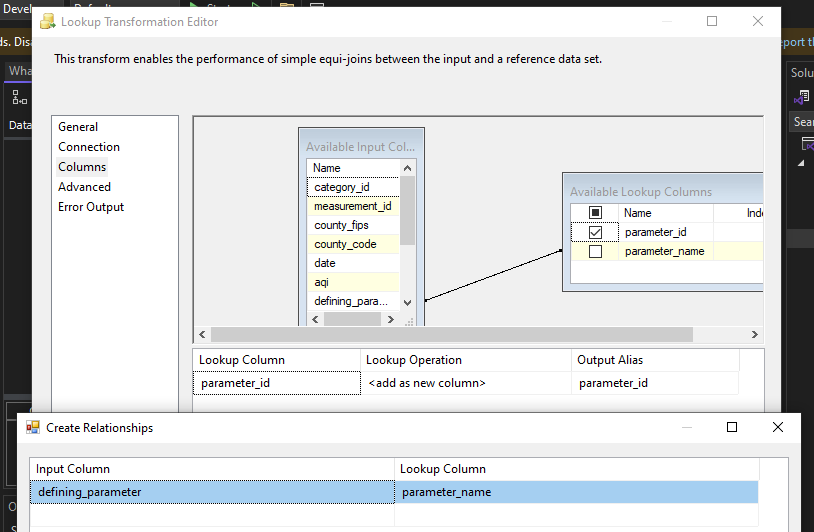


*Mapping thuộc tính và Lookup trên thuộc tính county\_id vì đây là khóa chính của bảng Dim\_County*

**Bước 3: Kiểm tra dòng tồn tại (parameter\_id) trong bảng Dim\_DefiningParameter DDS**

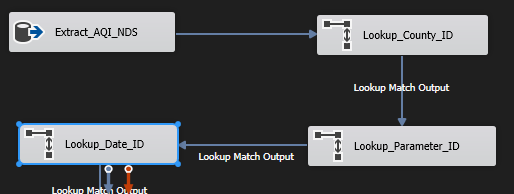


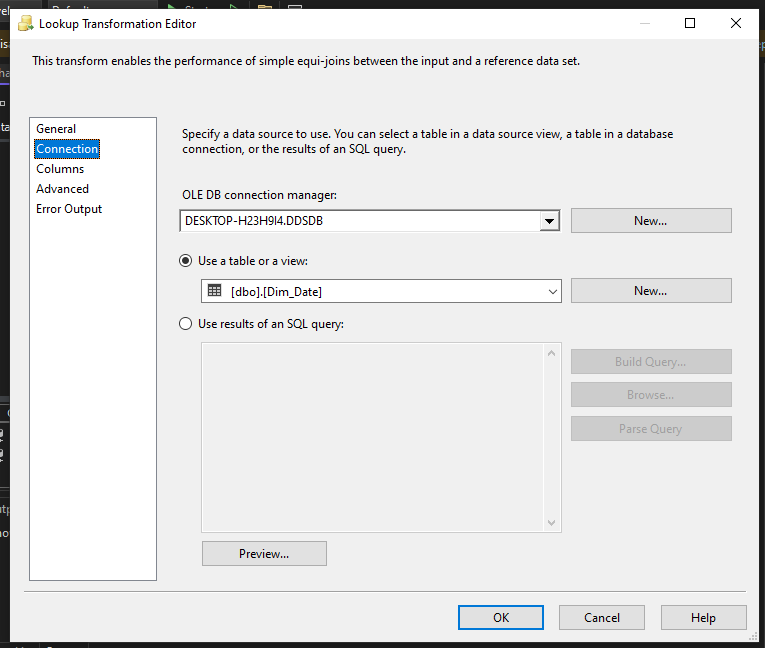


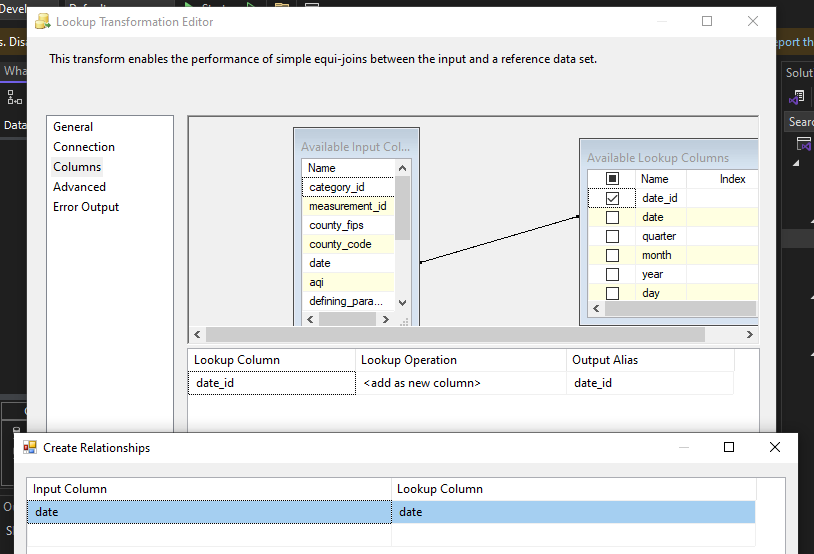


*Mapping thuộc tính và Lookup trên thuộc tính parameter\_id vì đây là khóa chính của bảng Dim\_DefiningParameter*

**Bước 4: Kiểm tra dòng tồn tại (date\_id) trong bảng Dim\_Date DDS**

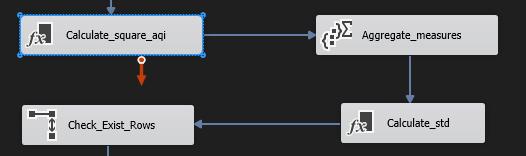


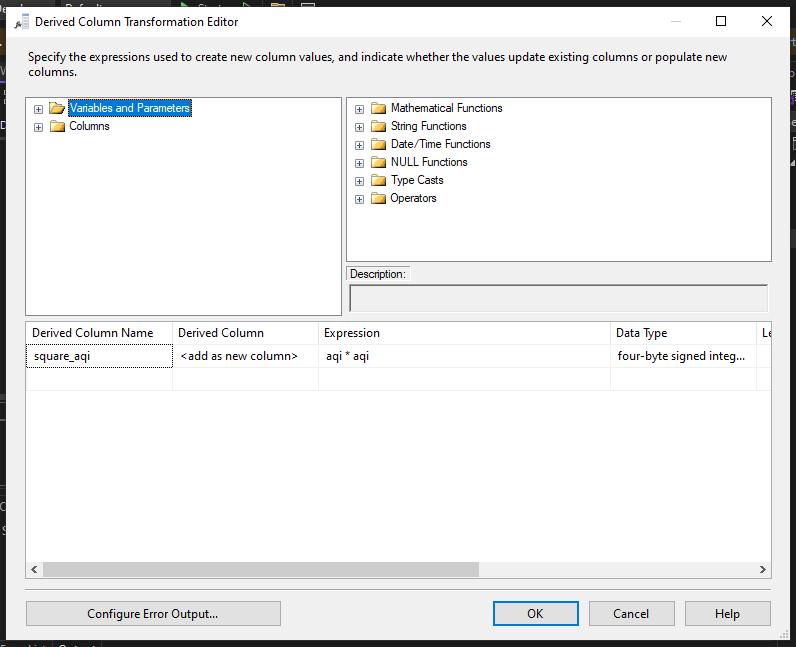




*Mapping thuộc tính và Lookup trên thuộc tính date\_id vì đây là khóa chính của bảng Dim\_Date*

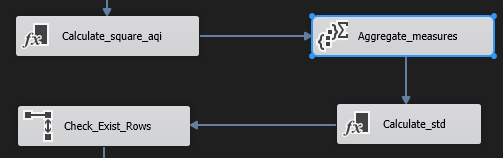
**Bước 5: Dẫn xuất. chuyển đổi và định dạng dữ liệu các cột**

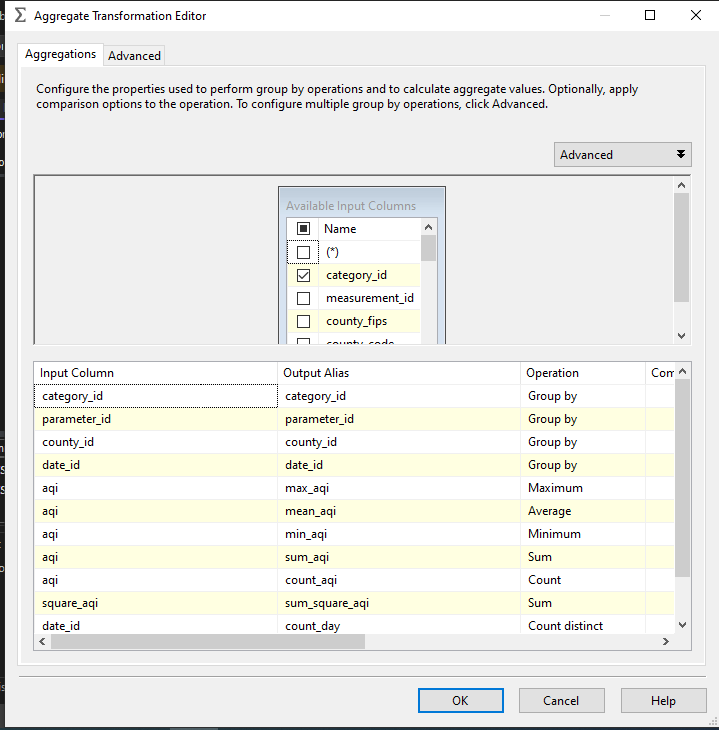


**

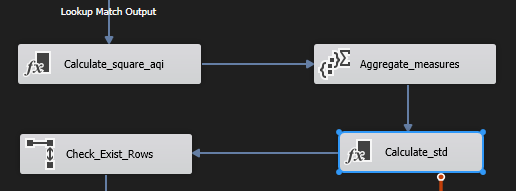
*Thêm cột square\_aqi với công thức tính toán square\_aqi = aqi \* aqi*

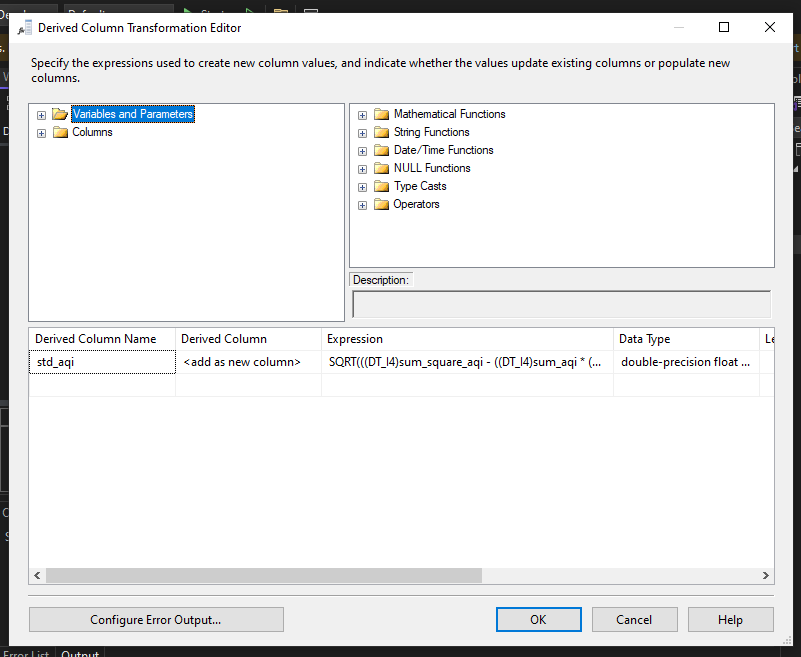
**Bước 6: Tính toán tổng hợp các giá trị measure**

****

****

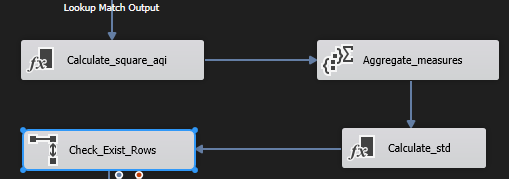
**Bước 7: Chuyển đổi và định dạng dữ liệu các cột**

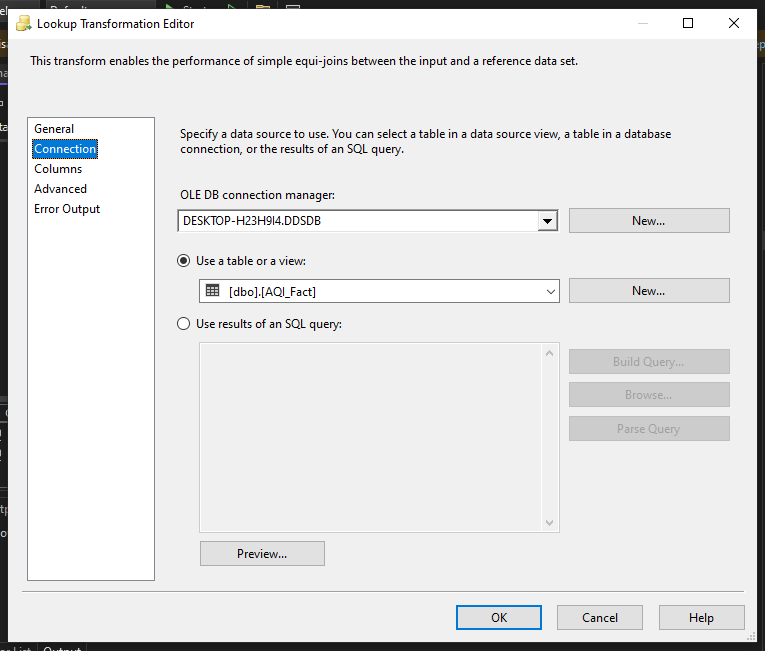


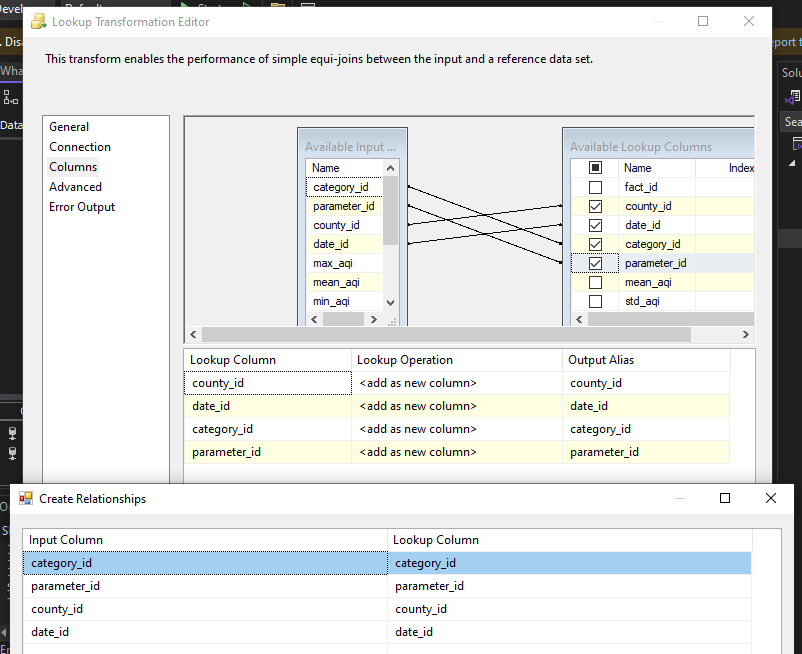
**

*Thêm cột std\_aqi được với công thức tính toán*

**Bước 8: Kiểm tra dòng tồn tại**

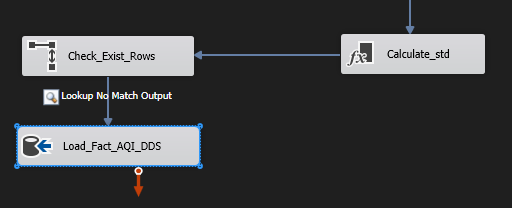


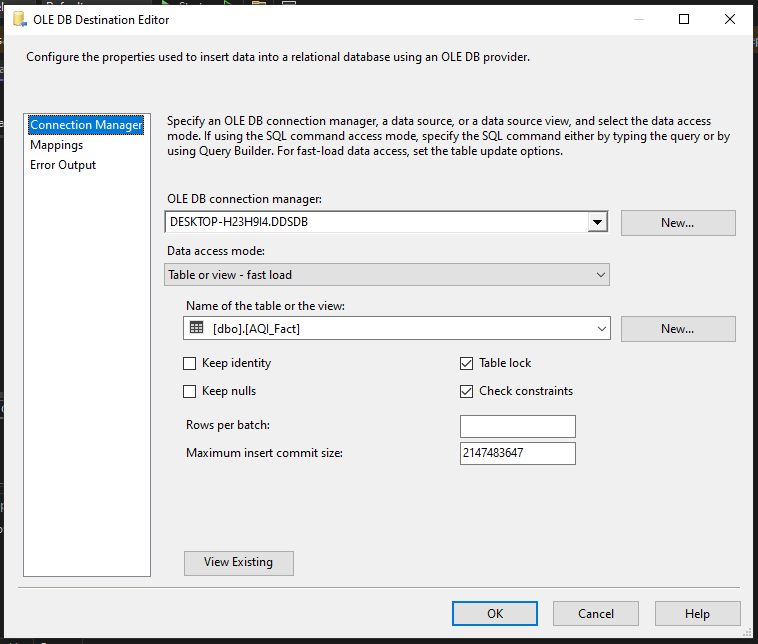




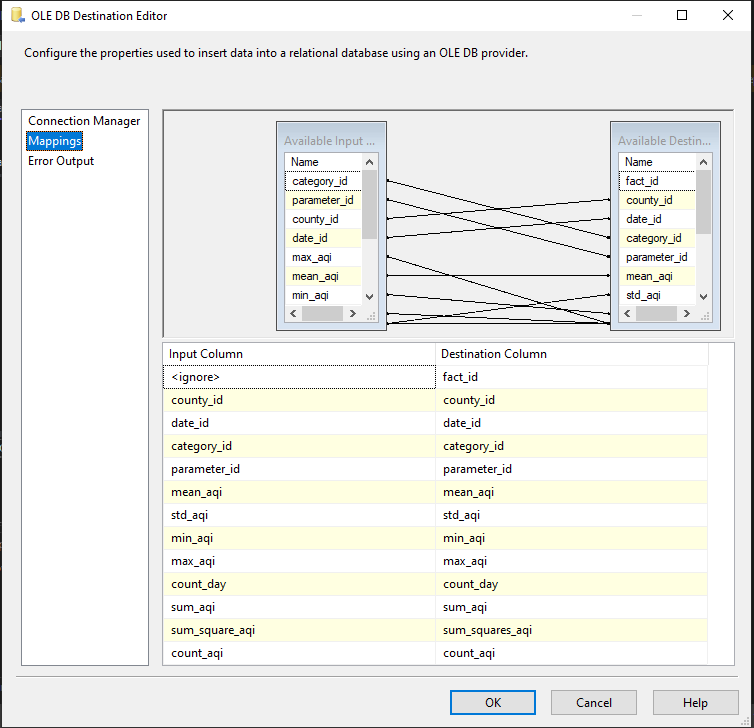
*Mapping thuộc tính và Lookup trên thuộc tính county\_id, date\_id, category\_id, parameter\_id,*

**Bước 4: Load vào bảng AQI\_Fact DDS**

****

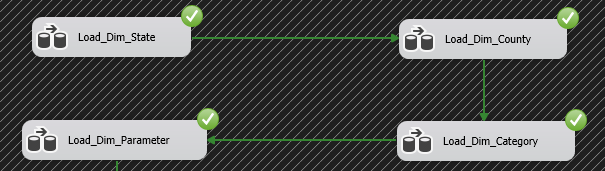


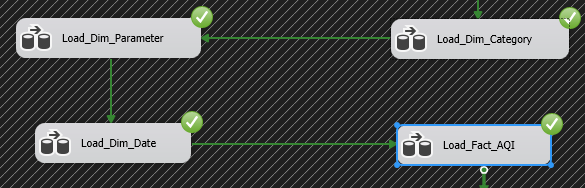
*Load dữ liệu vào bảng AQI\_Fact trong DDSDB*

****

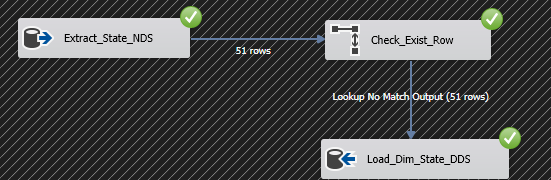
*Mapping (do fact\_id là khóa tự tăng nên không cần map)*

#### Kết quả chạy:

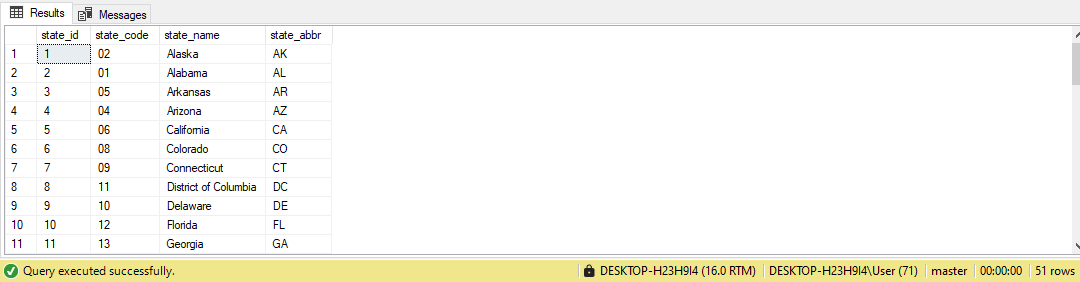
****

****

* **ETL dữ liệu vào bảng Dim\_State:**

****

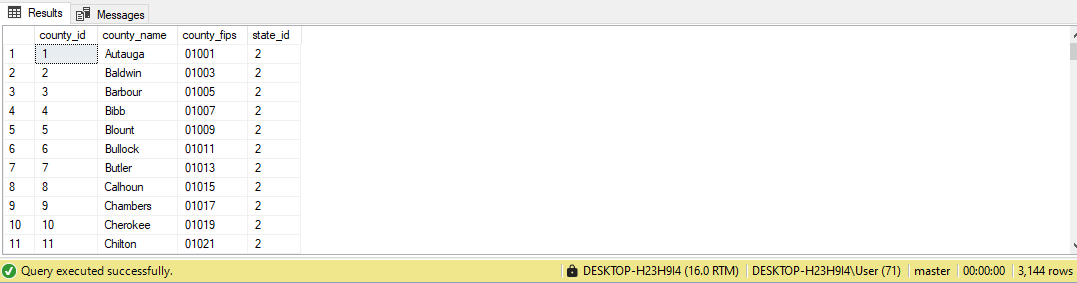
Vào DDSDB refresh và query bảng Dim\_State để kiểm tra kết quả:

****

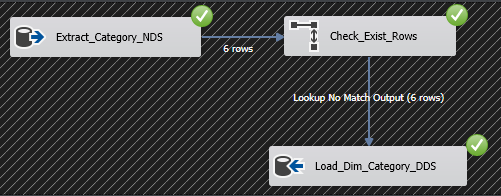
* **ETL dữ liệu vào bảng Dim\_ County:**

****

Vào DDSDB refresh và query bảng Dim\_County để kiểm tra kết quả

****

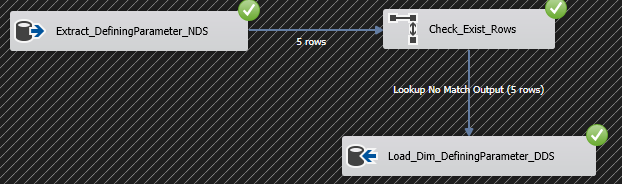
* **ETL dữ liệu vào bảng Dim\_Category:**

****

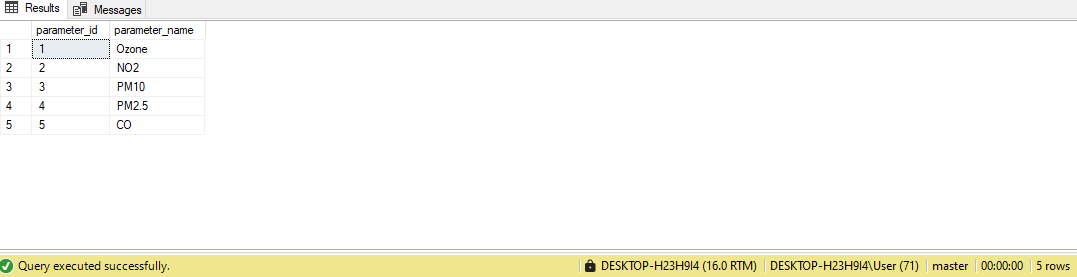
Vào DDSDB refresh và query bảng Dim\_Category để kiểm tra kết quả



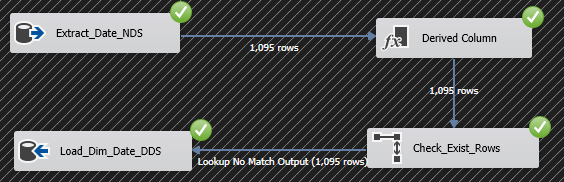
* **ETL dữ liệu vào bảng Dim\_DefiningParameter:**

****

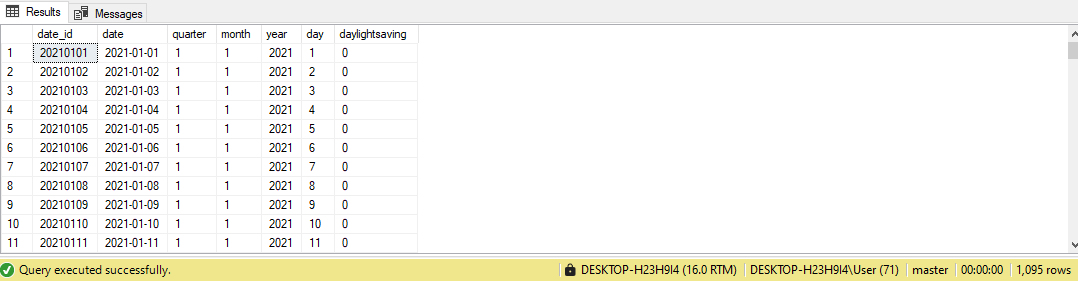
Vào DDSDB refresh và query bảng Dim\_DefiningParameter để kiểm tra kết quả



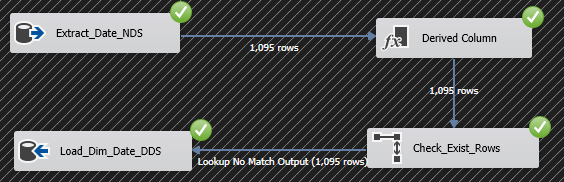
* **ETL dữ liệu vào bảng Dim\_Date:**



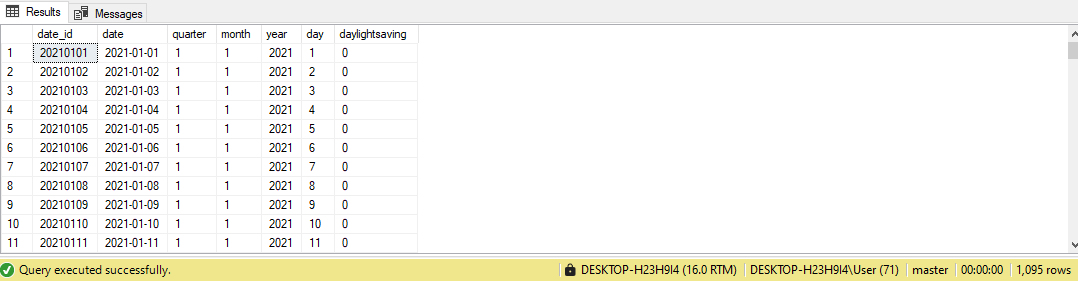
Vào DDSDB refresh và query bảng Dim\_Date để kiểm tra kết quả



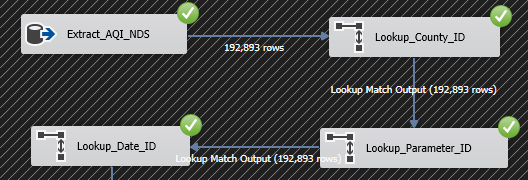
* **ETL dữ liệu vào bảng Dim\_Date:**



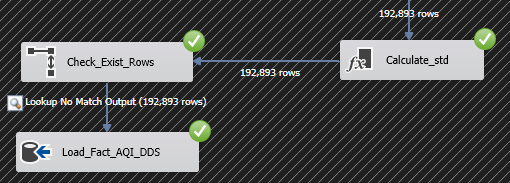
Vào DDSDB refresh và query bảng Dim\_Date để kiểm tra kết quả



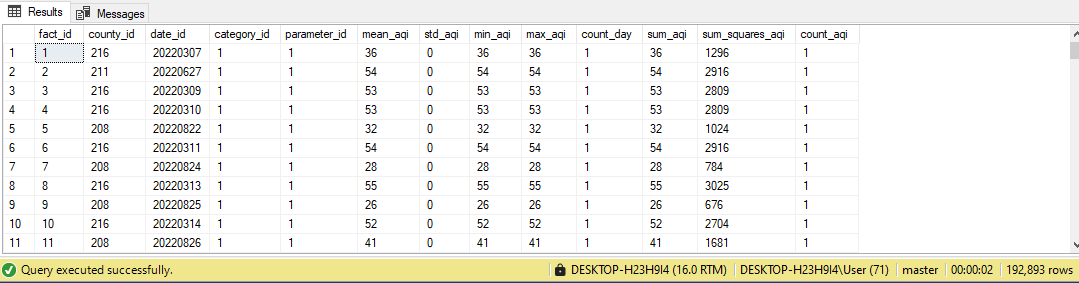
* **ETL dữ liệu vào bảng AQI\_Fact:**

****

****

****

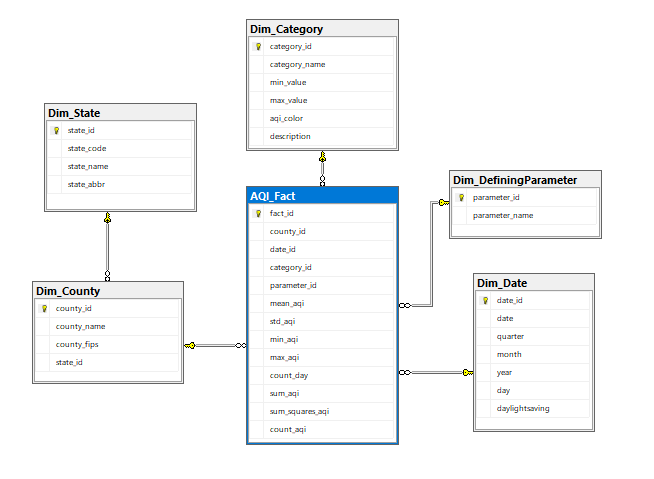
Vào DDSDB refresh và query bảng AQI\_Fact để kiểm tra kết quả



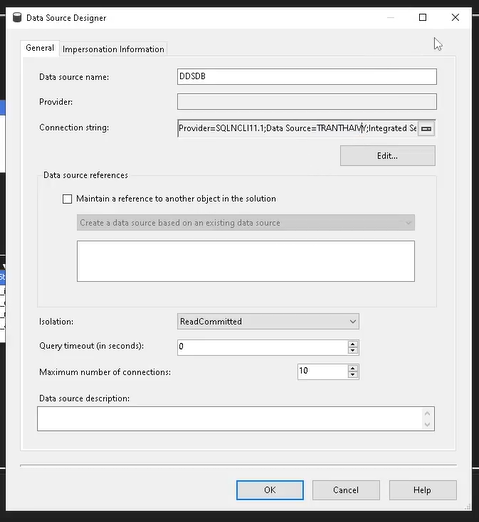
1. **OLAP:**

#### Tạo cube:

**Bước 1: Tạo csdl DDS trong SQL**

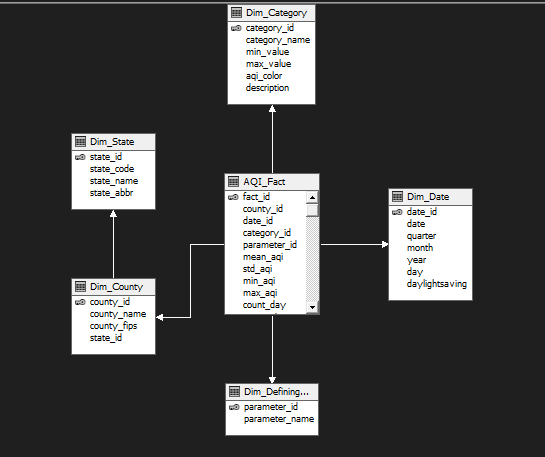
****

**Bước 2: Tạo data source**

****

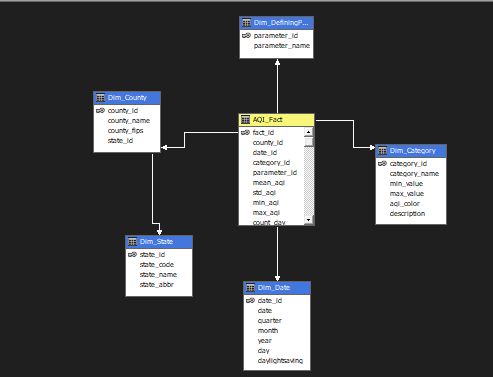
*Tạo kết nối với SQL server và load csdl DDS vừa tạo vào project*

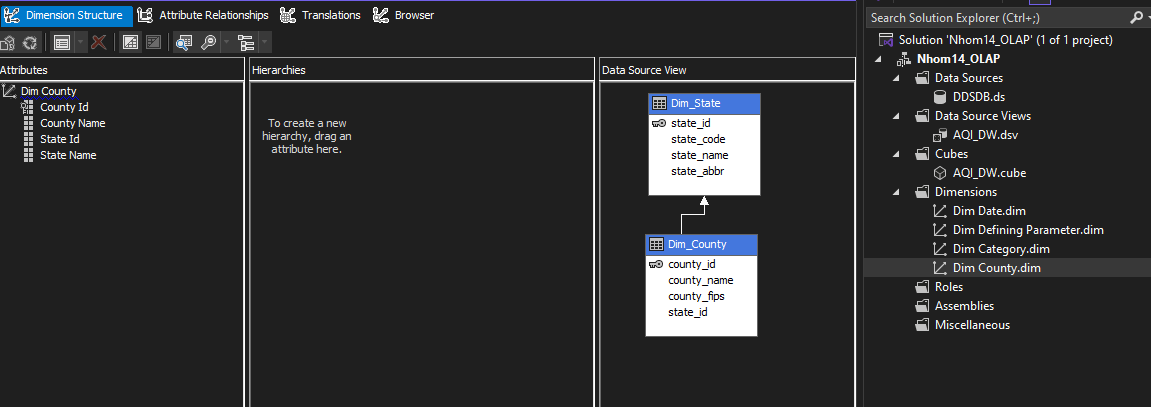
**Bước 2: Tạo data source view**

****

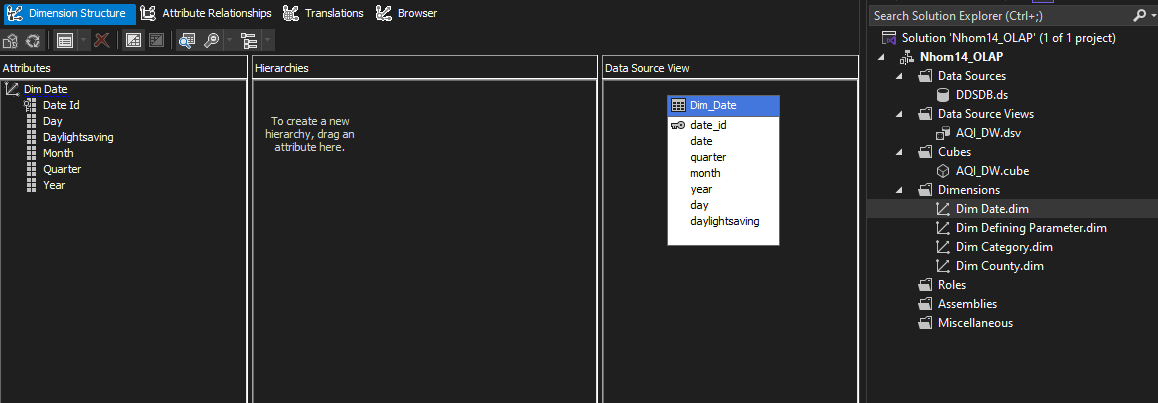
*Lấy những bảng cần sử dụng*

**Bước 3: Tạo cube, xác định measure và dimension:**

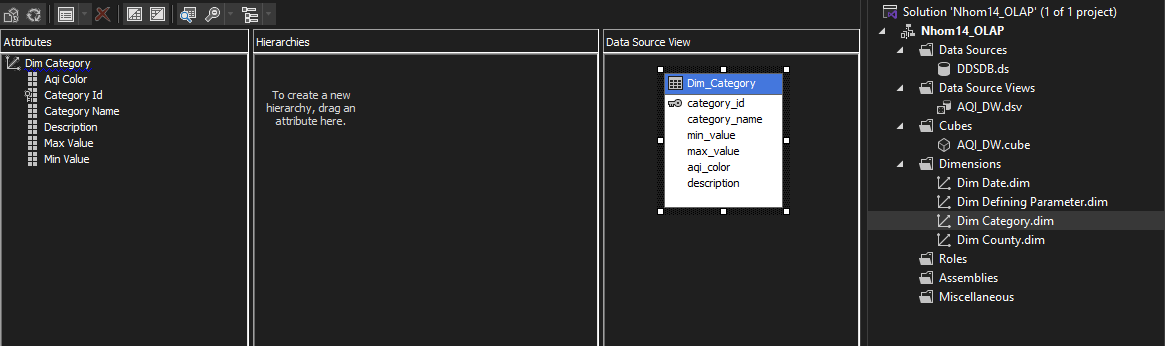
**Bước 4: Định nghĩa measure và dimension, chiều phân cấp từ các bảng đã chọn trong data source view**



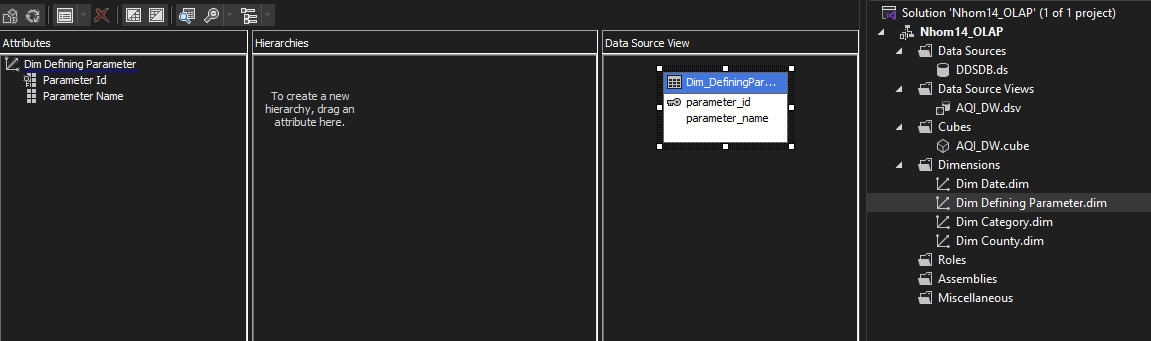
*Dim\_County*

**

*Dim\_Date*

**

*Category dim*

**

*Defining Parameter dim*

#### Truy vấn MDX:

**Câu 1: Đối với 4 tiểu bang Hawaii, Alaska, Illinois cà Delaware hãy tính giá trị AQI trung bình theo quý**

**Câu 2: Đối với 4 tiểu bang Hawaii, Alaska, Illinois cà Delaware hãy tính giá trị AQI trung bình theo quý**

**Câu 3: Đối với 4 tiểu bang Hawaii, Alaska, Illinois cà Delaware hãy tính giá trị AQI trung bình theo quý**

**Câu 4: Đối với 4 tiểu bang Hawaii, Alaska, Illinois cà Delaware hãy tính giá trị AQI trung bình theo quý**

**Câu 5: Đối với 4 tiểu bang Hawaii, Alaska, Illinois và Delaware hãy tính giá trị AQI trung bình theo quý. Đưa ra kết luận về những biến động**

**Truy vấn:**

**A screenshot of a computer

Description automatically generated**

**Nhận xét:**

* Hawaii:
  + Xu hướng: AQI thấp nhất vào quý 3 (~19.55), cao nhất vào quý 1 (~26.80).
  + Biến động nhỏ: Chất lượng không khí tốt, dao động nhẹ quanh mức 23.6.
  + Kết luận: Hawaii có chất lượng không khí ổn định, cải thiện rõ rệt vào quý 3.
* Alaska:
  + Xu hướng: AQI cao nhất vào quý 4 (~34.21), thấp nhất vào quý 2 (~25.89).
  + Biến động lớn hơn: Chất lượng không khí xấu đi vào mùa đông (quý 4).
  + Kết luận: Alaska có sự biến động mạnh, với mức ô nhiễm cao vào mùa đông.
* Illinois:
  + Xu hướng: AQI cao nhất vào quý 3 (~50.52), thấp nhất vào quý 4 (~41.68).
  + Biến động lớn: AQI tăng cao vào mùa hè do các hoạt động công nghiệp và giao thông.
  + Kết luận: Illinois có mức ô nhiễm tăng mạnh vào mùa hè và giảm vào mùa thu.
* Delaware:
  + Xu hướng: AQI tăng dần từ quý 1 (~42.44) đến quý 4 (~48.21).
  + Ổn định hơn: Không có biến động lớn giữa các quý.
  + Kết luận: Delaware có mức ô nhiễm tăng nhẹ về cuối năm, nhưng chất lượng không khí khá ổn định.

**Kết luận chung:**

* + Biến động lớn nhất: Alaska và Illinois có mức biến động rõ rệt, với AQI cao vào mùa đông (Alaska) và mùa hè (Illinois).
  + Ổn định nhất: Hawaii và Delaware có mức AQI ổn định, với Hawaii có chất lượng không khí tốt hơn.
  + Tác động mùa: Sự thay đổi AQI có liên quan đến khí hậu và các hoạt động của con người, đặc biệt là mùa đông ở Alaska và mùa hè ở Illinois.

**Câu 6: Thiết kế 1 báo cáo để chứng minh xu hướng biến động AQI trong năm cho bốn tiểu bang sau: Hawai, Alaska, Illinois và California. Đưa ra ý kiến đánh giá về những biến động AQI**

**Truy vấn:**

**A screenshot of a computer

Description automatically generated**

**Nhận xét về biến động AQI:**

* Hawaii: Chất lượng không khí tốt nhất với AQI trung bình thấp (~23.6), giảm vào mùa hè (tháng 8–9) và tăng nhẹ vào đầu và cuối năm.
* Alaska: Biến động lớn, AQI cao vào mùa đông (tháng 10–12, ~32–35) và giảm vào mùa hè (tháng 6–8, ~25–30). Có mức ô nhiễm cao vào mùa đông.
* Illinois: AQI tăng cao vào mùa hè (tháng 7, ~53.8) và giảm mạnh vào mùa thu (tháng 9–11). Mức ô nhiễm bị ảnh hưởng bởi hoạt động giao thông và công nghiệp.
* Delaware: Ổn định quanh năm (41–51), với AQI cao nhất vào tháng 12 (~51.2). Mức dao động thấp nhất trong các bang.

**Kết Luận:**

* Biến động AQI phụ thuộc vào mùa: mùa hè thường có AQI cao (Illinois), trong khi mùa đông AQI cao hơn ở Alaska và Delaware.
* Hawaii có chất lượng không khí tốt nhất, ổn định và ít biến động..

**Câu 8: Sử dụng bản đồ khu vực để biểu diễn trực quan (theo màu sắc) giá trị AQI trung bình ở các khu vực trong một năm**

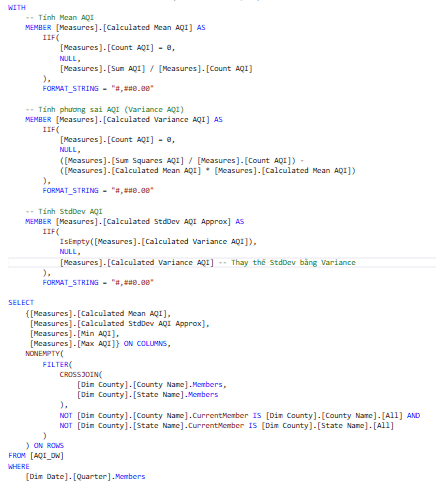
**Truy vấn:**

**A screenshot of a computer program

Description automatically generated**

**Câu 9: Báo cáo giá trị trung bình, độ lệch chuẩn, giá trị min và max của nhóm giá trị AQI theo tiểu bang và hạt trong mỗi quý của năm. So sánh các giá trị độ lệch chuẩn giữa câu hỏi 1 và 2, giải thích:**

**Truy vấn:**

****

A screenshot of a computer

Description automatically generated

**Câu 10: Tạo thêm 1 thuộc tính mới DayLightSaving trong 1 bảng phù hợp. DayLightSaving có 2 giá trị:**

* **Đúng: từ ngày 12/3/2023 đến ngày 5/11/2023**
* **Sai: ngược lại**

**Báo cáo giá trị AQI trung bình theo Tiểu bang, Loại, DayLightSaving qua mỗi năm. Có sự khác biệt đáng kể nào về chất lượng không khí trong day saving time (DST) so với thời gian khác không?**

**Truy vấn:**

**A screenshot of a computer

Description automatically generated**

**Nhận xét:**

* Tác động tích cực trong thời gian DST:
* Các danh mục nghiêm trọng như "Unhealthy", "Very Unhealthy", và "Hazardous" thường có AQI trung bình giảm trong giai đoạn DST.
* Điều này cho thấy chất lượng không khí có xu hướng cải thiện đối với các tình huống ô nhiễm nặng.
* Tác động tiêu cực trong thời gian DST:
* Các danh mục "Good" và "Moderate" lại có AQI trung bình tăng trong nhiều tiểu bang, đặc biệt ở Alabama, California và Texas.
* Điều này có thể phản ánh mức độ ô nhiễm không khí tăng nhẹ trong các điều kiện không nghiêm trọng.

**Kết luận chung:**

* Daylight Saving Time (DST) có tác động khác nhau đến chất lượng không khí, phụ thuộc vào tiểu bang và danh mục ô nhiễm.
* Chất lượng không khí thường cải thiện với các danh mục nghiêm trọng trong thời gian DST, nhưng ngược lại, chất lượng không khí ở các danh mục nhẹ hơn có thể xấu đi.

**Câu 11: Đếm số của Bang theo Loại AQI mỗi tháng**

**Truy vấn:**

A screenshot of a computer

Description automatically generated

**Câu 12:**

**Truy vấn:**

A screenshot of a computer

Description automatically generated

**Nhận xét:**

* Tình hình ô nhiễm không khí:
  + **Chất lượng tốt:** Phần lớn số ngày thuộc mức "Good" (CO: 232 ngày, Ozone: 72,522 ngày, PM2.5: 42,008 ngày).
  + **Mức trung bình và nhạy cảm:** PM2.5 và Ozone chiếm tỷ lệ cao ở mức "Moderate" và "Unhealthy for Sensitive Groups".
  + **Mức nguy hiểm:** Hiếm gặp, chỉ xuất hiện trong một số ngày.
* Nguyên nhân chính:
  + **PM2.5 và PM10:** Khí thải giao thông, đốt nhiên liệu hóa thạch, cháy rừng.
  + **Ozone:** Phản ứng VOC và NOx với ánh sáng mặt trời, tăng vào mùa hè.
  + **CO và NO2:** Khí thải từ giao thông và công nghiệp.
* Giải pháp cải thiện:
  + **Giao thông:** Khuyến khích xe điện, cải thiện giao thông công cộng.
  + **Cháy rừng:** Phòng chống và xử lý nhanh.
  + **Công nghiệp:** Tăng tiêu chuẩn kiểm soát khí thải.
  + **Giám sát:** Mở rộng mạng lưới giám sát không khí, cảnh báo sớm.
  + **Nhận thức:** Tuyên truyền để cộng đồng hiểu rõ và hành động bảo vệ môi trường.

**Kết luận:**

* Hoa Kỳ đã cải thiện chất lượng không khí đáng kể, nhưng cần tập trung kiểm soát PM2.5 và Ozone.
* Các giải pháp giảm phát thải từ giao thông và công nghiệp, cùng với phòng chống cháy rừng, sẽ giúp nâng cao chất lượng không khí và bảo vệ sức khỏe cộng đồng.

1. **Link Github và video demo:**

Link github nhóm: <https://github.com/BoundlessWay/HTTTPVTTKD/tree/stagging>

Video demo ETL Soure -> Stage -> NDS: <https://drive.google.com/file/d/1xFuDKqKqs4T4gS92m7tPY_n7pmesUHZn/view?usp=sharing>

Video demo ETL NDS -> DDS: <https://drive.google.com/file/d/1zwV1RZR0w7ulwsKombWFouomstzjHXig/view?usp=sharing>

Video demo OLAP:

<https://drive.google.com/file/d/1T3khjyCgY31DIaKW81km8VlzXiXdS93j/view?usp=sharing>