**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  OLAP: Tạo Cube | Dương Khánh An | 100% | 100% |
| Tạo ETL NDS -> DDS  OLAP: Tạo Cube | Trần Thái Vỹ | 100% | 100% |
| Tạo ETL Source -> Stage, Stage -> ND | Võ Đức Lợi | 100% | 100% |
| Xác định và phân tích cấu trúc CSDL NDS và cấu trúc metadata liên quan. Tổng hợp và viết báo cáo | Hồ Nguyễn Trâm Anh | 80% | 80% |

**MỤC LỤC**

[**A.** **Yêu cầu của Đồ án/Bài tập** 4](#_Toc184762580)

[**B.** **Kết quả** 5](#_Toc184762581)

[**1.** **Source:** 5](#_Toc184762582)

[**2.** **Stage:** 7](#_Toc184762583)

[**3.** **NDS:** 7](#_Toc184762584)

[**4.** **DDS:** 9](#_Toc184762585)

[**4.1.** **Các bảng dimension:** 9](#_Toc184762586)

[**4.2.** **Bảng Fact:** 10](#_Toc184762587)

[**4.3.** **Phân tích thiết kế các Measure:** 11](#_Toc184762588)

[**5.** **ETL:** 13](#_Toc184762589)

[**1.1.** **ETL từ Soure vào Stage:** 13](#_Toc184762590)

[a. ETL County từ Source vào Stage 13](#_Toc184762591)

[b. ETL AQI từ Source vào Stage 18](#_Toc184762592)

[**1.2.** **ETL từ Stage vào NDS:** 30](#_Toc184762593)

[a. Nạp dữ liệu vào bảng State NDS: 30](#_Toc184762594)

[b. Nạp dữ liệu vào bảng County NDS: 35](#_Toc184762595)

[c. Nạp dữ liệu vào bảng AQI NDS: 40](#_Toc184762596)

[**1.3.** **ETL từ NDS vào DDS:** 46](#_Toc184762597)

[a. Nạp dữ liệu vào bảng Dim\_State: 46](#_Toc184762598)

[b. Nạp dữ liệu vào bảng Dim\_County: 48](#_Toc184762599)

[c. Nạp dữ liệu vào bảng Dim\_ Category: 48](#_Toc184762600)

[d. Nạp dữ liệu vào bảng Dim\_Parameter: 49](#_Toc184762601)

[e. Nạp dữ liệu vào bảng Dim\_Date: 49](#_Toc184762602)

[f. Nạp dữ liệu vào bảng Dim\_FactAQI: 50](#_Toc184762603)

[**6.** **OLAP:** 51](#_Toc184762604)

[**7.** **Link Github và video demo:** 55](#_Toc184762605)

**YÊU CẦU ĐỒ ÁN- BÀI TẬP**

|  |  |
| --- | --- |
| **Loại bài tập** | **□ Lý thuyết 🗹 Thực hành 🗹 Đồ án □ Bài tập** |
| **Ngày bắt đầu** | **29/10/2024** |
| **Ngày kết thúc** | **03/11/2024** |

# **Yêu cầu của Đồ án/Bài tập**

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

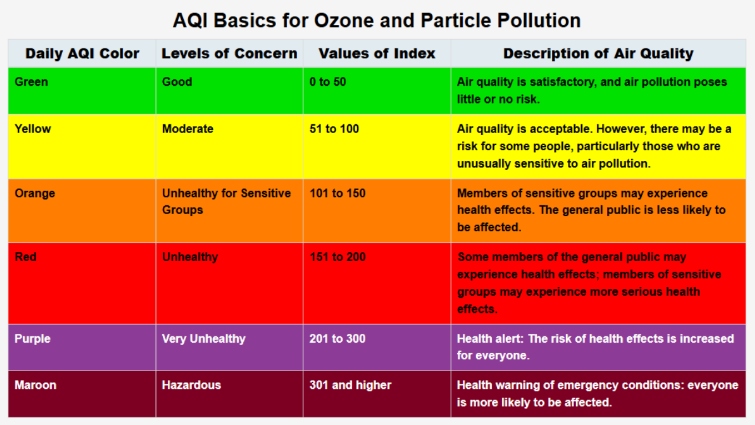
# **Kết quả**

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:

|  |  |  |  |
| --- | --- | --- | --- |
| **States** | **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\_abbreviation | Tên viết tắt của Bang | VARCHAR(2) | NOT NULL |

|  |  |  |  |
| --- | --- | --- | --- |
| **Counties** | **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) |  |
| populations | Dân số | INT |  |
| *state\_code* | *Mã code Bang. Khóa ngoại, tham chiếu đến state\_code của bảng States* | *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 Counties* | *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 |
| LowerBound | Mức AQI tối thiểu cho loại chất lượng không khí | INT | NULL |
| UpperBound | 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
* Mean AQI theo quý của 4 bang
* What: Giá trị trung bình AQI
* Where: 4 bang cụ thể
* When: Theo từng quý
* Why: Tìm hiểu sự dao động AQI trong năm
* Dimension cần: Dim\_Date, Dim\_County
* Measure: Mean(AQI\_Value)
* Granularity: Một bản ghi tương ứng với từng quận/ngày
* Xu hướng AQI trong năm
* What: Xu hướng AQI (Mean AQI).
* Where: Từng bang (4 bang cụ thể).
* When: Theo từng quý và năm.
* Why: Đánh giá sự thay đổi AQI theo thời gian.
* Dimension cần: Dim\_Date, Dim\_County.

Measure: Mean(AQI\_Value)

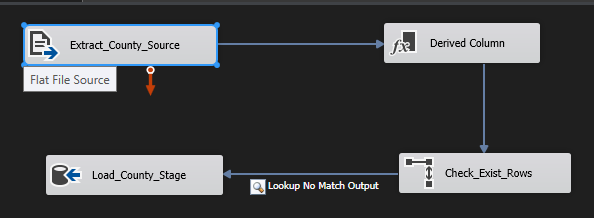
Granularity: Một bản ghi tương ứng với từng quận/ngày

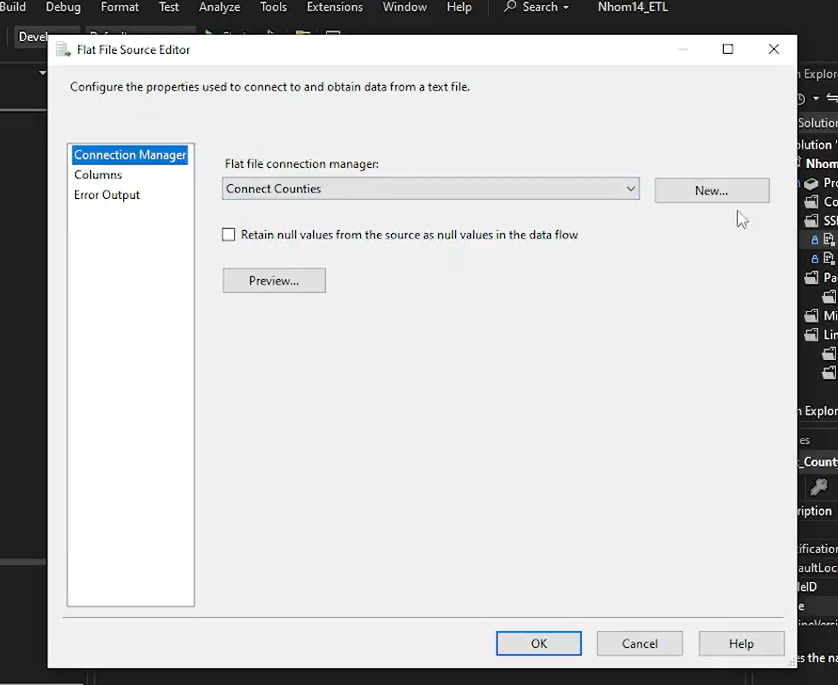
1. **ETL:**

### **ETL từ Soure vào Stage:**

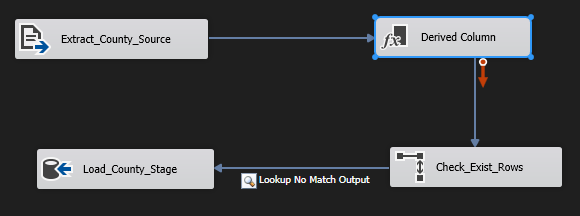
#### ETL County từ Source vào Stage

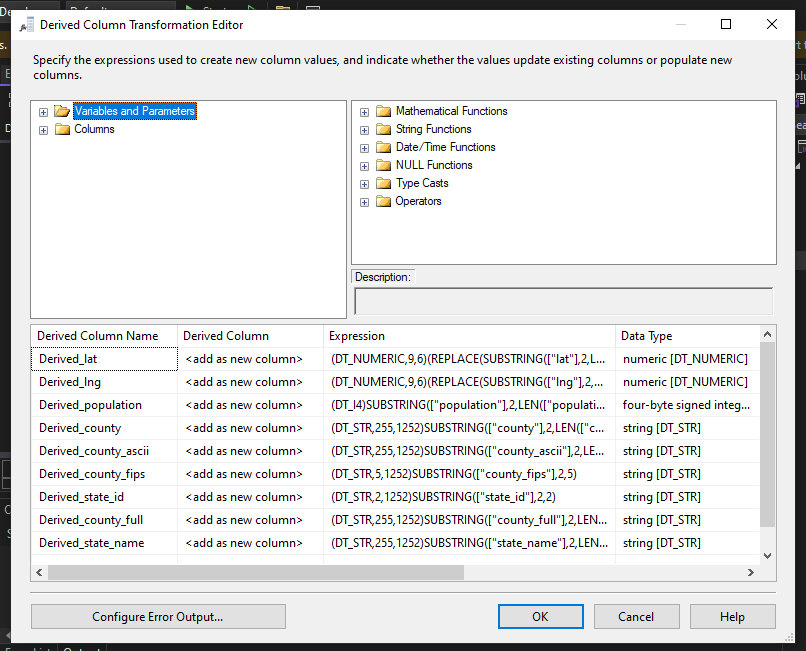
**Bước 1: Lấy dữ liệu County từ file (2B)uscounties.csv**



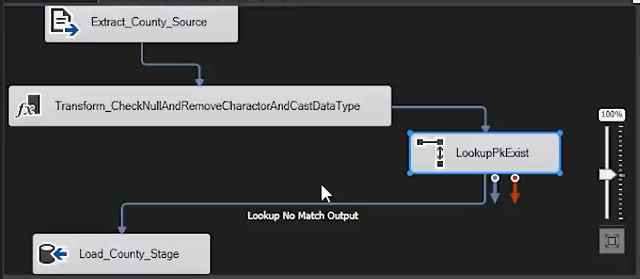


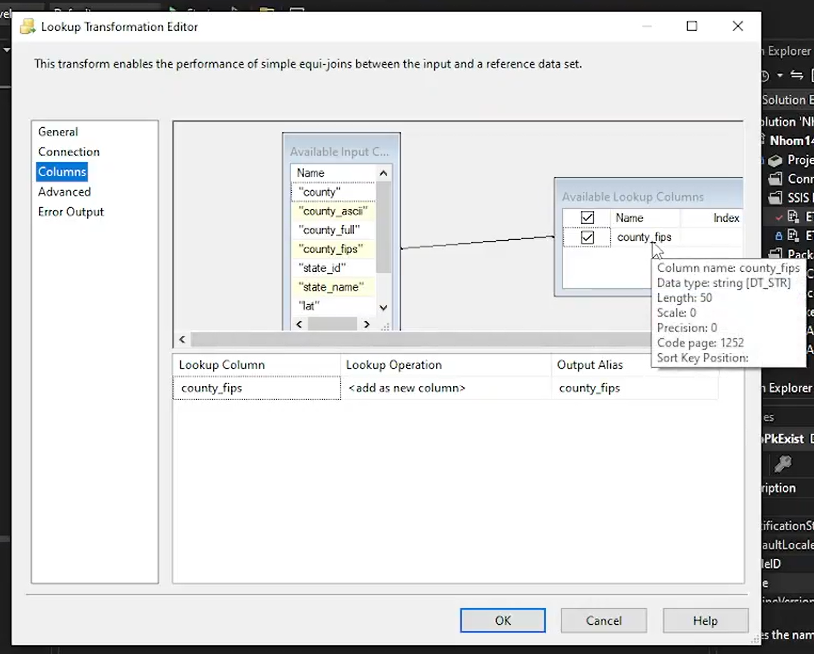
**Bước 2: Chuyển đổi và định dạng dữ liệu**



**

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



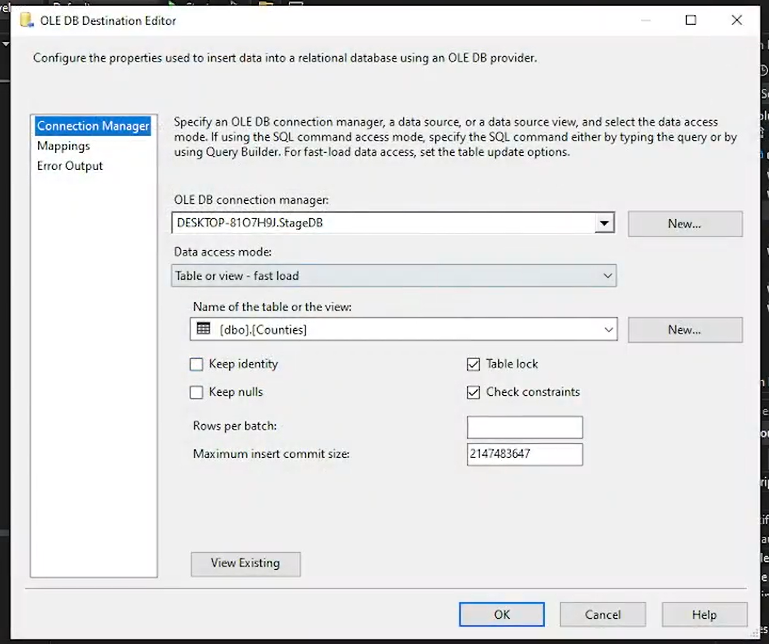


*Look up thuộc tính county\_fips vì đây là khóa định danh duy nhất của mỗi Hạt*

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

**Bước 4: Load vào Stage**

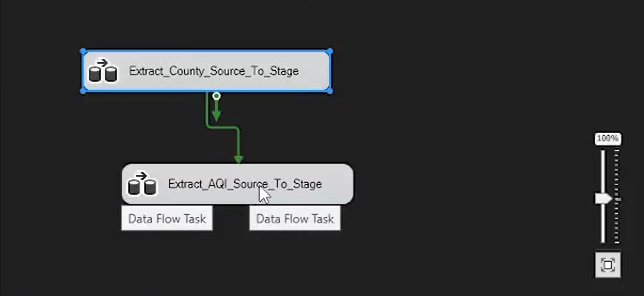




*Load vào Stage trong csdl*

#### ETL AQI từ Source vào Stage

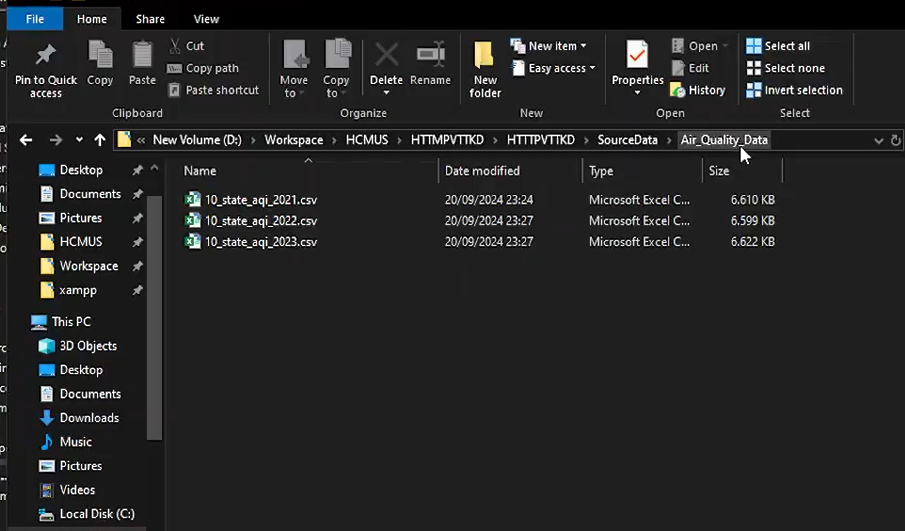
**Bước 1: Lấy dữ liệu từ các file AQI của từng năm**



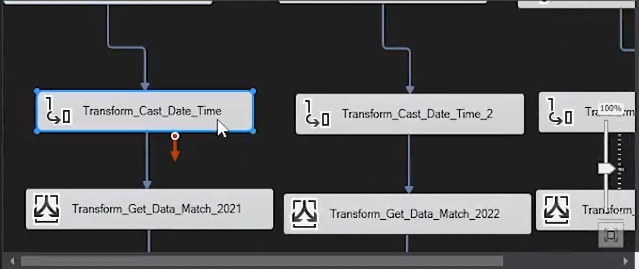
Vì có 3 file Soure là: 10\_state\_aqi\_2021, 10\_state\_aqi\_2022, 10\_state\_aqi\_2023 nên trước tiên sẽ tạo 3 source component

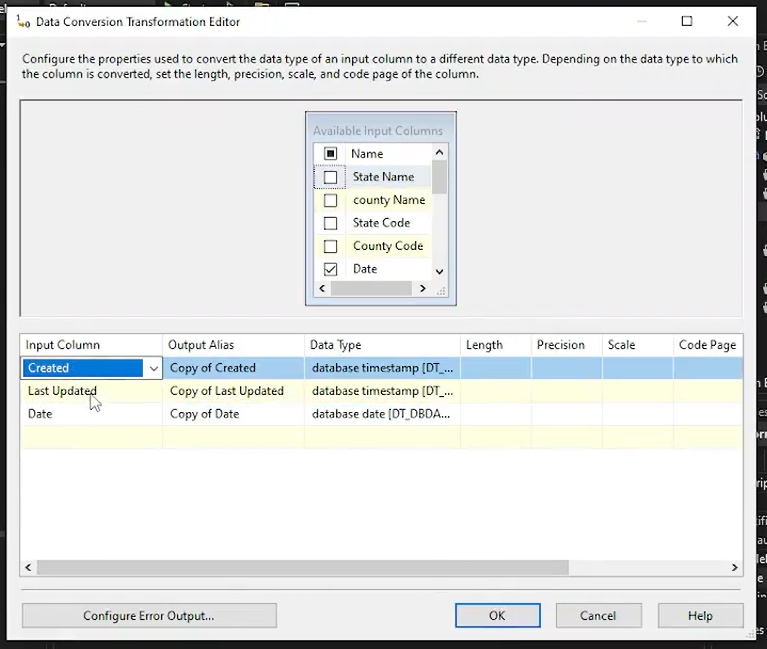


Mỗi source lấy dữ liệu từ các file tương ứng 10\_state\_aqi\_2021, 10\_state\_aqi\_2022, 10\_state\_aqi\_2023

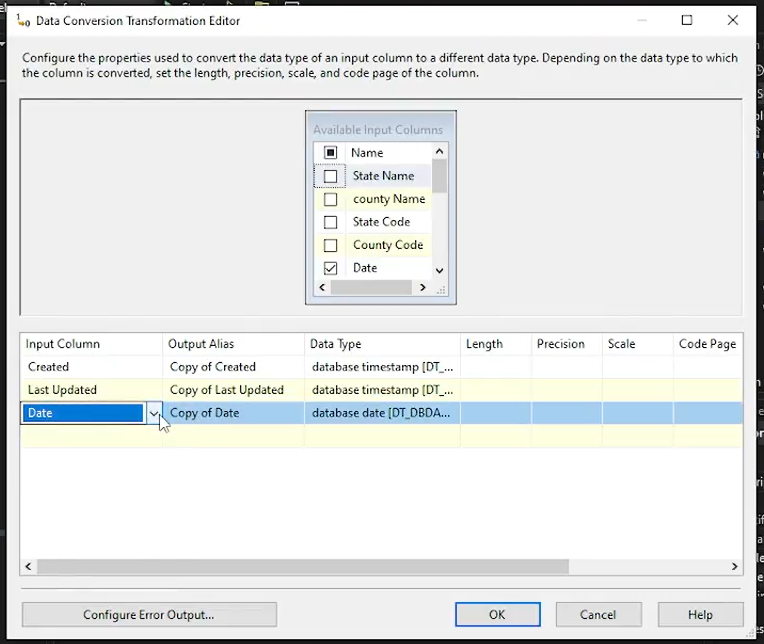


**Bước 2: Chuyển đổi và định dạng dữ liệu**



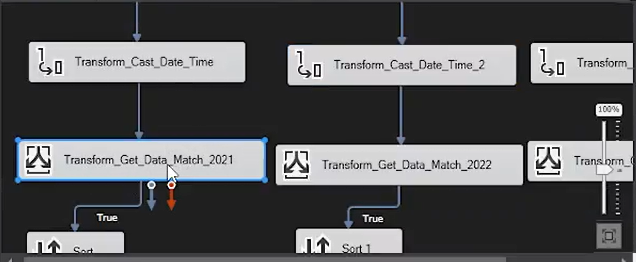


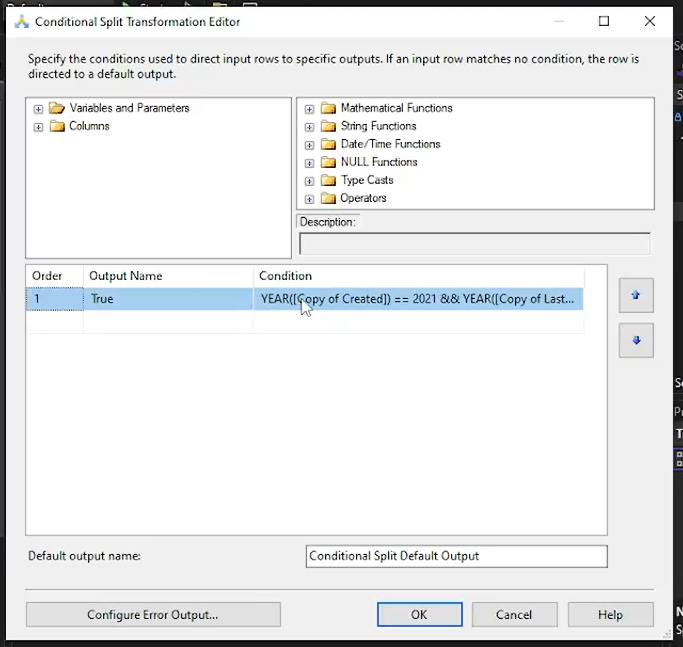
*Chuyển đổi kiểu dữ liệu của thuộc tính Created, Last Updated sang kiểu Timestamp để tương ứng với kiểu dữ liệu đã tạo trong SQL*

**

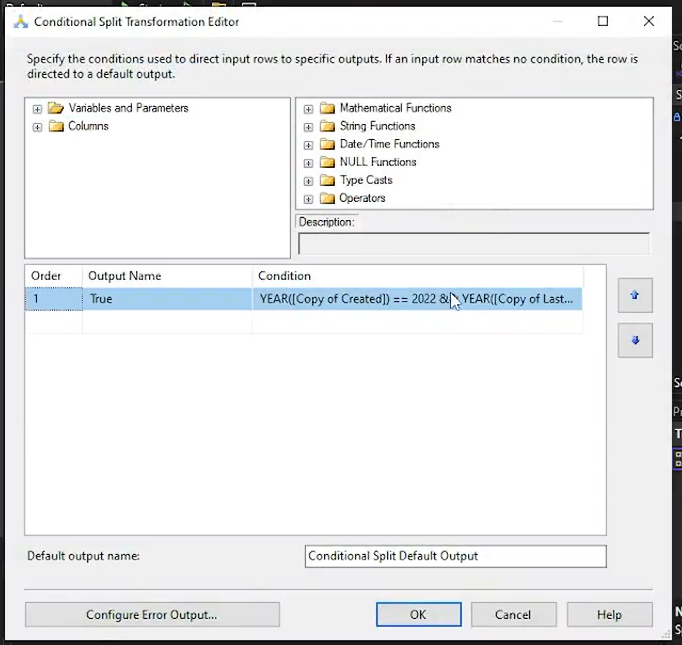
*Chuyển đổi kiểu dữ liệu của thuộc tính Date sang kiểu Date để tương ứng với kiểu dữ liệu đã tạo trong SQL*

**Bước 3: Lọc ra các dòng dữ liệu thỏa đúng là dữ liệu thuộc năm tương ứng**

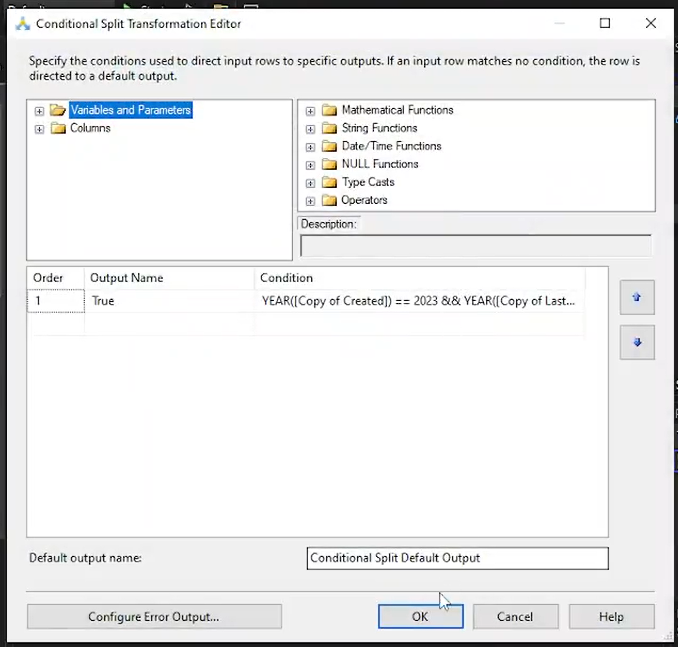




*Sử dụng câu điều kiện để kiểm tra xem các dòng dữ liệu được tạo có thuộc năm 2021 hay không*

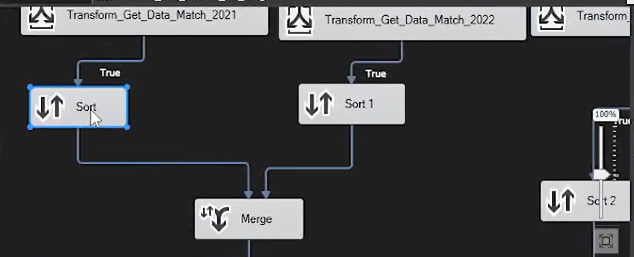


*Sử dụng câu điều kiện để kiểm tra xem các dòng dữ liệu được tạo có thuộc năm 2022 hay không*



*Sử dụng câu điều kiện để kiểm tra xem các dòng dữ liệu được tạo có thuộc năm 2023 hay không*

**Bước 4: Sắp xếp các dòng dữ liệu**



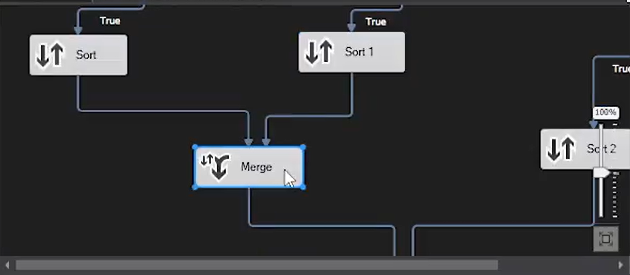
|  |  |
| --- | --- |
|  |  |

*Sắp xếp dữ liệu theo bộ khóa chính là StageCode, CountyCode, Created và LastUpdated*

Làm tương tự với dữ liệu của năm 2022 và 2023

**Bước 5: Merge dữ liệu**

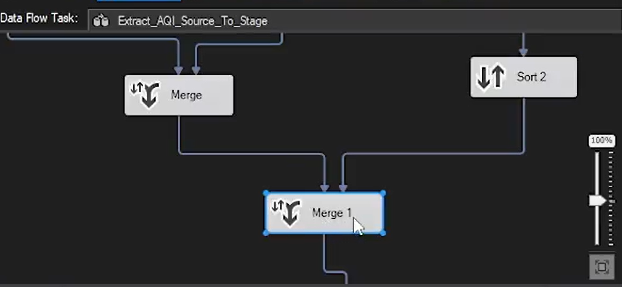
Lấy kết quả Sort của năm 2021 và 2022 Merge với nhau

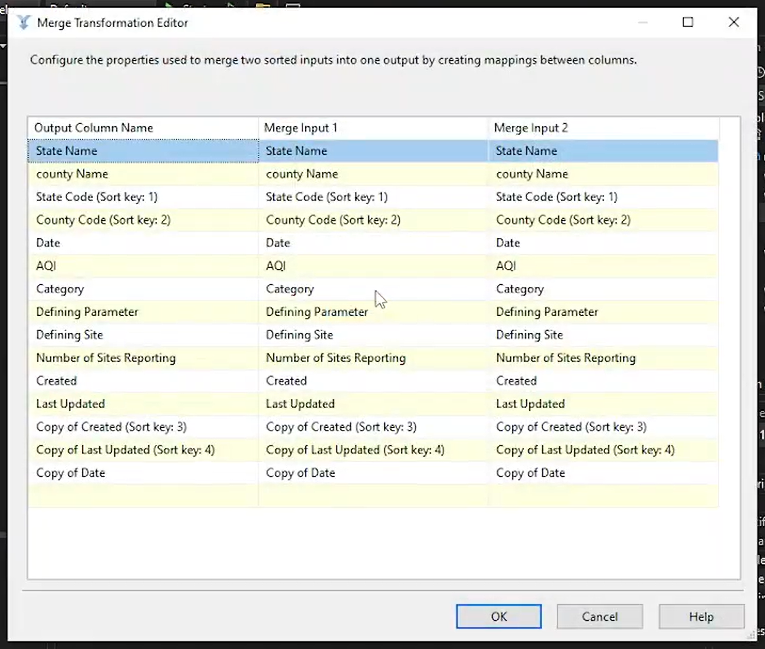


**

*Kết quả Sort của năm 2021 và 2022 Merge với nhau*

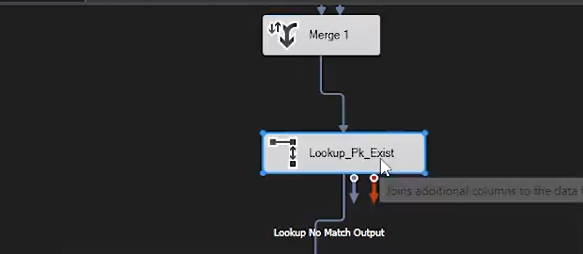
Lấy kết quả vừa Merge của năm 2021 và 2022 Merge với kết quả Sort của năm 2023

**

**

*Kết quả vừa Merge của năm 2021 và 2022 Merge với kết quả Sort của năm 2023*

**Bước 6: Kiểm tra Khóa chính tồn tại trong bảng để lọc ra các dòng đã tồn tại trong bảng**

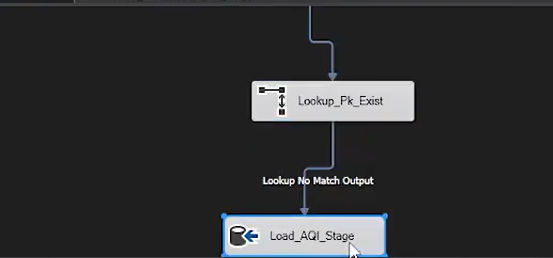




*Look up thuộc tính theo bộ khóa chính gồm county\_code, state\_code, last\_update, created*

Những hàng chưa tồn tại trong database sẽ được Load vào Stage

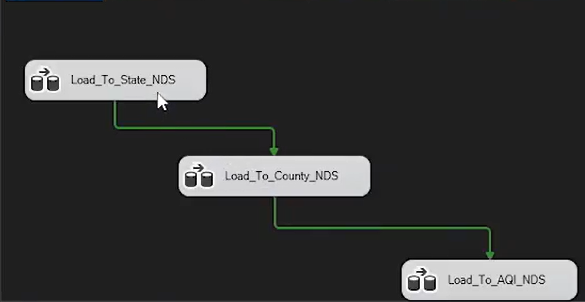
**Bước 7: Load vào Stage**



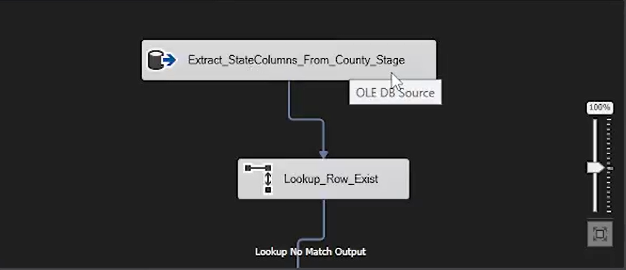
### **ETL từ Stage vào NDS:**

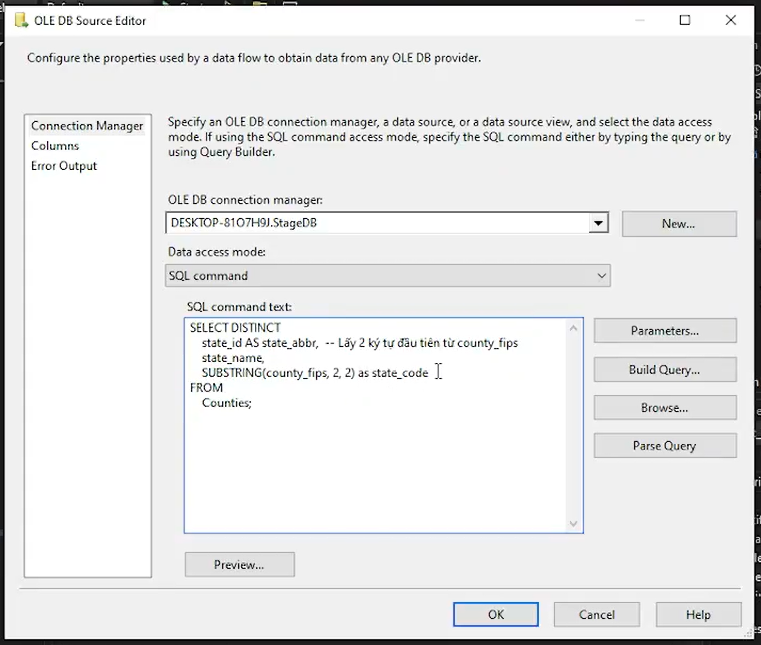
Trước khi nạp dữ liệu vào NDS thì cần thiết kế cơ sở dữ liệu cho NDS trước (các bảng dữ liệu thiết kế đã trình bày chi tiết ở mục 3.)

#### Nạp dữ liệu vào bảng State NDS:



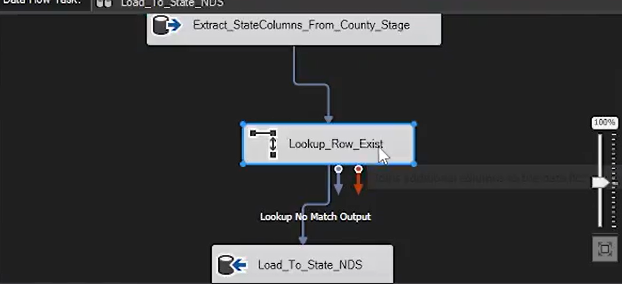
**Bước 1: Trích xuất các cột liên quan đến State trong bảng County trong Stage database**

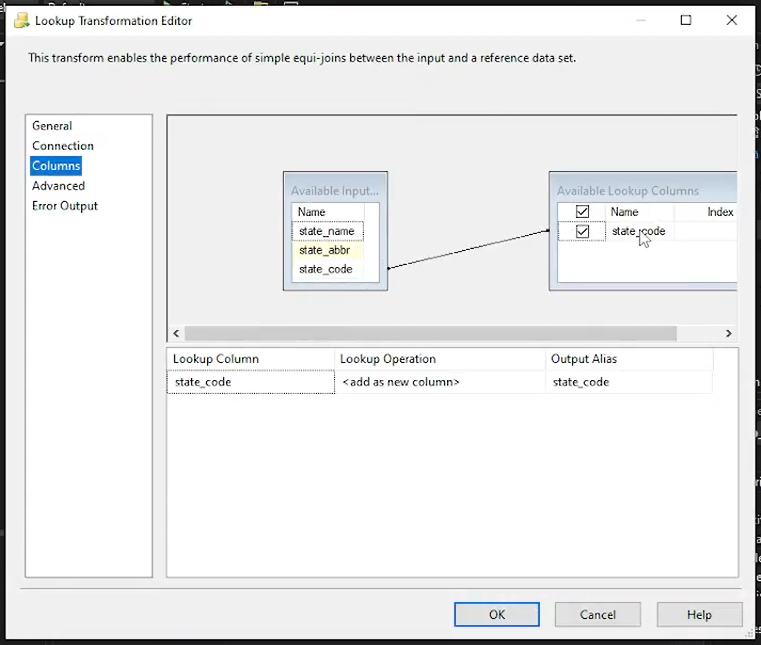
****

****

*Sử dụng code SQL để trích xuất thuộc tính liên quan đến Bang như tên Bang và mã Bang*

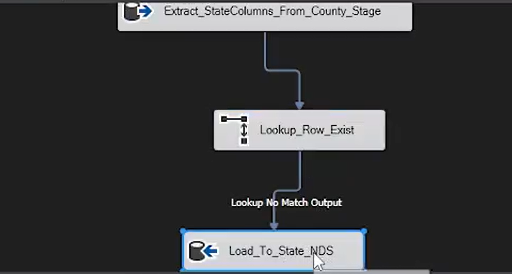
**Bước 2: Kiểm tra dòng tồn tại**

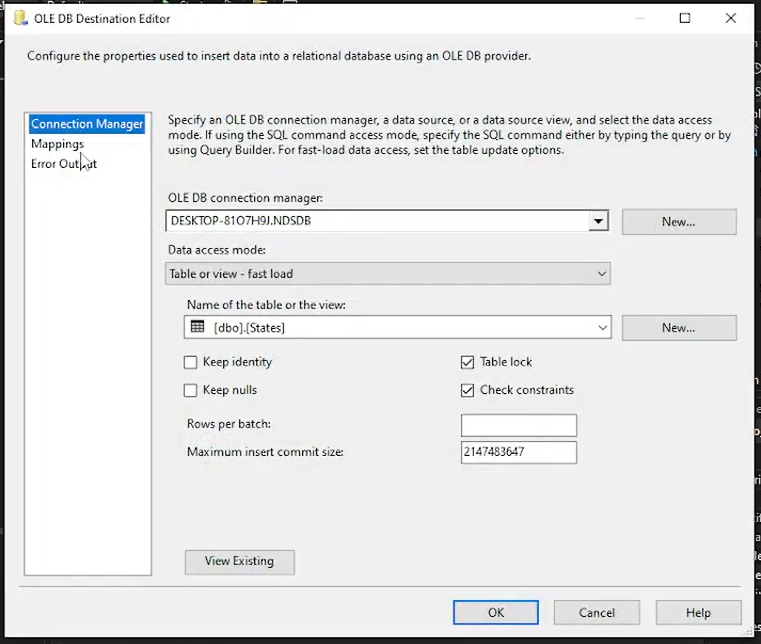


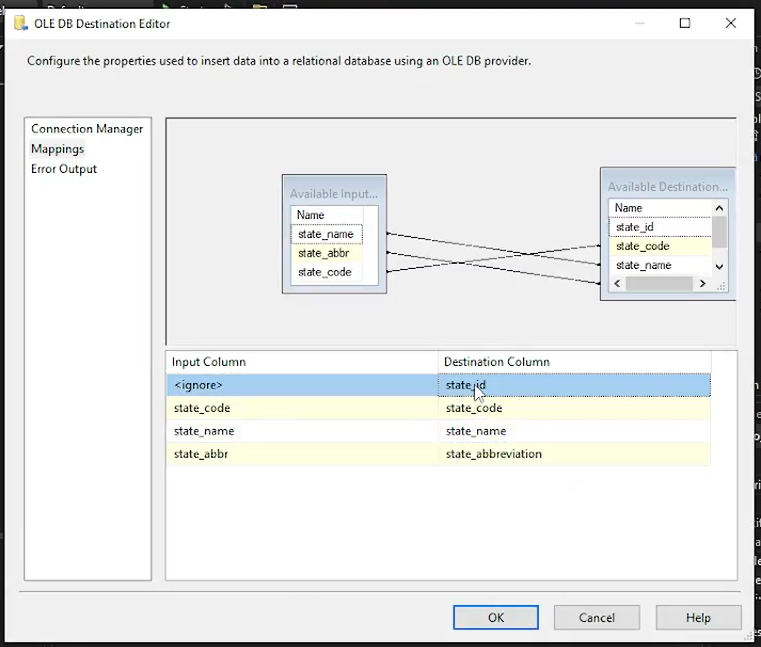


*Look up thuộc tính state\_code bởi vì trong bảng State thì stage\_code là unique*

**Bước 3: Load vào State NDS**

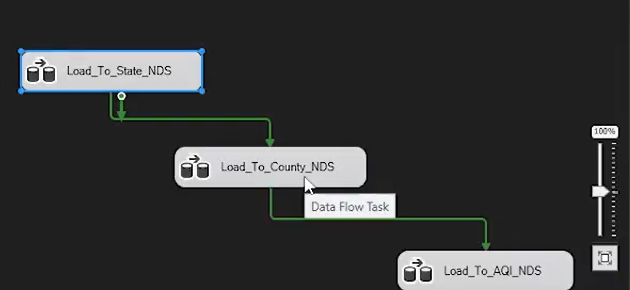




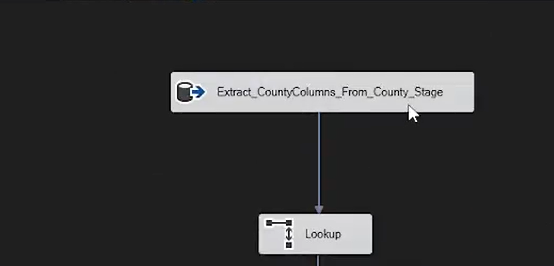


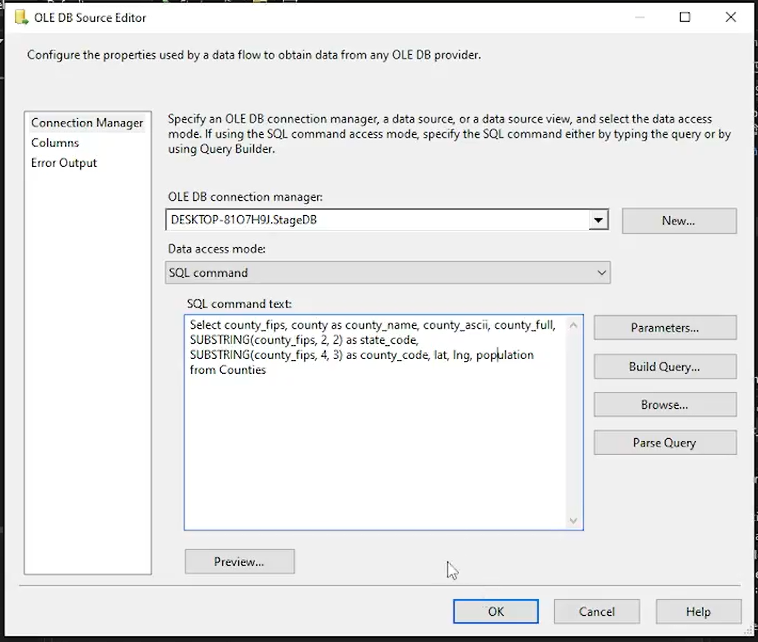
*Mapping (do state\_id là khóa tự tăng nên không cần map)*

#### Nạp dữ liệu vào bảng County NDS:



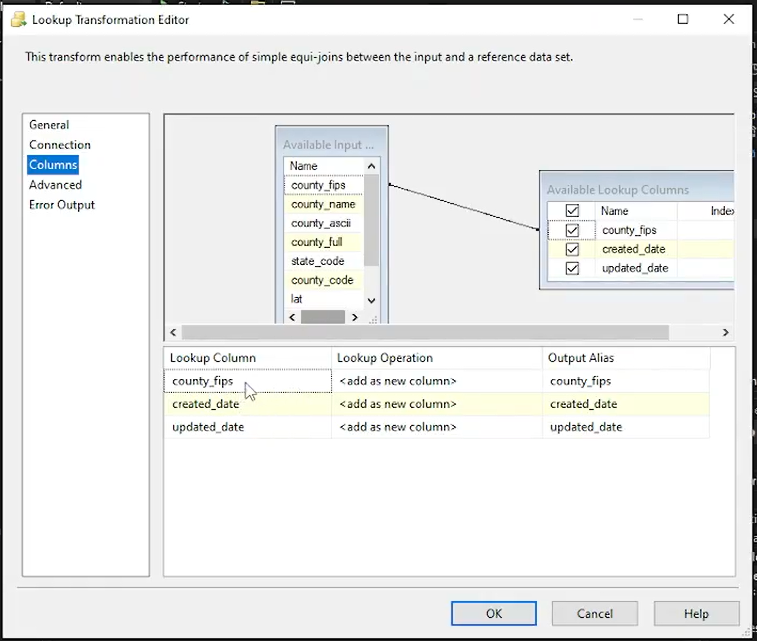
**Bước 1: Trích xuất các cột liên quan đến County trong bảng County trong Stage database**





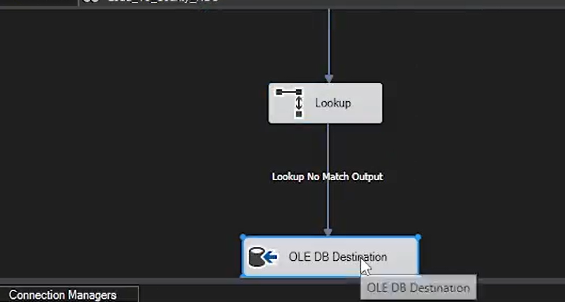
*Dùng code SQL trích xuất các thuộc tính liên quan đến County trong bảng Counties*

**Bước 2: Kiểm tra dòng tồn tại**

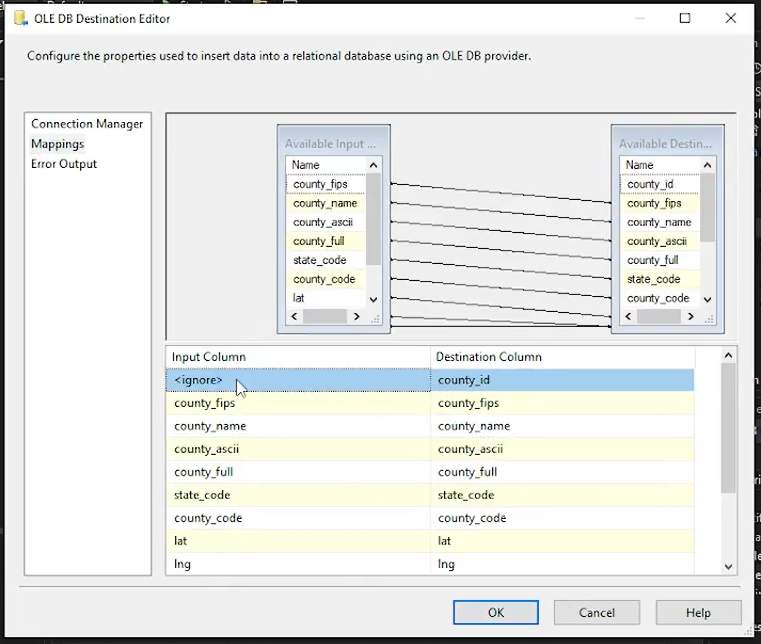
**

*Look up trên 3 thuộc tính county\_fips, created\_date, updates\_date*

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

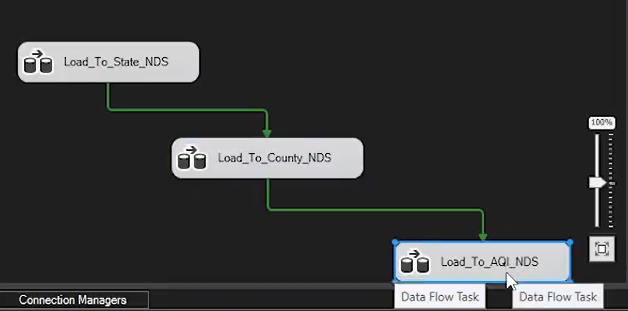
**

**

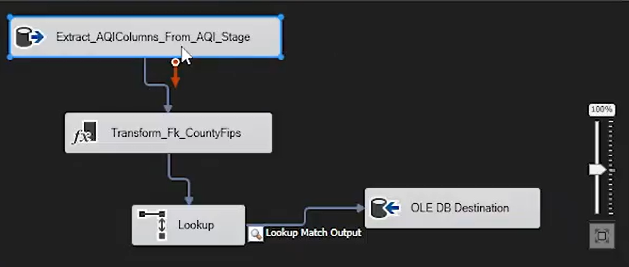
**

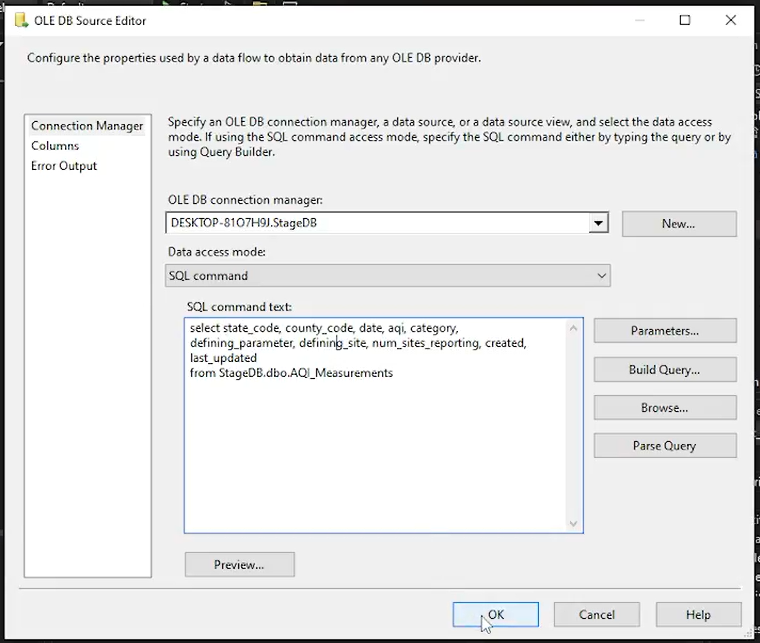
*Mapping*

#### Nạp dữ liệu vào bảng AQI NDS:



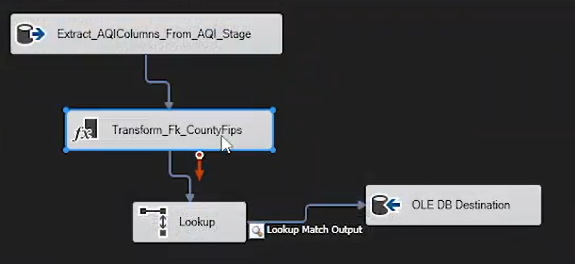
**Bước 1: Trích xuất các cột liên quan đến AQI trong bảng AQI trong Stage database**

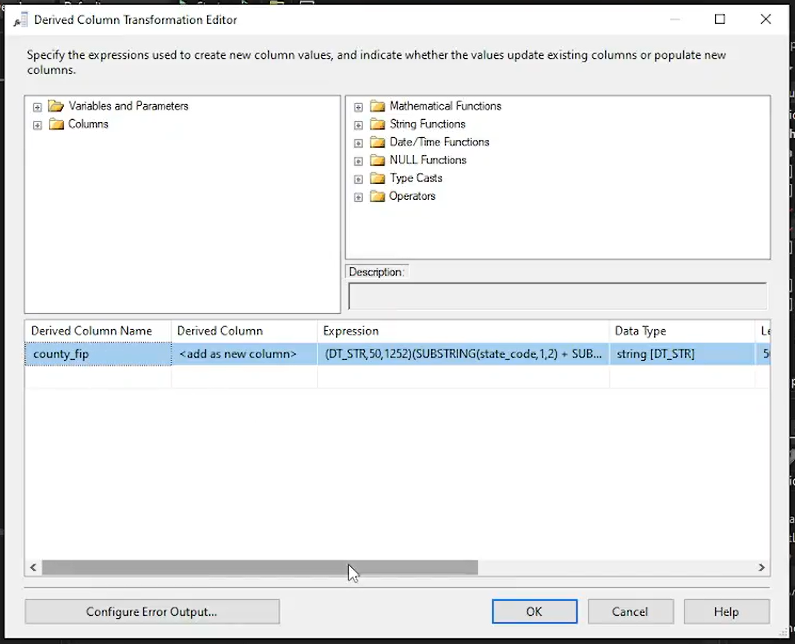




*Dùng code SQL trích xuất các thuộc tính liên quan đến AQI trong bảng AQI\_Measurements*

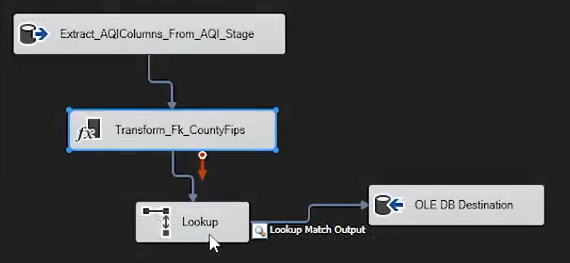
**Bước 2: Chuyển đổi dữ liệu**

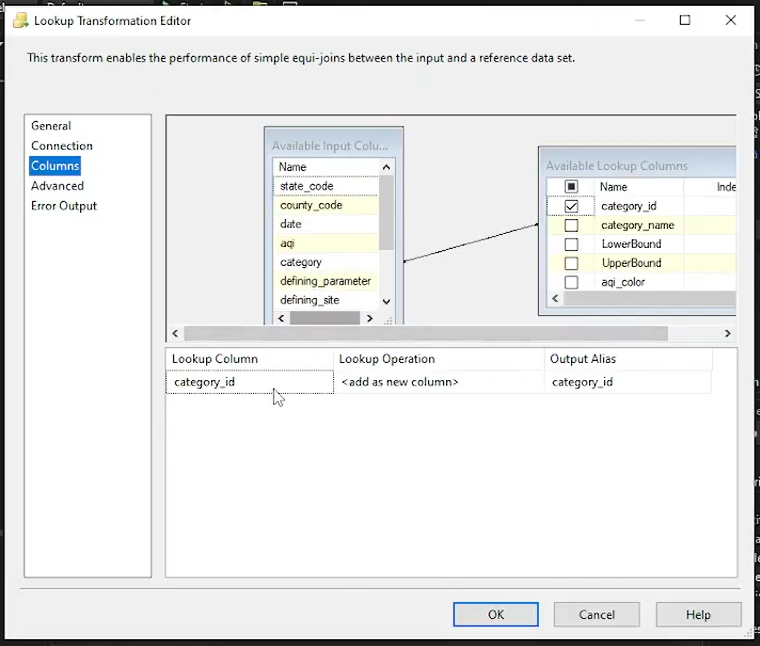




*Transform thuộc tính county\_fip*

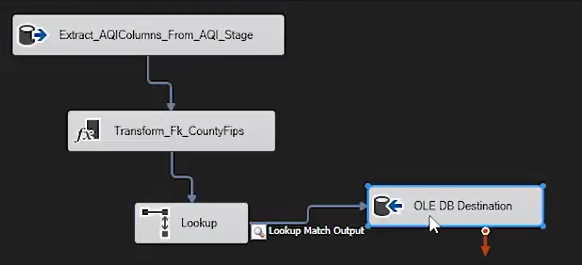
**Bước 3: Kiểm tra dòng tồn tại**

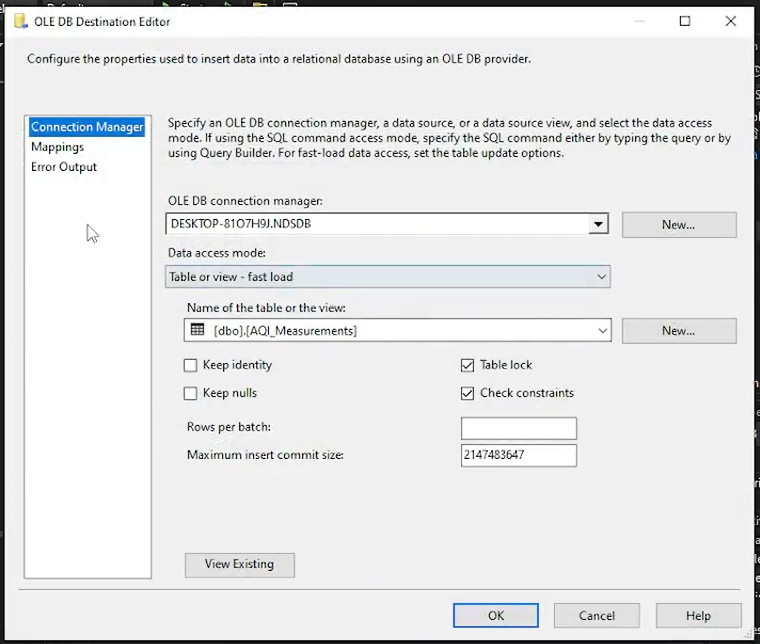
****

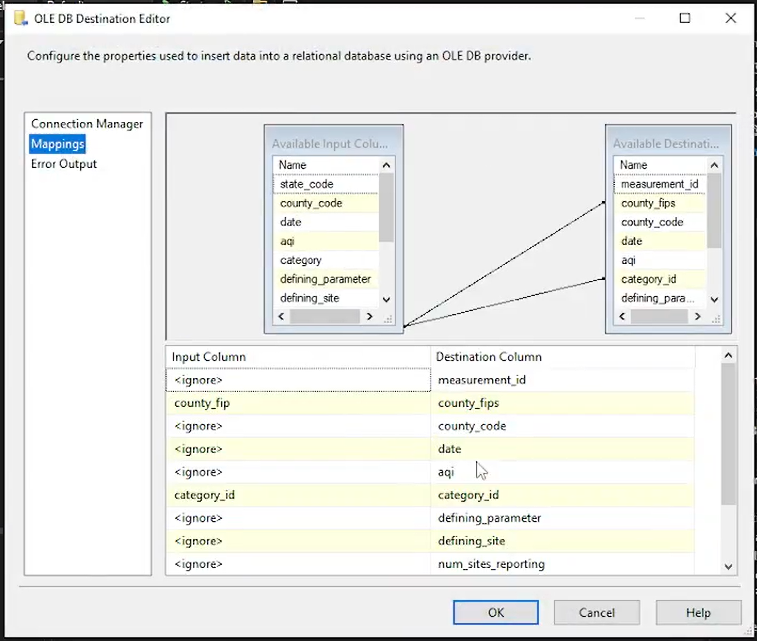
**

*Look up trên category\_id*

**Bước 4: Nạp vào bảng AQI NDS**

**

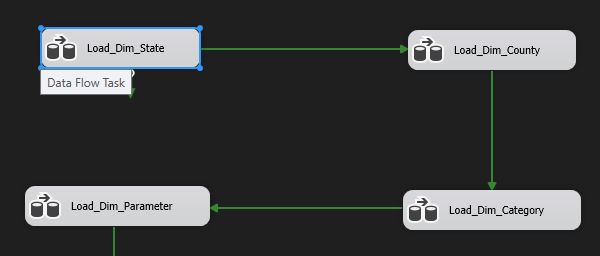
****

****

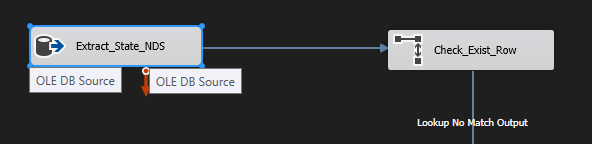
*Mapping*

### **ETL từ NDS vào DDS:**

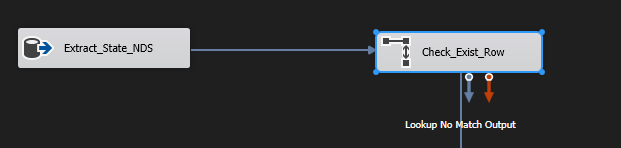
#### Nạp dữ liệu vào bảng Dim\_State:



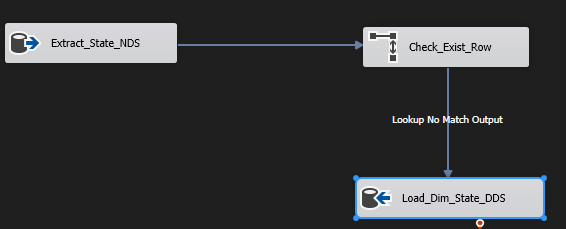
**Bước 1: Trích xuất dữ liệu từ bảng State trong NDS**

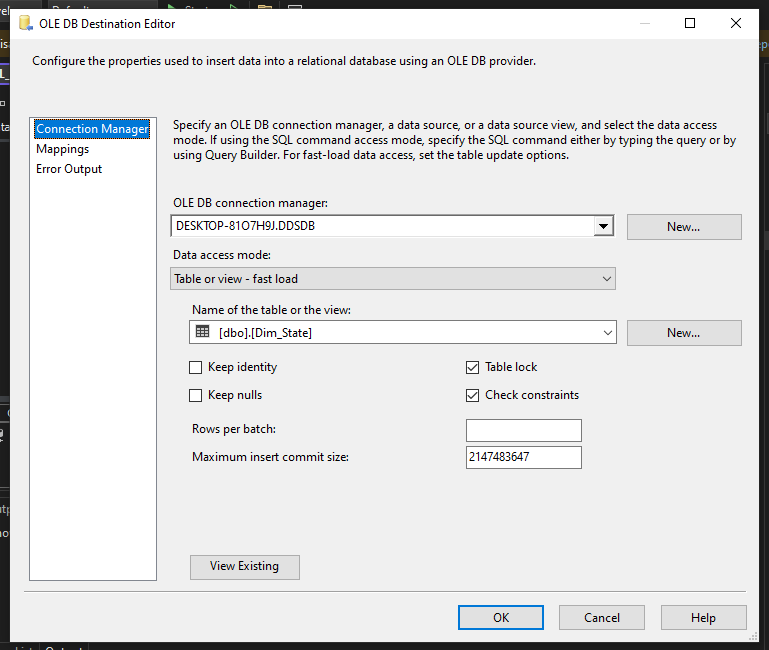
****

**Bước 2: Kiểm tra dòng tồn tại**

****

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

****

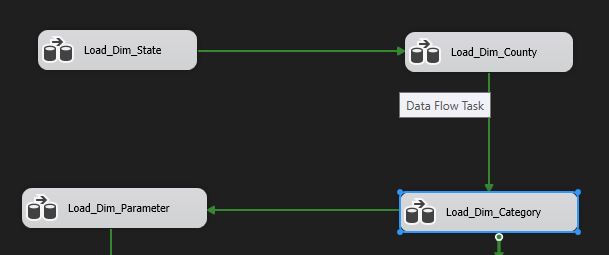
****

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



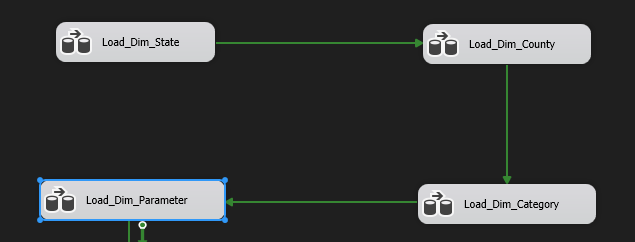
**Gồm 3 bước tương tự như trên**

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

****

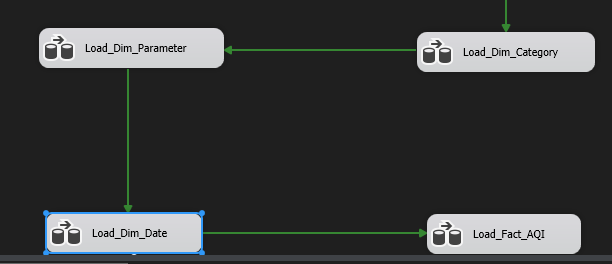
**Data flow gồm 3 bước tương tự như trên**

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

****

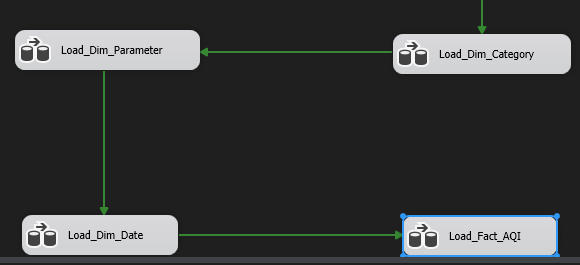
**Data flow gồm 3 bước tương tự như trên**

#### Nạp dữ liệu vào bảng Dim\_Date:

****

**Data flow gồm 3 bước tương tự như trên**

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

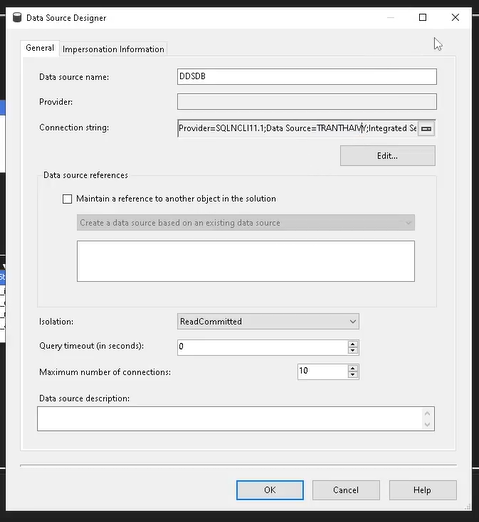
****

**Data flow gồm 3 bước tương tự như trên**

1. **OLAP:**

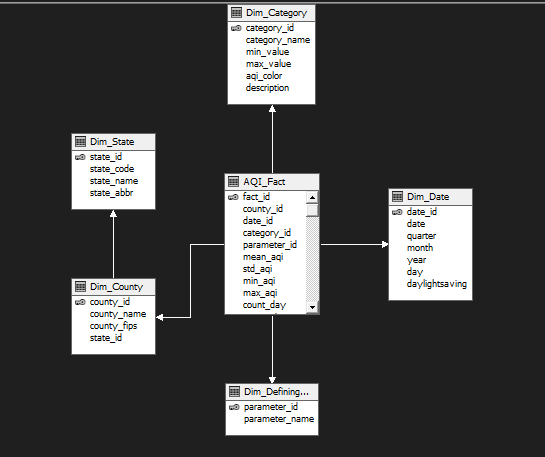
**Bước 1: Tạo csdl NDS trong SQL**

**Bước 2: Tạo data source**

****

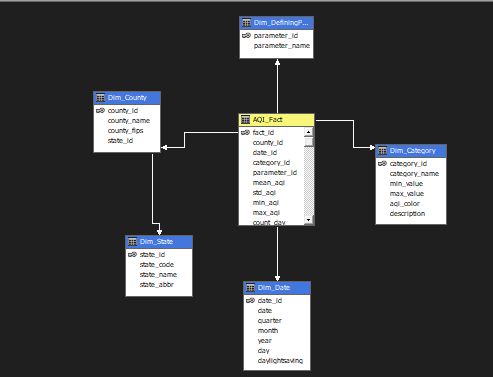
*Tạo kết nối với SQL server và load csdl DDS vừa tạo vào project*

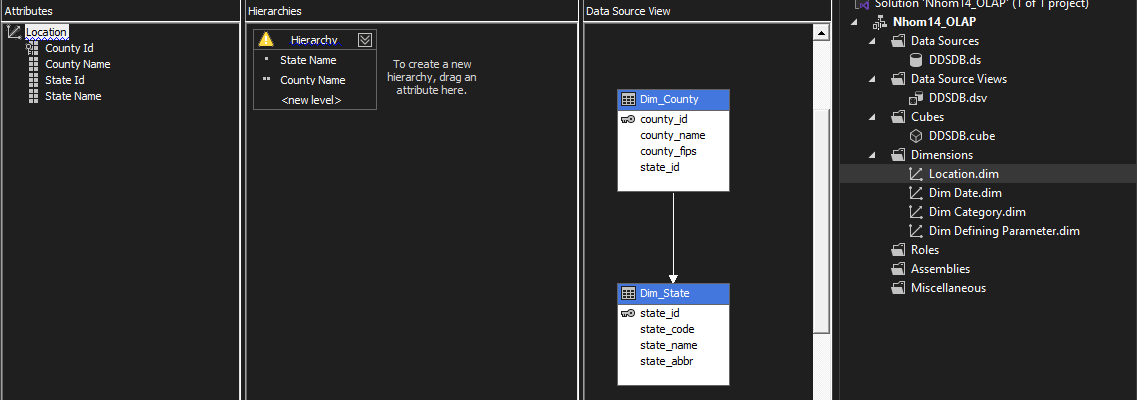
**Bước 2: Tạo data source view**

****

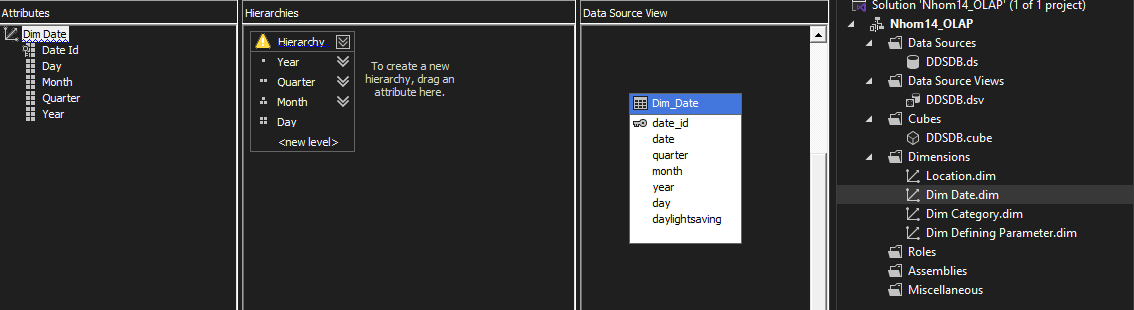
*Lấy những bảng cần sử dụng*

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

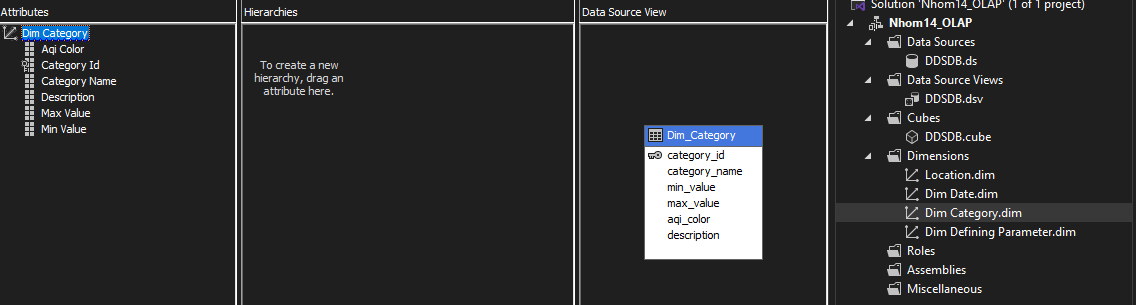
**Bước 4: Định nghĩa measure và dimension, chiều phân cấp từ các bảng đã chọn trong data source view**



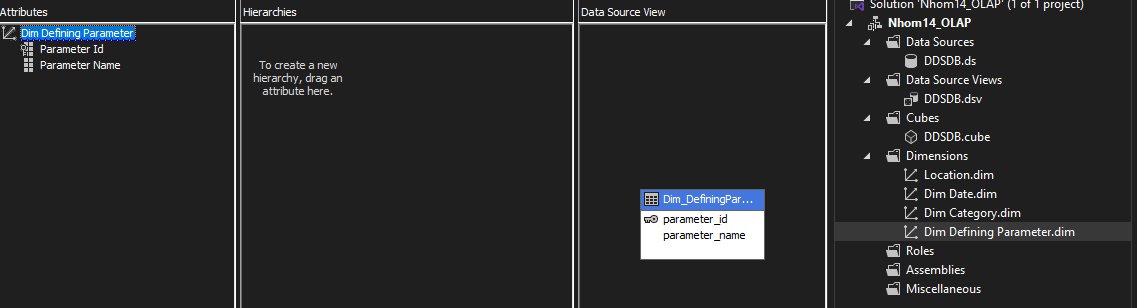
*Location dim, chiều phân cấp từ bang đến hạt*

**

*Date dim, chiều phân cấp từ năm đến quý đến tháng đến ngày*

**

*Category dim*

**

*Defining Parameter dim*

1. **Link Github và video demo:**

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

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

Video demo OLAP:

<https://drive.google.com/drive/folders/1rhqA4TR-g78XLzi2sUBmohhDBiFQfuSJ>