

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM  
KHOA CÔNG NGHỆ THÔNG TIN  
BỘ MÔN HỆ THỐNG THÔNG TIN

-----❖❖-----



MÔN: HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH

**BÁO CÁO ĐỒ ÁN THỰC HÀNH**

NHÓM THỰC HIỆN: 2020.CQ.BI.11

THÀNH PHỐ HỒ CHÍ MINH, 01/2024

## THÔNG TIN NHÓM

MSSV	HỌ TÊN	NỘI DUNG ĐÓNG GÓP	TỶ LỆ ĐÓNG GÓP	ĐÁNH GIÁ
19120114	Lê Bảo Chấn Phát	<ul style="list-style-type: none"> <li>- Đồng bộ hóa cấu trúc và định dạng dữ liệu gốc</li> <li>- Thiết kế DDS dựa trên NDS</li> </ul>	10%	100%
20120255	Phạm Mai Thiên Bảo	<ul style="list-style-type: none"> <li>- Thực hiện quy trình ETL, OLAP.</li> <li>- Xây dựng Dashboard Power BI.</li> </ul>	35%	100%
20120343	Trần Minh Nhựt	<ul style="list-style-type: none"> <li>- Đè xuất dataset.</li> <li>- Thực hiện quy trình ETL, OLAP.</li> <li>- Xây dựng Dashboard Power BI.</li> <li>- Mining Data.</li> </ul>	45%	100%
20120586	Ngô Lê Hưng Thịnh	<ul style="list-style-type: none"> <li>- Đồng bộ hóa cấu trúc và định dạng dữ liệu gốc</li> <li>- Thiết kế NDS dựa trên dataset</li> </ul>	10%	100%

# **MỤC LỤC**

I.	MÔ TẢ DỮ LIỆU NGUỒN .....	1
II.	PHÂN TÍCH KHO DỮ LIỆU .....	4
1.	NDS .....	4
2.	DDS .....	7
III.	PHÂN TÍCH QUY TRÌNH ETL .....	9
1.	Quy trình Source vào Stage.....	9
2.	Quy trình Stage vào NDS.....	12
3.	Quy trình NDS vào DDS.....	18
IV.	OLAP REPORT .....	30
V.	DASHBOARD .....	36
1.	Overview of Sales Dashboard.....	37
2.	Sales by Customers .....	38
3.	Sales Map .....	39
VI.	DATA MINING .....	39
VII.	TÀI LIỆU THAM KHẢO .....	42

# MỤC LỤC ẢNH

Hình 1. Thiết kế NDS .....	4
Hình 2. Thiết kế DDS .....	9
Hình 3. Source Central to Stage.....	10
Hình 4. Source East to Stage (1).....	10
Hình 5. Source East to Stage (2).....	11
Hình 6. Source South to Stage .....	11
Hình 7. Source West to Stage.....	12
Hình 8. Stage to NDS.....	12
Hình 9. Location Stage to Location NDS (1) .....	13
Hình 10. Location Stage to Location NDS (2) .....	13
Hình 11. Load dữ liệu từ Stage .....	14
Hình 12. Group by theo Postal Code, City, State .....	15
Hình 13. Tạo Derived Column.....	16
Hình 14. Sử dụng SCD để load dữ liệu vào NDS.....	17
Hình 15. Cập nhật lại NDS .....	17
Hình 16. Thêm dòng mới vào NDS .....	18
Hình 17. NDS to DDS .....	19
Hình 18. Customer DIM .....	19
Hình 19. Location DIM.....	20
Hình 20. Shipmode DIM.....	20
Hình 21. Category DIM .....	21
Hình 22. Segment DIM.....	21
Hình 23. Product DIM .....	21
Hình 24. Sub Category DIM .....	22
Hình 25. Load dữ liệu từ Customer NDS .....	22
Hình 26. Sử dụng SCD để đổ dữ liệu vào bảng DIM .....	23
Hình 27. Chọn loại SCD .....	23
Hình 28. Thêm dòng mới vào DIM .....	24
Hình 29. Cập nhật lại DIM .....	24
Hình 30. Đổ dữ liệu vào bảng FACT: Chưa tính toán .....	25
Hình 31. Đổ dữ liệu vào bảng FACT (1) .....	26
Hình 32. Đổ dữ liệu vào bảng FACT (2) .....	27
Hình 33. Đổ dữ liệu vào bảng FACT (3) .....	27
Hình 34. Đổ dữ liệu vào bảng FACT (4) .....	28
Hình 35. Đổ dữ liệu vào bảng FACT (5) .....	28

Hình 36. Đồ dữ liệu vào bảng FACT (6) .....	29
Hình 37. OLAP: Thống kê khách hàng mua hàng theo ngày .....	30
Hình 38. OLAP: Thống kê doanh thu theo từng Segment.....	31
Hình 39. OLAP: Thống kê doanh thu theo từng danh mục (Category) .....	32
Hình 40. OLAP: Thống kê số lượng sản phẩm đã mua của mỗi khách hàng .....	33
Hình 41. OLAP: Thống kê số sản phẩm bán được .....	34
Hình 42. OLAP: Thống kê doanh thu theo sản phẩm.....	35
Hình 43. OLAP: Thống kê doanh thu theo khách hàng .....	36
Hình 44. Overview of Sales Dashboard.....	37
Hình 45. Sales by Customers Dashboard.....	38
Hình 46. Sales Map.....	39
Hình 47. Dữ liệu mẫu Data Mining .....	40
Hình 48. Thông tin thống kê Data Mining.....	40
Hình 49. Thiết kế Process Mining .....	41
Hình 50. Kết quả Mining .....	41
Hình 51. Auto Model .....	42

# I. MÔ TẢ DỮ LIỆU NGUỒN

- Mô tả dữ liệu: Bộ dữ liệu mô tả chi tiết tình hình bán hàng của các cửa hàng trên toàn nước Mỹ gồm 4 vùng: Central, East, South, West trong toàn bộ năm 2018. Bộ dữ liệu được chia làm 4 nguồn khác nhau tương ứng với 4 vùng, được lưu trữ bằng file excel.
- **Source 1: Central**

- o Data2018\_Central:

Order ID	Mã đặt hàng, là duy nhất cho mỗi đơn hàng
Order Date	Ngày đặt hàng
Ship Date	Ngày giao hàng
Ship Mode	Mã hình thức giao hàng
Customer ID	Mã khách hàng
Segment	Nhóm khách hàng
City	Tên thành phố
State	Tên tiểu bang
Postal Code	Mã bưu chính
Product ID	Mã sản phẩm
Sales	Giá bán

- o Product:

Product ID	Mã sản phẩm, là duy nhất cho mỗi sản phẩm
Category	Mã danh mục sản phẩm
Sub-Category	Danh mục con
Product Name	Tên sản phẩm

- o Customer

Customer	Mã khách hàng, là duy nhất cho mỗi khách hàng
Customer	Tên Khách hàng

- o Shipmode

ShipModeKey	Mã loại giao hàng, là duy nhất của mỗi loại
Ship Mode	Loại hình thức giao hàng

- o Category

Cat_id	Mã danh mục, là duy nhất cho mỗi danh mục
Category	Tên danh mục

- **Source 2: East**

o Data2018\_East

Order ID: Mã đơn đặt hàng
Order Date : Ngày đặt hàng
Ship Date: Ngày giao
Ship Mode :Mã loại giao hàng
Customer ID : Mã khách hàng
Segment : Segment ID
Postal Code: Mã postal
Product ID: Mã mặt hàng
Product Name: Tên hàng
Product Category : mã loại hàng
Sales: Giá

o Product\_Category

Product_Catogory : Ứng với duy nhất (Category_ID,Sub-Category)
Category ID: Mã danh mục hàng

o Location

Postal Code: Duy nhất ứng với ( Country, City, State ) duy nhất
Country: Tên quốc gia
City: Thành phố
State: Bang

o Segment

Segment ID: Mã Segment ánh xạ duy nhất với 1 giá trị Segment
Segment : Tên của Segment

o Customer

Customer ID:	Mã Khách Hàng
Customer Name	Tên Khách Hàng

o ShipMode

Mode_ID: Mã Ship Mode ánh xạ duy nhất với mỗi Ship Mode
Ship Mode: Tên của Ship Mode

o Category

Category ID : Mã loại hàng ánh xạ duy nhất với Tên loại
Category: Tên loại hàng

- **Source 3: South**

o Data2018\_South

Order ID	Mã đơn đặt hàng
Order Date	Ngày đặt hàng
Ship Date	Ngày giao hàng
Ship Mode	Hình thức giao hàng
Customer ID	Mã khách hàng
Customer Name	Tên khách hàng
Segment	Mã loại người dùng
Postal Code	Mã postal
Product ID	Mã sản phẩm
Category	Mã danh mục
Sub-Category	Mã danh mục con
Product Name	Tên sản phẩm
Sales	Giá bán
Address	Địa chỉ bán

o Segment

SegmentID	Mã loại khách hàng
Segment	Phân khúc khách hàng

o ShipMode

ShipModeID	Mã hình thức giao hàng
Ship Mode	Hình thức giao hàng

o Category

CategoryID	Mã danh mục
Category	Tên danh mục

o Sub-Category

SubCategoryID	Mã danh mục con
SubCategory	Tên danh mục con

- **Source 4: West**

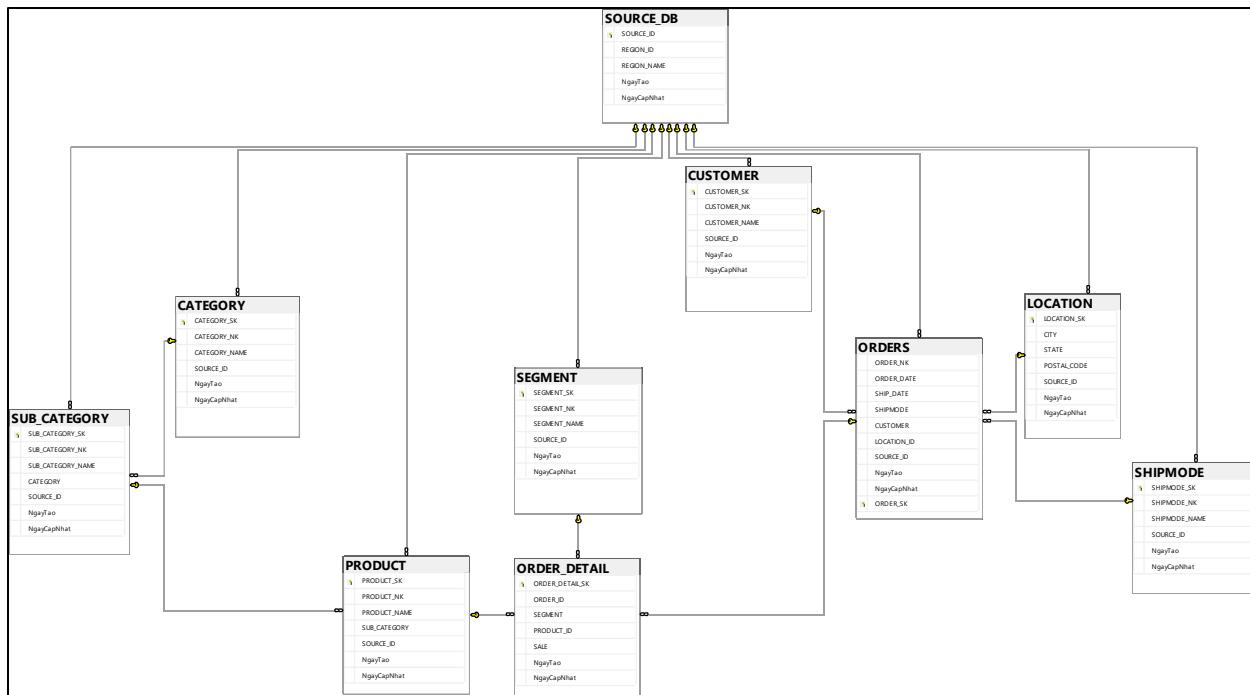
o Data2018\_West

Order ID	Mã đơn đặt hàng
Order Date	Ngày đặt hàng
Ship Date	Ngày giao hàng
Ship Mode	Hình thức giao hàng
Customer ID	Mã khách hàng
Customer Name	Tên khách hàng
Segment	Mã loại người dùng
Country	Tên quốc gia
City	Tên thành phố
State	Tên bang
Postal Code	Mã postal
Product ID	Mã sản phẩm
Category	Mã danh mục
Sub-Category	Mã danh mục con
Product Name	Tên sản phẩm
Sales	Giá bán

## II. PHÂN TÍCH KHO DỮ LIỆU

### 1. NDS

Diagram:



Hình 1. Thiết kế NDS

Chi tiết từng bảng trong NDS:

- Bảng CATEGORY

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[CATEGORY_SK]	Khóa đại diện	Int
2	[CATEGORY_NK]	Khóa nguồn	Nvarchar(255)
3	[CATEGORY_NAME]	Tên danh mục	Nvarchar(255)
4	[SOURCE_ID]	ID nguồn	Int
5	[NgayTao]	Ngày tạo	Datetime
6	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng CUSTOMER

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[CUSTOMER_SK]	Khoá đại diện	Int
2	[CUSTOMER_NK]	Khoá nguồn	Nvarchar(255)
3	[CUSTOMER_NAME]	Tên khách hàng	Nvarchar(255)
4	[SOURCE_ID]	ID nguồn	Int
5	[NgayTao]	Ngày tạo	Datetime
6	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng LOCATION

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[LOCATION_SK]	Khoá đại diện	Int
2	[CITY]	Tên thành phố	Nvarchar(255)
3	[STATE]	Tên bang	Nvarchar(255)
4	[POSTAL_CODE]	Mã Postal ( Khoá nguồn )	Float
5	[SOURCE_ID]	ID Nguồn	Int
6	[NgayTao]	Ngày tạo	Datetime
7	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng ORDER\_DETAIL

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[ORDER_DETAIL_SK]	Khoá đại diện	Int
2	[ORDER_ID]	Khoá nguồn	Int
3	[SEGMENT]	Tên nhóm khách hàng	Int
4	[PRODUCT_ID]	Mã mặt hàng	Int
5	[SALE]	Giá trị	Float
6	[NgayTao]	Ngày tạo	Datetime
7	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng ORDERS

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[ORDER_SK]	Khoá đại diện	Int
2	[ORDER_NK]	Khoá nguồn	Nvarchar(255)
3	[ORDER_DATE]	Ngày đặt hàng	Date
4	[SHIP_DATE]	Ngày giao hàng	Date
5	[SHIpmode]	Loại giao hàng	Int
6	[CUSTOMER]	Mã khách hàng	Int
7	[LOCATION_ID]	Mã địa điểm	Int
8	[SOURCE_ID]	ID nguồn	Int
9	[NgayTao]	Ngày tạo	Datetime
10	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng PRODUCT

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[PRODUCT_SK]	Khoá đại diện	Int
2	[PRODUCT_NK]	Khoá nguồn	Nvarchar(255)
3	[PRODUCT_NAME]	Tên mặt hàng	Nvarchar(255)
4	[SUB_CATEGORY]	Mã phân loại	Int
5	[SOURCE_ID]	ID nguồn	Int
6	[NgayTao]	Ngày tạo	Datetime
7	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng SEGMENT

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[SEGMENT_SK]	Khoá đại diện	Int
2	[SEGMENT_NK]	Khoá nguồn	Nvarchar(255)
3	[SEGMENT_NAME]	Tên nhóm khách hàng	Nvarchar(255)
4	[SOURCE_ID]	ID nguồn	Int
5	[NgayTao]	Ngày tạo	Datetime
6	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng SHIPMODE

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[SHIpmode_SK]	Khoá đại diện	Int
2	[SHIpmode_NK]	Khoá nguồn	Nvarchar(255)

			Int
3	[SHIPMODE_NAME]	Tên loại giao hàng	Nvarchar(255)
4	[SOURCE_ID]	ID nguồn	Int
5	[NgayTao]	Ngày tạo	Datetime
6	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng SOURCE\_DB

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[SOURCE_ID]	ID nguồn	Int
2	[REGION_ID]	Mã khu vực	Int
3	[REGION_NAME]	Tên khu vực	Nvarchar(255)
4	[NgayTao]	Ngày tạo	Datetime
5	[NgayCapNhat]	Ngày cập nhật	Datetime

- Bảng SUB\_CATEGORY

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	[SUB_CATEGORY_SK]	Khóa đại diện	Int
2	[SUB_CATEGORY_NK]	Khóa nguồn	Nvarchar(255)
3	[SUB_CATEGORY_NAME]	Tên danh mục con	Nvarchar(255)
4	[CATEGORY]	Khóa ngoại tham chiếu đến Category SK	Int
5	[SOURCE_ID]	ID nguồn	Int
6	[NgayTao]	Ngày tạo	Datetime
7	[NgayCapNhat]	Ngày cập nhật	Datetime

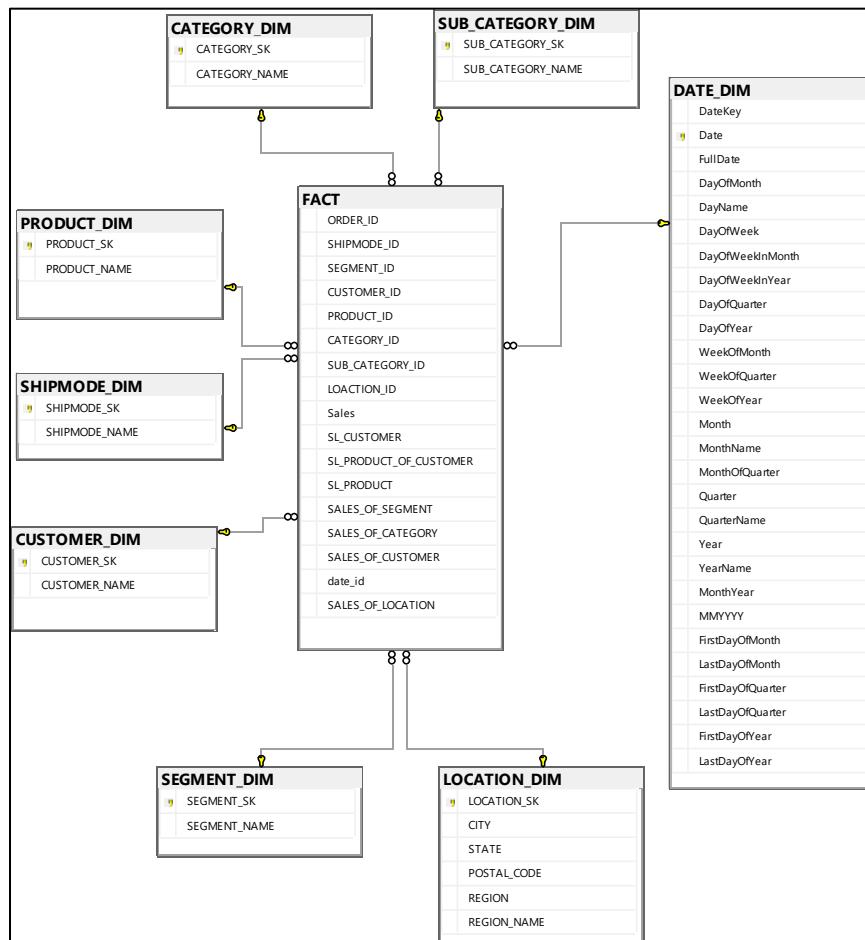
## 2. DDS

- Các nhu cầu cần phân tích:

- **Thống kê số khách hàng mua hàng theo ngày của mỗi vùng.**
  - Sự kiện: Khi thống kê số lượng khách hàng vào cuối ngày.
  - Bối cảnh:
    - Khi nào: cuối mỗi ngày.
    - Ở đâu: mỗi vùng.
  - Đo lường: Số lượng khách hàng.
- **Thống kê doanh thu của từng segment theo ngày của mỗi vùng.**
  - Sự kiện: Khi thống kê doanh thu của từng segment vào cuối ngày.
  - Bối cảnh:

- Khi nào: cuối mỗi ngày.
  - Ở đâu: mỗi vùng.
  - Cái gì: nhóm khách hàng.
  - Đo lường: Tổng doanh thu
- **Thống kê doanh thu của từng danh mục theo ngày theo mỗi vùng.**
  - Sự kiện: Khi thống kê doanh thu của từng danh mục vào cuối ngày.
  - Bối cảnh
    - Khi nào: cuối mỗi ngày.
    - Ở đâu: mỗi vùng
    - Cái gì: danh mục
    - Đo lường: tổng doanh thu
- **Thống kê số lượng sản phẩm đã mua của mỗi khách hàng trong mỗi ngày của mỗi vùng.**
  - Sự kiện: Khi thống kê số lượng sản phẩm của mỗi khách hàng vào cuối ngày
  - Bối cảnh
    - Khi nào: cuối mỗi ngày.
    - Ở đâu: mỗi vùng
    - Ai: khách hàng
    - Đo lường: tổng sản phẩm
- **Thống kê số lượng sản phẩm bán được theo ngày của từng vùng.**
  - Sự kiện: thống kê số lượng sản phẩm vào cuối ngày
  - Bối cảnh
    - Khi nào: cuối mỗi ngày.
    - Ở đâu: mỗi vùng
    - Đo lường: tổng sản phẩm
- **Thống kê doanh thu theo sản phẩm trong từng thành phố theo ngày.**
  - Sự kiện: thống kê doanh thu theo sản phẩm vào cuối ngày
  - Bối cảnh
    - Khi nào: cuối mỗi ngày.
    - Ở đâu: thành phố
    - Cái gì: sản phẩm
    - Đo lường: tổng doanh thu
- **Thống kê doanh thu theo khách hàng tại từng vùng theo từng ngày.**

- Sự kiện: Thống kê doanh thu theo khách hàng vào cuối ngày
- Bối cảnh
  - Khi nào: cuối mỗi ngày.
  - Ở đâu: vùng
  - Ai: Khách hàng
- Đo lường: tổng doanh thu
- **Lược đồ thiết kế:**



Hình 2. Thiết kế DDS

### III. PHÂN TÍCH QUY TRÌNH ETL

#### 1. Quy trình Source vào Stage

- Kỹ thuật: Incremental Extract.
- Các bước:
  - Cập nhật CET tại bảng Data\_flow trong database Metadata.

- Truncate bảng trong Stage.
  - Lấy LSET từ bảng Data\_flow trong database Metadata.
  - Đổ dữ liệu mới vào Stage.
  - Cập nhật lại LSET trong bảng Data\_flow.
- Source Central:

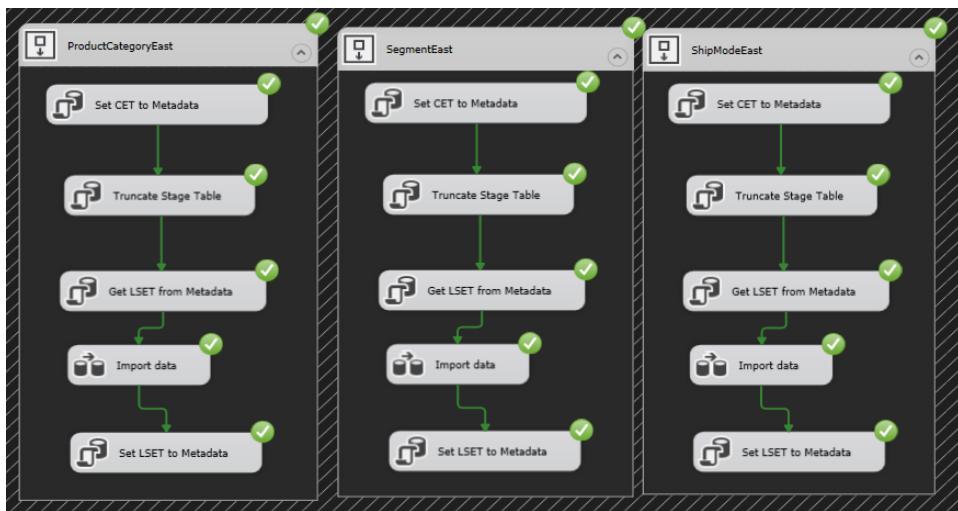


Hình 3. Source Central to Stage

- Source East:

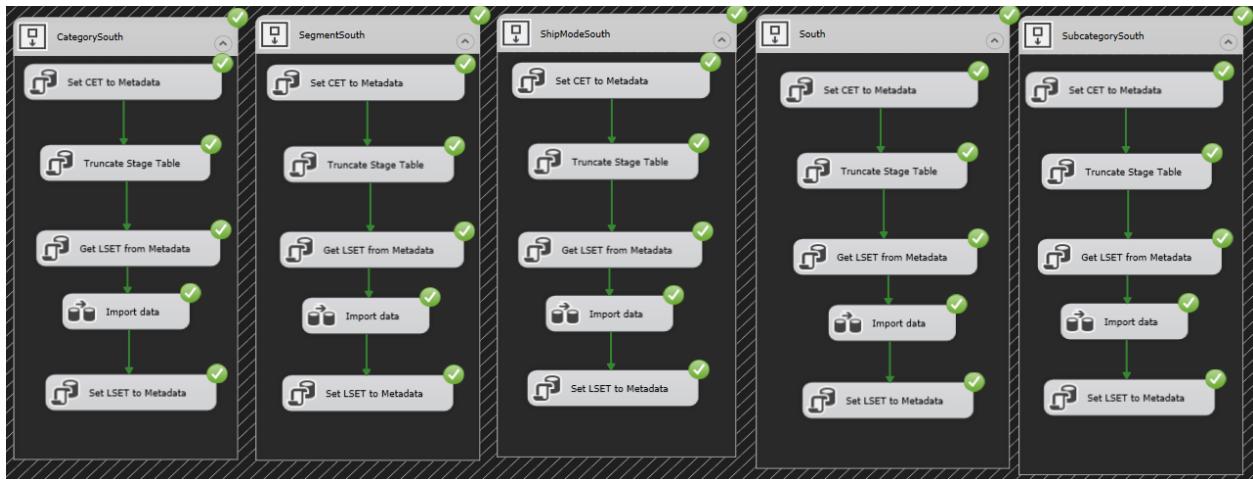


Hình 4. Source East to Stage (1)



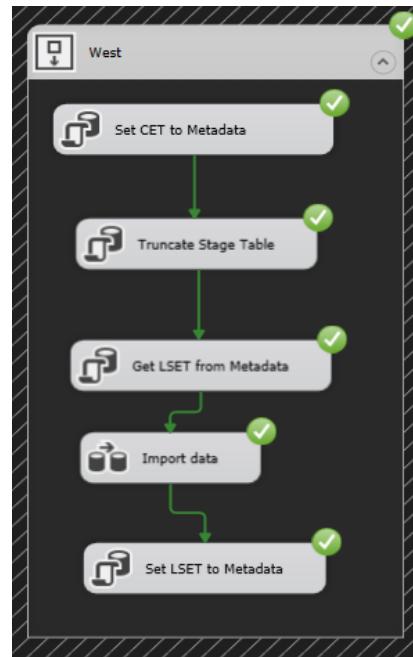
Hình 5. Source East to Stage (2)

- Source South:



Hình 6. Source South to Stage

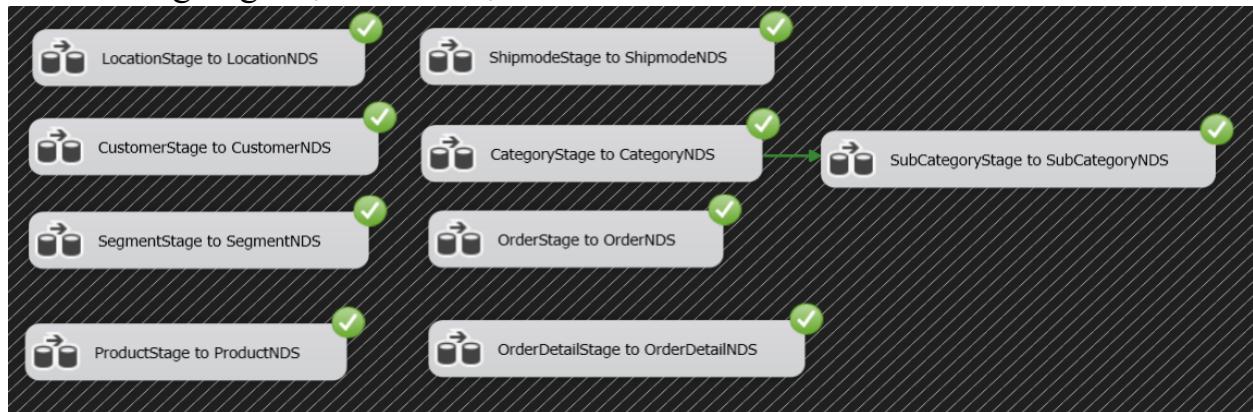
- Source West:



Hình 7. Source West to Stage

## 2. Quy trình Stage vào NDS

- Quá trình thực hiện: thực hiện load các dữ liệu mới từ stage và đổ vào bảng tương ứng được thiết kế tại NDS.



Hình 8. Stage to NDS

- Chi tiết:
  - o LocationStage to LocationNDS:

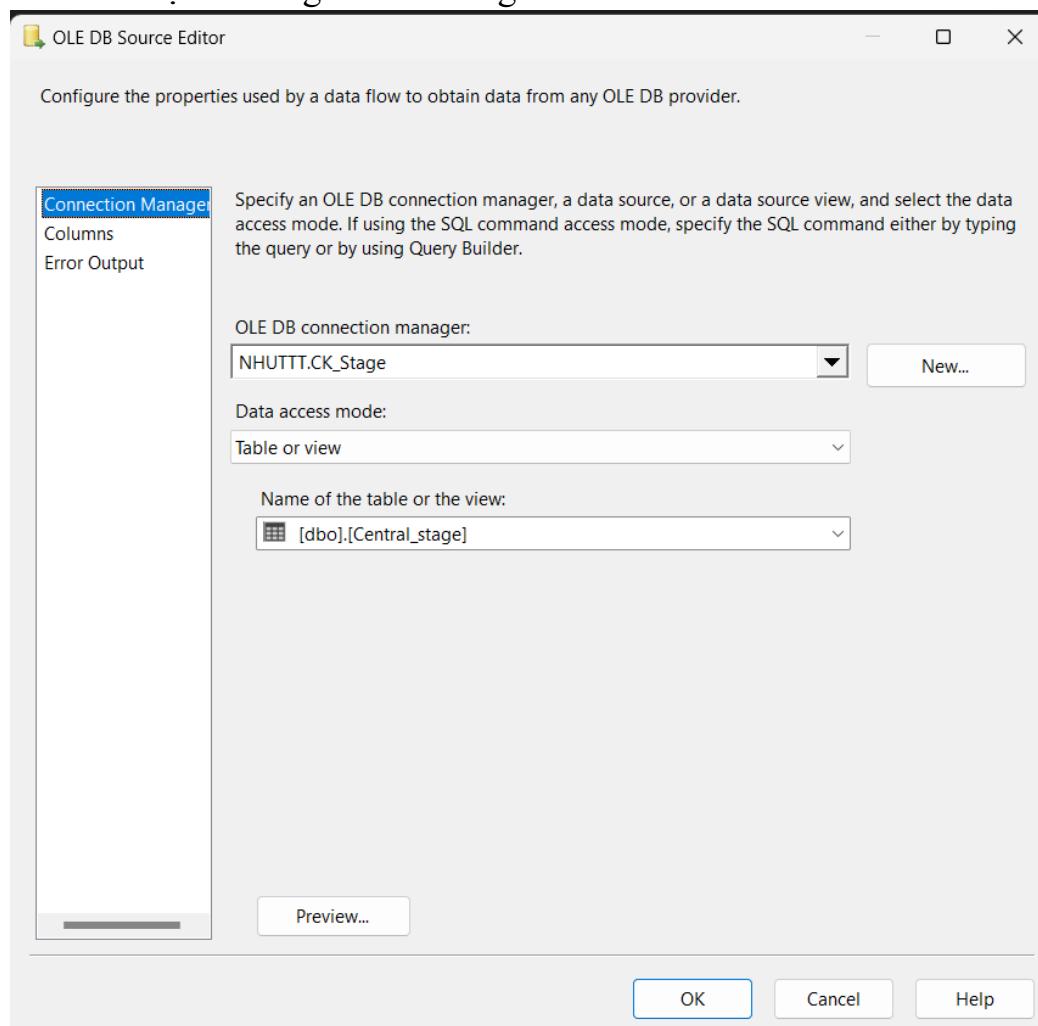


Hình 9. Location Stage to Location NDS (1)



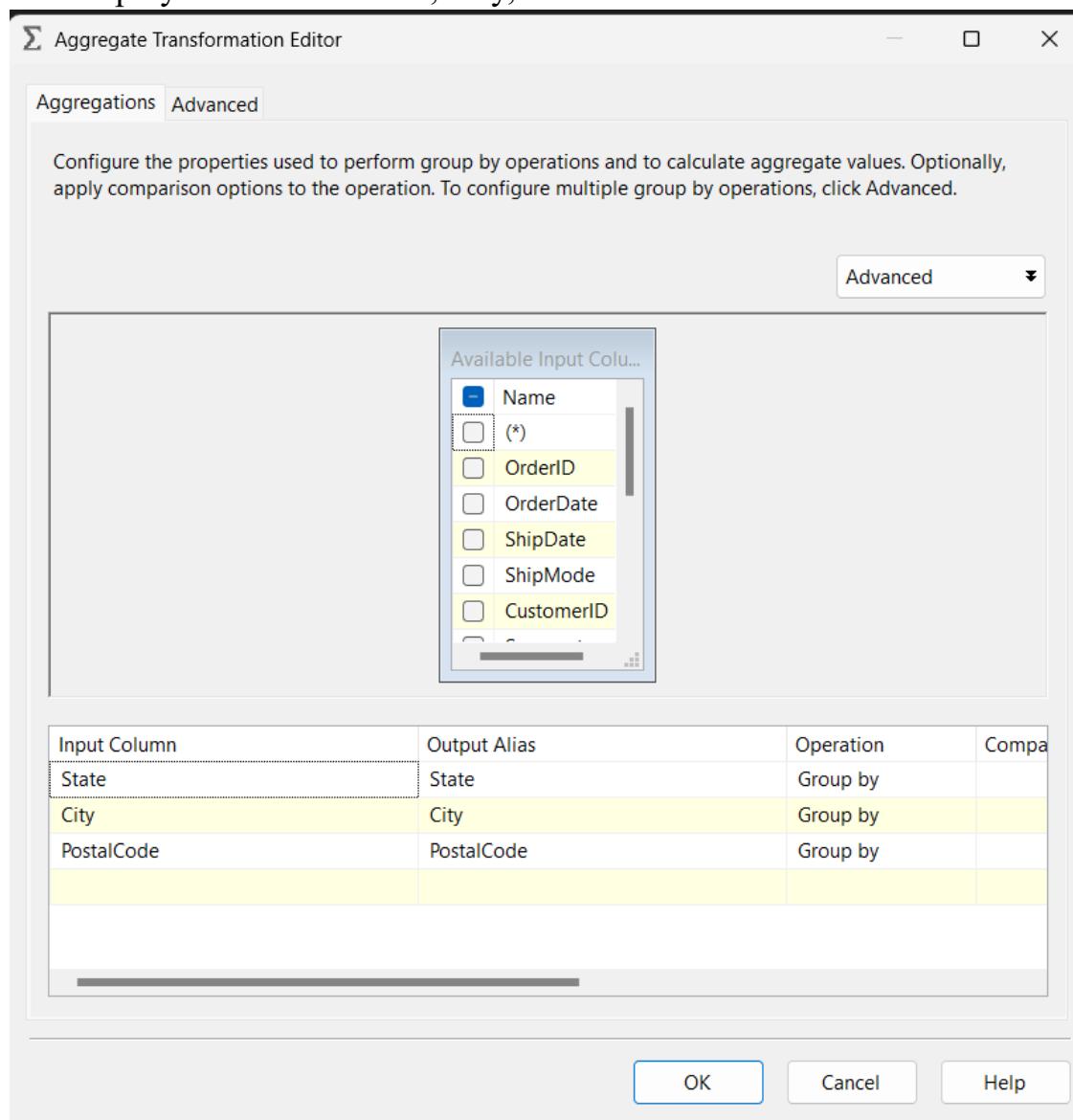
Hình 10. Location Stage to Location NDS (2)

- Load Dữ liệu từ bảng Central Stage:



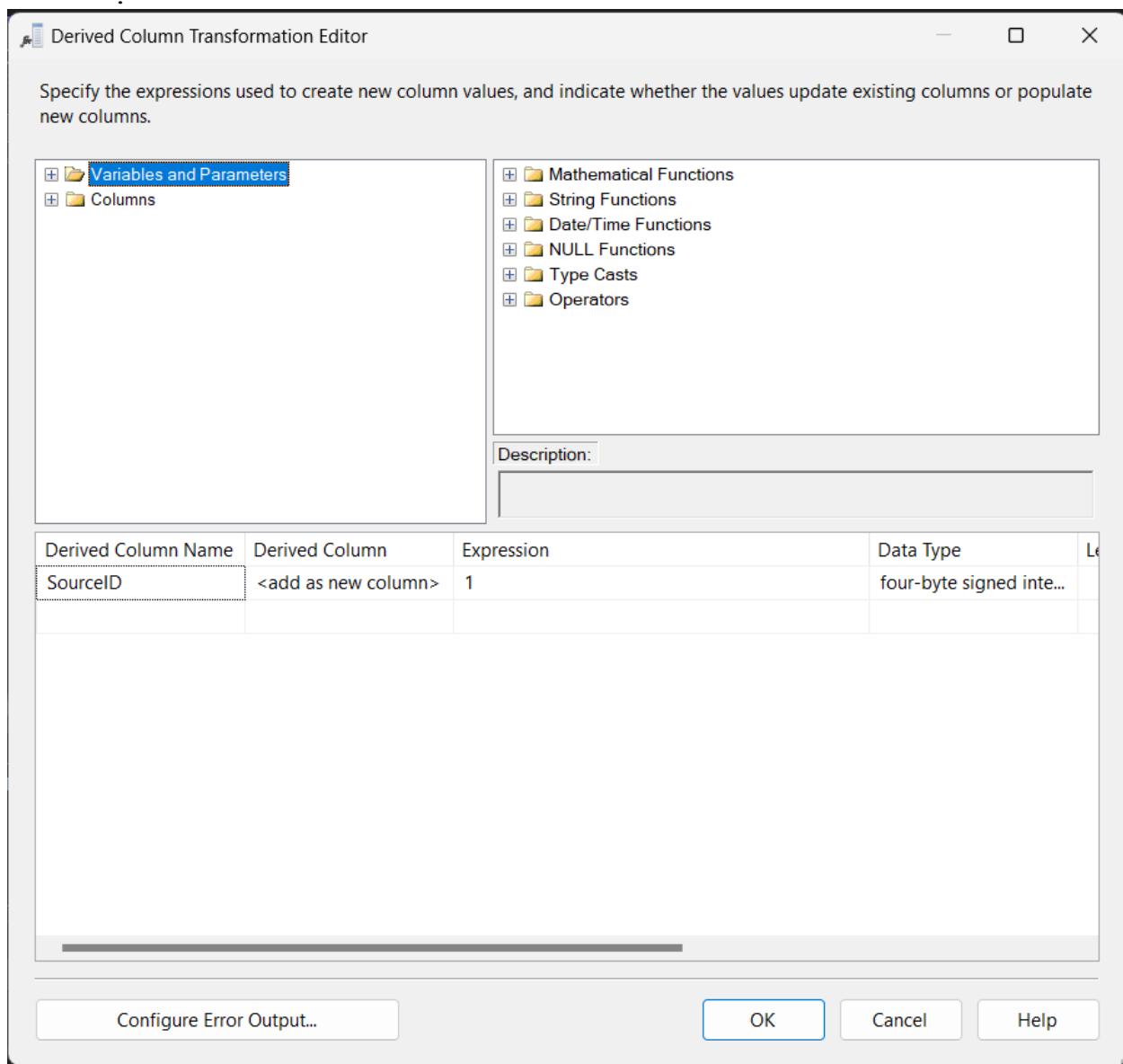
Hình 11. Load dữ liệu từ Stage

- Group by theo Postal Code, City, State:



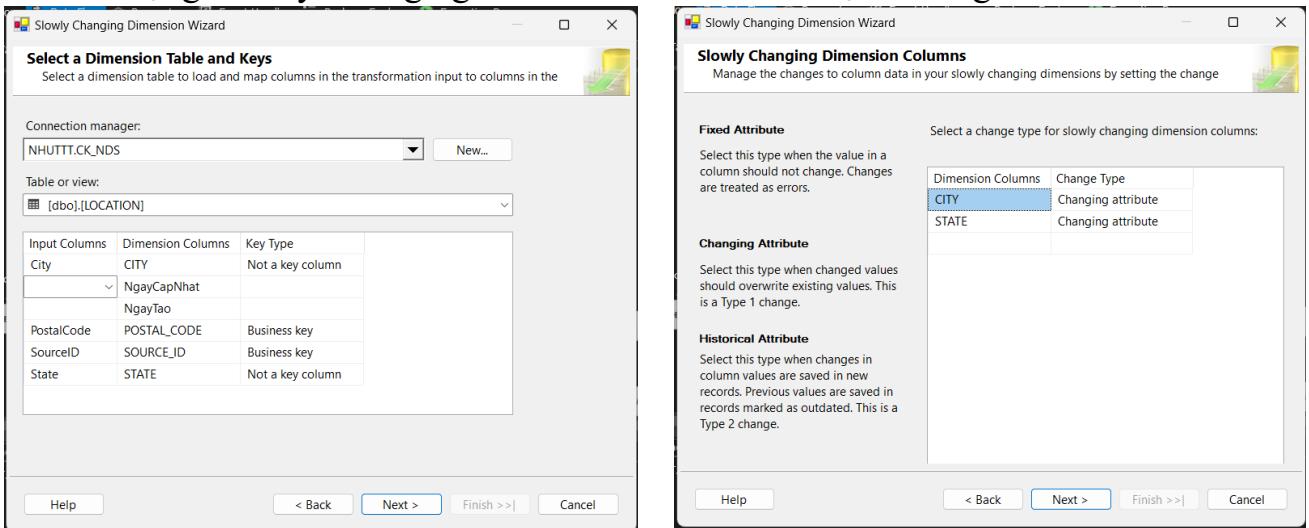
Hình 12. Group by theo Postal Code, City, State

- Tạo Derived Column Source ID



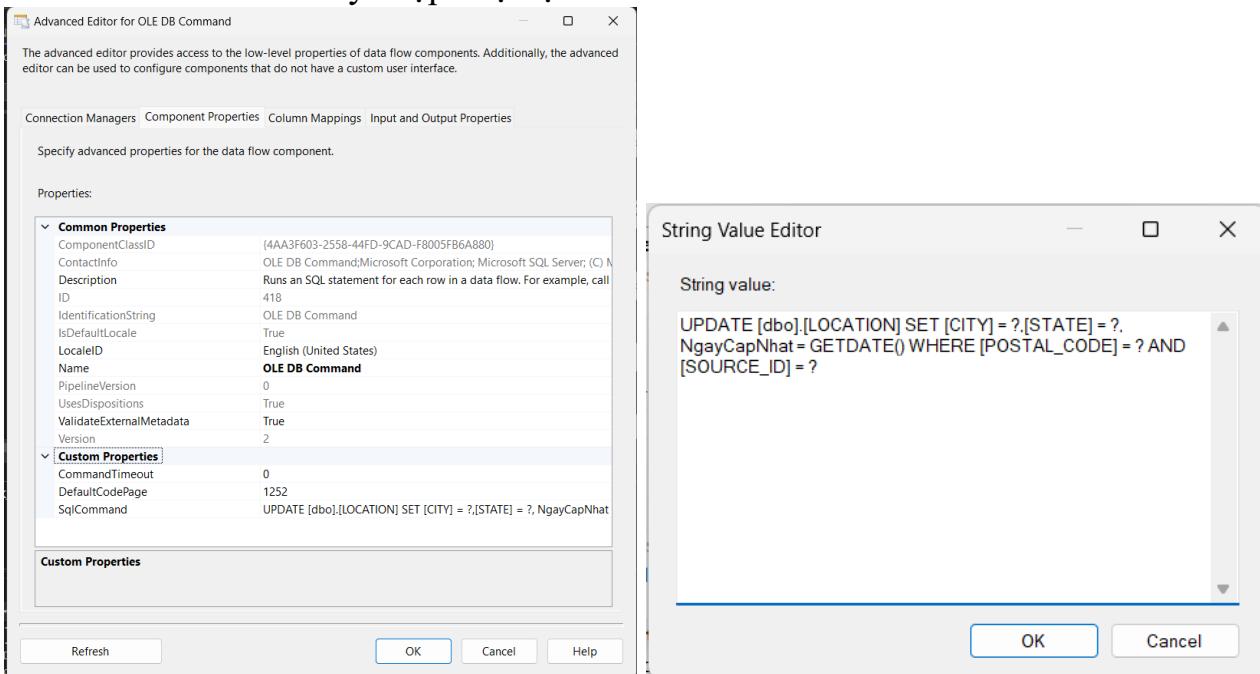
Hình 13. Tao Derived Column

- Sử dụng Slowly Changing Dimension để load dữ liệu từ Stage vào NDS



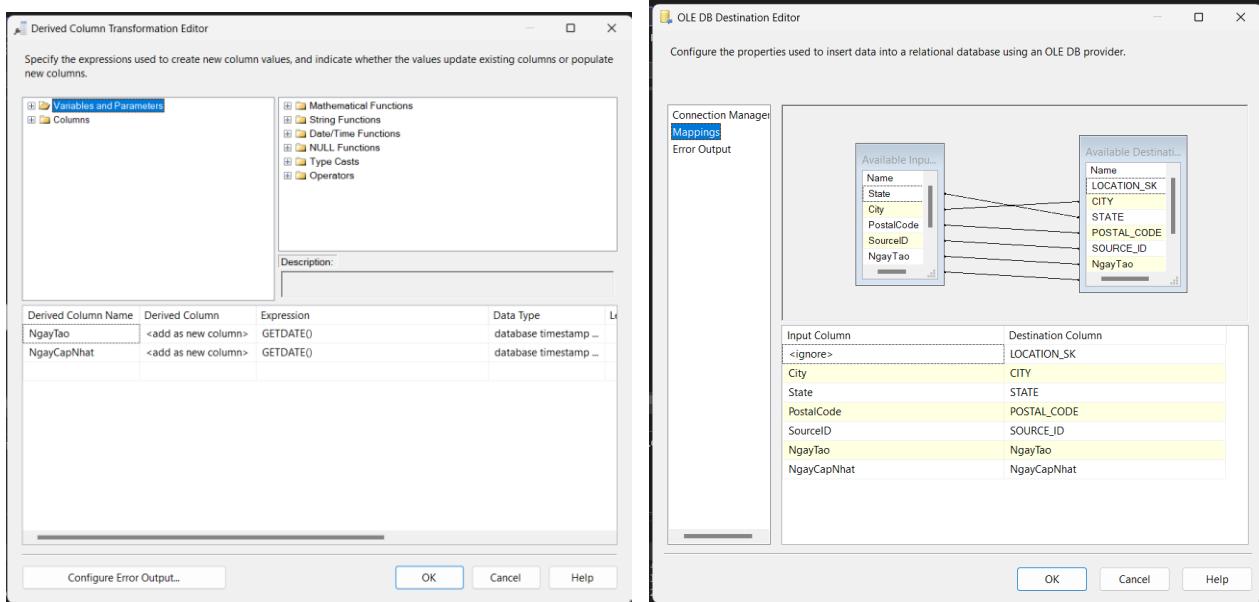
Hình 14. Sử dụng SCD để load dữ liệu vào NDS

### ○ Nếu tìm thấy: Cập nhật lại



Hình 15. Cập nhật lại NDS

- Nếu không tìm thấy: Thêm dòng mới



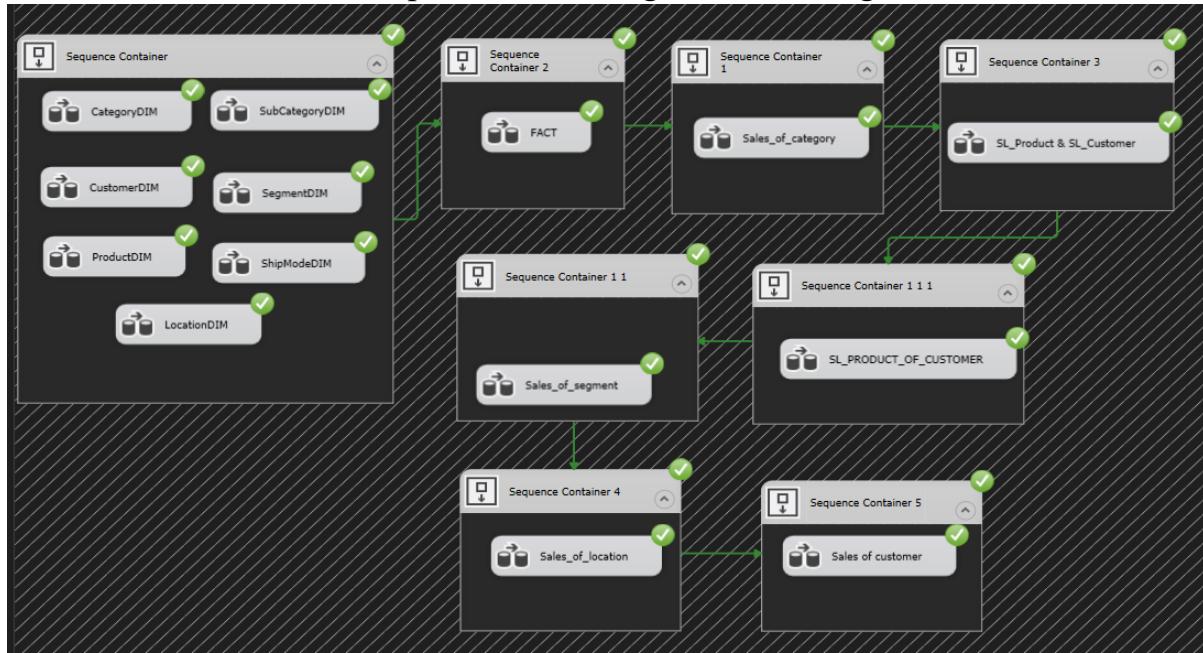
Hình 16. Thêm dòng mới vào NDS

- **Những bảng còn lại thực hiện tương tự.**

### 3. Quy trình NDS vào DDS

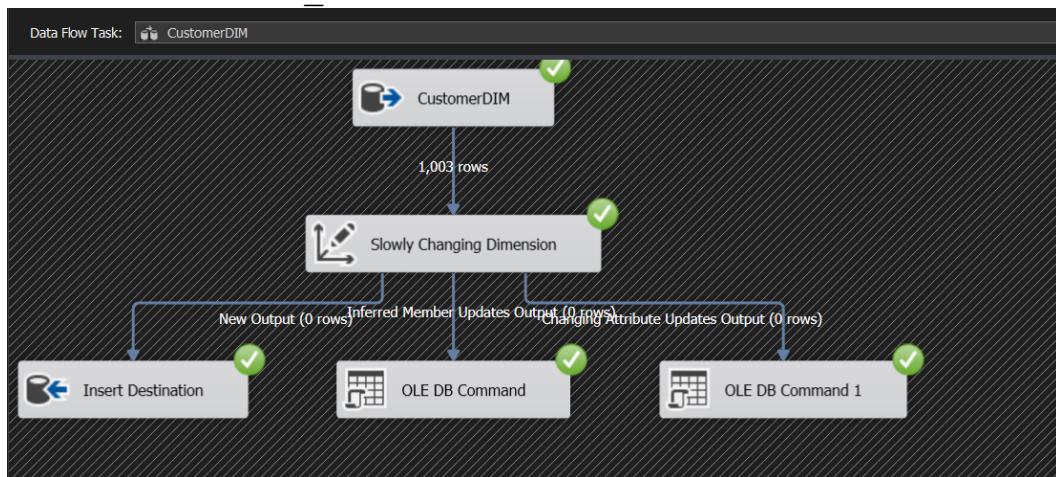
- Quy trình thực hiện:
  - Đỗ các bảng chiều: CUSTOMER\_DIM, LOCATION\_DIM, SHIPMODE\_DIM, CATEGORY\_DIM, SEGMENT\_DIM, DATE\_DIM, PRODUCT\_DIM, SUB\_CATEGORY\_DIM.
  - Đỗ dữ liệu vào bảng fact ứng với các nhu cầu cần phân tích.

- Ánh thực thi toàn bộ quá trình đổ bảng chiều và bảng fact:



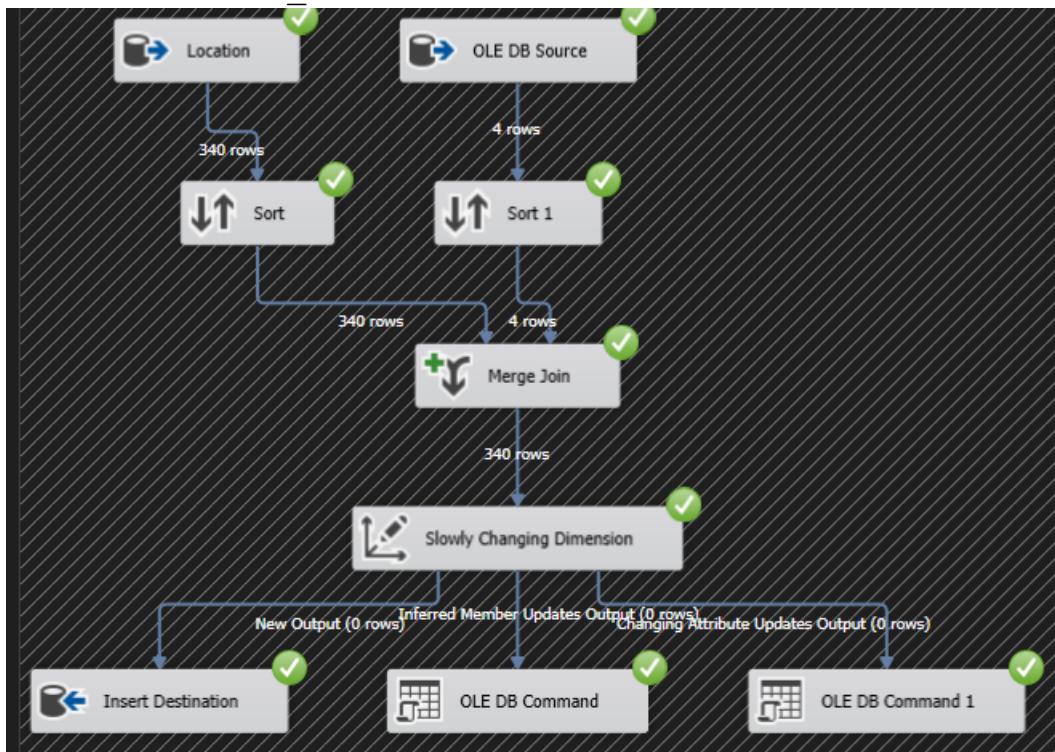
Hình 17. NDS to DDS

- Quá trình đổ các bảng chiều:
  - CUSTOMER\_DIM



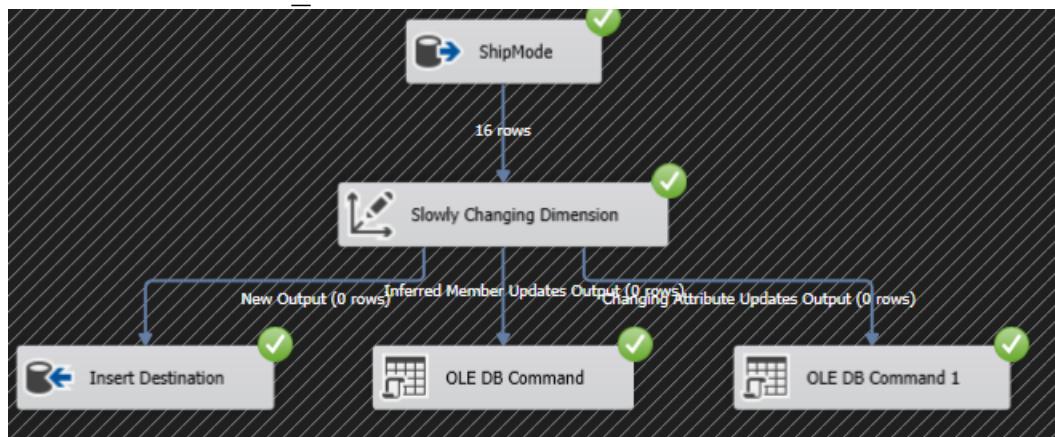
Hình 18. Customer DIM

- LOCATION\_DIM



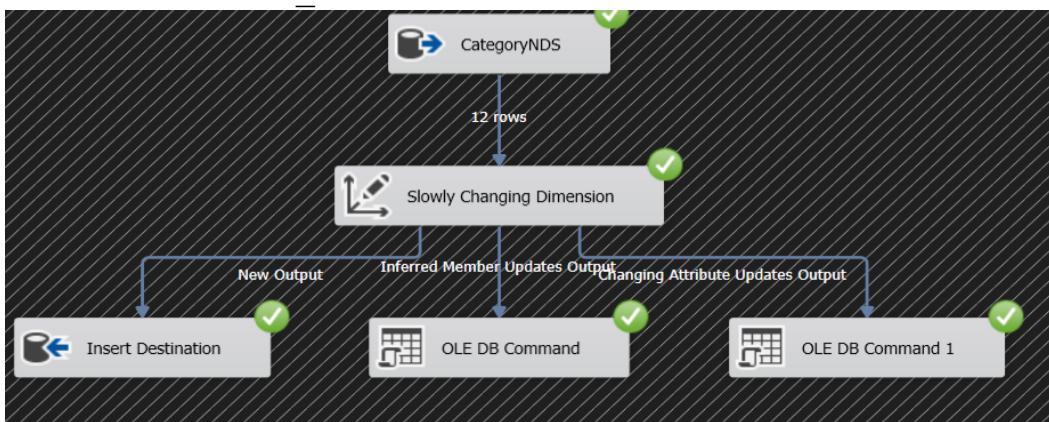
Hình 19. Location DIM

- SHIPMODE\_DIM



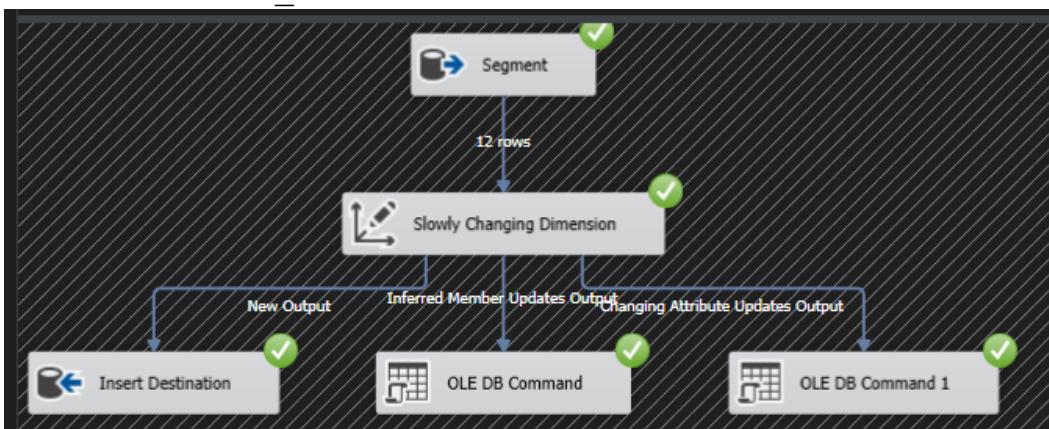
Hình 20. Shipmode DIM

- CATEGORY\_DIM



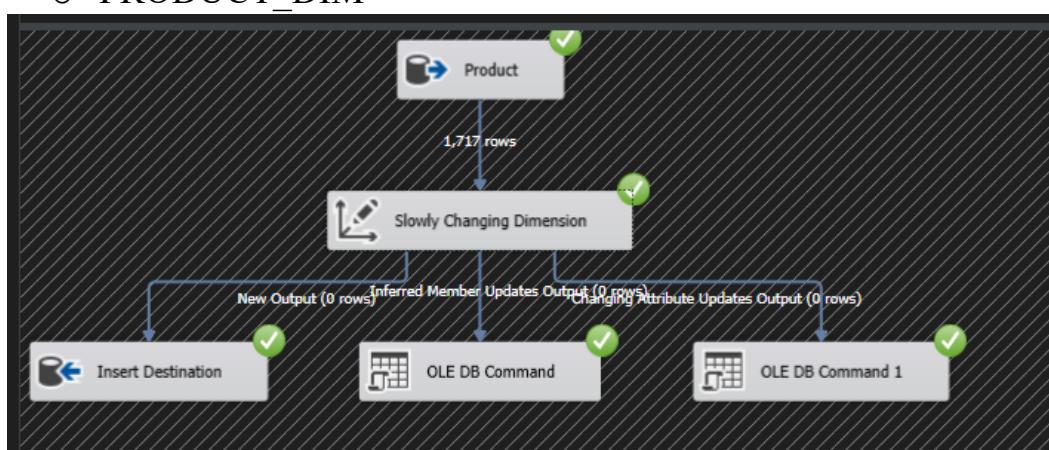
Hình 21. Category DIM

- SEGMENT\_DIM



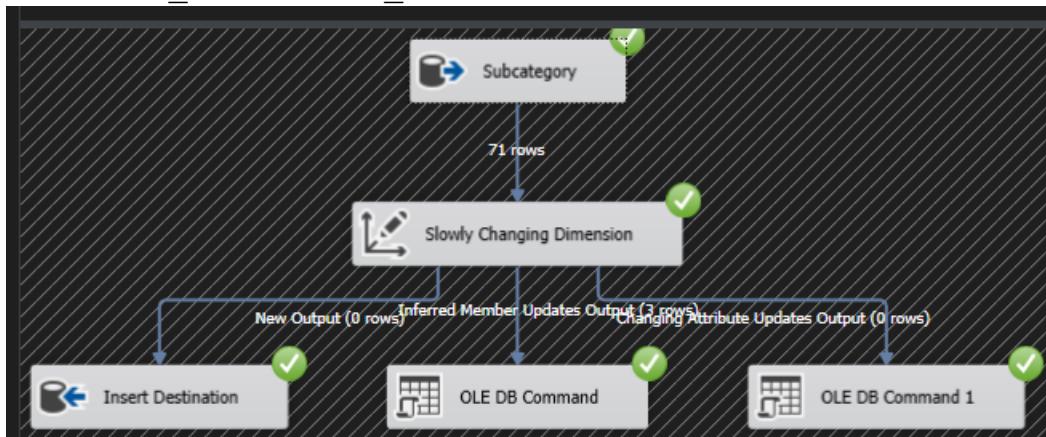
Hình 22. Segment DIM

- PRODUCT\_DIM



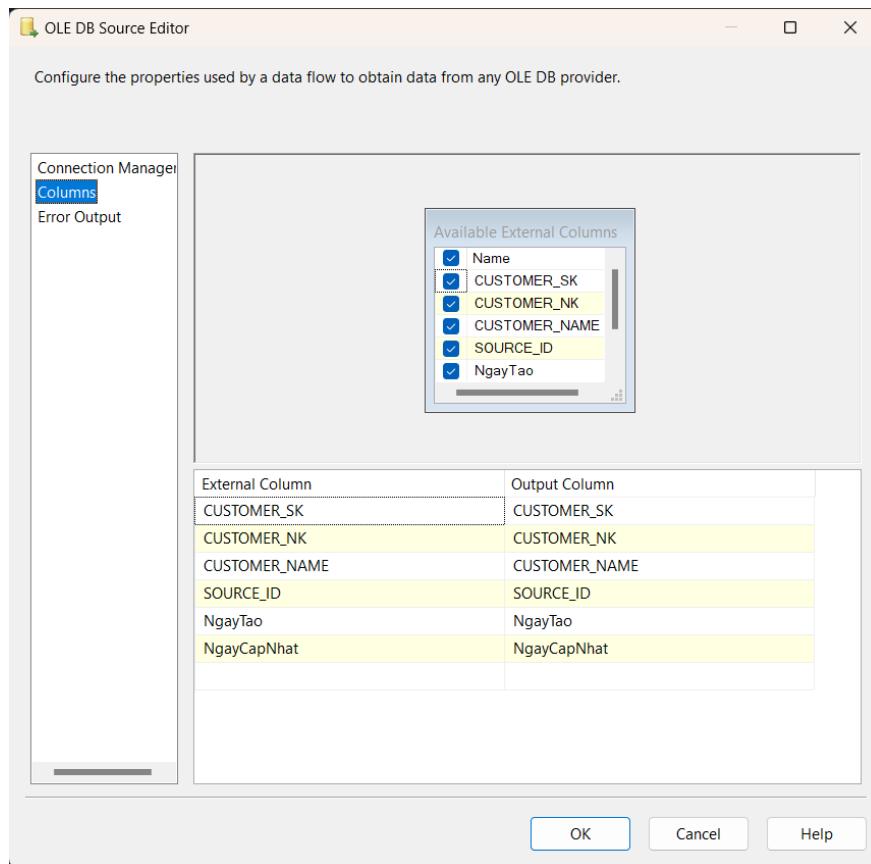
Hình 23. Product DIM

- SUB\_CATEGORY\_DIM



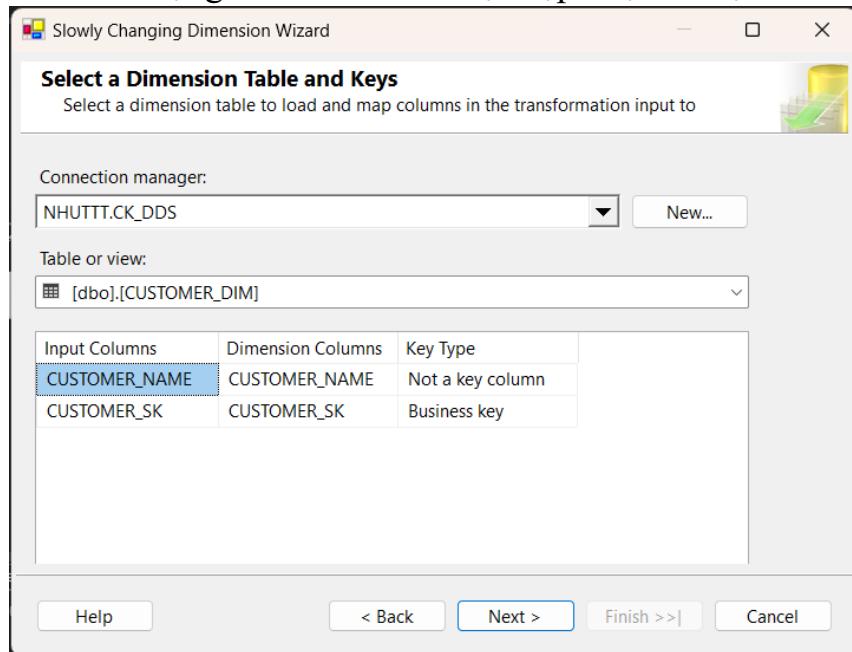
Hình 24. Sub Category DIM

- DATE\_DIM: thực thi script **DATE\_DIM.SQL**
- Chi tiết đỗ dữ liệu vào CUSTOMER\_DIM.
  - Load dữ liệu từ bảng Customer trong NDS

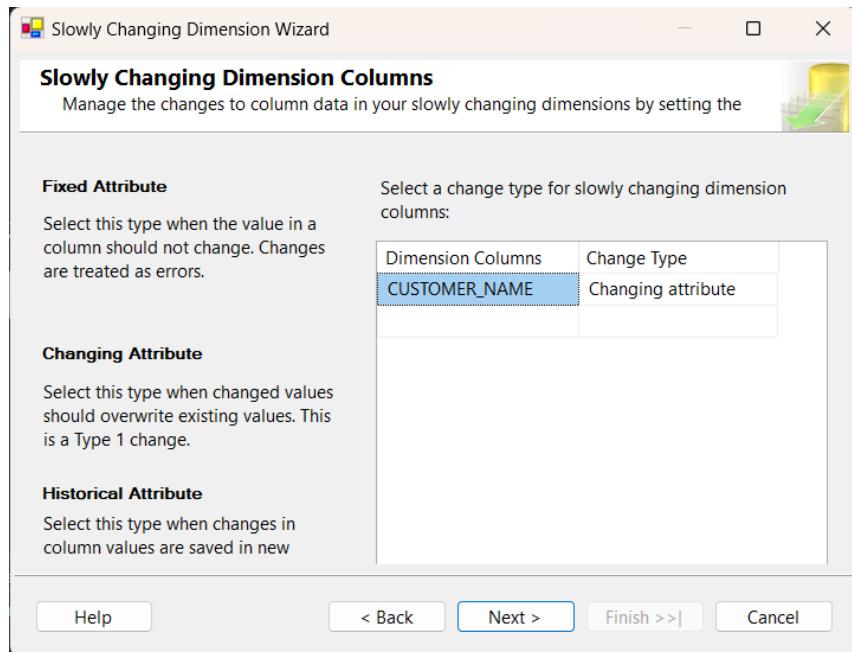


Hình 25. Load dữ liệu từ Customer NDS

- Sử dụng SCD để thêm hoặc cập nhật dữ liệu

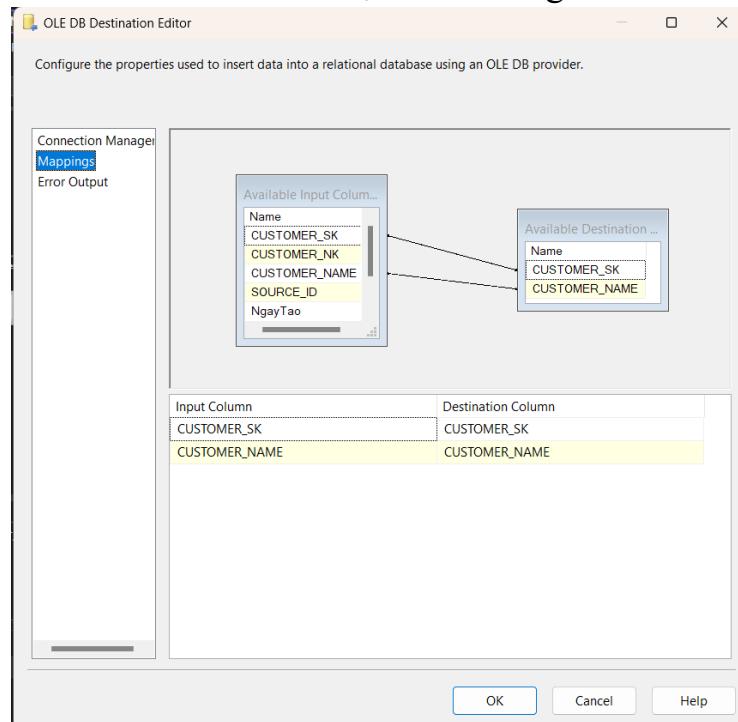


Hình 26. Sử dụng SCD để đỗ dữ liệu vào bảng DIM



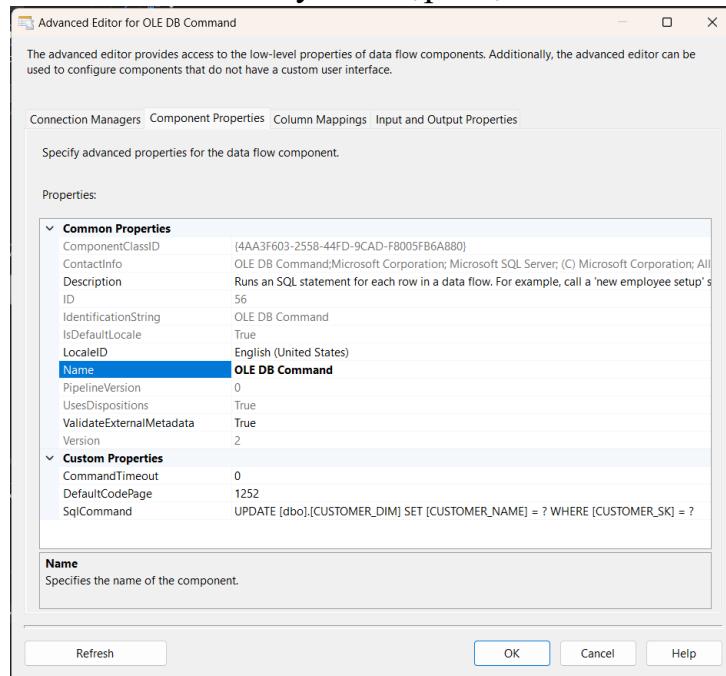
Hình 27. Chọn loại SCD

- Nếu chưa tồn tại: thêm dòng mới



Hình 28. Thêm dòng mới vào DIM

- Nếu có thay đổi: cập nhật



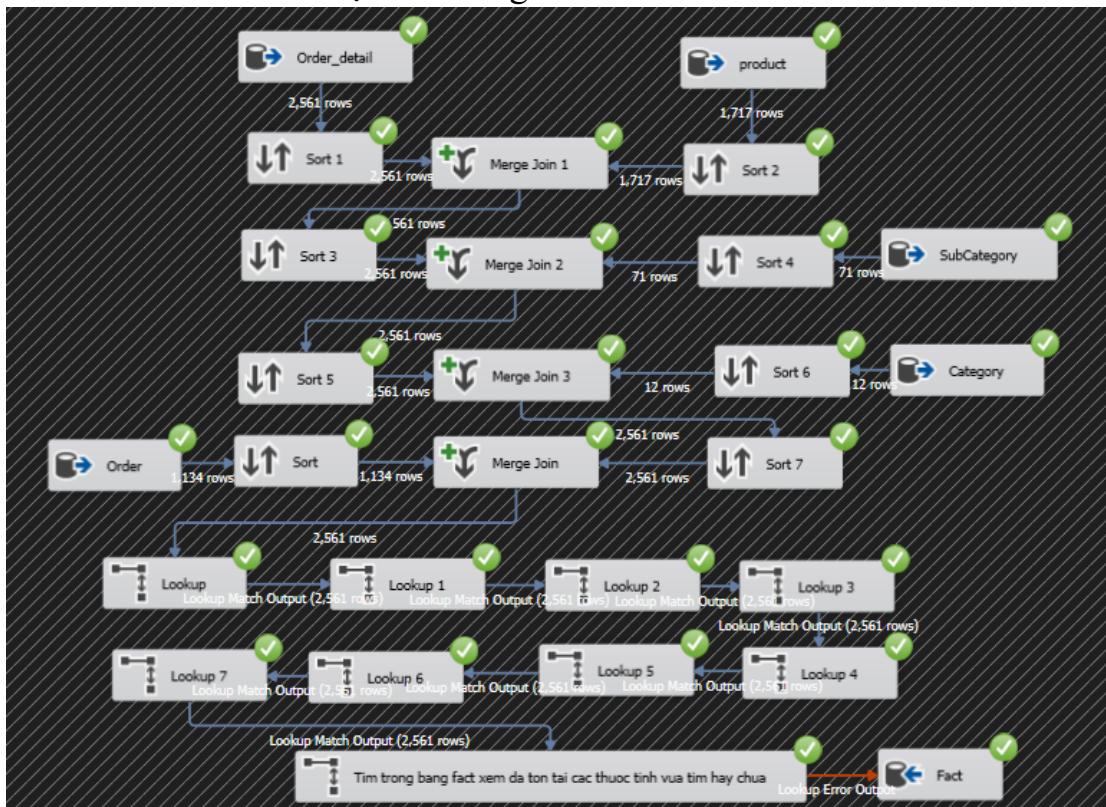
Hình 29. Cập nhật lại DIM

- Các bảng còn lại thực thi tương tự.

- **Đỗ bảng fact và các nhu cầu phân tích:**

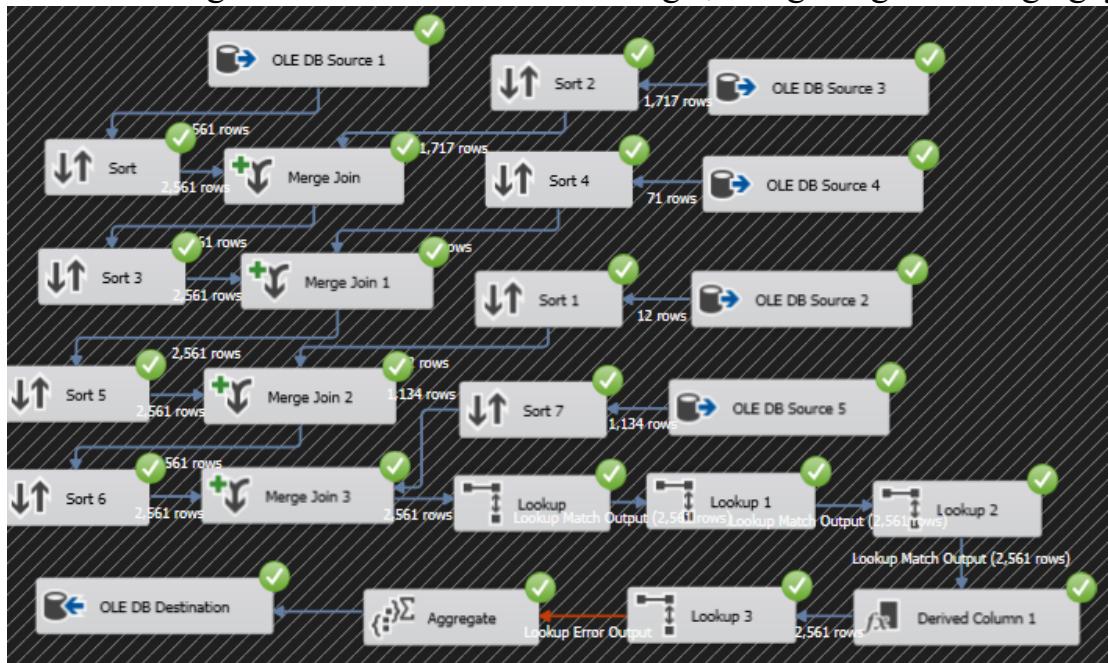
o Đỗ dữ liệu chưa tính toán:

- Tiến hành trích xuất dữ liệu từ NDS.
- Kết các bảng để lấy được thuộc tính cần thiết.
- Lookup để tìm kiếm sự tồn tại của các giá trị đó trong các bảng DIM tương ứng.
- Lookup để xác nhận các dòng vừa mới thêm vào chưa tồn tại trong bảng FACT (nghiêm cấm thay đổi lịch sử dữ liệu thống kê)
- Đỗ dữ liệu vào bảng FACT.



Hình 30. Đỗ dữ liệu vào bảng FACT: Chưa tính toán

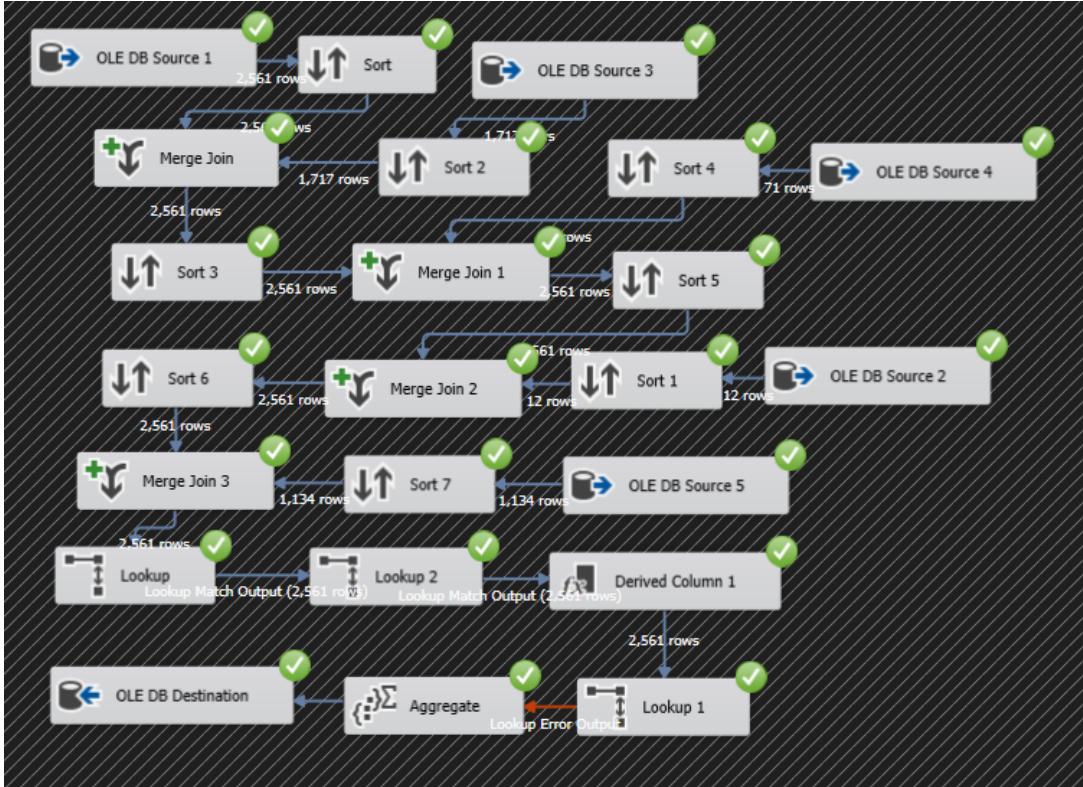
- Thống kê doanh thu theo khách hàng tại từng vùng theo từng ngày.



Hình 31. Đỗ dữ liệu vào bảng FACT (1)

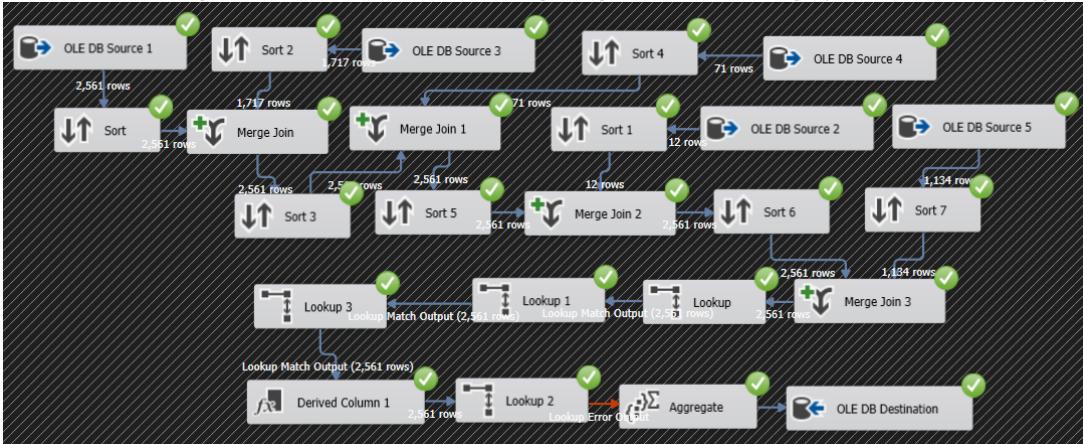
- Giải thích thực thi quá trình tính toán dữ liệu theo nhu cầu “Thống kê doanh thu theo khách hàng tại từng vùng theo từng ngày”.
  - Tiến hành trích xuất dữ liệu từ NDS.
  - Kết các bảng để lấy được thuộc tính cần thiết (CustomerID, Sales, RegionID, Date)
  - Lookup để tìm kiếm sự tồn tại của các giá trị đó trong các bảng DIM tương ứng.
  - Tạo một số cột mới trong Derived Column để phục vụ nhu cầu tìm kiếm, so sánh,... khi nạp dữ liệu vào bảng FACT.
  - Lookup để xác nhận các dòng vừa mới thêm vào chưa tồn tại trong bảng FACT (nghiêm cấm thay đổi lịch sử dữ liệu thống kê)
  - Tiến hành nhóm, tính toán các giá trị cần thống kê (Sales\_of\_Customer).
  - Đỗ dữ liệu vào bảng FACT.
- **Các nhu cầu thống kê còn lại được thực hiện với quá trình tương tự.**

- Thống kê số khách hàng mua hàng theo ngày của mỗi vùng; Thống kê số lượng sản phẩm bán được theo ngày của từng vùng.



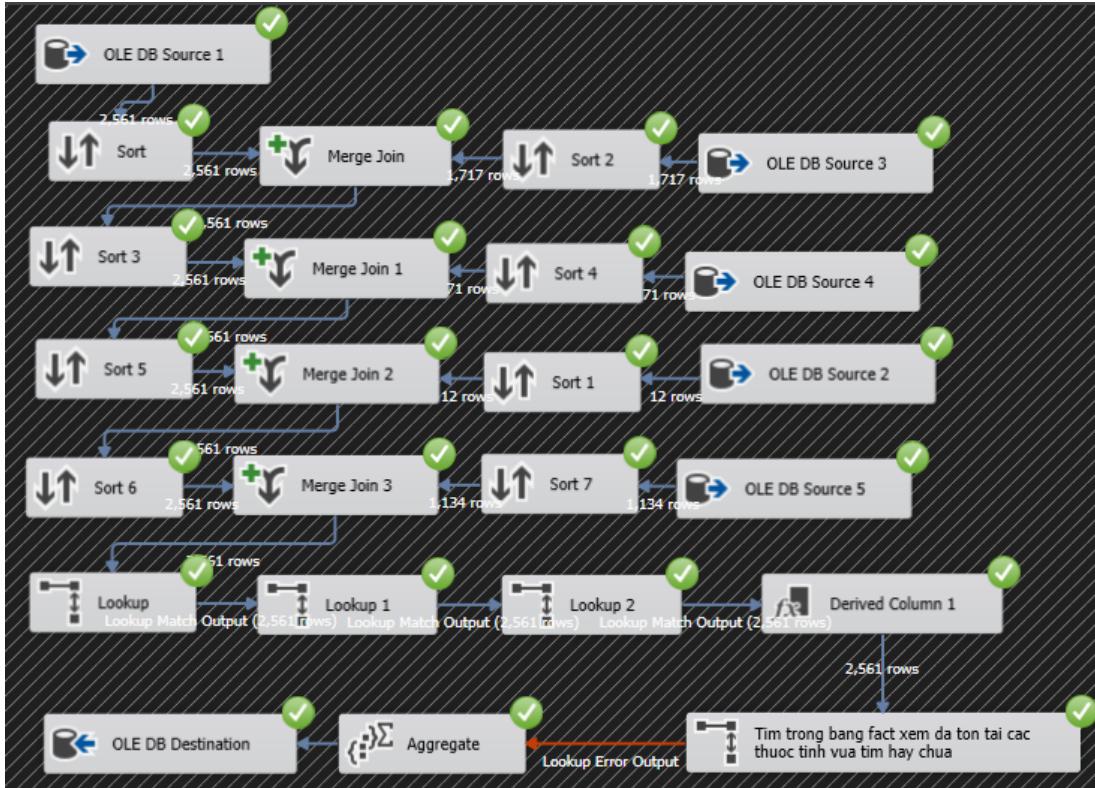
Hình 32. Đồđứklạivào bảng FACT (2)

- Thống kê doanh thu của từng segment theo ngày của mỗi vùng.



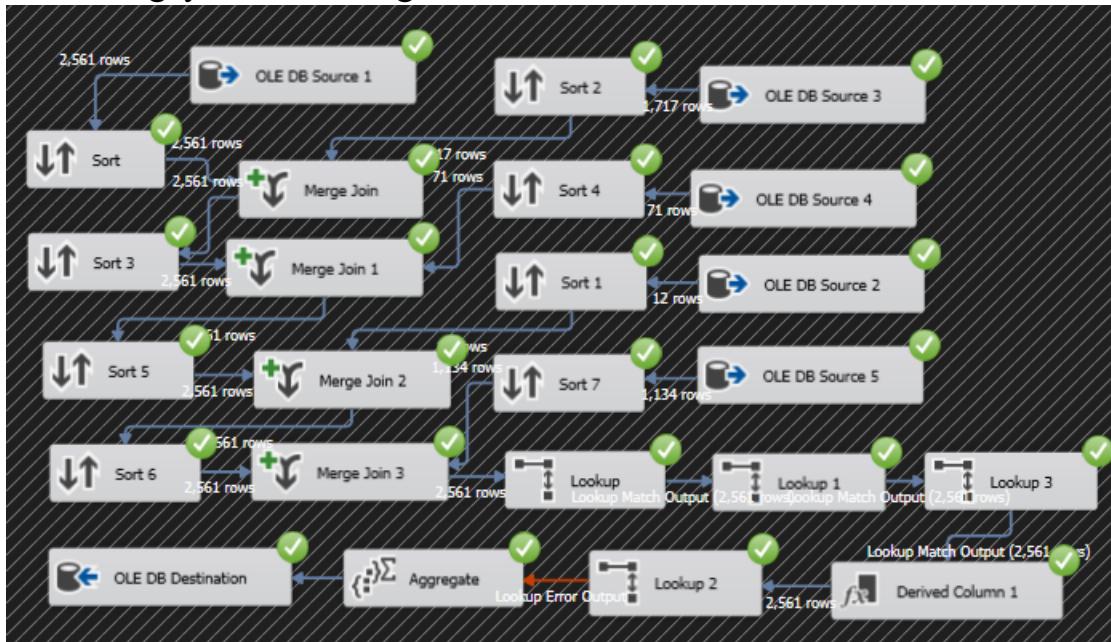
Hình 33. Đồđứklạivào bảng FACT (3)

- Thống kê doanh thu của từng danh mục theo ngày theo mỗi vùng.



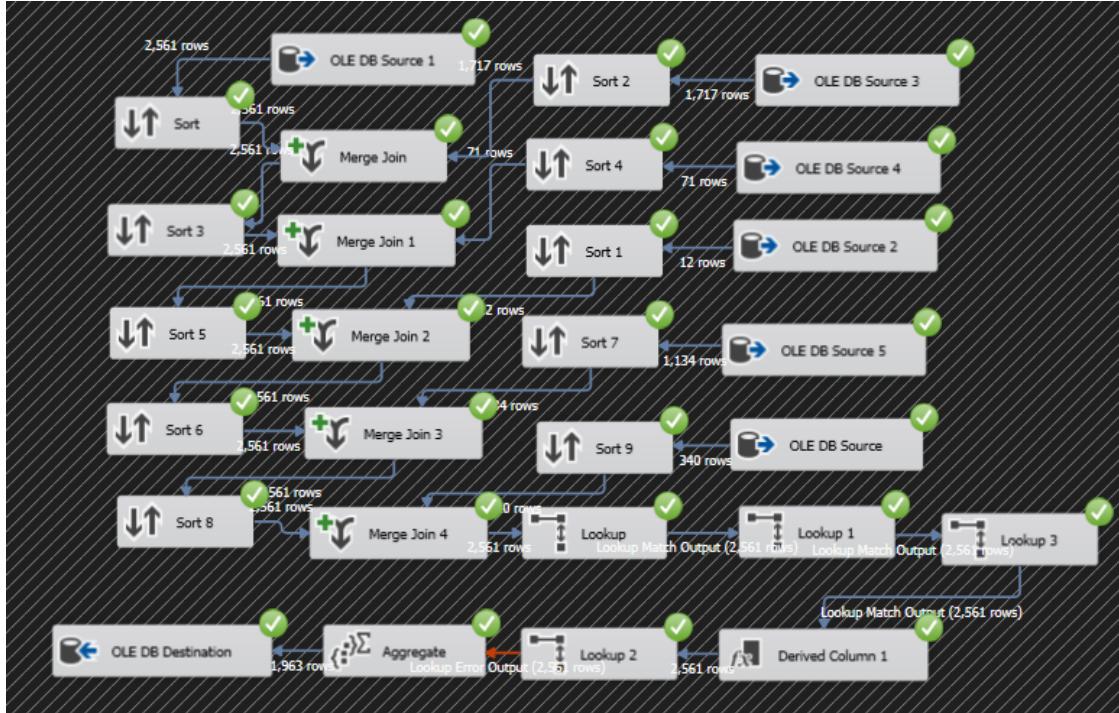
Hình 34. Đỗ dữ liệu vào bảng FACT (4)

- Thống kê số lượng sản phẩm đã mua của mỗi khách hàng trong mỗi ngày của mỗi vùng.



Hình 35. Đỗ dữ liệu vào bảng FACT (5)

- Thống kê doanh thu theo sản phẩm trong từng thành phố theo ngày.



Hình 36. Đồ dữ liệu vào bảng FACT (6)

## IV. OLAP REPORT

- Thống kê số khách hàng mua hàng theo ngày của mỗi vùng:
  - Vào ngày **<Date>**, vùng **<Region Name>** đã có **<SL Customer>** khách hàng mua hàng.

The screenshot shows the Microsoft Analysis Services (SSAS) Data Model Explorer interface. On the left, a tree view displays various dimensions: CATEGORY DIM 1, CUSTOMER DIM 1, DATE DIM 1 (with sub-nodes Date, Month, Quarter, Hierarchy), LOCATION DIM 1 (with sub-nodes CITY, LOCATION SK, REGION, REGION NAME, STATE, Hierarchy), and PRODUCT DIM 1. Below the tree is a section for Calculated Members, which is currently empty. On the right, a data grid displays a table with three columns: Date, REGION NAME, and SL CUSTOMER. The data shows daily customer counts across different regions from January 13 to January 23, 2018.

Date	REGION NAME	SL CUSTOMER
2018-01-13	Central	1
2018-01-14	South	2
2018-01-14	West	2
2018-01-15	Central	1
2018-01-15	East	2
2018-01-15	West	1
2018-01-16	Central	1
2018-01-16	West	1
2018-01-19	Central	1
2018-01-19	East	1
2018-01-20	Central	1
2018-01-20	East	2
2018-01-20	West	2
2018-01-21	Central	2
2018-01-21	South	1
2018-01-21	West	1
2018-01-22	Central	1
2018-01-22	West	1
2018-01-23	Central	1

Hình 37. OLAP: Thống kê khách hàng mua hàng theo ngày

- Thống kê doanh thu của từng segment theo ngày của mỗi vùng.
  - o Vào ngày <Date>, vùng <Region Name>, thu được doanh thu <Sale of Segment> từ nhóm khách hàng <Segment Name>

The screenshot shows a Microsoft Analysis Services cube interface. On the left, a tree view displays various dimensions: Quarter, Hierarchy, LOCATION DIM 1 (with CITY, LOCATION SK, REGION, REGION NAME, STATE, Hierarchy), PRODUCT DIM 1, SEGMENT DIM 1 (with SEGMENT NAME, SEGMENT SK, Hierarchy), and SHIPMODE DIM 1. Below the dimensions is a section for Calculated Members. To the right, a data grid titled '<Select dimension>' shows sales data across four columns: Date, REGION NAME, SEGMENT NAME, and SALES OF SEGMENT. The data spans from January 13 to January 22, 2018.

Date	REGION NAME	SEGMENT NAME	SALES OF SEGMENT
2018-01-13	Central	Corporate	4619.33
2018-01-14	South	Consumer	934.758
2018-01-14	West	Consumer	154.9
2018-01-14	West	Corporate	337.688
2018-01-15	Central	Consumer	147.636
2018-01-15	East	Consumer	1971.244
2018-01-15	West	Home Office	21.4
2018-01-16	Central	Consumer	5802.7
2018-01-16	West	Corporate	427.592
2018-01-19	Central	Consumer	6
2018-01-19	East	Consumer	1370.535
2018-01-20	Central	Consumer	11.52
2018-01-20	East	Consumer	3.52
2018-01-20	East	Corporate	207.846
2018-01-20	West	Consumer	544.952
2018-01-21	Central	Consumer	268.576
2018-01-21	Central	Home Office	1239.833
2018-01-21	South	Home Office	767.984
2018-01-21	West	Home Office	654.242
2018-01-22	Central	Consumer	22.000

Hình 38. OLAP: Thống kê doanh thu theo từng Segment

- Thống kê doanh thu của từng danh mục theo ngày theo mỗi vùng.
  - o Vào ngày **<Date>**, vùng **<Region Name>**, doanh thu thu được từ việc bán sản phẩm thuộc **<Category Name>** là **<Sale of Category>**

The screenshot shows a Microsoft Analysis Services (SSAS) cube interface. On the left, there is a navigation pane titled '<All>' containing a hierarchy tree. The tree includes categories like Month, Quarter, Hierarchy, LOCATION DIM 1 (which further branches into CITY, LOCATION SK, REGION, REGION NAME, STATE, and Hierarchy), and Calculated members. A vertical scroll bar is visible on the right side of the pane. On the right, there is a data grid with four columns: Date, REGION NAME, CATEGORY NAME, and SALES OF CATEGORY. The data consists of 18 rows, each representing a specific date, region, category, and sales value. The first row (2018-01-13) is highlighted with a blue selection bar.

Date	REGION NAME	CATEGORY NAME	SALES OF CATEGORY
2018-01-13	Central	Furniture	212.94
2018-01-13	Central	Office Supplies	4406.39
2018-01-14	South	Office Supplies	358.998
2018-01-14	South	Technology	575.76
2018-01-14	West	Office Supplies	323.524
2018-01-14	West	Technology	169.064
2018-01-15	Central	Office Supplies	37.06
2018-01-15	Central	Technology	110.576
2018-01-15	East	Office Supplies	1971.244
2018-01-15	West	Office Supplies	21.4
2018-01-16	Central	Furniture	302.67
2018-01-16	Central	Office Supplies	5500.03
2018-01-16	West	Furniture	427.592
2018-01-19	Central	Office Supplies	6
2018-01-19	East	Furniture	919.239
2018-01-19	East	Office Supplies	21.696

Hình 39. OLAP: Thống kê doanh thu theo từng danh mục (Category)

- Thống kê số lượng sản phẩm đã mua của mỗi khách hàng trong mỗi ngày của mỗi vùng.
  - o Vào ngày **<Date>**, vùng **<Region Name>**, đã bán được cho khách hàng **<Customer name>** **<SL Product of customer>** sản phẩm.

The screenshot shows a data analysis interface with a sidebar on the left containing a tree view of dimensions and measures. The tree includes KPIs, Category Dim 1 (Category Name, Category SK, Hierarchy), Customer Dim 1 (Customer Name, Customer SK, Hierarchy), and Date Dim 1 (Date, Month, Quarter, Hierarchy). The main area displays a data grid with the following columns: Date, Region Name, Customer Name, and SL Product of Cus... (partially visible). The data consists of 18 rows of sales data for January 2018, categorized by date, region, customer, and quantity.

Date	REGION NAME	CUSTOMER NAME	SL PRODUCT OF CUS...
2018-01-13	Central	Cathy Prescott	4
2018-01-14	South	Brosina Hoffman	25
2018-01-14	South	Rob Beeghly	4
2018-01-14	West	Sean O'Donnell	1
2018-01-14	West	Tracy Poddar	2
2018-01-15	Central	Tim Brockman	4
2018-01-15	East	Robert Waldorf	2
2018-01-15	East	Sanjit Chand	1
2018-01-15	West	Tonja Turnell	1
2018-01-16	Central	Andy Reiter	3
2018-01-16	West	Jack O'Briant	2
2018-01-19	Central	Dan Lawera	1
2018-01-19	East	Neola Schneider	4
2018-01-20	Central	Maris LaWare	1
2018-01-20	East	Cynthia Voltz	1
2018-01-20	East	Jeremy Farry	1
2018-01-20	West	Tamara Manning	2
2018-01-20	West	Thea Hendricks	1
2018-01-21	Central	Bobby Elias	1

Hình 40. OLAP: Thống kê số lượng sản phẩm đã mua của mỗi khách hàng

- Thống kê số lượng sản phẩm bán được theo ngày của từng vùng.
  - Vào ngày <Date>, vùng <Region Name>, bán được <SL Product> sản phẩm.

The screenshot shows a cube browser interface with two main sections: a dimension tree on the left and a data grid on the right.

**Dimension Tree (Left):**

- <All>
- CATEGORY DIM 1
- CUSTOMER DIM 1
- DATE DIM 1
  - Date
  - Month
  - Quarter
  - Hierarchy
- LOCATION DIM 1
  - CITY
  - LOCATION SK
  - REGION
  - REGION NAME
  - STATE
  - Hierarchy
- PRODUCT DIM 1

**Data Grid (Right):**

Date	REGION NAME	SL PRODUCT
2018-01-13	Central	4
2018-01-14	South	29
2018-01-14	West	3
2018-01-15	Central	4
2018-01-15	East	3
2018-01-15	West	1
2018-01-16	Central	3
2018-01-16	West	2
2018-01-19	Central	1
2018-01-19	East	4
2018-01-20	Central	1
2018-01-20	East	2
2018-01-20	West	3
2018-01-21	Central	6
2018-01-21	South	4
2018-01-21	West	6
2018-01-22	Central	2
2018-01-22	West	4
2018-01-23	Central	1

Hình 41. OLAP: Thống kê số sản phẩm bán được

- Thống kê doanh thu theo sản phẩm trong từng thành phố theo ngày
  - Vào ngày **<Date>**, thành phố **<City>**, doanh thu của sản phẩm **<Product Name>** là **<Sales of location>**

The screenshot shows a business intelligence tool's interface. On the left, a tree view displays the cube structure under 'All'. It includes 'Measures' (with 'Sales' selected), 'FACT' (containing 'ORDER ID', 'Sales', 'SALES OF CATEG...', 'SALES OF CUSTOM...', 'SALES OF LOCATIO...', 'SALES OF SEGMENT...', 'SL CUSTOMER', 'SL PRODUCT', 'SL PRODUCT OF CU...'), and 'KPIs'. Below these are dimension tables: 'CATEGORY DIM 1' and 'CUSTOMER DIM 1'. A 'Calculated Members' section is also present. On the right, a data grid displays sales data with columns: Date, CITY, PRODUCT NAME, and SALES OF LOCATION. The data spans from January 13 to January 15, 2018, across various cities like Springfield, Aurora, Jacksonville, Johnson City, Los Angeles, and Austin, listing products like Hon Metal Bookcases, Hunt Boston Vacuum, Martin Yale Chadless, Xerox 23, Imation 8GB Mini Tr..., Mobile Personal File C..., Sannysis Cute Owl D..., Tyvek Top-Opening ..., Avery Durable Slant ..., Ibico Presentation In..., iHome FM Clock Radi..., Xerox 1931, Xerox 19, Acco D-Ring Binder ..., Logitech 910-00297..., Maxell 74 Minute CD..., Wilson Jones Ledger..., and Newell 347.

Date	CITY	PRODUCT NAME	SALES OF LOCATION
2018-01-13	Springfield	Hon Metal Bookcases...	212.94
2018-01-13	Springfield	Hunt Boston Vacuum...	209.94
2018-01-13	Springfield	Martin Yale Chadless ...	4164.05
2018-01-13	Springfield	Xerox 23	32.4
2018-01-14	Aurora	Imation 8GB Mini Tr...	169.064
2018-01-14	Aurora	Mobile Personal File C...	168.624
2018-01-14	Jacksonville	Sannysis Cute Owl D...	15.84
2018-01-14	Jacksonville	Tyvek Top-Opening ...	43.488
2018-01-14	Johnson City	#10- 4 1/8" x 9 1/2"...	91.68
2018-01-14	Johnson City	Avery Durable Slant ...	12.54
2018-01-14	Johnson City	Ibico Presentation In...	29.85
2018-01-14	Johnson City	iHome FM Clock Radi...	559.92
2018-01-14	Johnson City	Xerox 1931	181.44
2018-01-14	Los Angeles	Xerox 19	154.9
2018-01-15	Austin	Acco D-Ring Binder ...	4.276
2018-01-15	Austin	Logitech 910-00297...	47.984
2018-01-15	Austin	Maxell 74 Minute CD...	62.592
2018-01-15	Austin	Wilson Jones Ledger...	32.784
2018-01-15	Los Angeles	Newell 347	21.4

Hình 42. OLAP: Thống kê doanh thu theo sản phẩm

- Thống kê doanh thu theo khách hàng tại từng vùng của từng ngày.
  - Vào ngày <Date>, vùng <Region Name>, doanh thu thu được từ khách hàng <Customer Name> là <Sales of Customer> USD

The screenshot shows a Microsoft Analysis Services (MAS) cube interface. On the left, a tree view displays dimensions: CUSTOMER SK, DATE DIM 1, LOCATION DIM 1, and PRODUCT DIM 1. The DATE DIM 1 node is expanded, showing Date, Month, Quarter, and Hierarchy. The LOCATION DIM 1 node is also expanded, showing CITY, LOCATION SK, REGION, REGION NAME, STATE, and Hierarchy. Below these are Calculated Members. On the right, a data grid titled '<Select dimension>' shows sales data for January 2018. The columns are Date, REGION NAME, CUSTOMER NAME, and SALES OF CUSTOMER. The data includes rows for various dates, regions, customers, and their corresponding sales amounts.

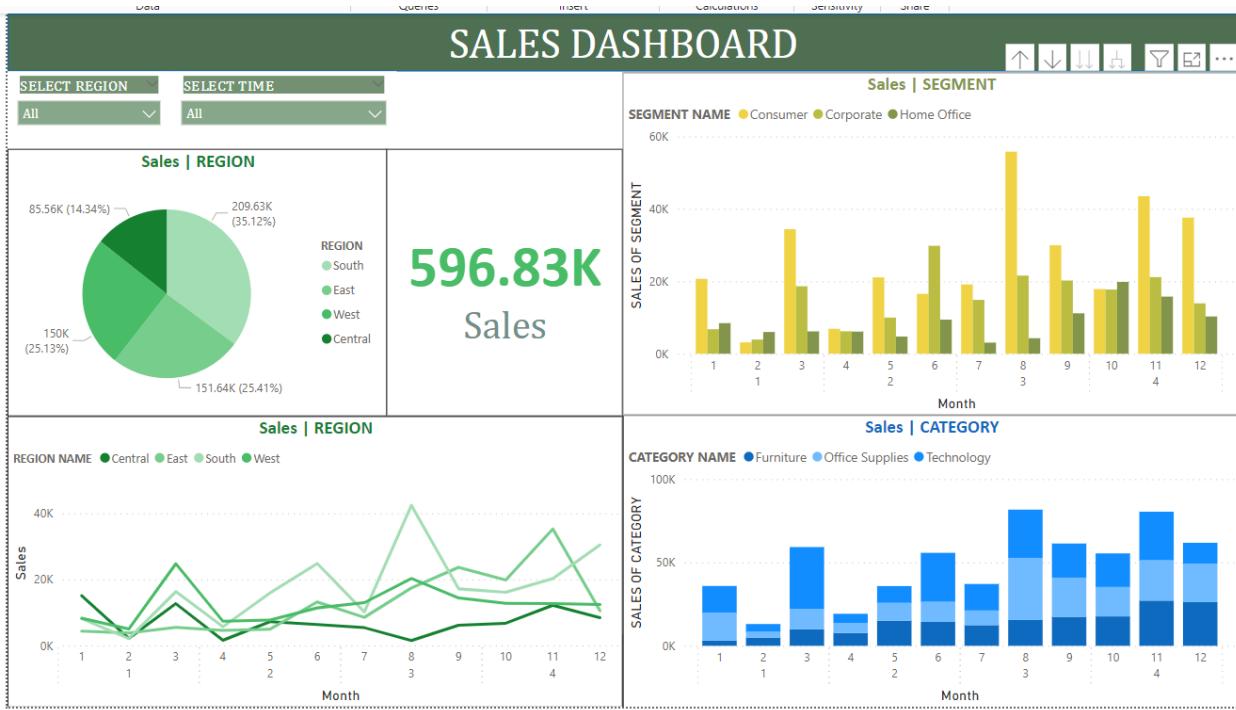
Date	REGION NAME	CUSTOMER NAME	SALES OF CUSTOMER
2018-01-13	Central	Cathy Prescott	4619.33
2018-01-14	South	Brosina Hoffman	875.43
2018-01-14	South	Rob Beeghly	59.328
2018-01-14	West	Sean O'Donnell	154.9
2018-01-14	West	Tracy Poddar	337.688
2018-01-15	Central	Tim Brockman	147.636
2018-01-15	East	Robert Waldorf	1958.544
2018-01-15	East	Sanjit Chand	12.7
2018-01-15	West	Tonja Turnell	21.4
2018-01-16	Central	Andy Reiter	5802.7
2018-01-16	West	Jack O'Briant	427.592
2018-01-19	Central	Dan Lawera	6
2018-01-19	East	Neola Schneider	1370.535
2018-01-20	Central	Maris LaWare	11.52
2018-01-20	East	Cynthia Voltz	207.846
2018-01-20	East	Jeremy Farry	3.52
2018-01-20	West	Tamara Manning	384.176
2018-01-20	West	Thea Hendricks	160.776
2018-01-21	Central	Bobby Elias	268.576
2018-01-21	Central		1020.000

Hình 43. OLAP: Thống kê doanh thu theo khách hàng

## V. DASHBOARD

Ứng dụng nhóm dùng để thực hiện trực quan hóa các kết quả phân tích là **Power BI**.

# 1. Overview of Sales Dashboard



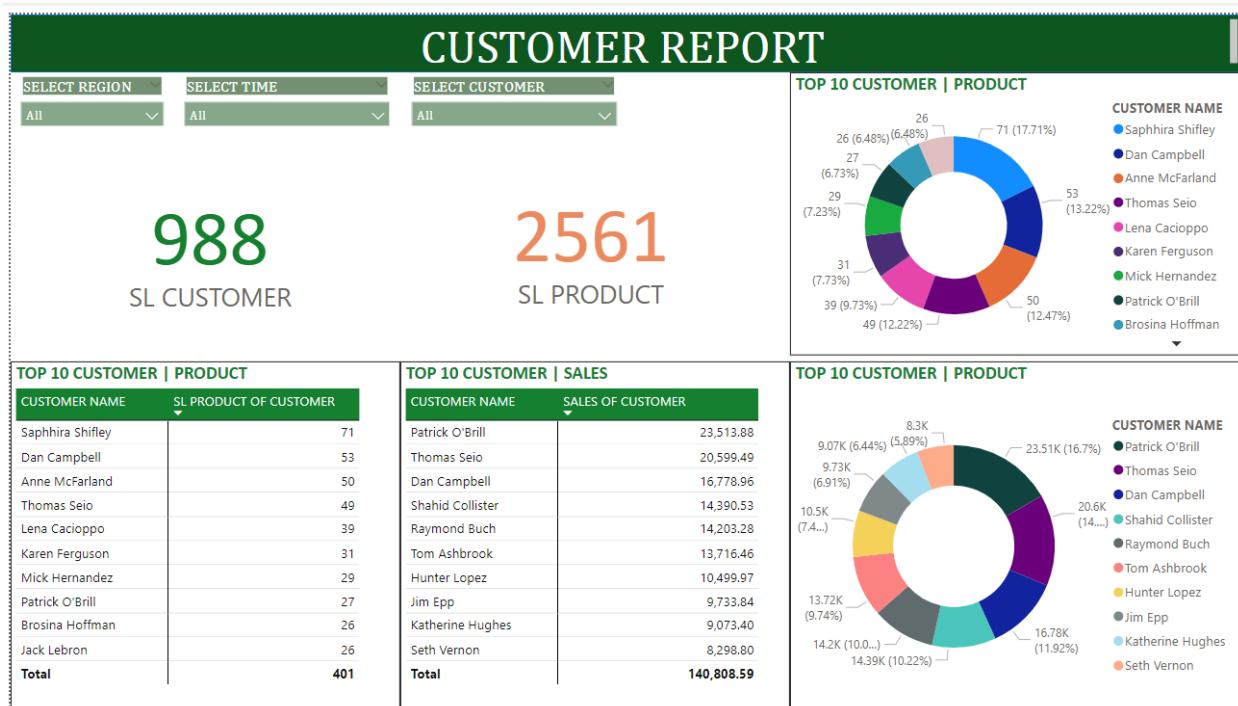
Hình 44. Overview of Sales Dashboard

- Kết quả phân tích: Biểu đồ cho thấy kết quả thống kê tổng quan tình hình bán hàng cơ bản trên 4 vùng, có thể lọc theo khu vực (vùng, bang, thành phố) và thời gian (tháng, quý) để xem chi tiết.
  - Tổng quan:
    - Doanh thu trong năm đạt \$596,83K. Trong đó, vùng chiếm tỷ lệ đóng góp cao nhất là South (209,6K – 35,12%), tiếp đến là East (151,64K – 25,41%), thứ ba là West (150K – 25,13%), thấp nhất là Central (85,56K – 14,34%)
    - Doanh thu có biến động khá nhiều qua các tháng tại tất cả các vùng, đa số có xu hướng giảm vào tháng 2, 4. Tháng 8 có doanh thu cao nhất tại vùng South.
    - Nhìn tổng quan, phân khúc Consumer có sức mua mạnh nhất, tiếp đến là Corporate, cuối cùng là Home Office nhưng cũng chưa có sự khác biệt rõ ràng của phân khúc Corporate và Home Office.
    - Loại hàng về Office Supplies và Technology được khách hàng ưa chuộng với sức mua khá mạnh.

- Có thể thấy nhu cầu mua sắm tăng lên nhiều tại 2 quý cuối năm, nhiều nhất là quý 4.

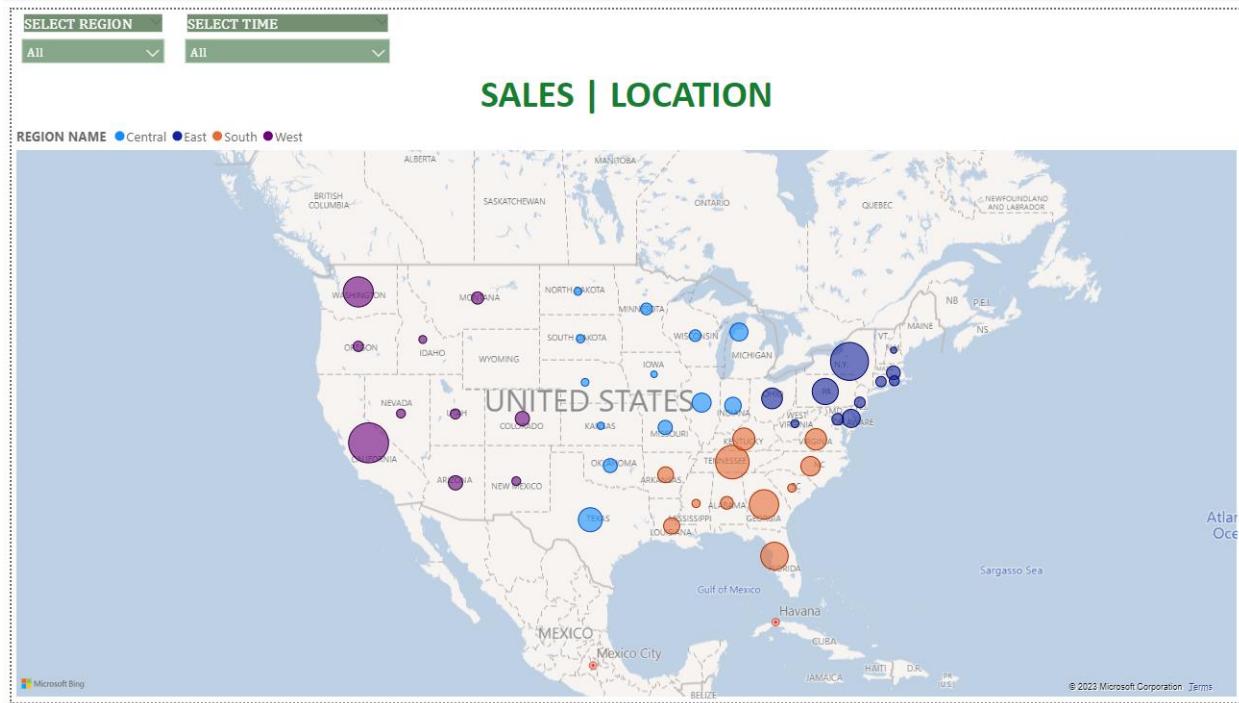
## 2. Sales by Customers

- Biểu đồ cho phép xem chi tiết số lượng khách hàng, số lượng sản phẩm, chi tiết top những khách hàng theo doanh thu và số lượng sản phẩm đã mua.
- Theo số lượng sản phẩm đã mua: khách hàng Saphhira Shifley mua nhiều sản phẩm nhất (71 sản phẩm)
- Theo doanh thu thu được từ khách hàng: Patrick O'Brill đã chi \$23.513,88 cho việc mua hàng trong năm 2018



Hình 45. Sales by Customers Dashboard

### 3. Sales Map



Hình 46. Sales Map

- Kết quả phân tích: Doanh thu chủ yếu tập trung tại bờ Nam và Đông nước Mỹ, một vài khu vực trọng điểm tại bờ Tây như California và Washington, tại khu vực trung tâm tình hình mua sắm khá thưa thớt.  
⇒ **Kết luận:** Nhìn chung, trong năm 2018, tổng doanh thu từ các vùng có xu hướng tăng dần đến cuối năm, tuy nhiên, sự phân bố doanh thu chưa đồng đều, khu vực trung tâm và bờ Tây còn khá ít, bờ Nam và Đông mang đến lượng doanh thu khá cao và đồng đều giữa các bang. Do vậy, cần tập trung phát triển các sản phẩm phục vụ nhu cầu dành cho các khách hàng thuộc khu vực trung tâm và bờ Tây, giữ ổn định và phát triển nguồn doanh thu tại bờ Nam và Đông.

## VI. DATA MINING

- Công cụ nhóm lựa chọn để thực hiện data mining là **RapidMiner**.
- Nhóm sử dụng model Decision Tree để dự đoán người dùng sẽ mua sản phẩm dựa trên các thuộc tính: Segment, Age, Gender, Payment Method

Result History      ExampleSet (ILocal Repository\data\sample-mining)      Tree (Decision Tree)

Open in: Turbo Prep   Auto Model   Interactive Analysis      Filter (497 / 497 examples): all

Data      Statistics      Visualizations      Annotations

Row No.	Customer ID	Customer N...	Segment	Age	Gender	Paymet Met...	Buy "Xerox ...
1	AA-10480	Andrew Allen	Consumer	80	female	credit card	yes
2	SF-20085	Sandra Flanagan	Consumer	33	female	credit card	no
3	MA-17560	Matt Abelman	Home Office	54	female	credit card	yes
4	ES-14080	Erin Smith	Corporate	53	male	credit card	no
5	TB-21520	Tracy Blumstein	Consumer	63	male	credit card	no
6	CS-12400	Christopher Steward	Home Office	91	male	credit card	no
7	PO-18865	Patrick O'Donnell	Consumer	22	male	credit card	no
8	PG-18895	Paul Gonzalez	Consumer	25	male	credit card	yes
9	KD-16345	Katherine Duckett	Consumer	46	male	cheque	no
10	JM-15250	Janet Martin	Consumer	28	male	credit card	yes
11	CV-12805	Cynthia Voltz	Corporate	20	male	credit card	yes
12	SH-19975	Sally Hughsby	Corporate	57	female	cash	no
13	SG-20080	Sandra Glassco	Consumer	24	male	cash	no
14	SJ-20500	Shirley Jackson	Consumer	55	female	cash	no
15	MB-17305	Maria Bertelson	Consumer	42	female	credit card	yes
16	LC-17140	Logan Currie	Consumer	22	male	credit card	yes
17	CB-12535	Claudia Bergman	Corporate	58	female	credit card	yes
18	KH-16690	Kristen Hastings	Corporate	35	male	credit card	no

ExampleSet (497 examples, 0 special attributes, 7 regular attributes)

Hình 47. Dữ liệu mẫu Data Mining

Result History      ExampleSet (ILocal Repository\data\sample-mining)      Tree (Decision Tree)

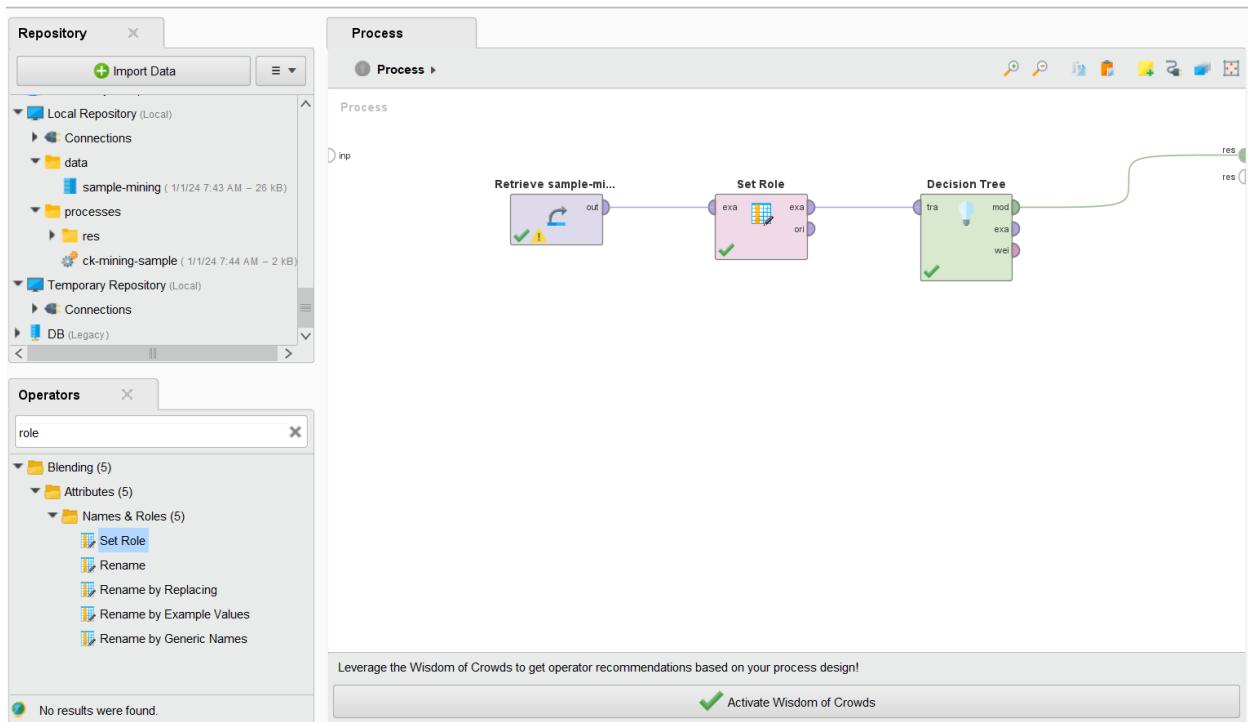
Filter (7 / 7 attributes): Search for Attributes

Data      Statistics      Visualizations      Annotations

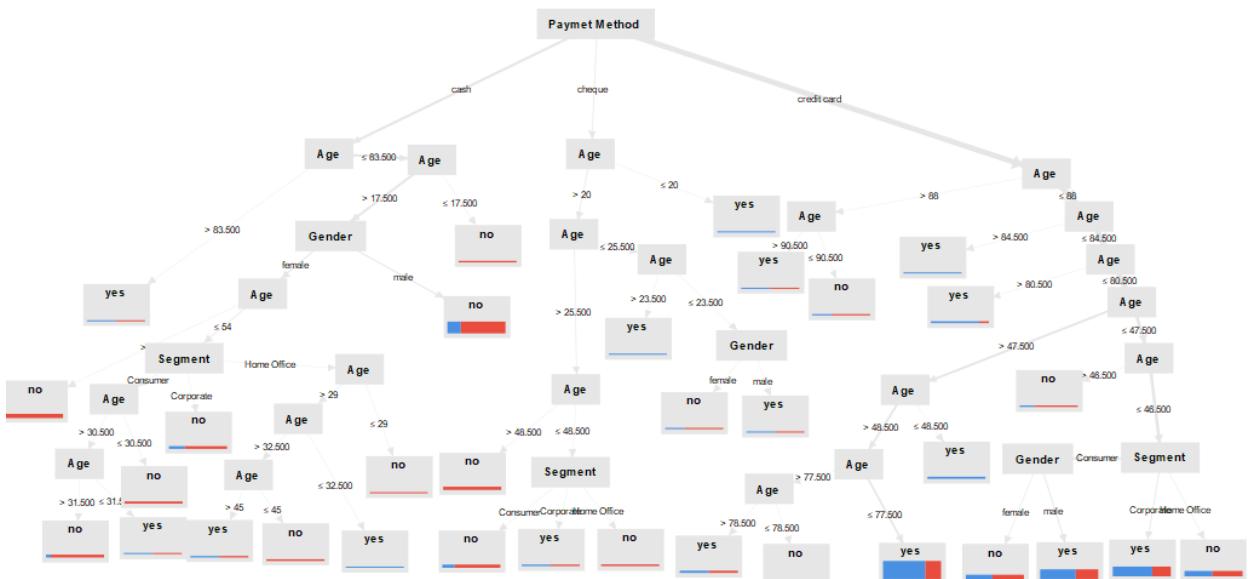
Name	Type	Missing	Statistics	Values
Customer ID	Nominal	0	Least ZC-21910 (1)      Most AA-10315 (1)      Values AA-10315 (1), AA-10480 (1), ...[495 more]	
Customer Name	Nominal	0	Least Zuschuss Carroll (1)      Most Aaron Hawkins (1)      Values Aaron Hawkins (1), Adam Bellavance (1), ...[495 more]	
Segment	Nominal	0	Least Home Office (92)      Most Consumer (262)      Values Consumer (262), Corporate (143), ...[1 more]	
Age	Integer	0	Min 17      Max 91      Average 44.755	
Gender	Nominal	0	Least Female (2)      Most male (284)      Values male (284), female (211), ...[1 more]	
Paymet Method	Nominal	0	Least cheque (46)      Most credit card (314)      Values credit card (314), cash (137), ...[1 more]	
Buy "Xerox 1967?"	Nominal	0	Least yes (236)      Most no (261)      Values no (261), yes (236)	

Showing attributes 1 - 7      Examples: 497 Special Attributes: 0 Regular Attributes: 7

Hình 48. Thông tin thống kê Data Mining



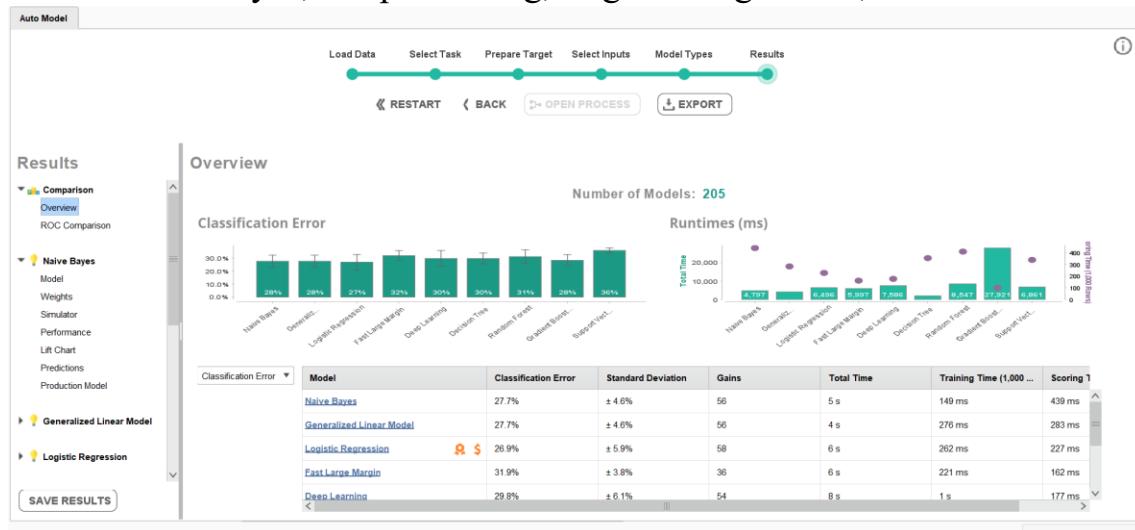
Hình 49. Thiết kế Process Mining



Hình 50. Kết quả Mining

⇒ **Kết luận:** Dựa vào mô hình đã áp dụng để phân tích kết quả mua hàng dựa trên các thuộc tính, người dùng có thể dự đoán kết quả mua hàng của các khách hàng trong tương lai dựa trên các nhánh của Decision Tree. Bên

cạnh đó, RapidMining cũng cung cấp các mô hình dự đoán khác như Navies Bayes, Deep Learning, Logistic Regression,...



Hình 51. Auto Model

## VII. TÀI LIỆU THAM KHẢO

Easy, S. I. [@sqliseeasy]. (2023, September 23). SSAS - Creating your First Cube with MDX. Youtube. [https://www.youtube.com/watch?v=wD\\_SMyyFXpI](https://www.youtube.com/watch?v=wD_SMyyFXpI)

glamb. (2023, June 30). How to use rapidminer for data mining? Google Lambda. <https://googlelambda.com/ai-faq/how-to-use-rapidminer-for-data-mining>

Rainardi, V. (2008). Building a data warehouse: With examples in SQL server. Apress.

Ssis, L. [@learnssis]. (2021, December 20). 37 How to load DimDate table in SQL | date dimension table example. Youtube.

<https://www.youtube.com/watch?v=5OieIJeNXZA>