**Thiết kế DDS**

**An, Vỹ**

**1. Hướng dẫn chi tiết thực hiện tiếp trong Visual Studio (SSIS)**

* Tạo thêm một file DTSX mới: **ETL\_NDS\_DDS.dtsx**.
* Bố cục Control Flow:
  + **Task 1**: Tải dữ liệu vào các Dimension Tables (Dim\_State, Dim\_County, Dim\_Date).
  + **Task 2**: Tải dữ liệu vào Fact Table (Fact\_AQI).
* Data Flow cho từng Dimension và Fact sẽ được thiết kế riêng.

**2. Tạo ETL\_NDS\_DDS.dtsx:** chuột phải vào **SSIS Packages** > **New SSIS Package** > Đặt tên **ETL\_NDS\_DDS.dtsx**.

**3. Thiết kế Control Flow**

**3.1. Tải Dimension Tables**

1. **Tạo 3 Data Flow Tasks**:
   * **Load Dim\_State**
   * **Load Dim\_County**
   * **Load Dim\_Date**
2. Kết nối **Data Flow Tasks** với nhau nếu cần (nếu các bảng phụ thuộc vào nhau).

**3.2. Tải Fact Table**

1. Thêm một **Data Flow Task**:
   * **Load Fact\_AQI**.

**4. Thiết kế Data Flow Tasks**

**4.1. Load Dim\_State**

1. **Source**:
   * Sử dụng **OLE DB Source**:
     + Kết nối tới NDS database.
     + Truy vấn:

SELECT DISTINCT state\_name FROM NDSDB.dbo.States;

1. **Transform**:
   * **Sort**:
     + Loại bỏ các bản ghi trùng lặp.
2. **Destination**:
   * Sử dụng **OLE DB Destination**:
     + Kết nối tới DDS database.
     + Nạp vào **Dim\_State**.

**4.2. Load Dim\_County**

1. **Source**:
   * Sử dụng **OLE DB Source**:
     + Kết nối tới NDS database.
     + Truy vấn:

SELECT DISTINCT county\_name, state\_name FROM NDSDB.dbo.Counties;

1. **Transform**:
   * **Lookup**:
     + Lookup state\_name trong **Dim\_State** để lấy StateKey.
     + Xử lý lỗi nếu không tìm thấy (ghi vào file log hoặc bỏ qua).
2. **Destination**:
   * Sử dụng **OLE DB Destination**:
     + Nạp vào **Dim\_County**.

**4.3. Load Dim\_Date**

1. **Source**:
   * Sử dụng **OLE DB Source**:
     + Kết nối tới NDS database.
     + Truy vấn:

SELECT DISTINCT date FROM NDSDB.dbo.AQI\_Measurements;

1. **Transform**:
   * **Derived Column**:
     + Tạo các cột mới:
       - DateKey: YEAR(date) \* 10000 + MONTH(date) \* 100 + DAY(date)
       - Year: YEAR(date)
       - Quarter: DATEPART(QUARTER, date)
       - Month: MONTH(date)
       - Day: DAY(date).
2. **Destination**:
   * Sử dụng **OLE DB Destination**:
     + Nạp vào **Dim\_Date**.

**4.4. Load Fact\_AQI**

1. **Source**:
   * Sử dụng **OLE DB Source**:
     + Kết nối tới NDS database.
     + Truy vấn:

SELECT

aqi, county\_name, state\_name, date, category\_name

FROM NDSDB.dbo.AQI\_Measurements;

1. **Transform**:
   * **Lookup**:
     + Lookup state\_name trong **Dim\_State** để lấy StateKey.
     + Lookup county\_name trong **Dim\_County** để lấy CountyKey.
     + Lookup date trong **Dim\_Date** để lấy DateKey.
   * **Derived Column**:
     + Tạo cột Is\_VeryUnhealthy: aqi >= 201 ? 1 : 0.
   * **Aggregate**:
     + Tính Days\_Count: Đếm số ngày theo StateKey, CountyKey, và Quarter.
     + Tính Mean\_AQI và StdDev\_AQI:
       - **Mean\_AQI**: Trung bình AQI.
       - **StdDev\_AQI**: Độ lệch chuẩn.
2. **Destination**:
   * Sử dụng **OLE DB Destination**:
     + Nạp vào **Fact\_AQI**.

* **5. Kết nối các bước với nhau**
* Trong **Control Flow**:
  1. Kết nối các **Data Flow Tasks** theo trình tự: Load Dim\_State → Load Dim\_County → Load Dim\_Date → Load Fact\_AQI.
  2. Sử dụng **Success Constraints** để đảm bảo mỗi bước chỉ chạy khi bước trước thành công.

**6. Xử lý lỗi và kiểm tra**

1. **Data Viewer**:
   * Thêm Data Viewer vào các bước quan trọng trong **Data Flow Tasks** để kiểm tra dữ liệu tại từng giai đoạn.
2. **Error Output**:
   * Cấu hình Error Output cho các bước **Lookup** và **Data Conversion** để ghi lại lỗi vào bảng log hoặc file CSV.
3. **Logging**:
   * Kích hoạt SSIS Logging để theo dõi tiến trình và lỗi.

**7. Kết quả:**

1. Kiểm tra dữ liệu trong các bảng trong SSMS:
   * **Dim\_State**, **Dim\_County**, **Dim\_Date**, và **Fact\_AQI**.
2. Thực hiện các truy vấn phân tích (1-6) để xác minh dữ liệu.