**DEMO CÁC BƯỚC XÂY DỰNG 1 DATA WAREHOUSE CƠ BẢN SỬ DỤNG SSIS TRONG VISUAL STUDIO**

# 1) Chuẩn bị Data Source

- Ở đây, em tải 2 dataset trên Kaggle:

+ Bank Customer Segmentation: <https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation>

+ Customer Segmentation Data

<https://www.kaggle.com/datasets/ravalsmit/customer-segmentation-data>

- Vốn không có dataset hoàn chỉnh để phục vụ demo, em đã lấy Customer\_ID của `Bank Customer Segmentation` để thế vào Customer\_ID của `Customer Segmentation Data`.

- Tiến hành xử lí dữ liệu:

+ Loại bỏ các record trùng nhau, blank record, giá trị null và giá trị sai.

+ Chuyển đổi các cột có dữ liệu Date về dạng Month/Day/Year (vì lí do nào đó khi chuyển dữ liệu vào Database trong SQL Server lại báo lỗi định dạng ngày, phải đổi về kiểu trên mới được).

- Tạo 1 Database tên Bank\_Transactions, tạo 2 bảng Transactions và Customers.

- Cuối cùng, Insert dữ liệu từ 2 file CSV đã tải vào 2 bảng tương ứng.

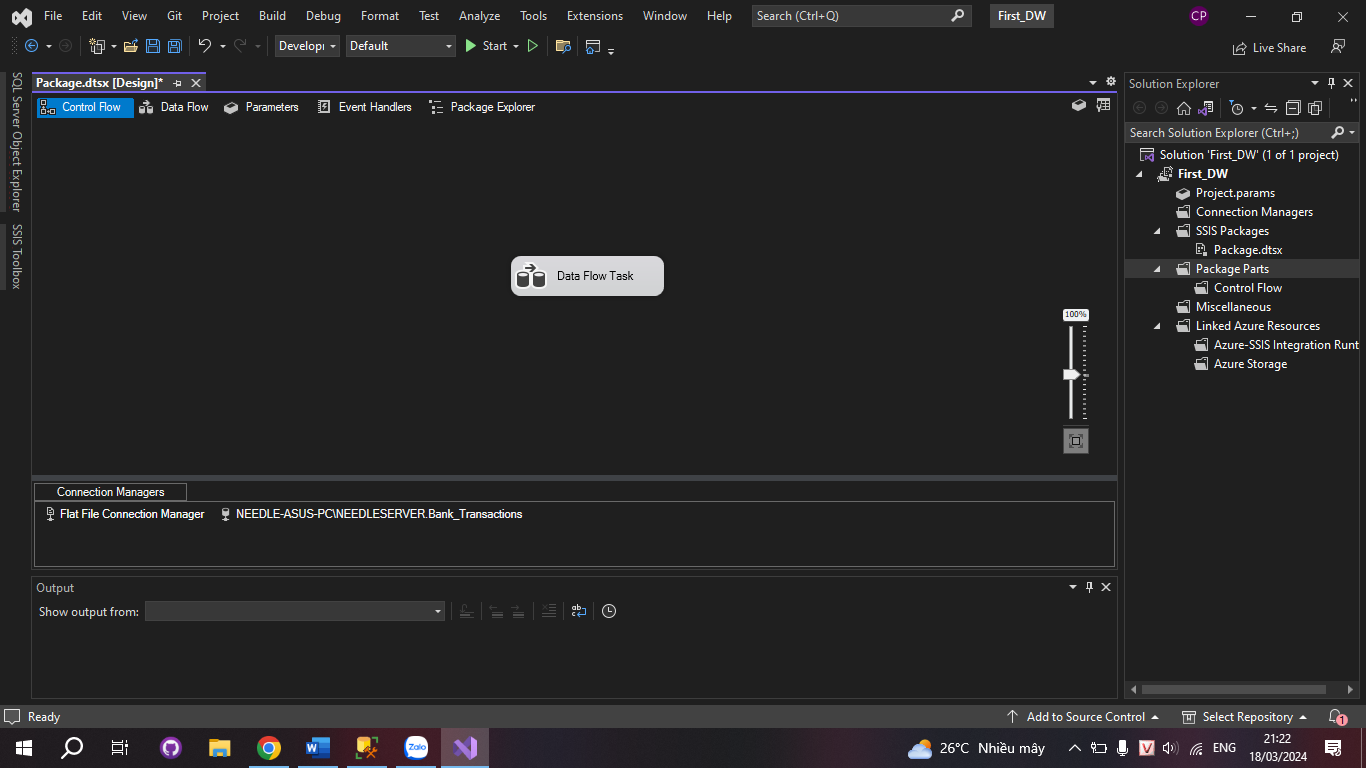
Vậy là ta đã có 1 Data Source đơn giản là 1 Relational Database (OLTP).

# 2) Tạo Data Pipeline, sử dụng ETL process để đưa dữ liệu từ Data Sourse sang Data Warehouse

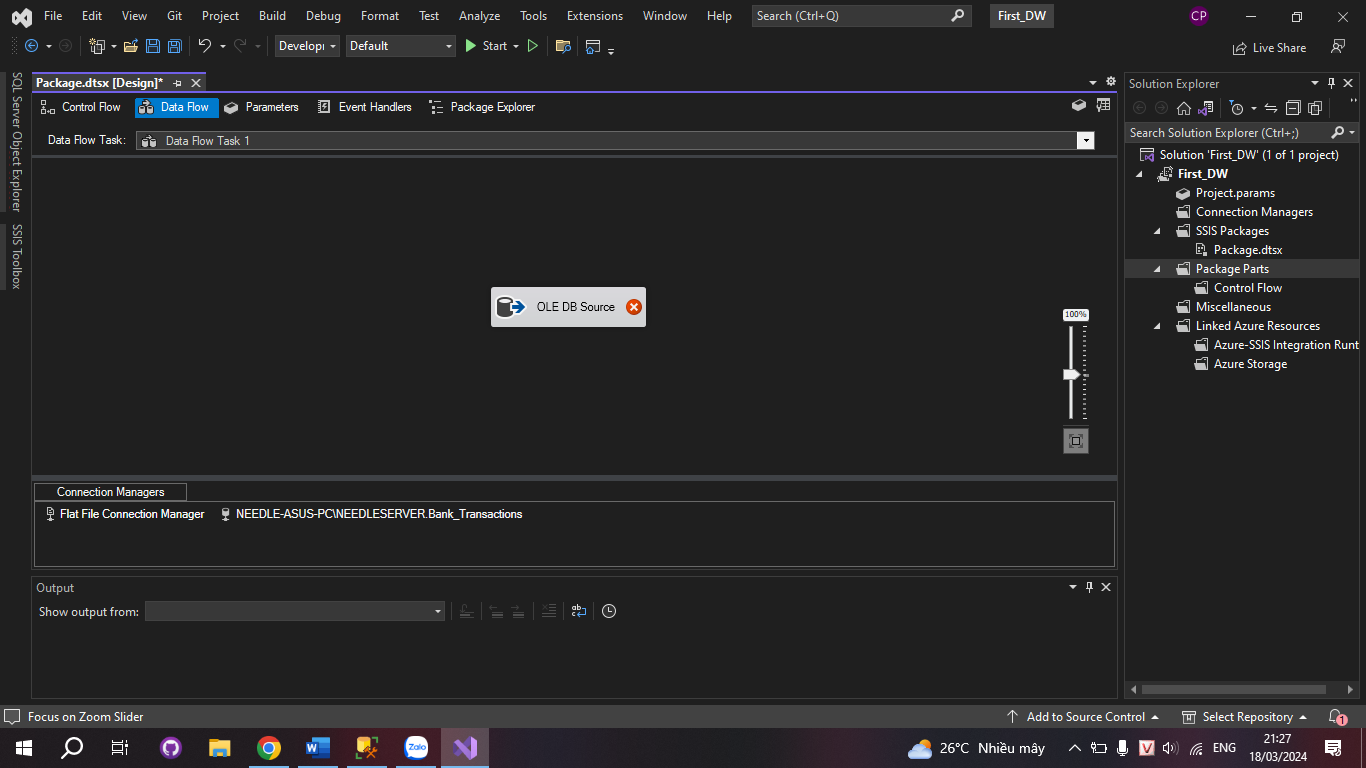
- Data Warehouse là nơi chứa Historical Data, vì vậy ta cần xây dựng 1 schema khác với RDB, đó là Star Schema và Snowflake Schema (trên lý thuyết là vậy, nhưng do Database của em không quá phức tạp, nên em không thực hiện bước này).

- Sau khi xây dựng schema của Data Warehouse, ta tiến hành bước **Data Extraction**.

+ Trong SSIS, ta tạo một Data Flow Task:



+ Trong Data Flow Task, để kết nối với Database, sử dụng thẻ OLE DB Source



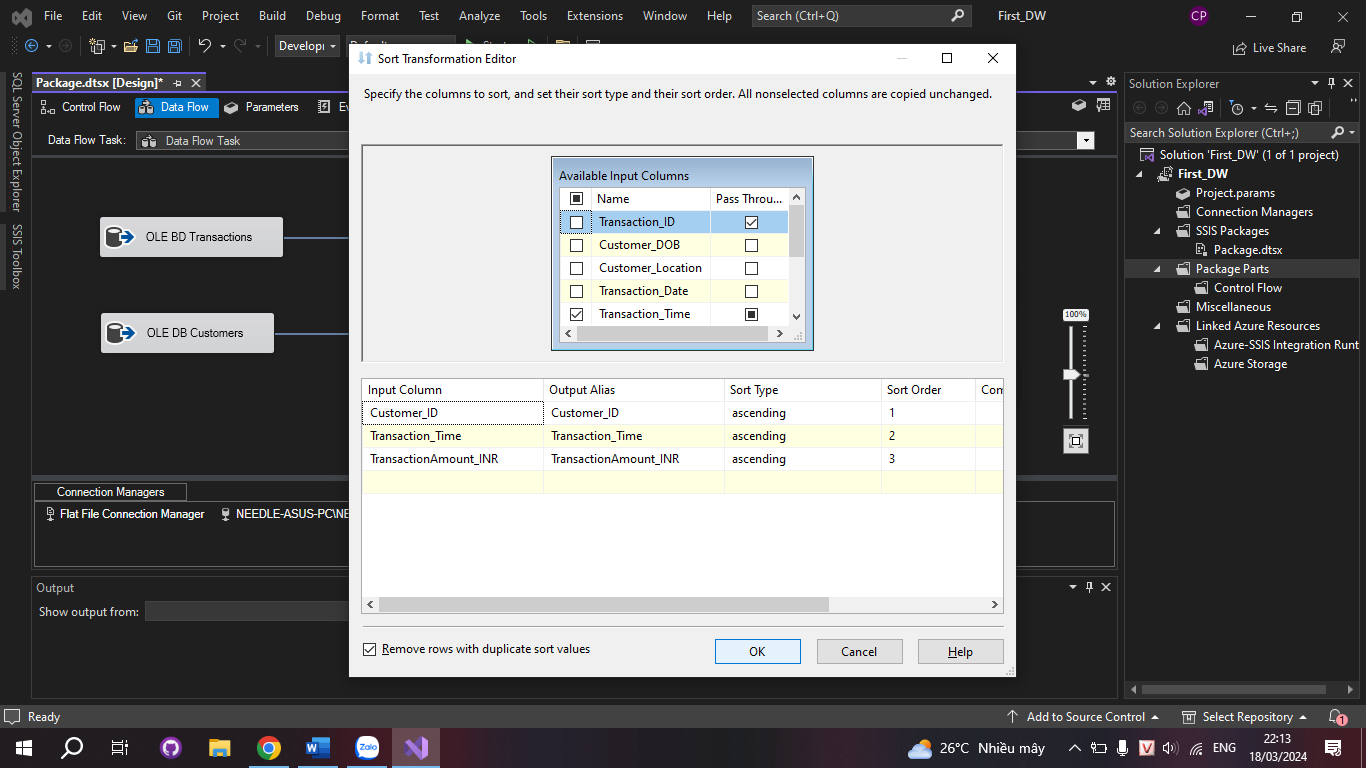
+ Khi kết nối được với Data source, ta sử dụng ETL tools hoặc viết scripts để lấy dữ liệu từ Data source. Sử dụng tool Filter và Validation data để đảm bảo dữ liệu được chính xác và đồng bộ.

- **Data Transformation**:

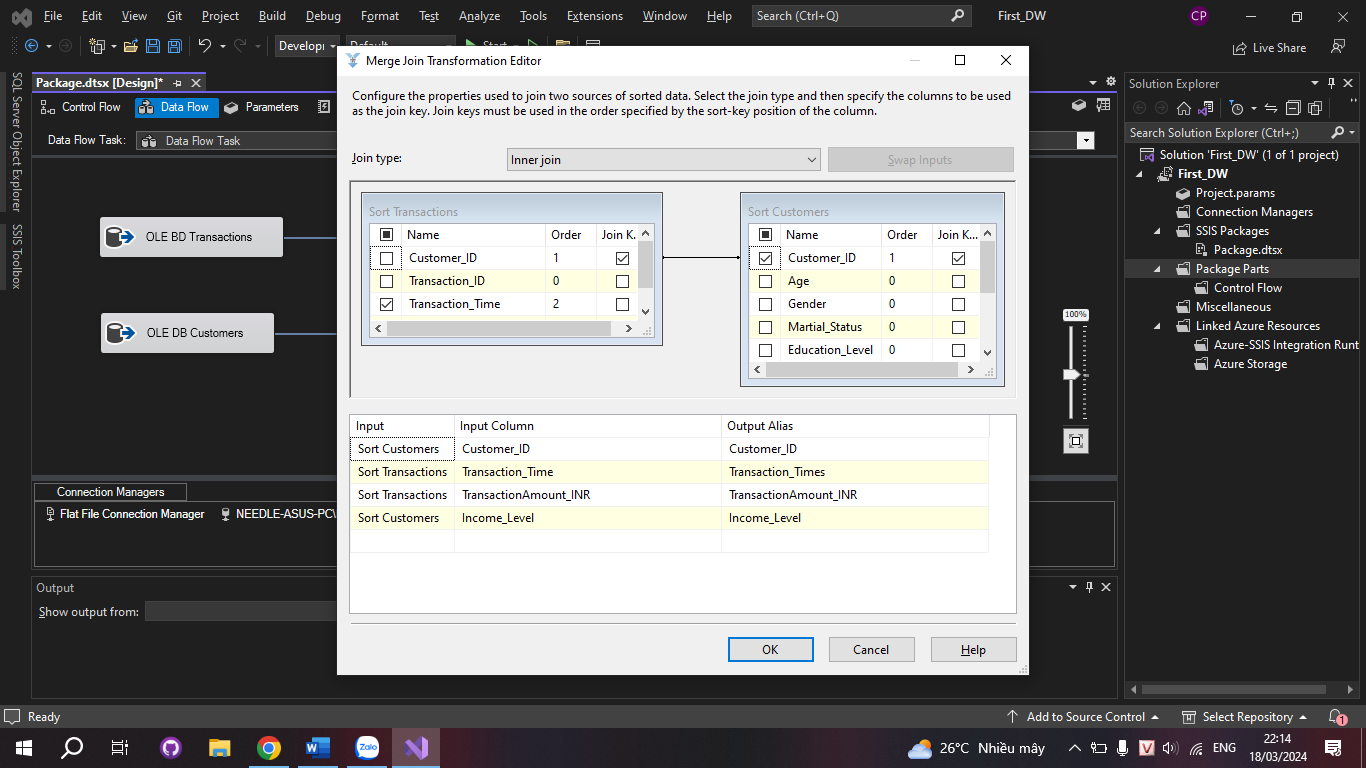
+ Ở bước này, ta tiến hành làm sạch dữ liệu, chuẩn hóa mô hình dữ liệu.

+ Thực hiện các bước Transformation như Aggregate, Merge, các phép tính toán và chuyển kiểu dư liệu cho phù hợp với thiết kế của DWH.

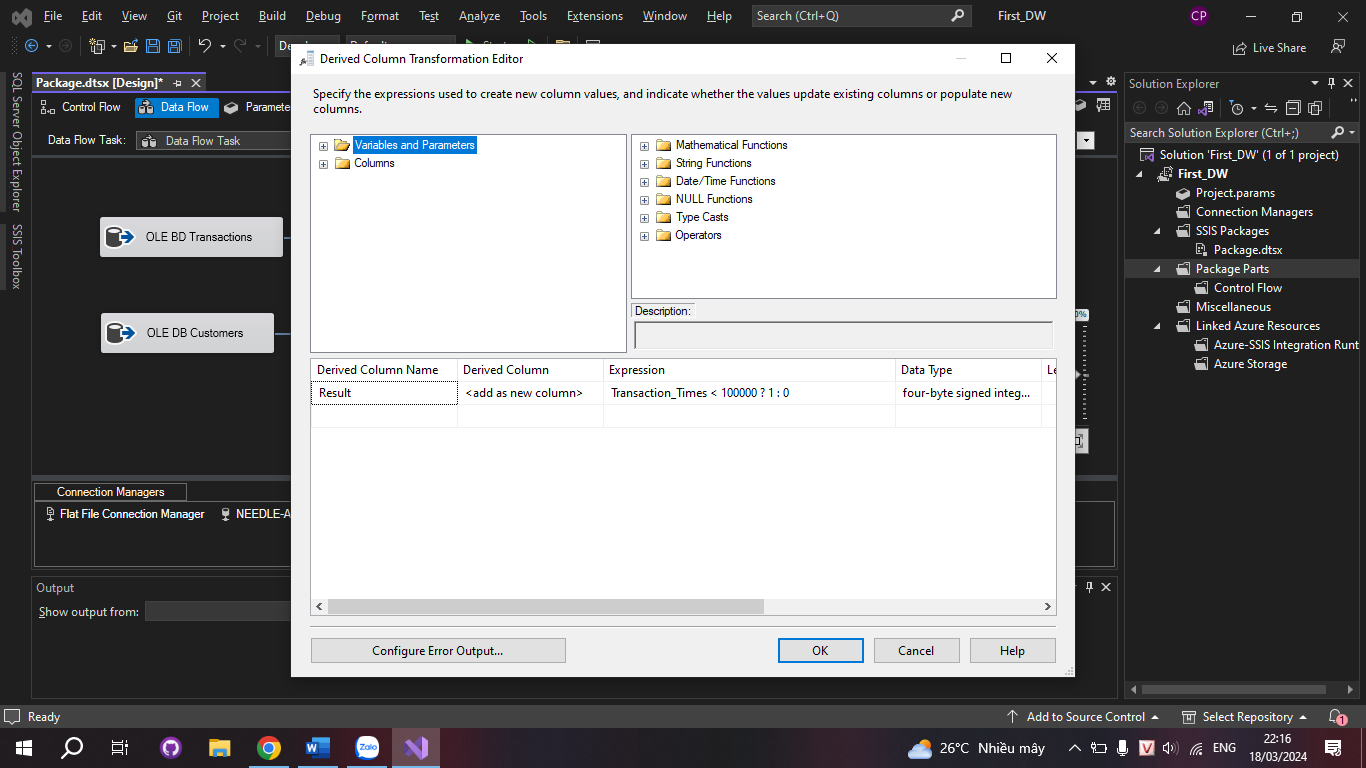
* Ở đây, em sắp xếp dữ liệu theo Customer\_ID.



* Sau đó Merge 2 bảng lại, lựa chọn các cột cần thiết cho bảng Fact.



* Cuối cùng, em hiển thị thêm 1 cột Result, với những người có lượng Transaction\_Times nhỏ hơn 100000.



- **Data Load:**

+ Chọn kĩ thuật load: full load hoặc increment load. Với bản Demo, em dùng `full load` (fast load).

+ Chọn phương pháp xử lí lỗi khi load. Ở đây em chọn `Ignore failure`.

+ Cuối cùng load vào DWH. Với bản Demo, em lưu dữ liệu vào bảng Fact trong Database Bank\_Transactions.

