

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN**

**ĐỒ ÁN MÔN HỌC
KHO DỮ LIỆU VÀ OLAP**



ĐỀ TÀI

PHÂN TÍCH BỘ DỮ LIỆU DOANH SỐ BÁN PHỤ KIỆN XE ĐẠP

Lớp: IS217.N22.HTCL

Giảng viên hướng dẫn:

Ths. Nguyễn Thị Kim Phụng

Sinh viên thực hiện:

Bùi Đức Duy – 20521228

Bùi Duy Thiện – 20520772

TP HCM, tháng 05 năm 2023

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM, đặc biệt là quý thầy cô khoa Hệ thống thông tin của trường đã giúp cho chúng em trang bị các kiến thức cơ bản, các kỹ năng thực tế và tạo điều kiện để chúng em có thể hoàn thành đồ án môn học của mình.

Đặc biệt chúng em xin chân thành cảm ơn cô Nguyễn Thị Kim Phụng đã nhiệt tình hướng dẫn hướng dẫn, quan tâm truyền đạt những kiến thức và kinh nghiệm, trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp ý kiến quý báu cho chúng em trong suốt thời gian học tập để chúng em có thể hoàn thành tốt môn học này.

Trong thời gian một học kỳ thực hiện đề tài, nhóm tác giả đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới. Từ đó, nhóm tác giả vận dụng tối đa những gì đã thu thập được để hoàn thành một báo cáo đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm tác giả không tránh khỏi những thiếu sót. Chính vì vậy, nhóm tác giả rất mong nhận được những sự góp ý từ phía Thầy/Cô nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Nhóm chúng em xin chân thành cảm ơn!

Nhóm thực hiện

NHẬN XÉT CỦA GIẢNG VIÊN

MỤC LỤC

LỜI CẢM ƠN	1
CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN	7
1. Phát biểu về dữ liệu	7
1.1 Mô tả dataset.....	7
1.2 Thuộc tính kho dữ liệu	7
1.3 Mô tả chi tiết thuộc tính	10
1.4 Kho dữ liệu đã thực hiện tiền xử lý.....	13
2. Xây dựng kho dữ liệu	16
2.1 Lược đồ kho dữ liệu (Star Schema)	16
2.2 Chi tiết các bảng.....	16
CHƯƠNG 2. QUÁ TRÌNH SSIS	19
1. Chuẩn bị công cụ và Data Warehouse.....	19
2. Tạo new database trong Microsoft SQL Server Management Studio	19
3. Quá trình SSIS.....	22
3.1 Khởi tạo project.....	22
3.2 Khởi tạo Sequence Container	23
3.3 Import Data	23
3.4 Tạo bảng DIM_DATE	30
3.5 Tạo bảng DIM_CUS_CATEGORY	34
3.6 Tạo bảng DIM_CUSTOMER	39
3.7 Tạo bảng DIM_PRODUCT	45
3.8 Tạo bảng DIM_TYPE_ORDER	50
3.9 Tạo bảng DIM_STATUS	55
3.10 Tạo bảng FACT	60

3.11 Import dữ liệu vào Database	73
CHƯƠNG 3. QUÁ TRÌNH SSAS.....	75
1. Tạo Project SSAS trong Visual Studio 2019.....	75
2. Connect đến data source	77
3. Tạo Data Source View.....	79
4. Tạo Cube và Dimensions.....	82
5. Thao tác trên các bảng Dim	85
5.1 Bảng DIM_DATE.....	85
5.2 Bảng DIM_CUSTOMER.....	86
5.3 Bảng DIM_PRODUCT.....	87
5.4 Bảng DIM_TYPE_ORDER	88
5.5 Bảng DIM_STATUS	89
6. Deploy và process project.....	90
7. Thực hiện 10 câu truy vấn	91
7.1 Thống kê tổng số giao dịch của tiểu bang NWS	91
7.2 Thống kê doanh số của từng thương hiệu qua các quý	95
7.3 Thống kê tổng đơn đặt hàng online của từng dòng sản phẩm trong tháng 4 năm 2017.....	98
7.4 Thống kê tổng hóa đơn của từng nhóm khách hàng theo tháng trong năm	101
7.5 Thống kê top 5 nhóm khách hàng thuộc bang VIC mua hàng nhiều nhất	104
7.6 Thống kê những khách hàng mua hàng hơn 9 lần và tổng số tiền mua hơn 10000\$	110
7.7 Thống kê danh sách và số lượng đơn hàng khách hàng nữ đã mua và đã thanh toán tại cửa hàng ngày 8/3/2017 không thuộc bang VIC	114
7.8 Thống kê top 5 sản phẩm bán chạy nhất trong từng bang	118

7.9 Thống kê những khách hàng tiềm năng thuộc nhóm khách hàng từ 20 đến dưới 40 tuổi mua sản phẩm trong quý 1	123
7.10 Thống kê top 3 tháng có đơn hàng bị hủy nhiều nhất.....	127
CHƯƠNG 4. REPORT BẰNG CÔNG CỤ POWER BI	131
1. Đỗ dữ liệu vào project Power BI.....	131
2. Thực hiện Report.....	133
2.1 Thống kê tổng số giao dịch ở tiểu bang NWS	133
2.2 Thống kê doanh số của từng thương hiệu qua các quý	135
2.3 Thống kê tổng đơn đặt hàng online của từng dòng sản phẩm trong tháng 4 năm 2017.....	137
2.4 Thống kê tổng hóa đơn của từng nhóm khách hàng theo tháng trong năm	140
2.5 Thống kê top 5 nhóm khách hàng thuộc bang VIC mua hàng nhiều nhất	143
2.6 Thống kê những khách hàng mua hàng hơn 9 lần và tổng số tiền mua hơn 10000\$	145
2.7 Thống kê danh sách và số lượng đơn hàng khách hàng nữ đã mua và đã thanh toán tại cửa hàng ngày 8/3/2017 không thuộc bang VIC	148
2.8 Thống kê top 5 sản phẩm bán chạy nhất trong từng bang	150
2.9 Thống kê những khách hàng tiềm năng thuộc nhóm khách hàng từ 20 đến dưới 40 tuổi mua sản phẩm trong quý 1	153
2.10 Thống kê top 3 tháng có đơn hàng bị hủy nhiều nhất.....	155
CHƯƠNG 5. DATA MINING.....	157
1.Phần mềm và ngôn ngữ sử dụng	157
2.Các bước xử lý chung.....	157
3. Chuẩn bị dữ liệu	161
4. Train model	161

4.1 Random Forest Classifier model	161
4.2 Support Vector Machine model.....	166
5. Kết luận.....	170

CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN

1. Phát biểu về dữ liệu

1.1 Mô tả dataset

- Dữ liệu về doanh số bán phụ kiện xe đạp tại một số cửa hàng từ nước Úc được tổng hợp trong năm 2017.
- Thông qua dữ liệu, người dùng có thể biết được thông tin của khách hàng, lợi nhuận, số hóa đơn đã đặt theo các hình thức khác nhau, số đơn hàng bị hủy, tỉ lệ khách hàng trờ lại mua hàng, phân loại khách hàng,...
- Dữ liệu bao gồm 30 cột, 19445 dòng.
- Link dữ liệu gốc: [Bike Accessories Sales | Kaggle](#)

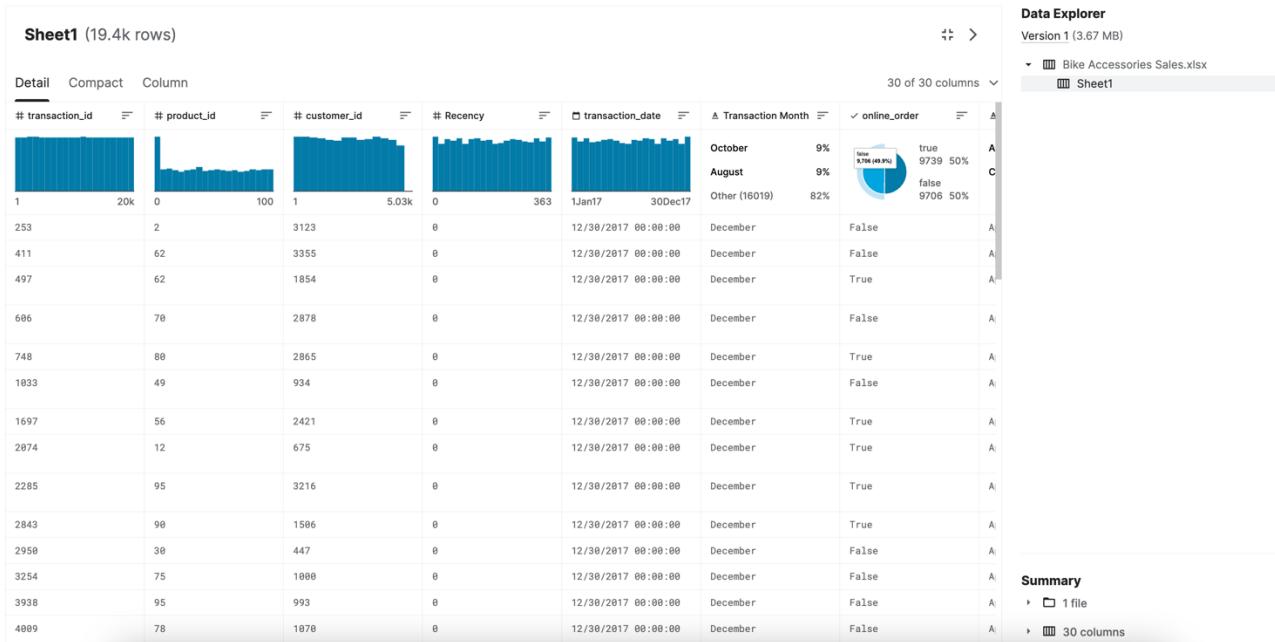
1.2 Thuộc tính kho dữ liệu

STT	Thuộc tính	Kiểu dữ liệu	Ý nghĩa	Giá trị
1	transaction_id	int	Mã giao dịch	Giá trị ngẫu nhiên
2	product_id	int	Mã sản phẩm	Giá trị ngẫu nhiên
3	customer_id	int	Mã khách hàng	Giá trị ngẫu nhiên
4	transaction_date	date	Ngày giao dịch được thực hiện.	
5	transaction_mon th	nvarchar(40)	Tháng giao dịch được thực hiện	Trong khoảng [1, 12]
6	online_order	Boolean	Cho biết liệu đơn hang này có được đặt trực tuyến	[True, False]
7	order_status	nvarchar(40)	Trạng thái của đơn hang	['Approved', 'Can celled']

8	brand	nvarchar(255)	Thương hiệu của sản phẩm.	
9	product_line	nvarchar(255)	Dòng sản phẩm.	
10	product_class	nvarchar(255)	Phân loại hạng sản phẩm.	
11	product_size	nvarchar(255)	Kích cỡ sản phẩm.	
12	list_price	float	Giá niêm yết của sản phẩm	
13	standard_cost	float	Chi phí sản xuất sản phẩm.	
14	profit	float	Lợi nhuận thu được từ mỗi sản phẩm.	
15	gender	nvarchar(10)	Giới tính của khách hàng.	Giới hạn [Male, Female]
16	past_3_years_bike_related_purchases	int	Số lượng sản phẩm liên quan đến xe đạp được khách hàng mua trong 3 năm trước đó.	
17	DOB	date	Ngày sinh của khách hàng.	
18	job_industry_category	nvarchar(255)	Ngành nghề của khách hàng.	
19	tenure	int	Thời gian khách hàng đã trở thành khách hàng của cửa hàng.	Trong khoảng [0,100]
20	customer_title	nvarchar(255)	Nhóm khách hàng	

21	state	nvarchar(20)	Tiêu bang hoặc khu vực của khách hàng.	
22	recency	int	Chỉ thời gian trôi qua kể từ lần cuối cùng khách hàng mua sản phẩm hoặc sử dụng dịch vụ	
23	Recency Base Date	date		
24	product_first_sold_date	date	Chỉ thời điểm sản phẩm đó được bán lần đầu tiên.	
25	age	int	Chỉ độ tuổi của khách hàng	
26	age_group	int	Thuộc tính phân loại khách hàng theo từng khoảng độ tuổi	
27	job_title	nvarchar(255)	Nghề nghiệp của khách hàng	
28	wealth_segment	nvarchar(255)	Là thuộc tính phân loại khách hàng theo mức thu nhập, tài sản và chi tiêu	Giới hạn [High Net Worth, Affluent, Mass Market]
29	deceased_indicator	boolean	Chỉ định xem khách hàng có còn sống hay không	True: Còn sống False: Đã chết
30	owns_car	boolean	Chỉ định liệu khách hàng có sở hữu	True: Có False: Không

			chiếc xe hơi nào hay không	
--	--	--	-------------------------------	--



Hình 1.1 Kho dữ liệu khi chưa xử lý

1.3 Mô tả chi tiết thuộc tính

- Dictionary thuộc tính **online_order**
- Thuộc tính **online_order** thể hiện trạng thái đặt hàng trực tuyến

online_order		
STT	Tên	Ý Nghĩa
1	True	Khách hàng đã đặt hàng online
2	False	Khách hàng không đặt hàng online

- Dictionary thuộc tính **order_status**
- Thuộc tính **order_status** thể hiện tình trạng đơn hàng

order_status		
STT	Tên	Ý Nghĩa
1	Approved	Đơn hàng đã được chấp thuận
2	Cancelled	Đơn hàng đã bị hủy

- Dictionary thuộc tính **product_line**
- Thuộc tính **product_line** thể hiện dòng sản phẩm

product_line		
STT	Tên	Ý Nghĩa
1	Moutain	Dòng sản phẩm leo núi
2	Road	Dòng sản phẩm đường phố
3	Standard	Dòng sản phẩm tiêu chuẩn, phổ biến
4	Touring	Dòng sản phẩm đi đường trường hoặc các chuyến đi dài

- Dictionary thuộc tính **product_class**
- Thuộc tính **product_class** thể hiện phân loại hạng sản phẩm

product_class		
STT	Tên	Ý Nghĩa
1	High	Sản phẩm hạng cao cấp
2	Medium	Sản phẩm hạng trung bình
3	Low	Sản phẩm hạng thấp

- Dictionary thuộc tính **product_size**
- Thuộc tính **product_size** thể hiện kích cỡ sản phẩm

product_size		
STT	Tên	Ý Nghĩa
1	Large	Sản phẩm kích cỡ lớn
2	Medium	Sản phẩm kích cỡ trung bình
3	Small	Sản phẩm kích cỡ nhỏ

- Dictionary thuộc tính **customer_title**
- Thuộc tính **customer_title** thể hiện khảo sát nhóm khách hàng sử dụng sản phẩm

customer_title		
STT	Tên	Ý Nghĩa
1	High risk customer	Khách hàng có rủi ro cao là khách hàng có tiềm năng để gây thiệt hại cho doanh nghiệp.
2	Late Bloomer	Khách hàng trễ là khách hàng không quan tâm đến sản phẩm hoặc dịch vụ ban đầu, nhưng sau đó thay đổi ý kiến và quyết định mua sau một thời gian.
3	Lost customer	Khách hàng đã mất là khách hàng đã ngừng mua hàng hoặc sử dụng dịch vụ.
4	Losing customer	Khách hàng đang mất là khách hàng đang bị đe dọa bởi các yếu tố như giá cả cạnh tranh, chất lượng sản phẩm, dịch vụ khách hàng kém, sự cạnh tranh từ các đối thủ cạnh tranh và nhiều yếu tố khác.
5	Almost lost customer	Khách hàng suýt bị mất là khách hàng đã có xu hướng chuyển sang dùng sản phẩm hoặc dịch vụ từ đối thủ cạnh tranh, nhưng sau đó đã quay trở lại và tiếp tục mua hàng.
6	Potential customer	Khách hàng tiềm năng là khách hàng có khả năng trở thành khách hàng thực sự trong tương lai.
7	Recent customer	Khách hàng mới là khách hàng vừa mới bắt đầu sử dụng sản phẩm hoặc dịch vụ và chưa có sự trung thành hoặc quan tâm lâu dài.
8	Becoming loyal	Khách hàng trở thành khách hàng trung thành là khách hàng đã mua hàng một số lần và có xu hướng tiếp tục sử dụng sản phẩm hoặc dịch vụ trong tương lai.
9	Platinum customer	Khách hàng bạch kim là khách hàng trung thành, mua hàng thường xuyên và chi tiêu nhiều tiền trong các giao dịch.
10	Very loyal	Khách hàng rất trung thành là khách hàng đã mua hàng trong một thời gian dài và có xu hướng tiếp tục mua hàng trong tương lai. Họ thường có mức độ trung thành cao và sẽ giới thiệu sản phẩm hoặc dịch vụ của bạn cho người khác.

11	Evasive Customer	Khách hàng hoặc người tiêu dùng có xu hướng tránh, trì hoãn hoặc trả lời một cách không rõ ràng các câu hỏi hoặc yêu cầu thông tin.
----	------------------	---

- Dictionary thuộc tính **age_group**
- Thuộc tính **age_group** thể hiện khảo sát nhóm tuổi khách hàng sử dụng sản phẩm

age_group		
STT	Tên	Ý Nghĩa
1	U20	Nhóm khách hàng có độ tuổi từ 0 đến nhỏ hơn 20 tuổi
2	U40	Nhóm khách hàng có độ tuổi từ 20 đến nhỏ hơn 40 tuổi
3	U60	Nhóm khách hàng có độ tuổi từ 40 đến nhỏ hơn 60 tuổi
4	U100	Nhóm khách hàng có độ tuổi lớn hơn 60 tuổi

1.4 Kho dữ liệu đã thực hiện tiền xử lý

- Nhóm thực hiện các bước tiền xử lý
- Sau khi tiền xử lý, ta được 12051 dòng và 21 cột để sử dụng cho việc phân tích đề tài

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	transaction_id	product_id	customer_id	transaction_date	transaction_month	online_order	transaction_dt	product_brand	product_line	product_class	product_size	list_price	standard_cont	profit	gender	past_3_years_bike_related_purchases	customer_dsgnage_group	job_industry_c	
2	253	8123	30/12/2017	December	FALSE	Approved	Giant Bicycles	Road	low	small	590.26	525.33	64.93	Female	71	17/08/1976	50	Health	
3	1697	56	2421	30/12/2017	December	TRUE	Approved	OHM Cycles	Standard	medium	183.86	137.90	45.96	Female	54	16/08/1954	70	Health	
4	2074	12	675	30/12/2017	December	TRUE	Approved	WearA2B	Standard	medium	1231.15	161.60	1069.55	Female	37	19/12/1976	50	Financial Servic	
5	2285	95	3216	30/12/2017	December	TRUE	Approved	Giant Bicycles	Standard	large	569.56	528.43	41.13	Male	59	08/05/1999	30	Manufacturing	
6	2843	90	1508	30/12/2017	December	TRUE	Approved	Norco Bicycles	Standard	low	medium	363.01	290.41	72.60	Male	21	21/02/1958	0	Property
7	3254	75	1000	30/12/2017	December	FALSE	Approved	Giant Bicycles	Touring	medium	1873.97	863.95	1010.02	Female	44	07/08/1978	50	Financial Servic	
8	3938	81	993	30/12/2017	December	TRUE	Approved	OHM Cycles	Touring	low	medium	933.00	512.00	139.23	Female	1	18/06/1980	50	Argentina
9	4009	70	1520	30/12/2017	December	FALSE	Approved	Giant Bicycles	Standard	medium	1762.50	120.48	1642.02	Male	5	18/06/1995	30	Health	
10	4669	29	1052	30/12/2017	December	FALSE	Approved	WearA2B	Standard	medium	1065.03	330.09	834.94	Female	31	02/02/1986	40	Financial Servic	
11	4942	72	3293	30/12/2017	December	TRUE	Approved	OHM Cycles	Standard	medium	912.52	141.40	771.12	Female	98	23/01/1986	40	n/a	
12	5042	90	1985	30/12/2017	December	FALSE	Approved	Norco Bicycles	Standard	low	medium	363.01	290.41	72.60	Female	1	02/12/1978	50	n/a
13	5651	25	2713	30/12/2017	December	TRUE	Approved	Giant Bicycles	Road	medium	1538.94	829.65	709.34	Male	55	03/11/1954	70	Retail	
14	7267	57	1972	30/12/2017	December	TRUE	Approved	WearA2B	Touring	medium	1890.39	260.14	1630.25	Male	12	14/07/1989	0	IT	
15	7311	1	1328	30/12/2017	December	FALSE	Approved	Giant Bicycles	Standard	medium	1403.50	954.82	448.68	Male	19	02/09/1981	40	IT	
16	7518	65	3433	30/12/2017	December	FALSE	Approved	WearA2B	Standard	medium	1807.43	778.69	1028.76	Female	10	28/04/1999	30	Property	
17	7547	21	1970	30/12/2017	December	TRUE	Approved	Norco Bicycles	Standard	small	685.00	312.00	372.88	Female	1	18/06/1990	50	Manufacturing	
18	8111	79	1970	30/12/2017	December	TRUE	Approved	Norco Bicycles	Standard	medium	1555.58	618.01	737.57	Male	51	23/08/1982	40	Manufacturing	
19	9158	88	86	30/12/2017	December	TRUE	Approved	Norco Bicycles	Standard	high	1661.92	1479.11	182.81	Male	75	04/03/1989	40	Health	
20	9820	31	1425	30/12/2017	December	TRUE	Approved	WearA2B	Standard	medium	752.64	205.36	547.28	Male	20	06/07/1986	40	Retail	
21	11701	93	1453	30/12/2017	December	TRUE	Approved	OHM Cycles	Standard	high	1458.17	874.90	583.27	Male	54	30/07/1977	50	Health	
22	12453	12	151	30/12/2017	December	FALSE	Approved	WearA2B	Standard	medium	1231.15	161.60	1069.55	Male	36	31/01/1965	60	Manufacturing	
23	12667	89	1594	30/12/2017	December	TRUE	Approved	WearA2B	Touring	medium	1362.99	57.74	1305.25	Male	62	26/11/1981	40	Property	
24	15522	19	2434	30/12/2017	December	FALSE	Approved	OHM Cycles	Road	high	12.01	4.80	8.40	Male	40	17/11/1956	70	n/a	
25	16366	89	2345	30/12/2017	December	FALSE	Approved	WearA2B	Touring	medium	1386.99	57.74	1305.25	Female	17	20/04/1996	30	Retail	
26	17192	26	3459	30/12/2017	December	TRUE	Approved	WearA2B	Standard	medium	1993.53	292.63	1230.41	Male	96	12/01/1964	60	n/a	
27	19346	4	1453	30/12/2017	December	FALSE	Approved	Giant Bicycles	Standard	high	1129.13	677.48	451.65	Male	9	03/08/1973	50	Retail	
28	19907	41	1721	30/12/2017	December	FALSE	Approved	Norco Bicycles	Standard	low	958.74	748.90	209.84	Female	82	22/10/1980	50	Financial Servic	
29	2257	31	2908	29/12/2017	December	TRUE	Approved	Giant Bicycles	Standard	medium	230.91	173.18	57.73	Female	42	21/09/1955	70	Health	
30	4066	81	888	29/12/2017	December	FALSE	Approved	Norco Bicycles	Standard	small	586.45	521.94	64.51	Female	53	15/11/1974	50	Health	
31	6089	89	1390	29/12/2017	December	FALSE	Approved	Giant Bicycles	Standard	large	1812.79	582.48	1230.27	Female	74	25/01/2002	30	Retail	
32	6110	94	884	29/12/2017	December	FALSE	Approved	Giant Bicycles	Standard	medium	1635.50	993.00	641.64	Male	11	30/04/1978	50	Property	
33	64779	29	2865	29/12/2017	December	TRUE	Approved	Norco Bicycles	Standard	medium	547.00	407.54	139.46	Female	1	18/06/1988	50	Argentina	
34	7139	9	1453	29/12/2017	December	FALSE	Approved	Norco Bicycles	Standard	small	1216.14	1082.36	133.78	Male	54	30/07/1977	50	Health	
35	8352	99	3262	29/12/2017	December	TRUE	Approved	OHM Cycles	Standard	medium	1273.34	770.89	456.45	Male	65	14/10/1959	70	Entertainment	
36	10996	61	856	29/12/2017	December	TRUE	Approved	OHM Cycles	Standard	low	71.16	56.93	14.23	Male	63	01/10/1984	40	Financial Servic	
37	11679	32	1991	29/12/2017	December	FALSE	Approved	Giant Bicycles	Standard	medium	642.70	211.37	431.33	Male	35	27/09/1993	30	Health	
38	13208	57	3374	29/12/2017	December	FALSE	Approved	WearA2B	Touring	medium	1890.39	260.14	1630.25	Female	77	17/11/1959	70	Manufacturing	
39	13894	15	224	29/12/2017	December	TRUE	Approved	Norco Bicycles	Standard	low	958.74	748.90	209.84	Male	2	03/12/1967	60	Financial Servic	
40	14491	90	1538	29/12/2017	December	TRUE	Approved	Norco Bicycles	Standard	low	363.01	290.41	72.60	Male	89	12/01/1976	50	Health	
41	15080	66	2869	29/12/2017	December	TRUE	Approved	Giant Bicycles	Standard	low	958.74	522.46	436.28	Male	62	12/01/1976	50	Health	
42	15159	64	1053	29/12/2017	December	TRUE	Approved	Giant Bicycles	Standard	high	1577.36	1281.85	247.51	Female	31	02/02/1986	40	Financial Servic	
43	15877	77	2561	29/12/2017	December	FALSE	Approved	WearA2B	Standard	medium	1769.64	108.76	1660.88	Female	44	06/09/1963	40	Manufacturing	
44	16358	10	703	29/12/2017	December	TRUE	Approved	WearA2B	Touring	medium	1465.68	363.25	1103.43	Female	66	31/10/1966	0	Property	
45	16362	0	1470	29/12/2017	December	TRUE	Approved	WearA2B	Standard	medium	175.89	131.92	43.97	Male	62	04/12/1995	30	Financial Servic	

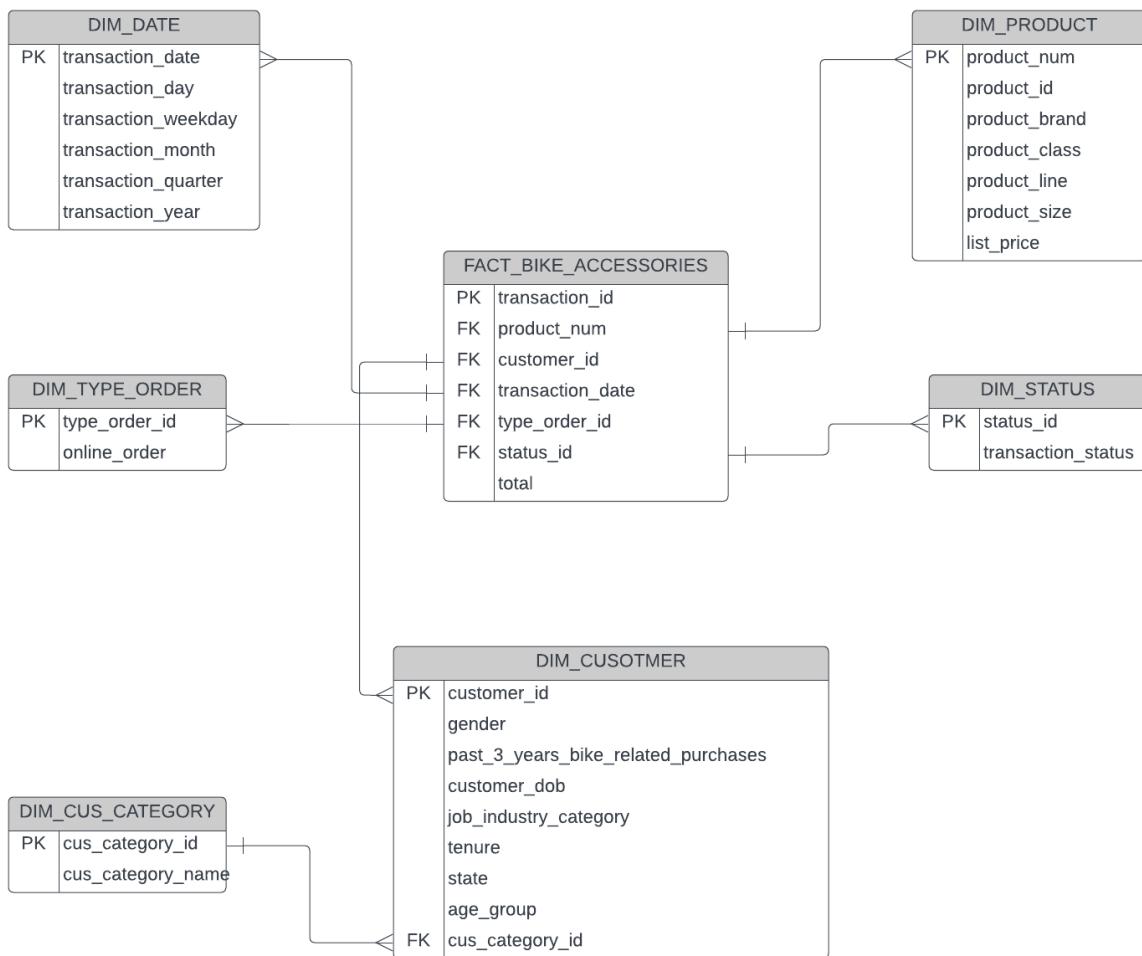
Hình 1.2 Kho dữ liệu sau khi xử lý

STT	Thuộc tính	Kiểu dữ liệu	Ý nghĩa	Giá trị
1	transaction_id	int	Mã giao dịch	Giá trị ngẫu nhiên
2	product_id	int	Mã sản phẩm	Giá trị ngẫu nhiên
3	customer_id	int	Mã khách hàng	Giá trị ngẫu nhiên
4	transaction_date	date	Ngày giao dịch được thực hiện.	
5	transaction_month	nvarchar(40)	Tháng giao dịch được thực hiện	Trong khoảng [1, 12]
6	online_order	Boolean	Cho biết liệu đơn hàng này có được đặt trực tuyến	[True, False]
7	transaction_status	nvarchar(40)	Trạng thái của đơn hàng	['Approved', 'Cancelled']
8	product_brand	nvarchar(255)	Thương hiệu của sản phẩm	
9	product_line	nvarchar(255)	Dòng sản phẩm	
10	product_class	nvarchar(255)	Phân loại hạng sản phẩm	
11	product_size	nvarchar(255)	Kích cỡ sản phẩm.	
12	list_price	Float	Giá niêm yết của sản phẩm	
13	standard_cost	float	Chi phí sản xuất sản phẩm.	
14	profit	float	Lợi nhuận thu được từ mỗi sản phẩm.	
15	gender	nvarchar(10)	Giới tính của khách hàng.	Giới hạn [Male, Female]

16	past_3_years_bike_related_purchases	int	Số lượng sản phẩm liên quan đến xe đạp được khách hàng mua trong 3 năm trước đó	
17	customer_dob	date	Ngày sinh của khách hàng.	
18	age_group	nvarchar(255)	Thuộc tính phân loại khách hàng theo từng khoảng độ tuổi	
19	job_industry_category	nvarchar(255)	Ngành nghề của khách hàng	
20	tenure	int	Thời gian khách hàng đã trở thành khách hàng của cửa hàng	Trong khoảng [0,100]
21	customer_title	nvarchar(255)	Nhóm khách hàng	
22	state	nvarchar(20)	Tiểu bang hoặc khu vực của khách hàng	

2. Xây dựng kho dữ liệu

2.1 Lược đồ kho dữ liệu (Star Schema)



2.2 Chi tiết các bảng

- Bảng **FACT_BIKE_ACCESSORIES**

Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
transaction_id	int	Khóa chính
product_num	int	Khóa ngoại
customer_id	int	Khóa ngoại
transaction_date	date	Khóa ngoại
type_order_id	int	Khóa ngoại
status_id	int	Khóa ngoại

total	float	
-------	-------	--

- Bảng DIM_DATE

Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
transaction_date	date	Khóa chính
transaciont_day	int	
transacion_weekday	int	
transaction_month	int	
transaction_quarter	int	
transaction_year	int	

- DIM_CUSTOMER

Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
customer_id	int	Khóa chính
gender	nvarchar(255)	
past_3_years_bike_related_purchases	int	
customer_dob	date	
job_industry_category	nvarchar(255)	
tenure	int	
state	nvarchar(255)	
age_group	nvarchar(255)	
cus_category_id	int	

- DIM_PRODUCT

Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
product_num	int	Khóa chính
product_id	int	
product_brand	nvarchar(255)	
product_line	nvarchar(255)	
product_class	nvarchar(255)	
product_size	nvarchar(255)	
list_price	float	

- DIM_TYPE_ORDER

Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
type_order_id	int	Khóa chính
online_order	nvarchar(255)	

- DIM_STATUS

Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
status_id	int	Khóa chính
transaction_status	nvarchar(255)	

CHƯƠNG 2. QUÁ TRÌNH SSIS

1. Chuẩn bị công cụ và Data Warehouse

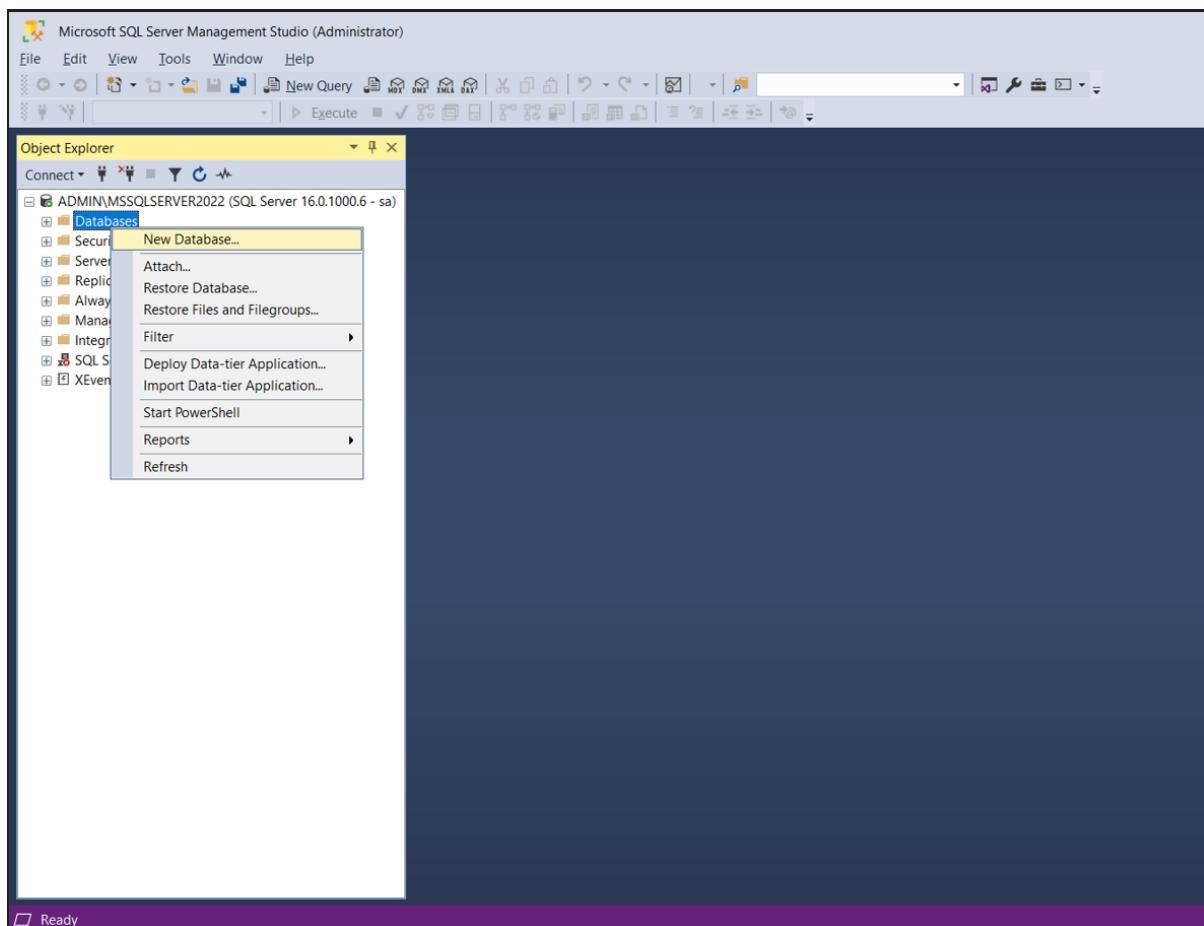
Tải Microsoft Visual Studio 2019.

Tải Microsoft SQL Server Management Studio.

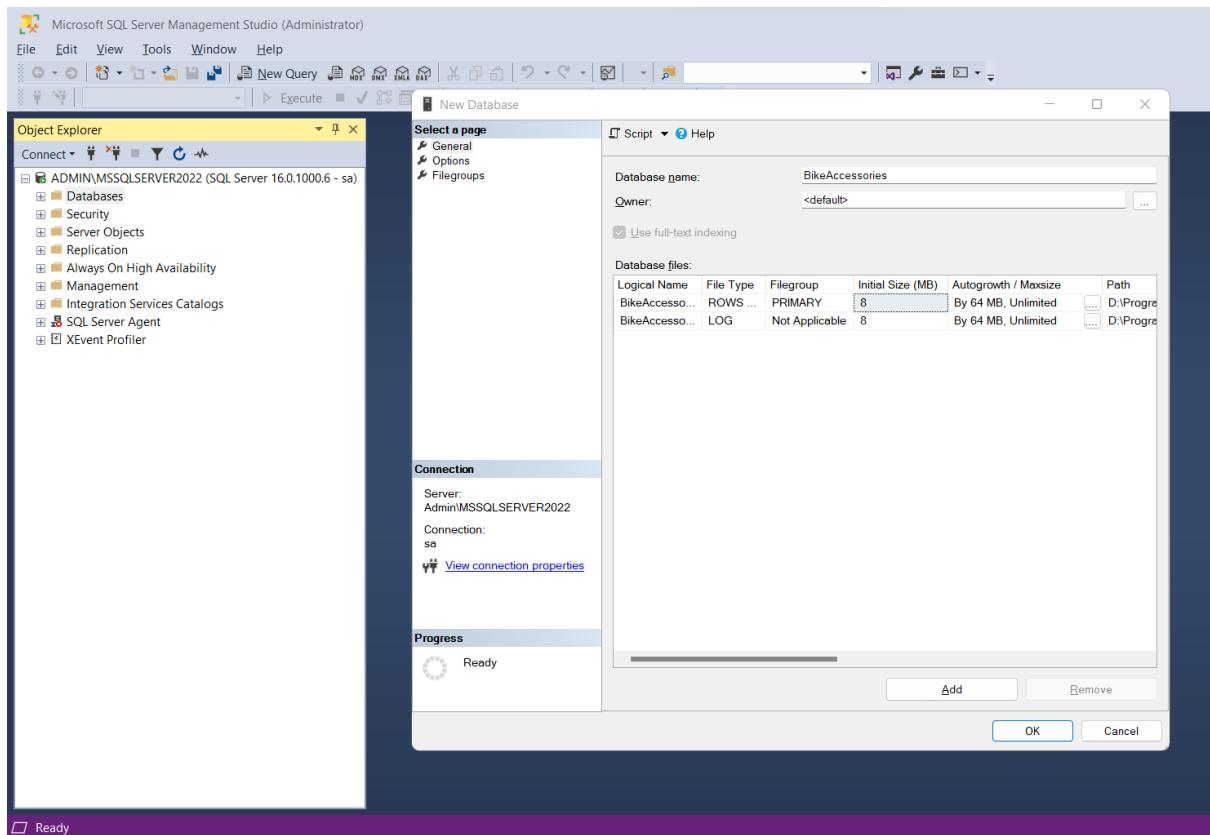
Tải công cụ SQL Server Data Tools cho phiên bản Visual Studio 2019.

2. Tạo new database trong Microsoft SQL Server Management Studio

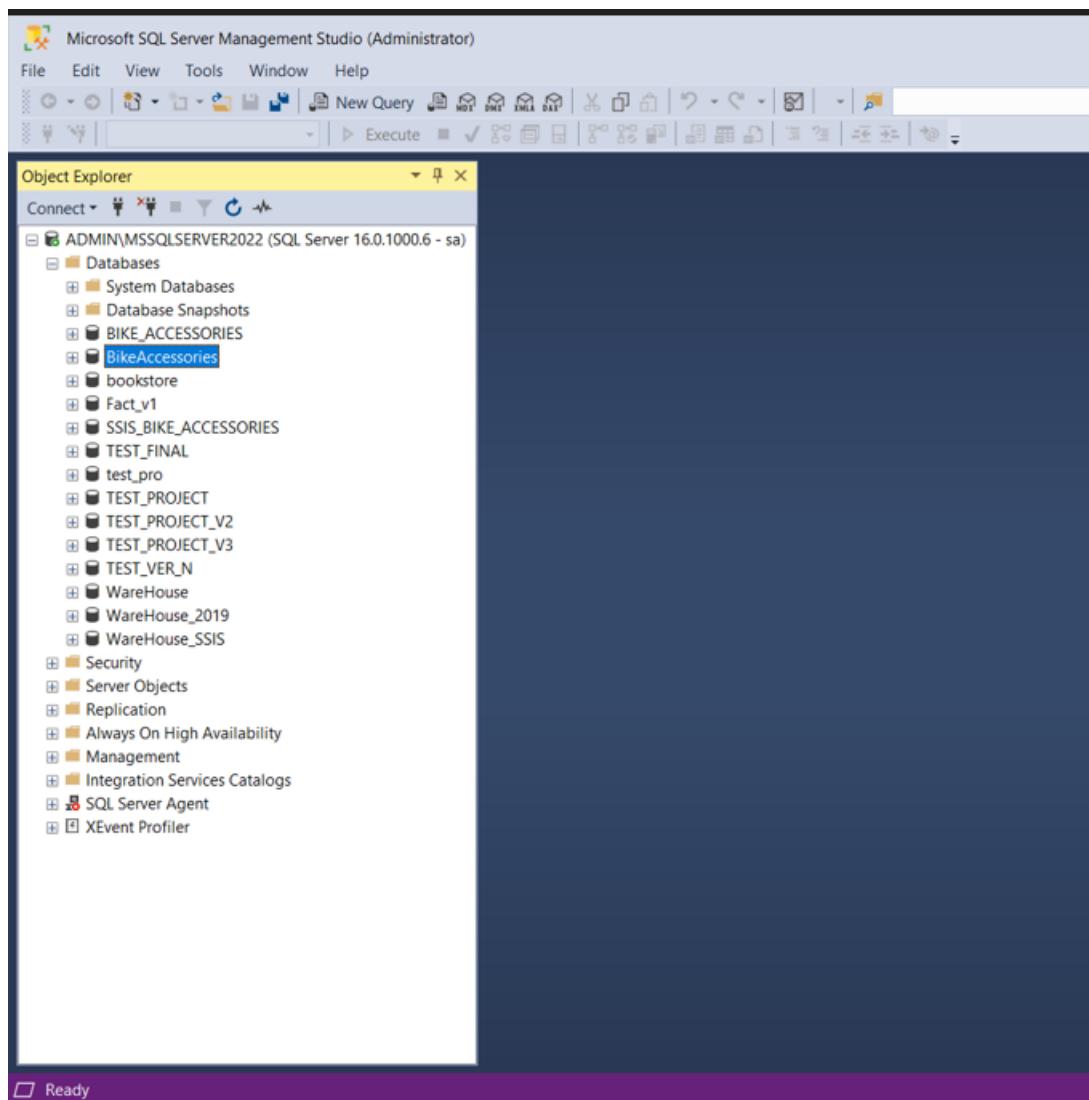
Bước 1: Nhấn chuột phải vào Database, sau đó chọn New Database để thực hiện tạo.



Bước 2: Đặt tên cho cơ sở dữ liệu mới tạo, sau đó nhấn OK.



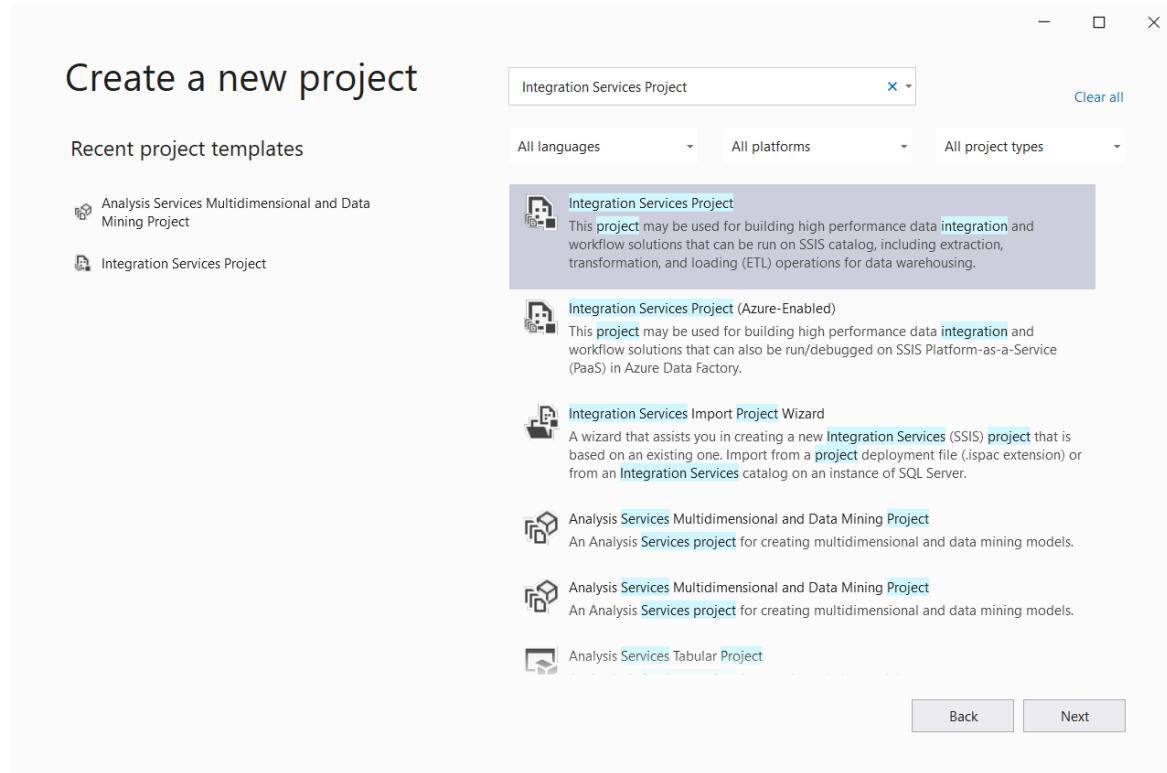
Kết quả:



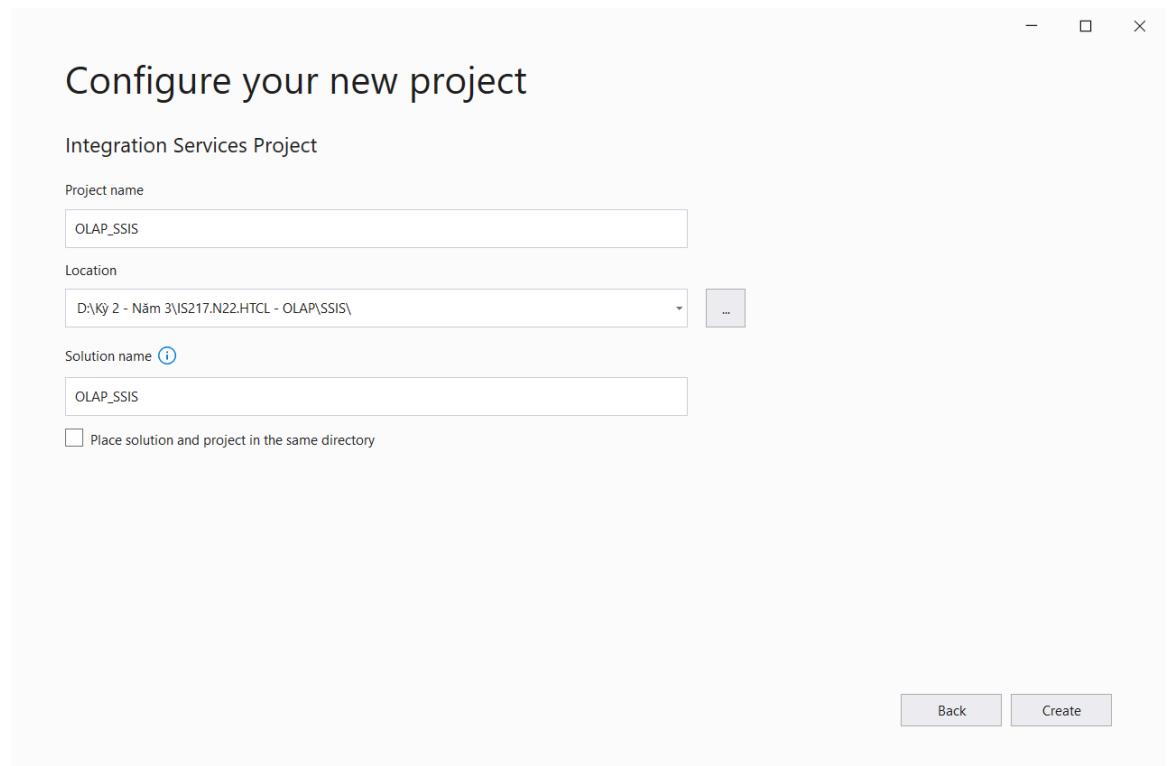
3. Quá trình SSIS

3.1 Khởi tạo project

Bước 1: Tìm từ khóa integration, sau đó chọn **Integration Services Project**, chọn tiếp **Next**.



Bước 2: Điền tên project SSIS, sau đó nhấn **Create**.

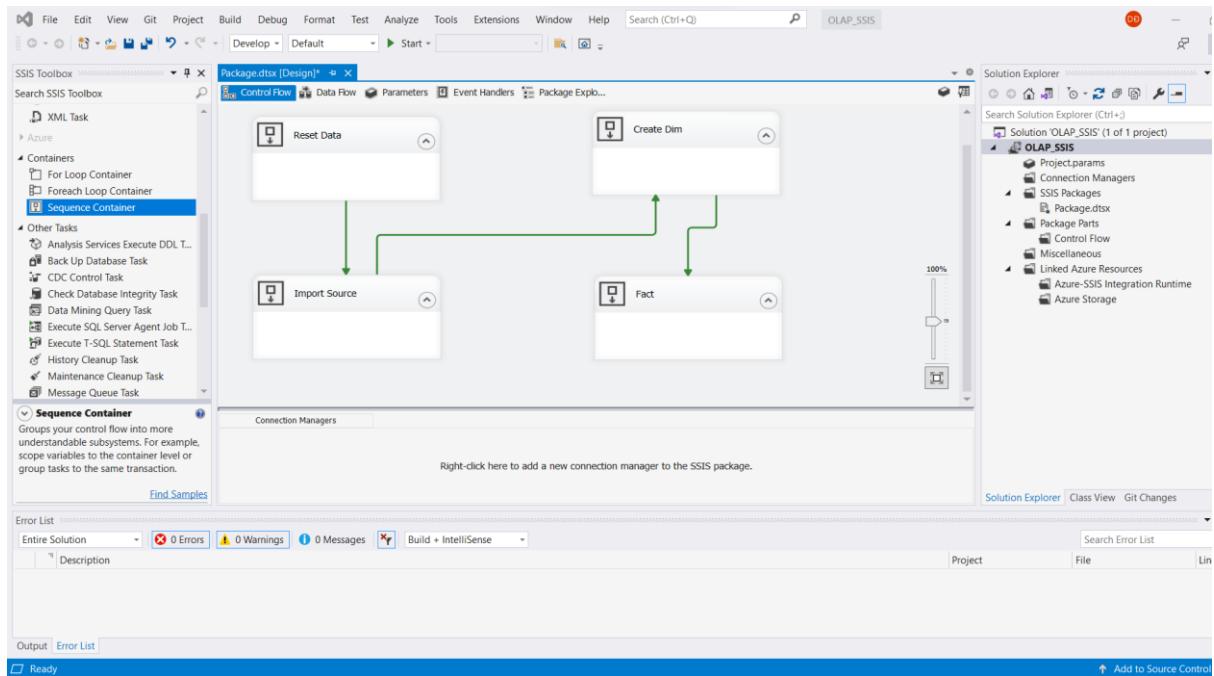


3.2 Khởi tạo Sequence Container

Tạo các Sequence Container:

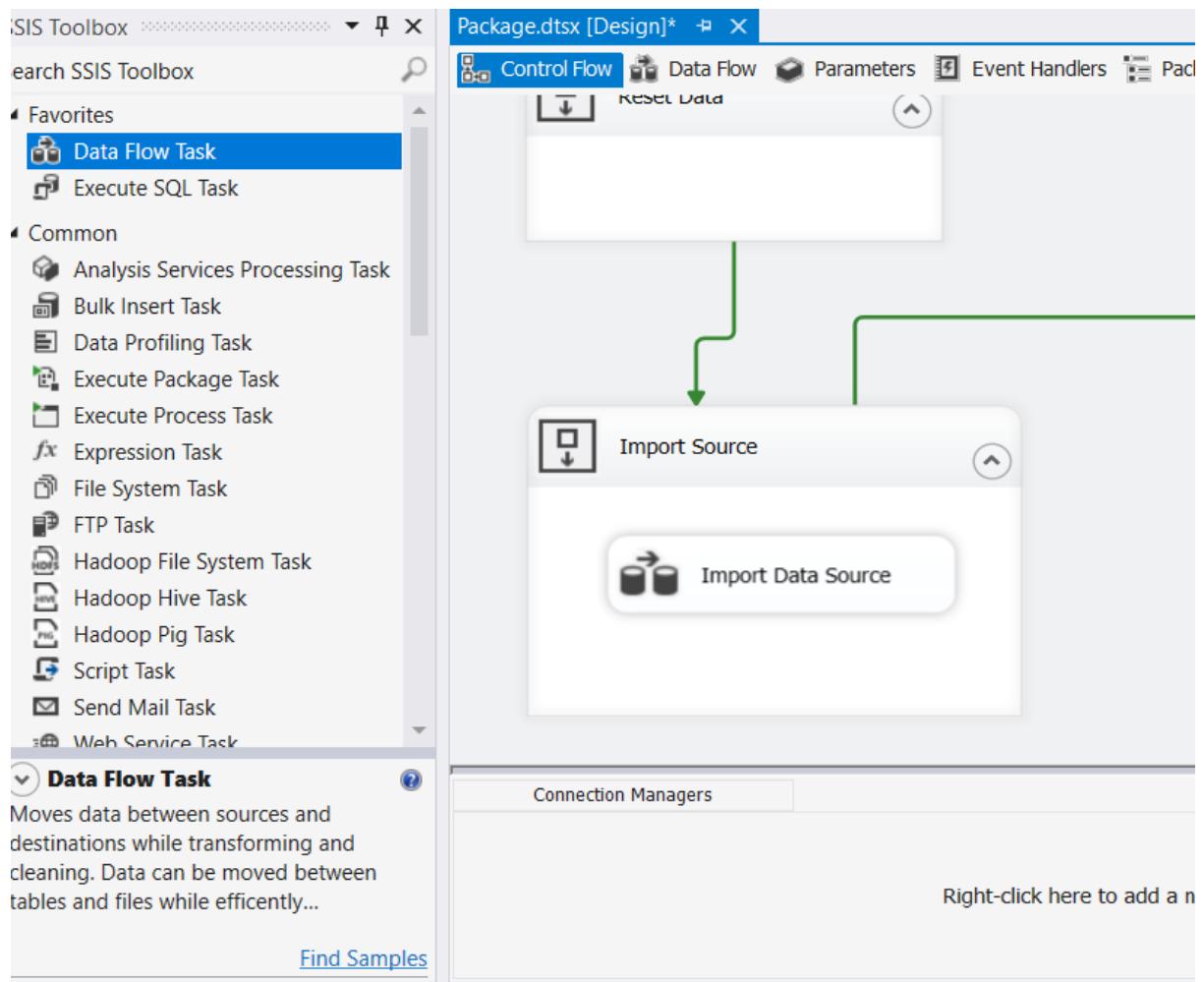
- Reset Data
- Create Dim
- Import Source
- Fact

Tạo các Sequence Container:

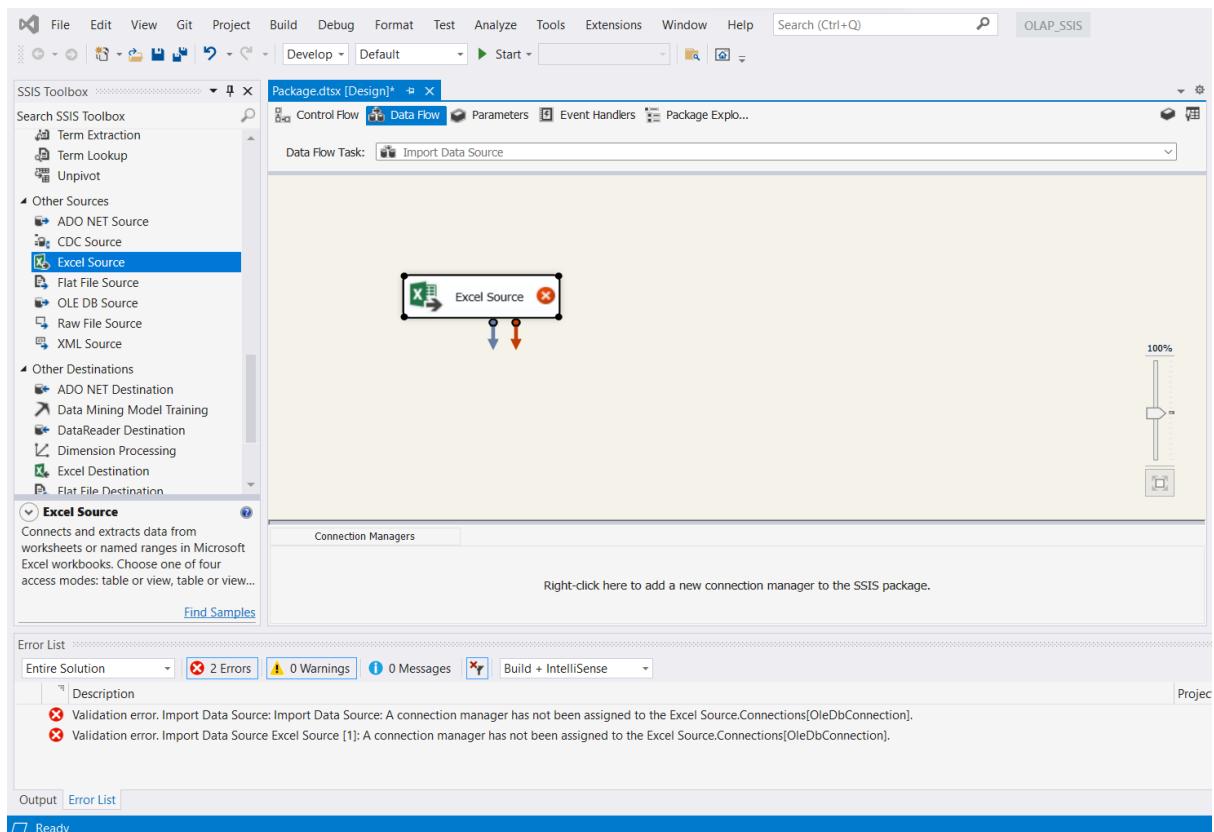


3.3 Import Data

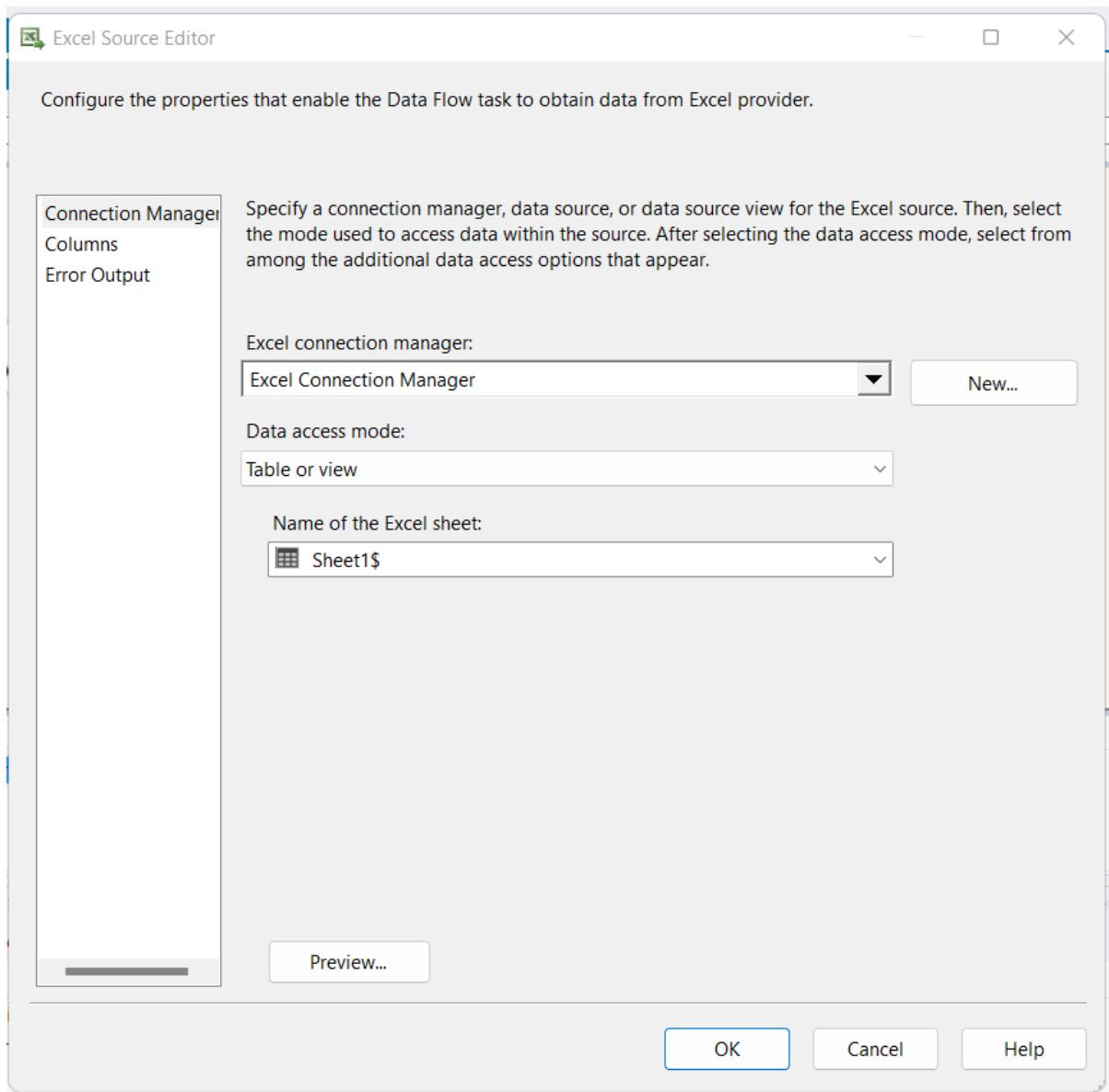
Bước 1: Tạo một “Data Flow Task” và đổi tên thành Import Data Source.



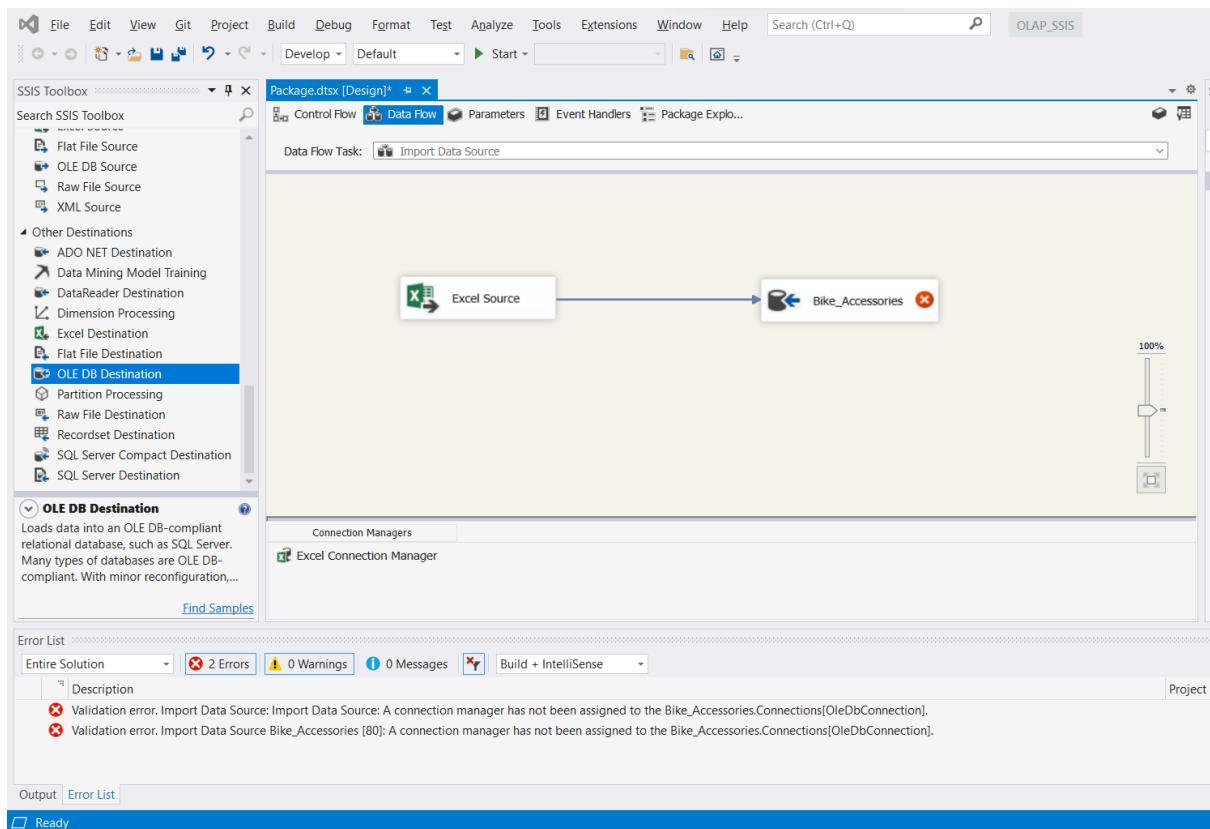
Bước 2: Tạo “Excel Source” và để chuẩn bị đỗ dữ liệu file dataset.xlsx vào Excel Source đó.



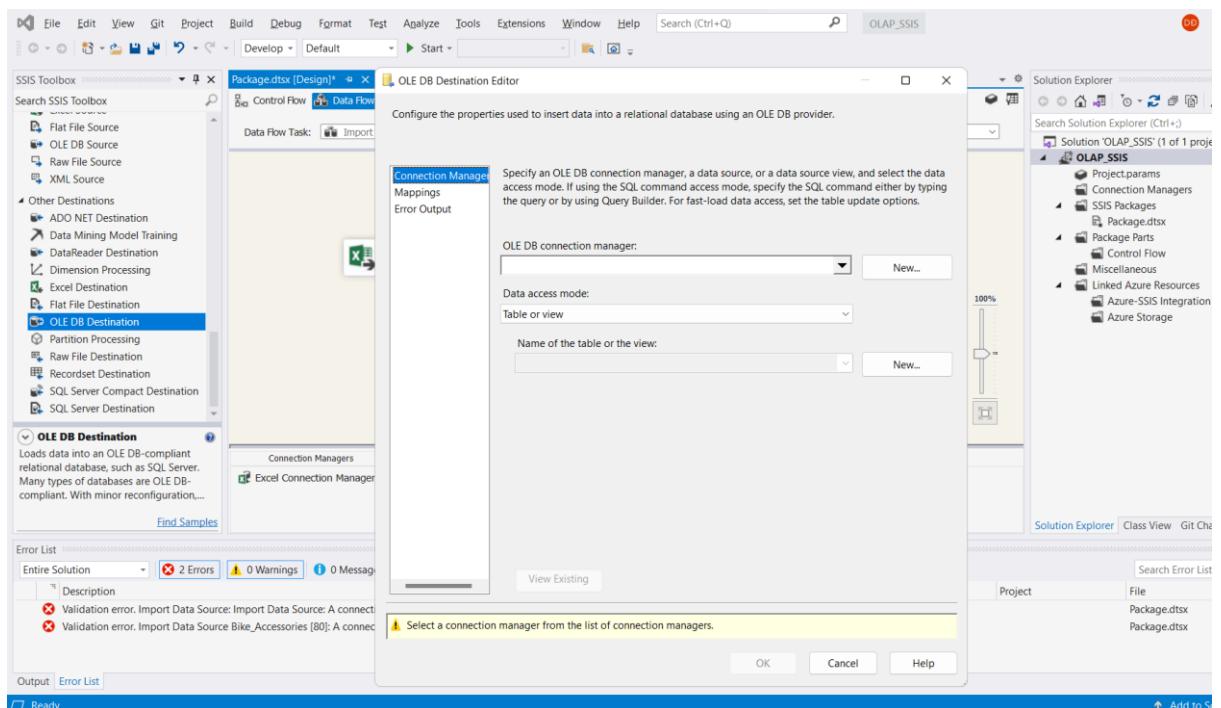
Bước 3: Nhấn đúp vào “Excel Source” vừa tạo để chọn sheet excel mình muốn.



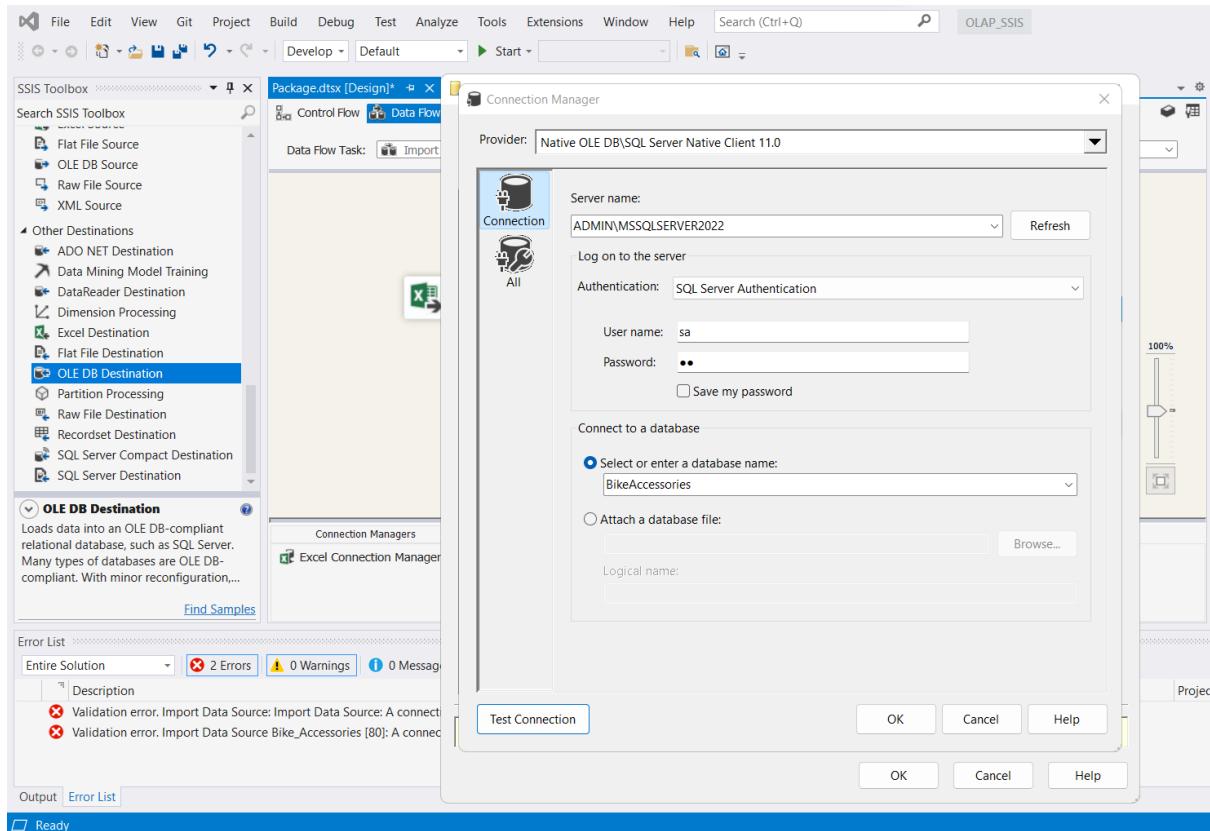
Bước 4: Tạo một “OLE DB Destination” và đổi tên Bike_Accessories để chứa dữ liệu đó từ Excel Source.



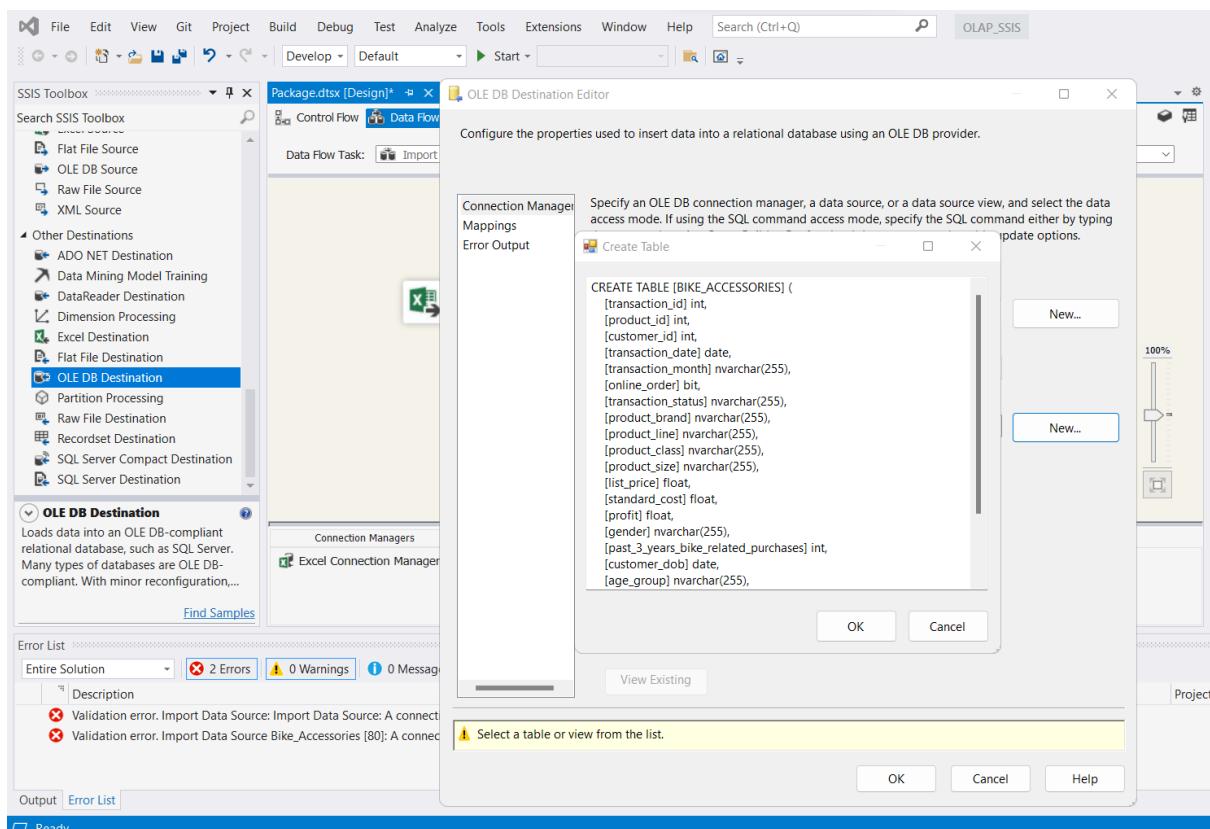
Bước 5: Nhấn đúp vào OLE DB Destination để bắt đầu tạo một OLE DB Connection manager mới.



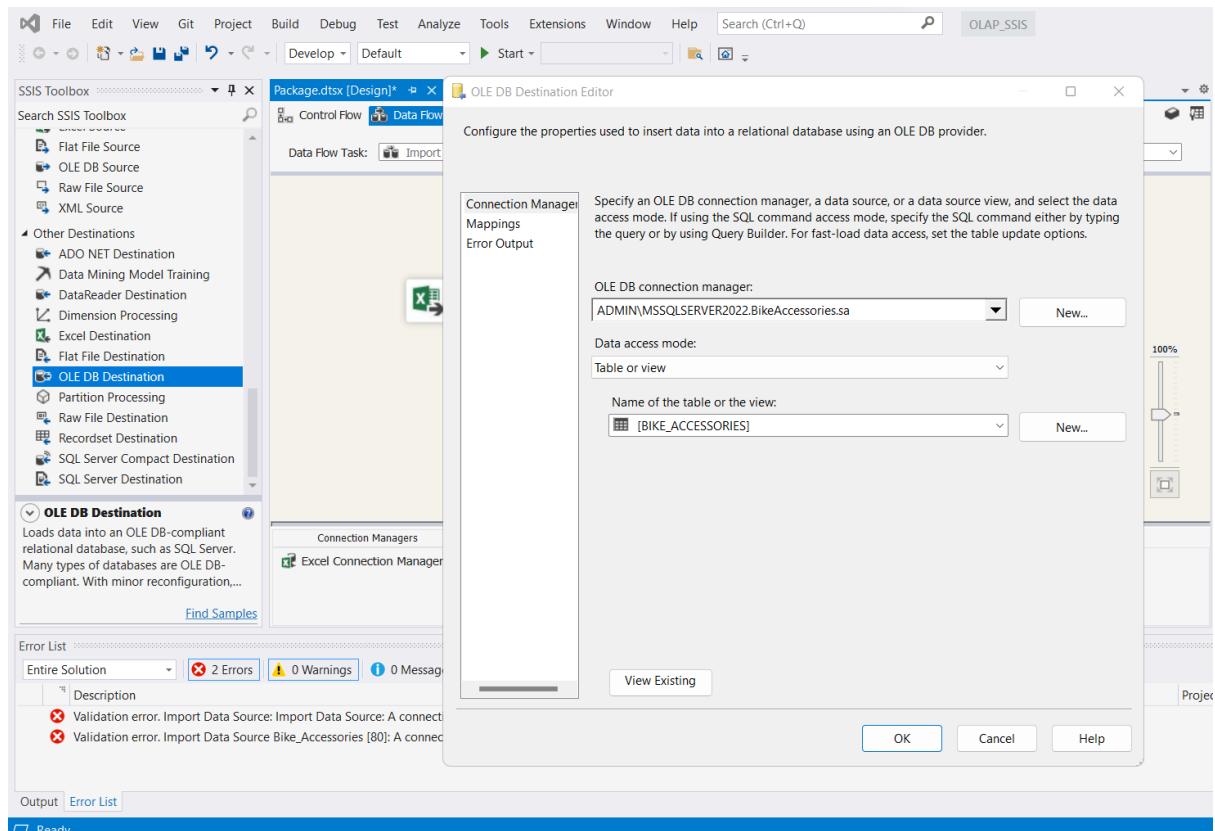
Bước 6: Chọn “New” để tạo Connection manager mới, điền Server name sau đó chọn database muốn đổ vào, sau đó chọn OK.



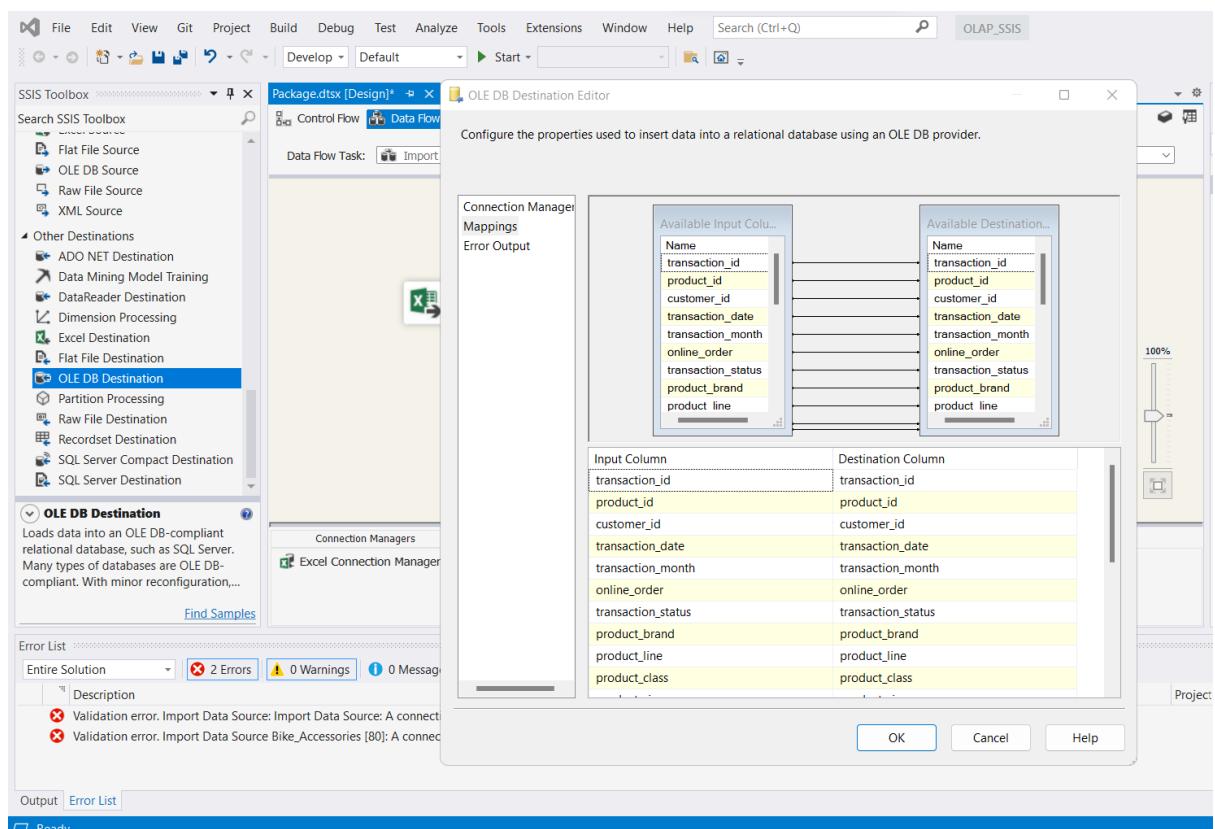
Bước 7: Chọn “New” ở Name of table or view để tạo table sau đó nhấn OK.



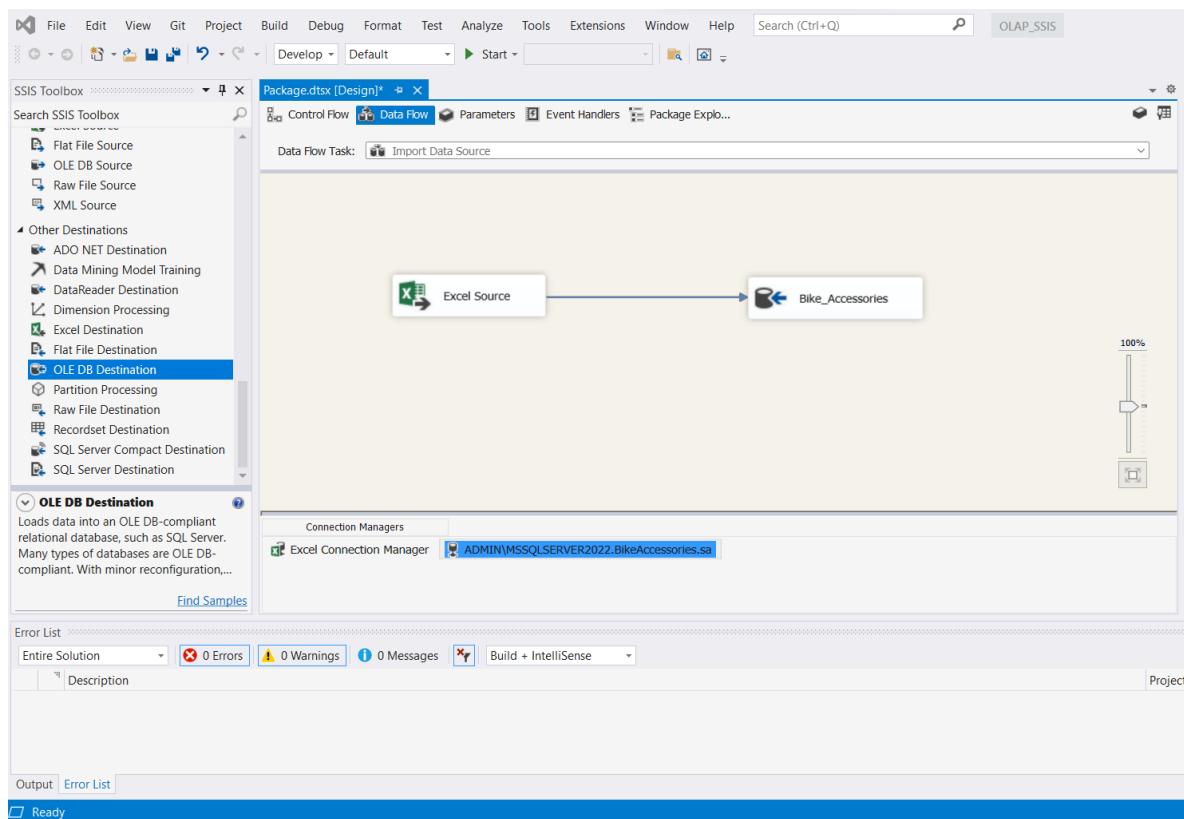
Kết quả:



Bước 8: Qua tab “Mappings” để kiểm tra lại các columns sau đó nhấn OK.

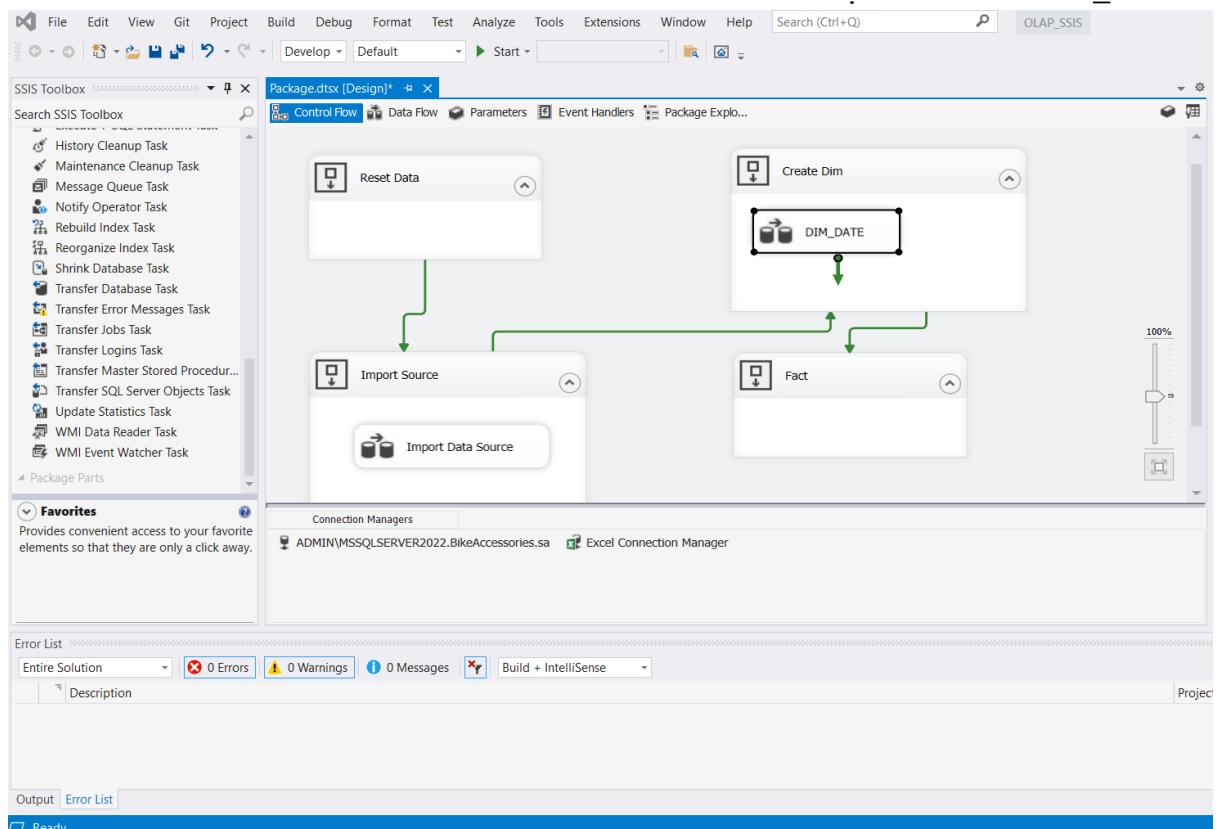


Kết quả:

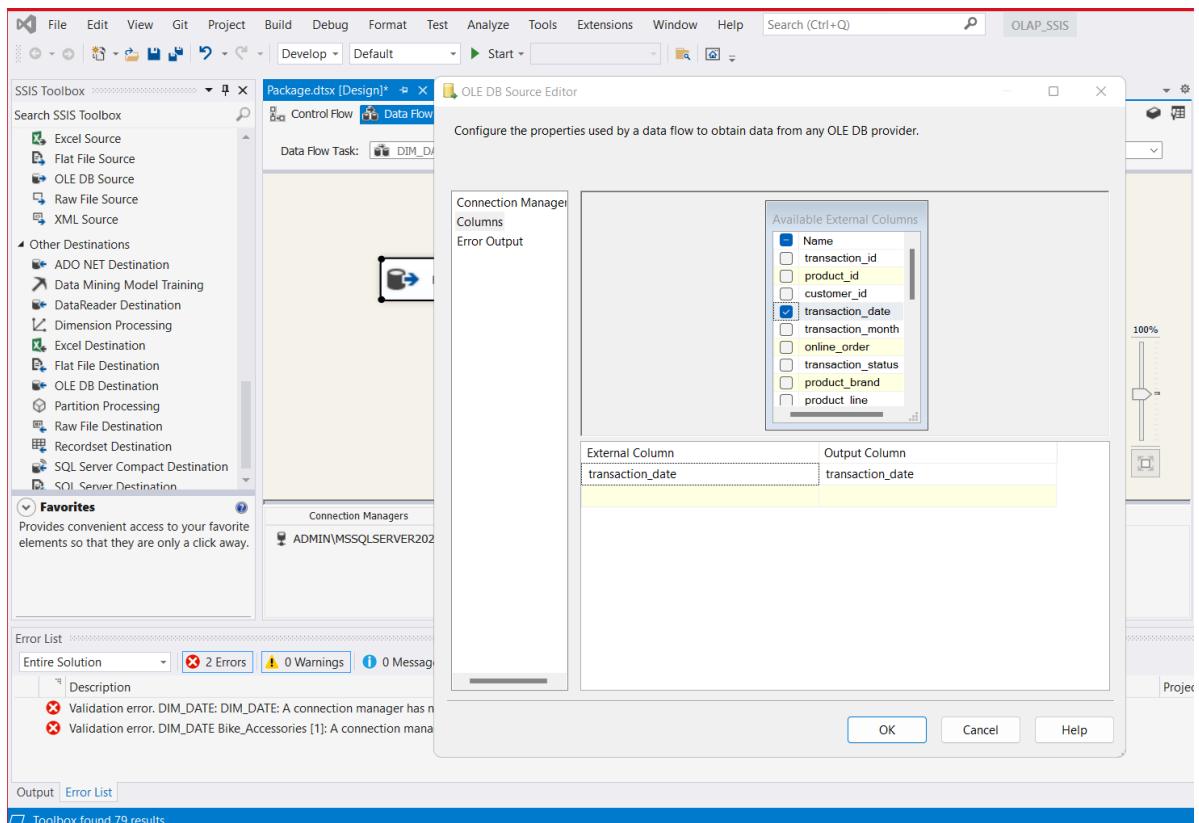


3.4 Tạo bảng DIM_DATE

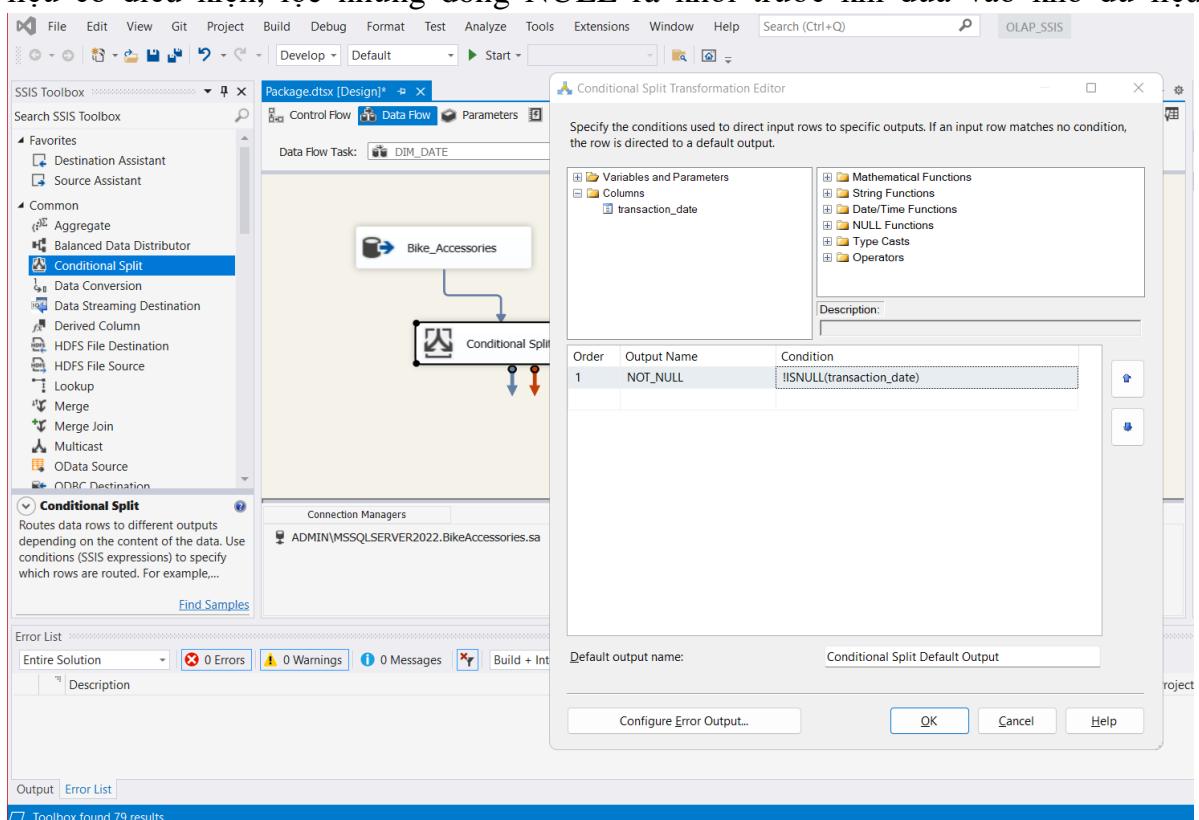
Bước 1: Kéo “Data Flow Task” vào Container đặt tên là DIM_DATE.



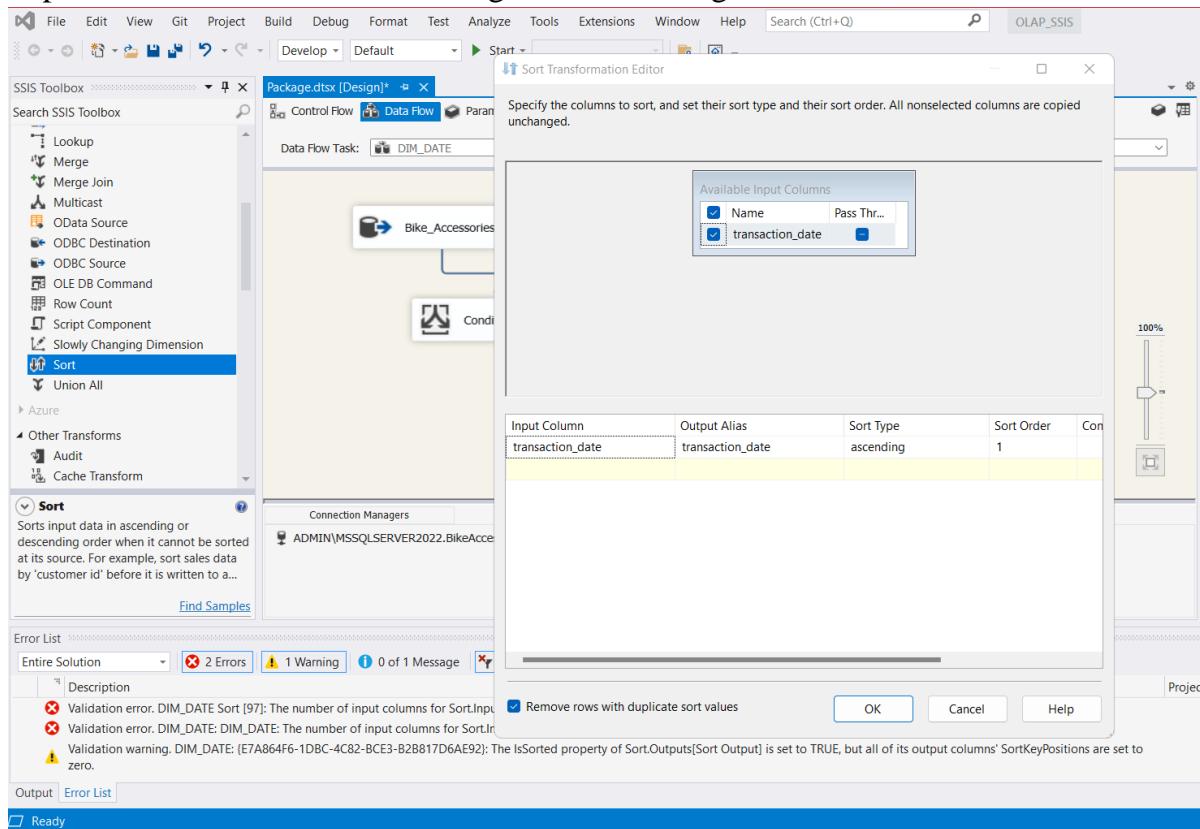
Bước 2: Khởi tạo “OLE DB Source” chứa Dataset và chọn thuộc tính cho bảng DIM_DATE.



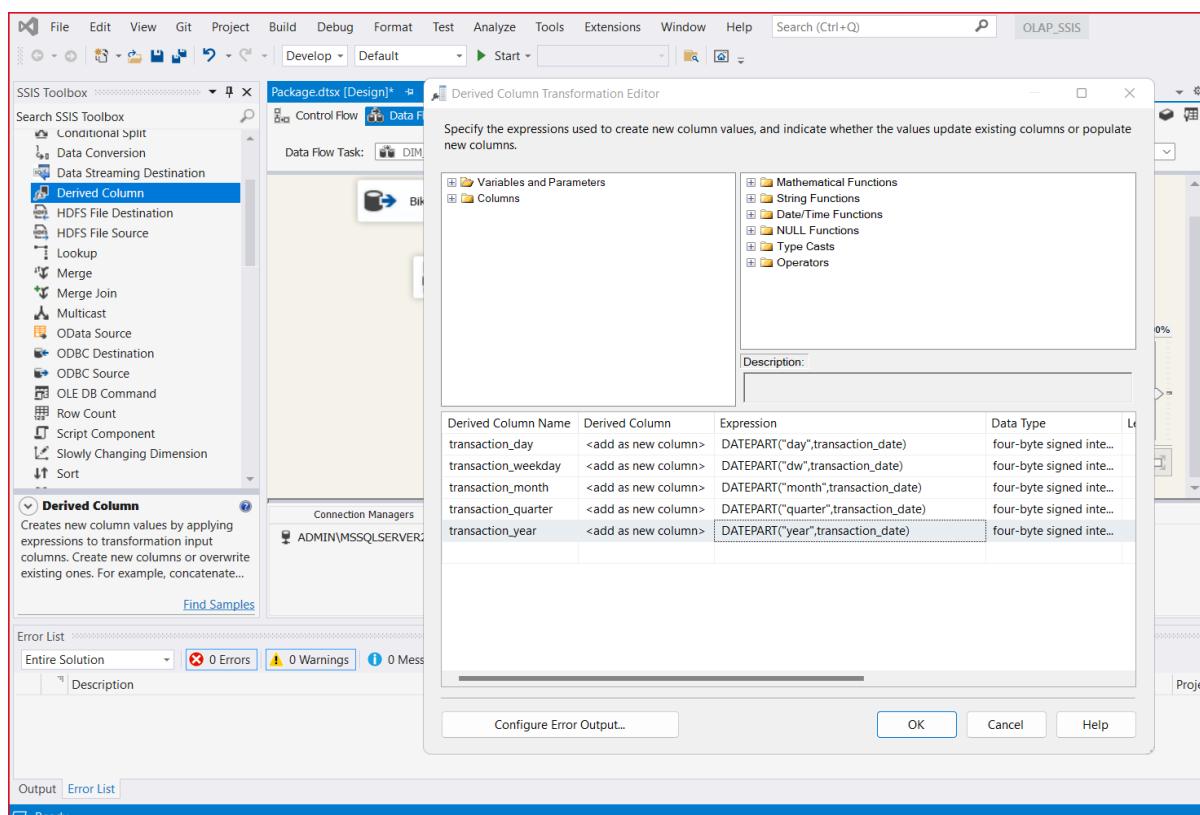
Bước 3: Dùng công cụ “Conditional Split” và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.



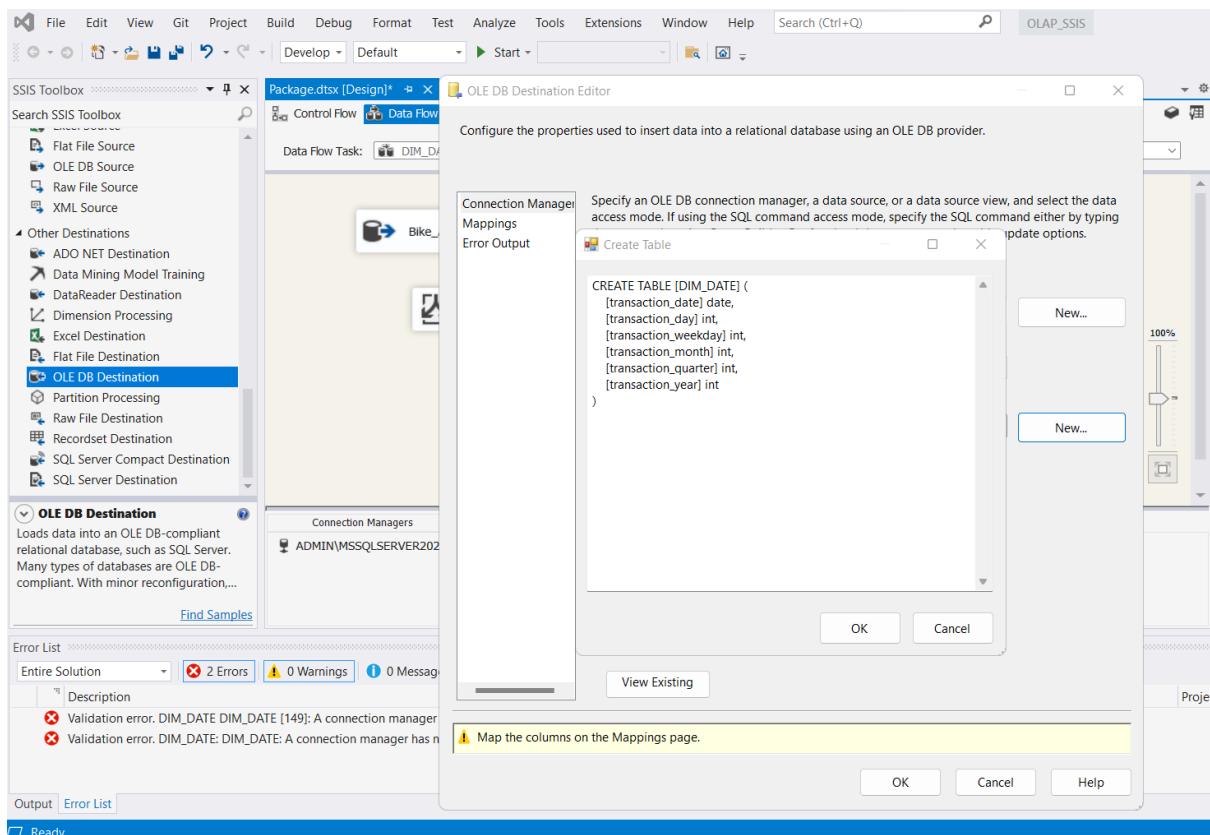
Bước 4: Dùng “Sort” để sắp xếp lại dữ liệu và tick vào ô “Remove rows with duplicate sort values” để xóa những dữ liệu bị trùng.



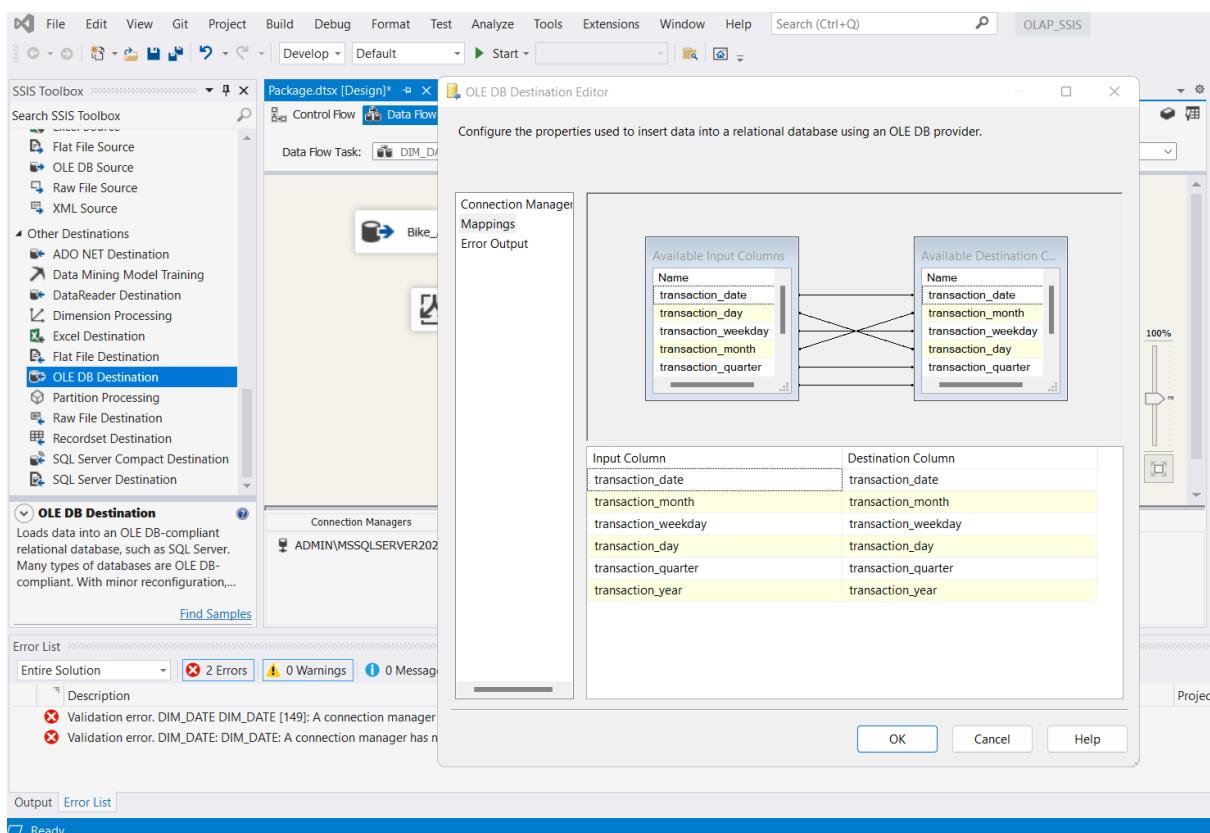
Bước 5: Dùng “Derived Column” để tách phân cấp dữ liệu theo ngày, tuần, tháng, quý,... của bảng DIM_DATE.



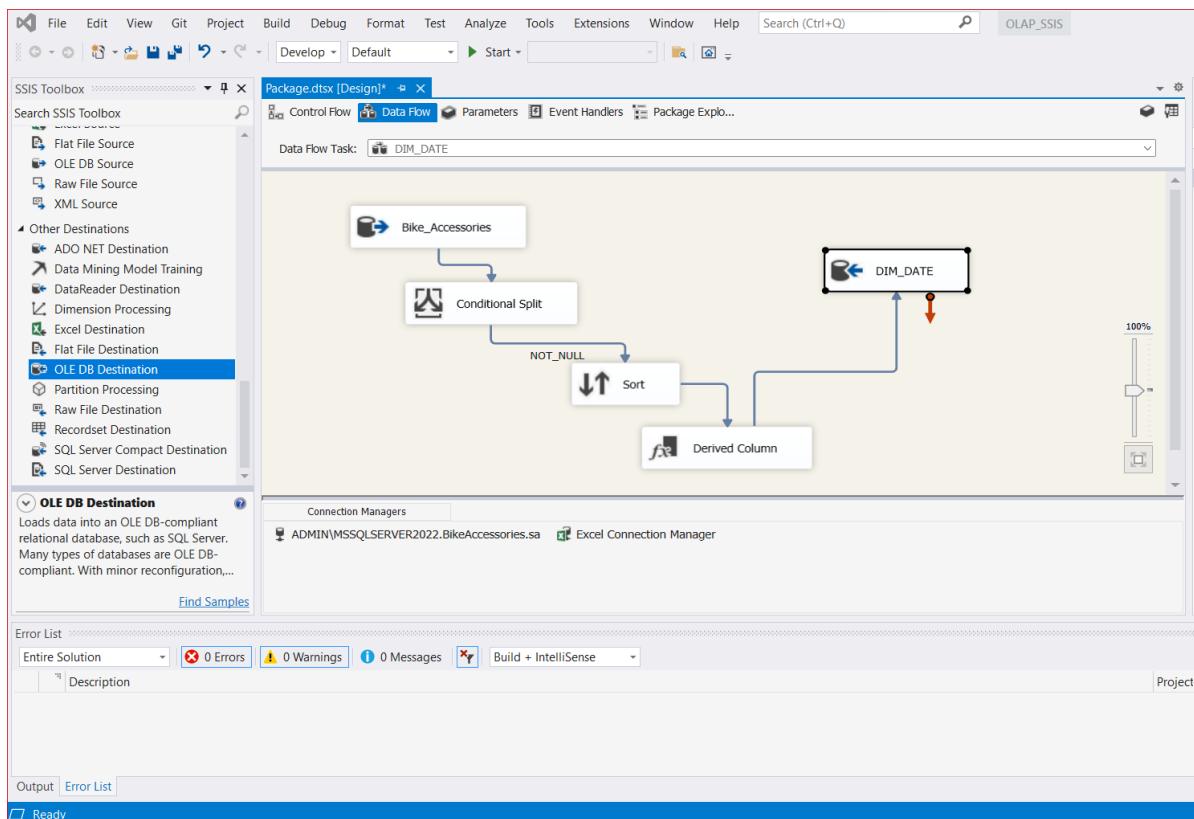
Bước 6: Tiếp theo khởi tạo “OLE DB Destination”, sau đó tạo bảng DIM_DATE.



Bước 7: Qua tab Mappings để kiểm tra.

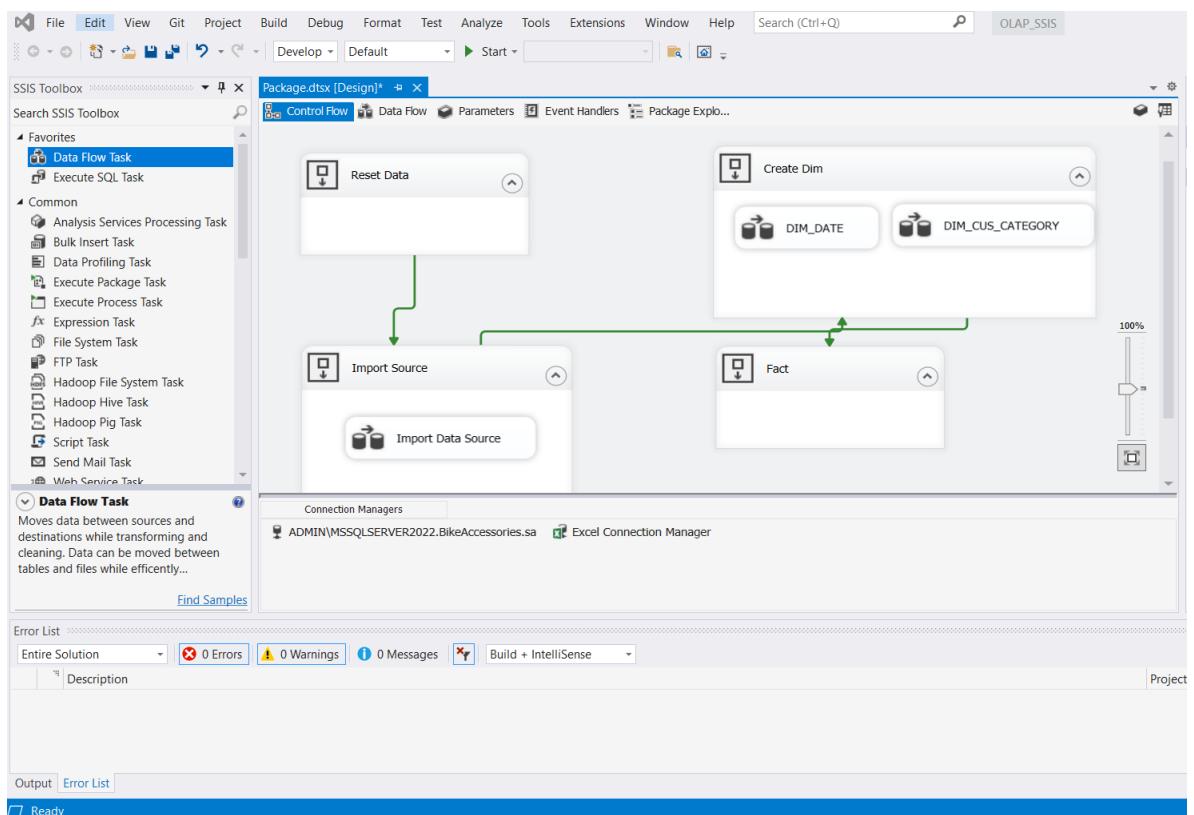


Kết quả luồng thực hiện của bảng DIM_DATE.

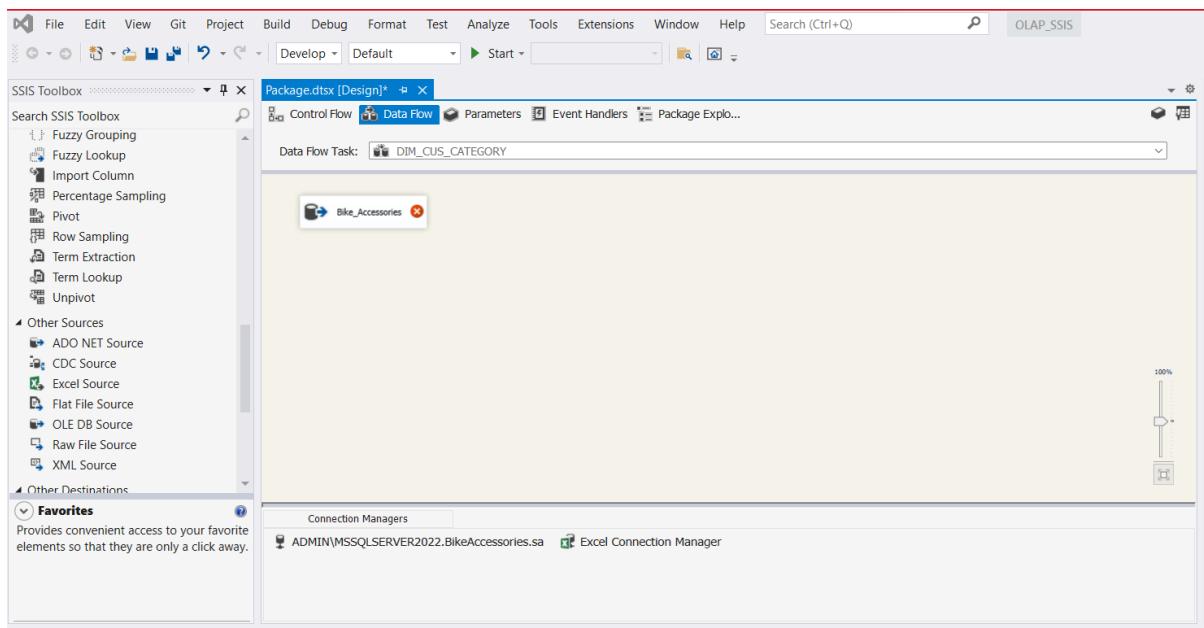


3.5 Tạo bảng DIM_CUS_CATEGORY

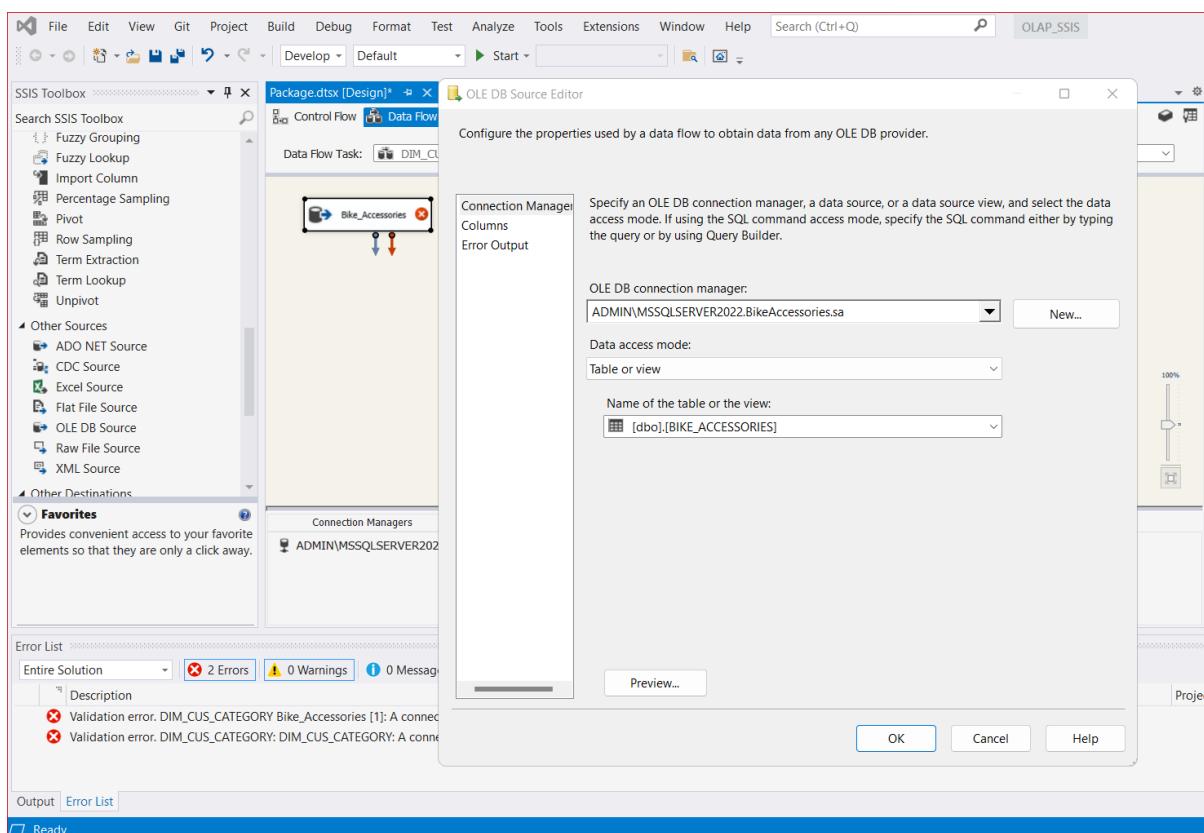
Bước 1: Kéo “Data Flow Task” vào Container đặt tên là DIM_CUS_CATEGORY.



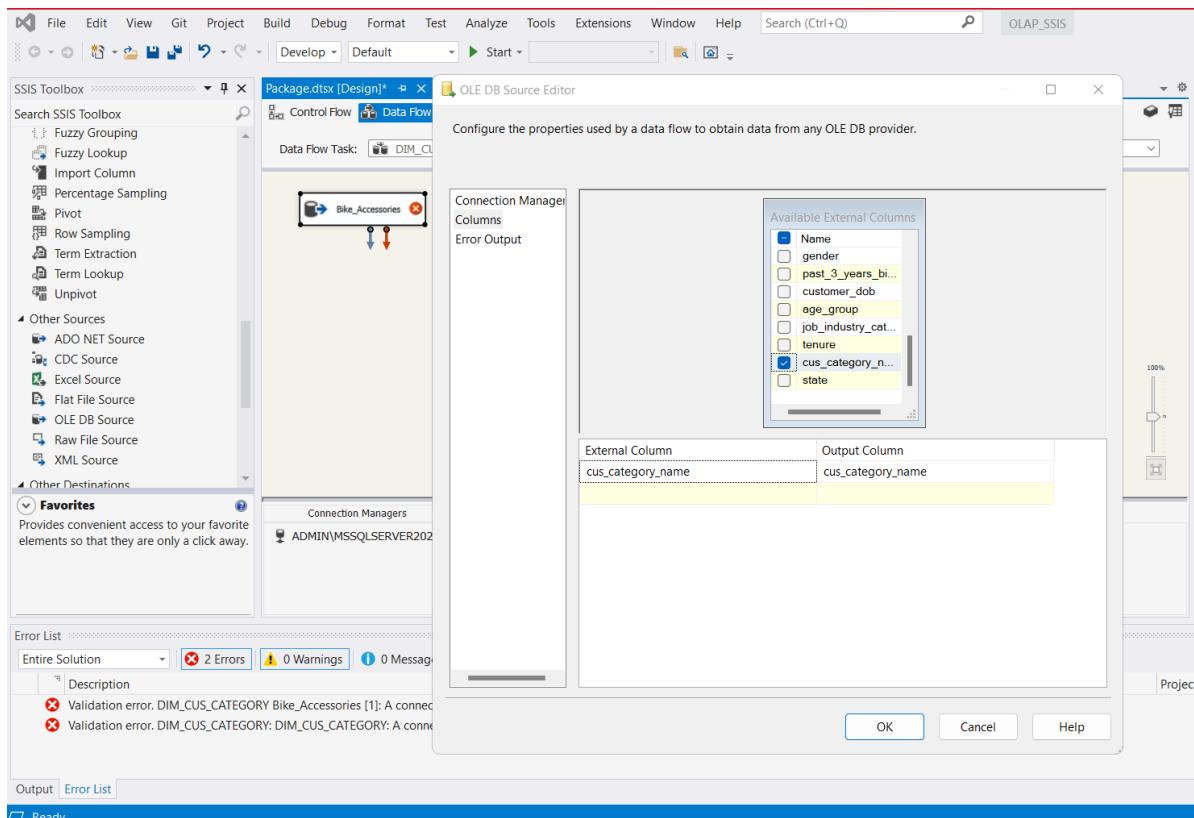
Bước 2: Khởi tạo “OLE DB Source” chứa Dataset.



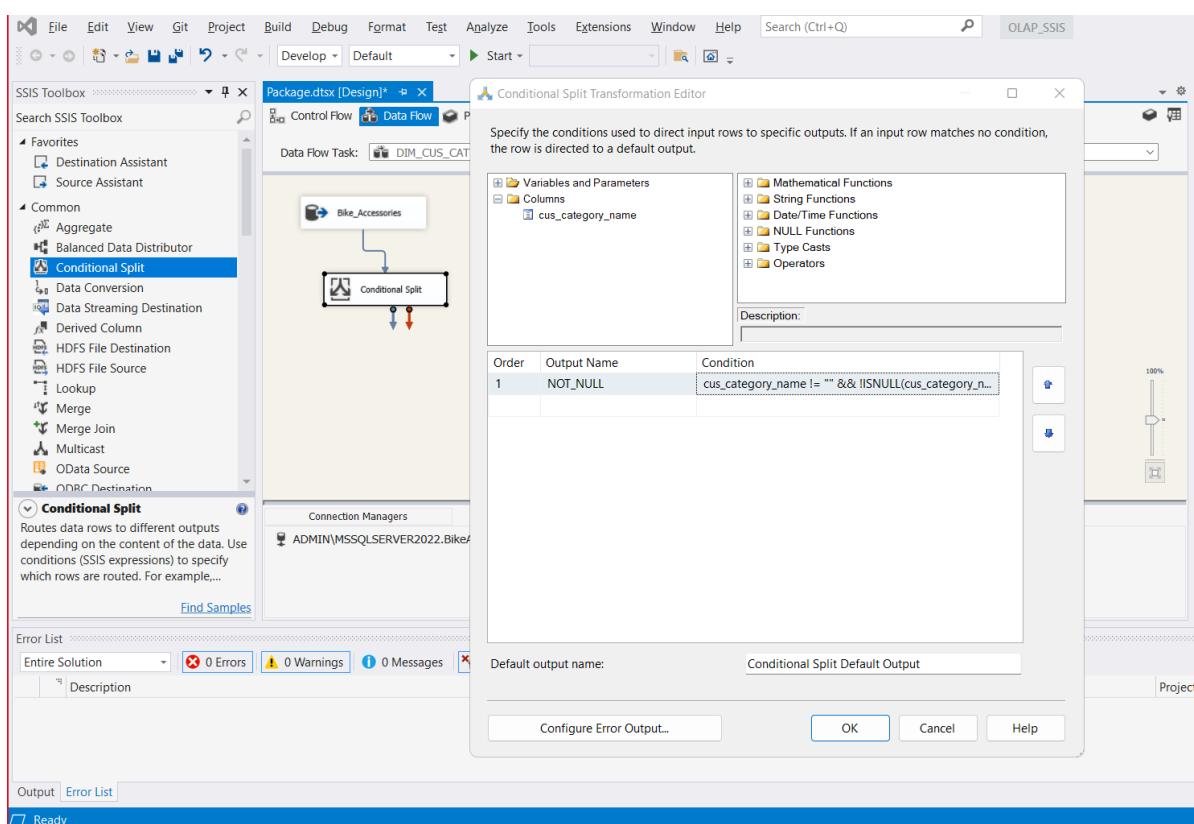
Bước 3: Chọn “Connection BikeAccessories” và chọn bảng BIKE_ACCESSORIES.



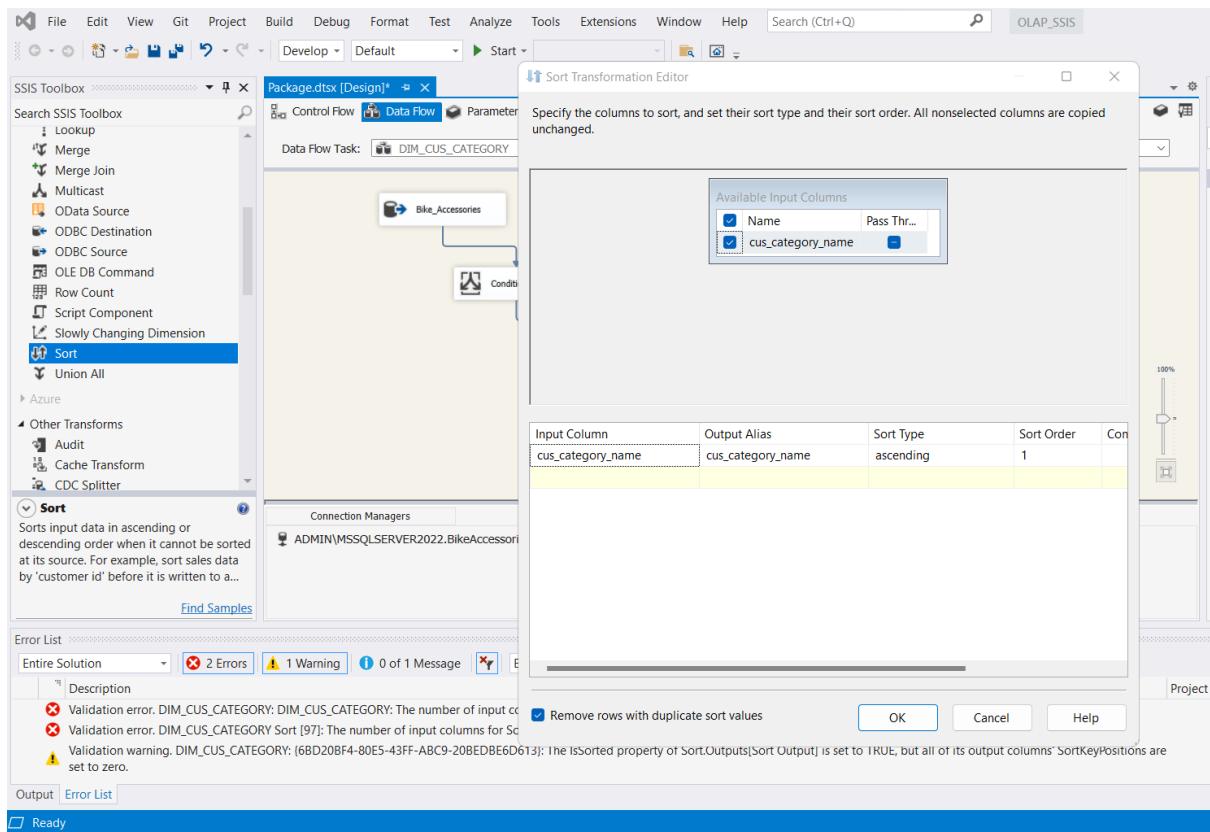
Bước 4: Preview bảng. Ở mục tab Column chọn thuộc tính thuộc bảng DIM_CUS_CATEGORY sau đó chọn OK.



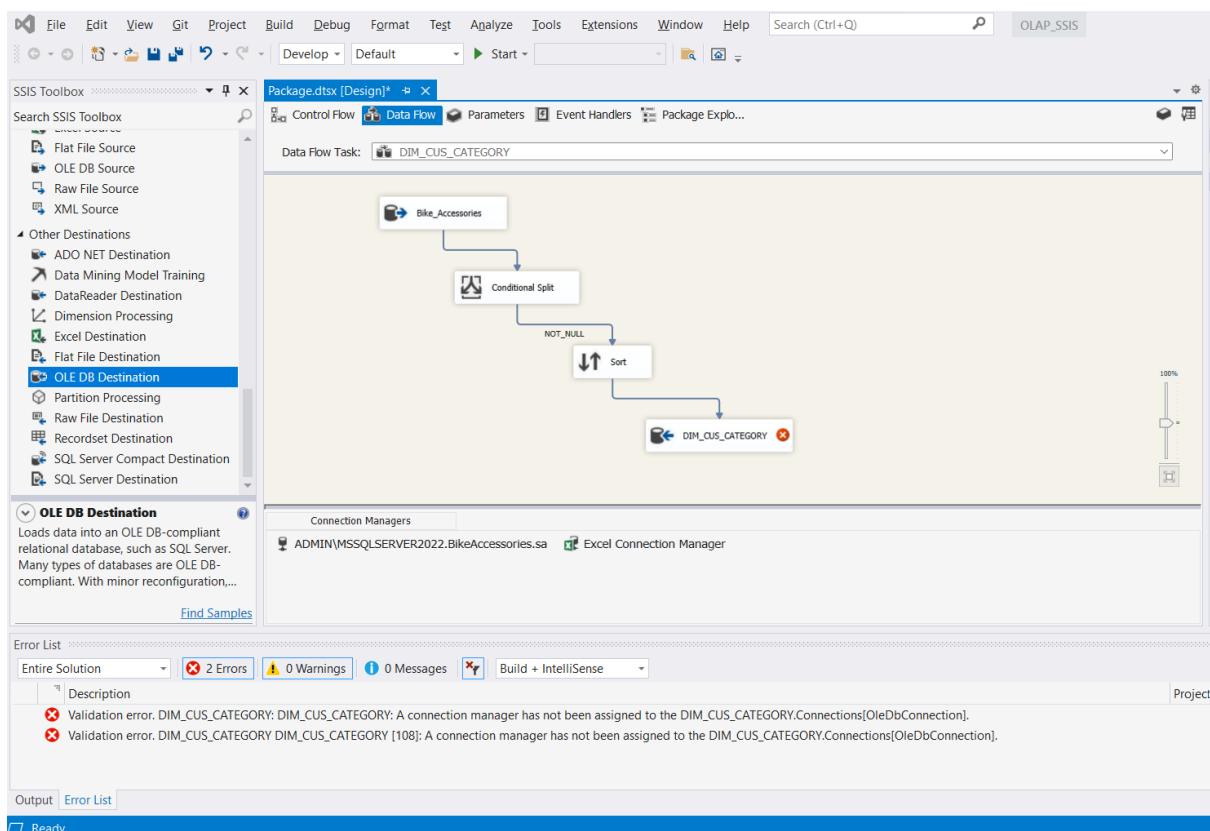
Bước 5: Dùng công cụ “Conditional Split” và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.

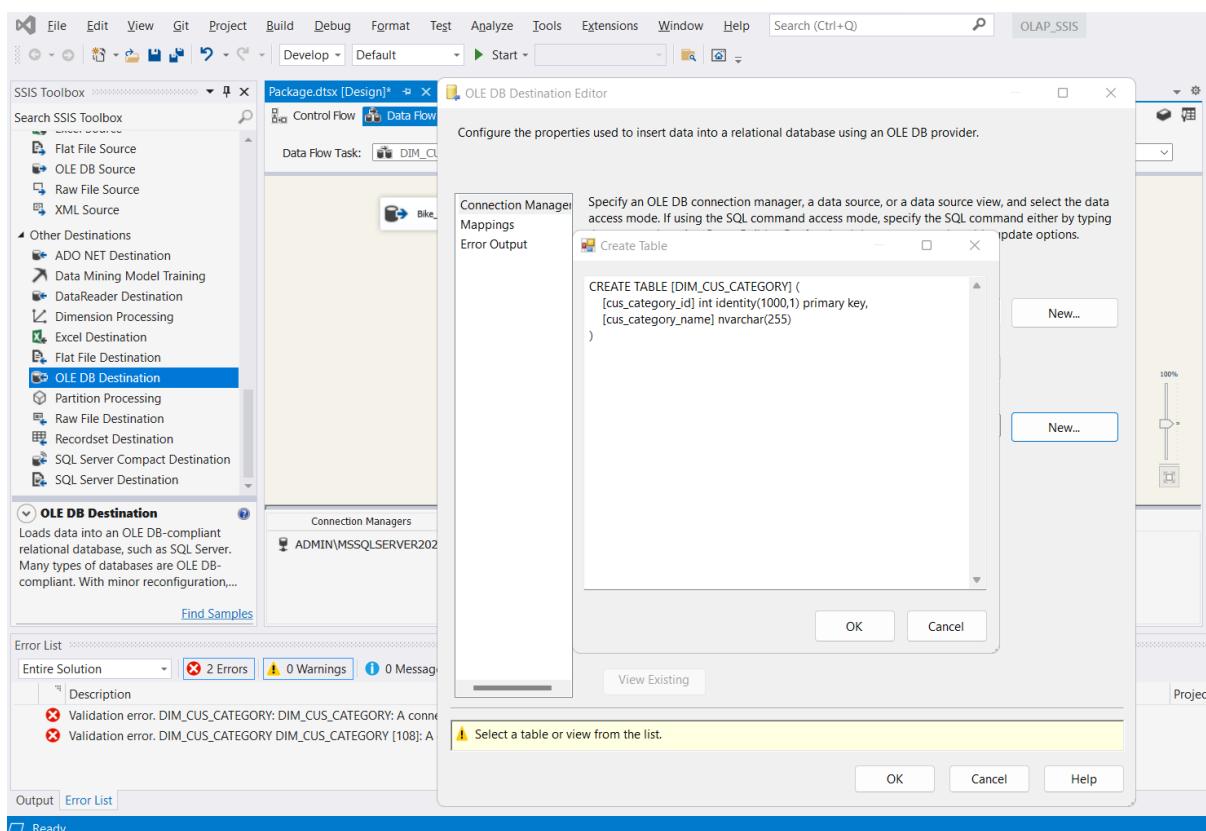


Bước 6: Dùng “Sort” để sắp xếp lại dữ liệu và tick vào ô “Remove rows with duplicate sort values” để xóa những dữ liệu bị trùng.

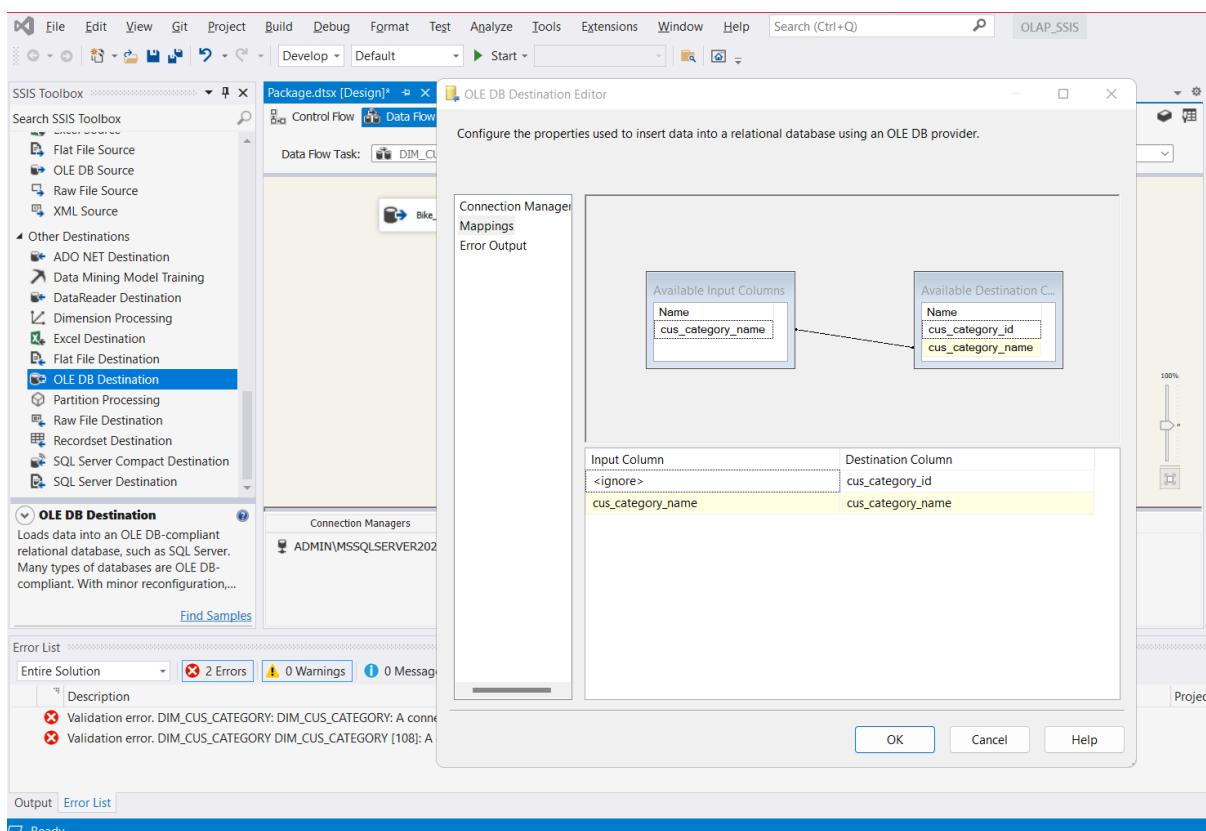


Bước 7: Tạo một “OLE Destination” sau đó tạo bảng DIM_CUS_CATEGORY.

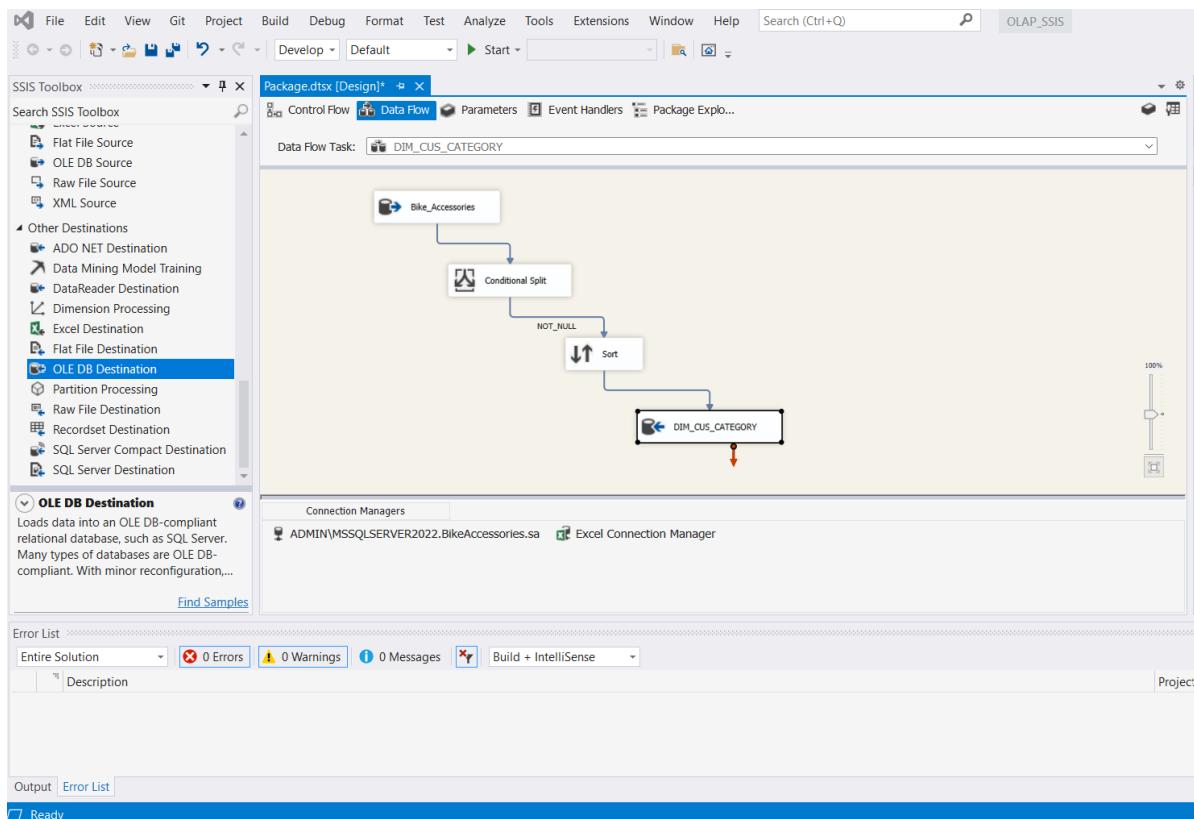




Bước 8: Qua tab “Mappings” để kiểm tra.

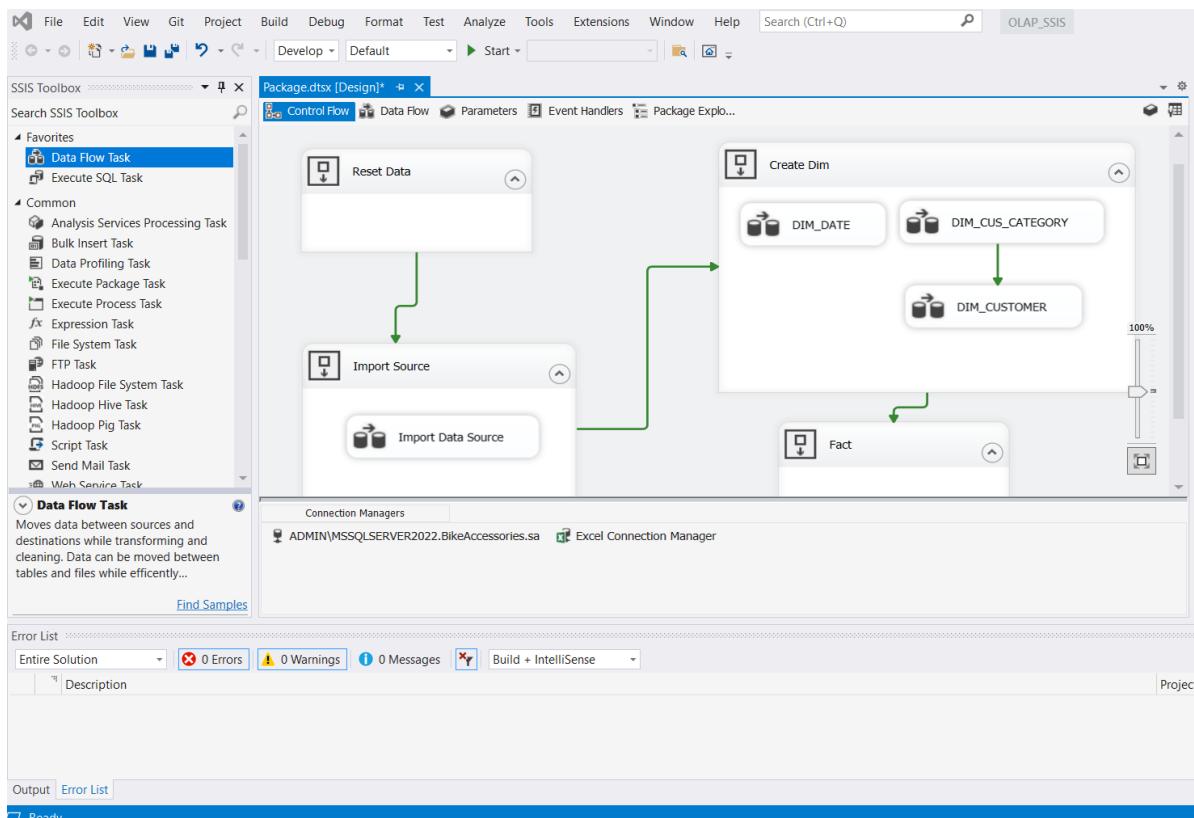


Kết quả luồng thực hiện của bảng DIM_CUS_CATEGORY.

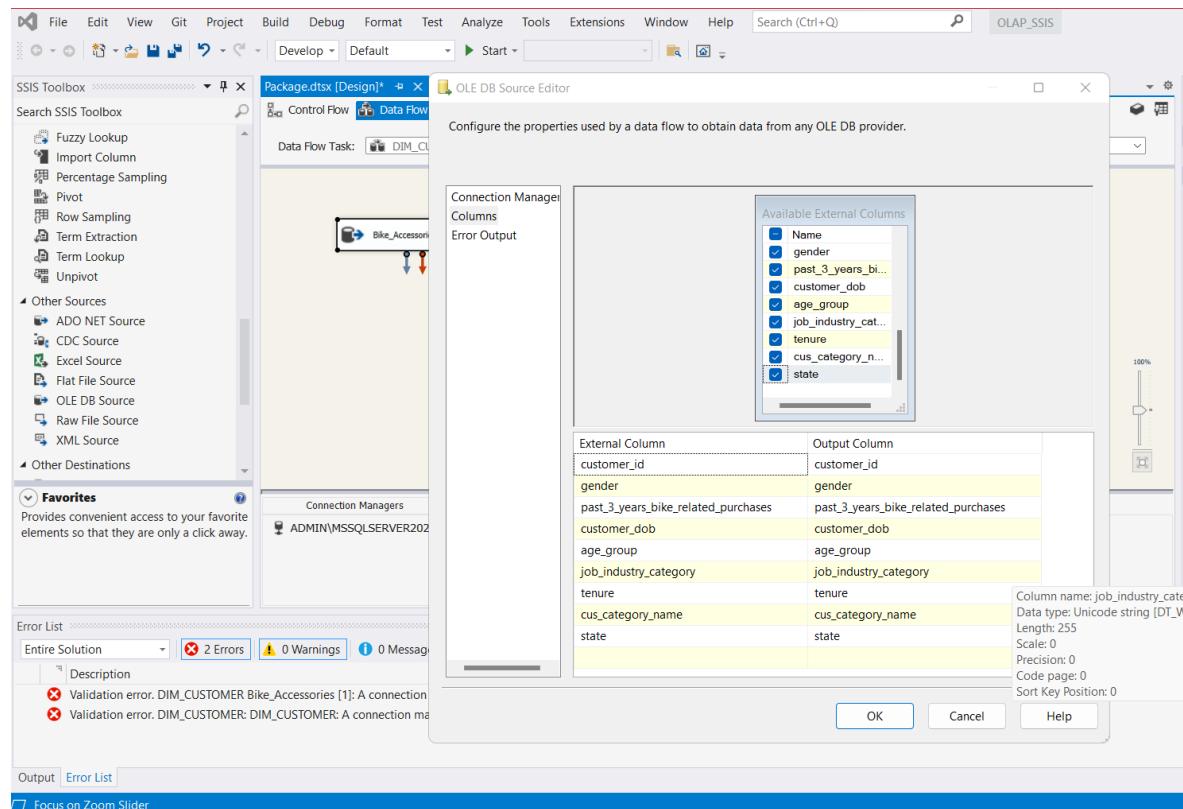


3.6 Tạo bảng DIM_CUSTOMER

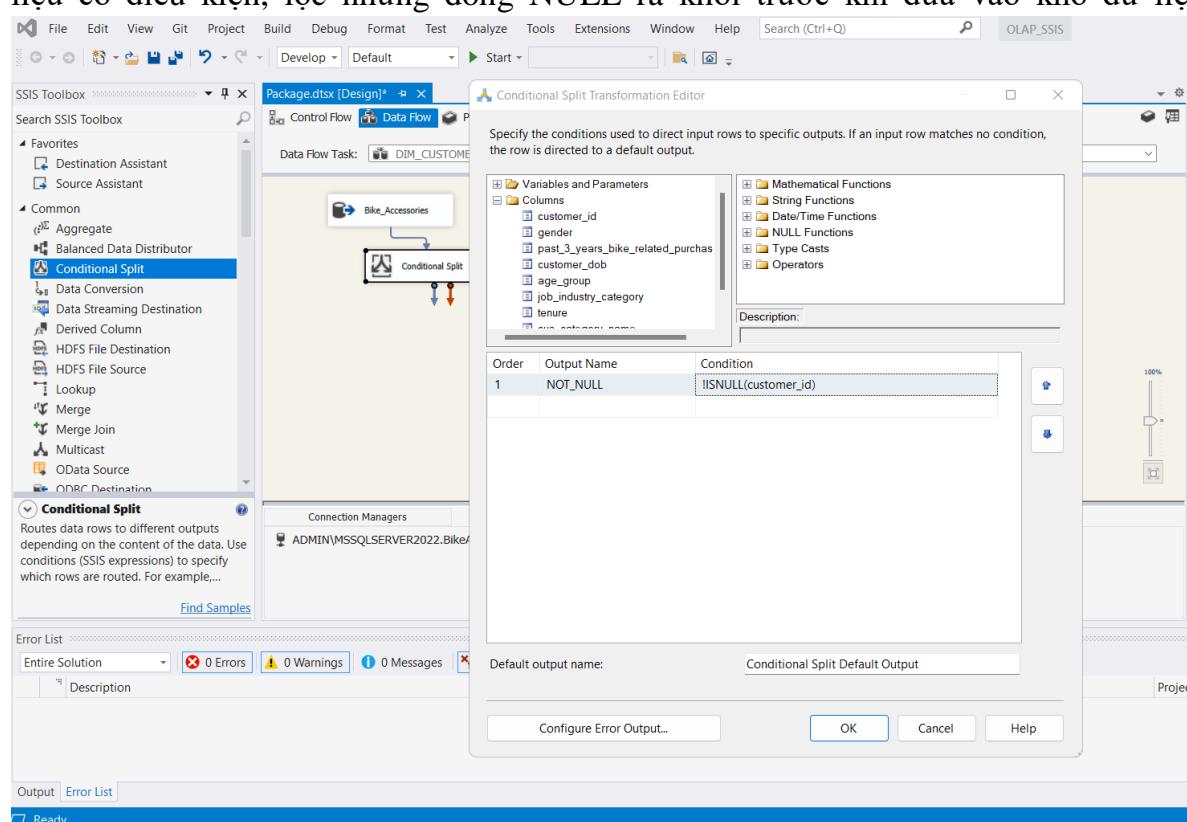
Bước 1: Kéo “Data Flow Task” vào Container đặt tên là DIM_CUSTOMER.



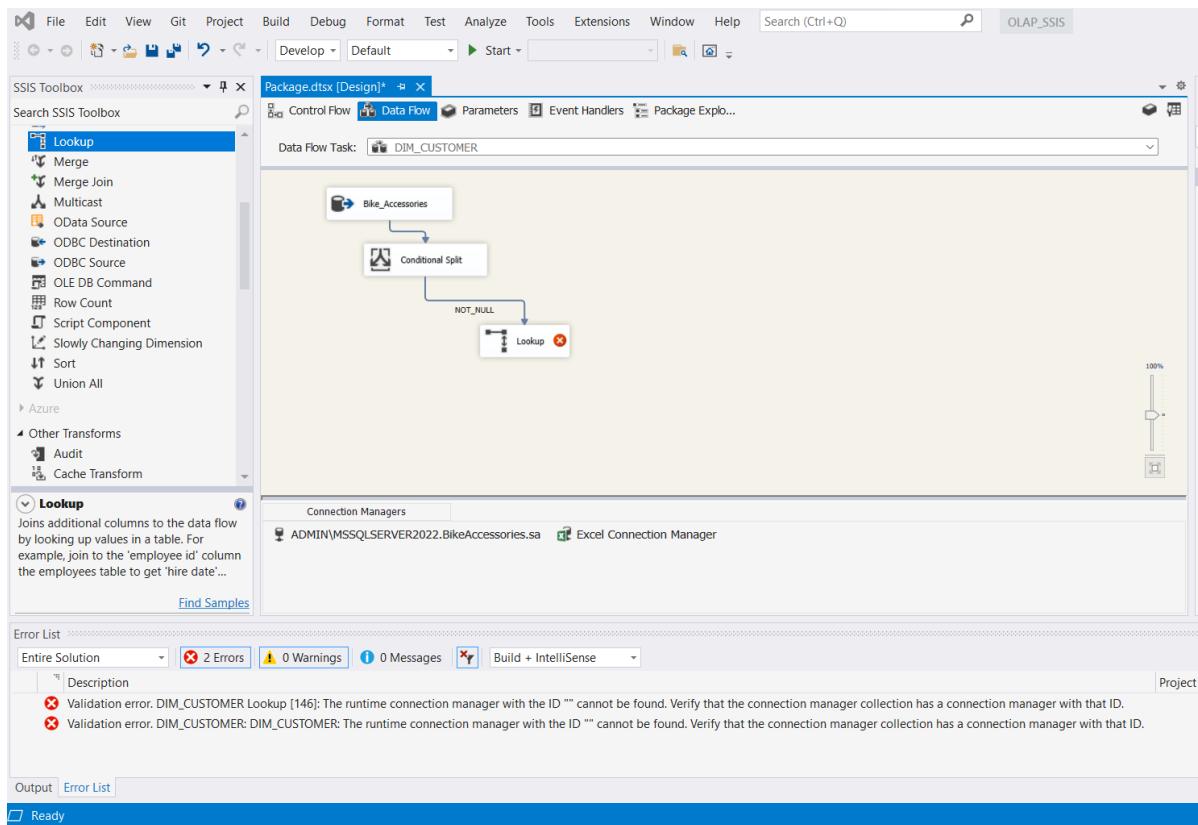
Bước 2: Khởi tạo “OLE DB Source” chứa Dataset và chọn thuộc tính cho bảng DIM_CUSTOMER.



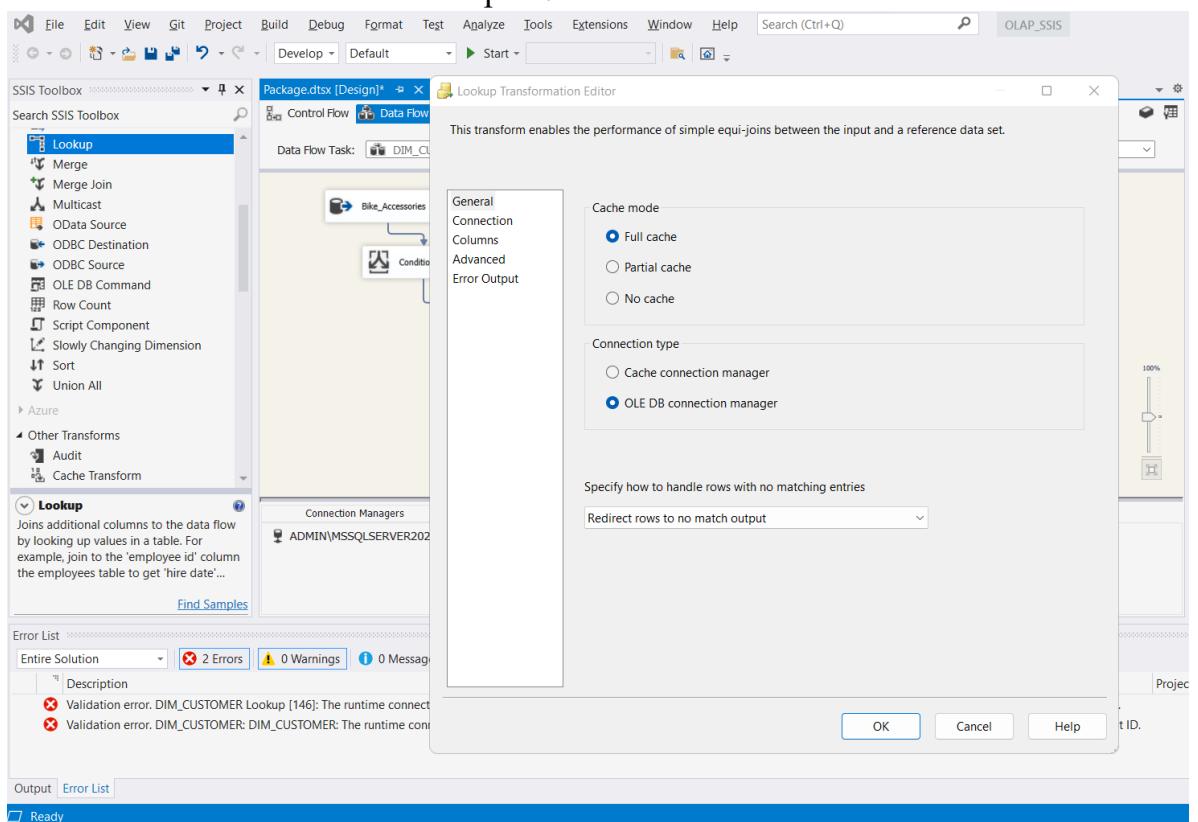
Bước 3: Dùng công cụ “Conditional Split” và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.



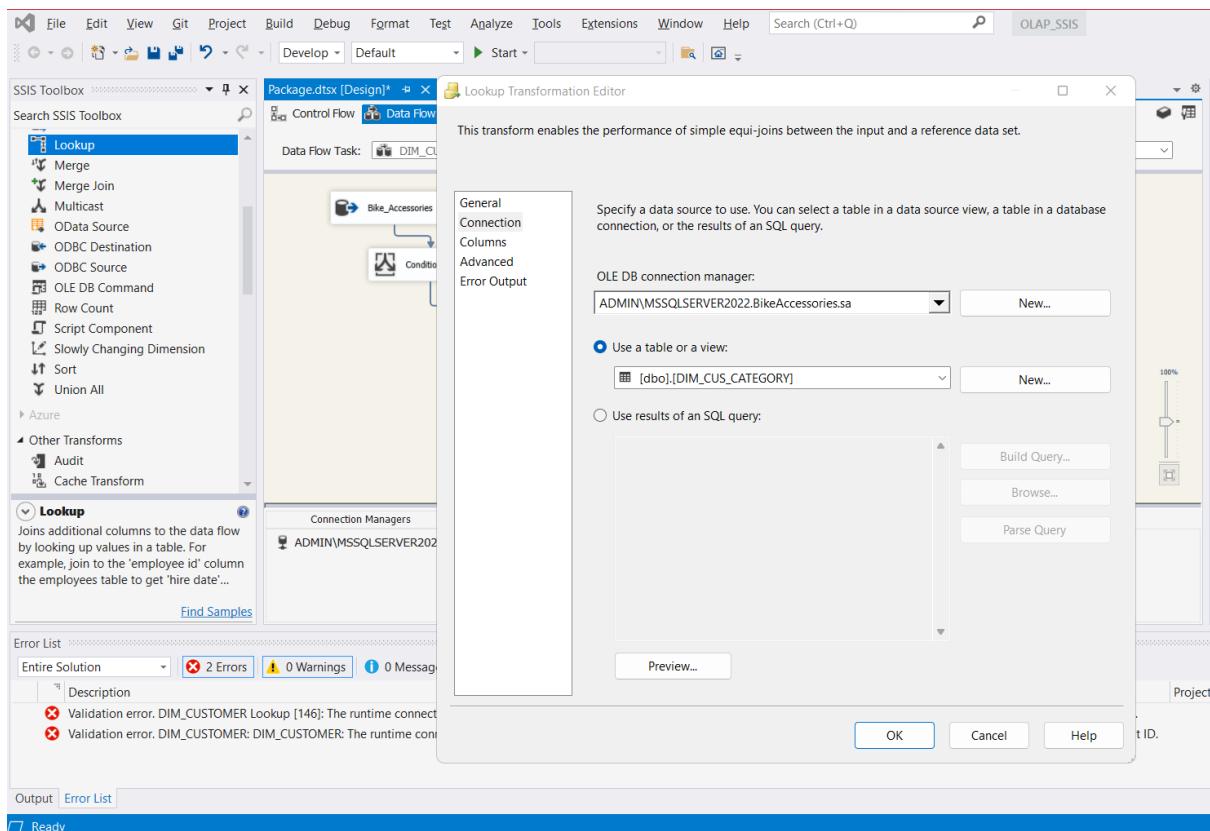
Bước 4: Dùng “Lookup” để kết bảng BIKE_ACCESSORIES và bảng DIM_CUS_CATEGORY thông qua thuộc tính “cus_category_name”.



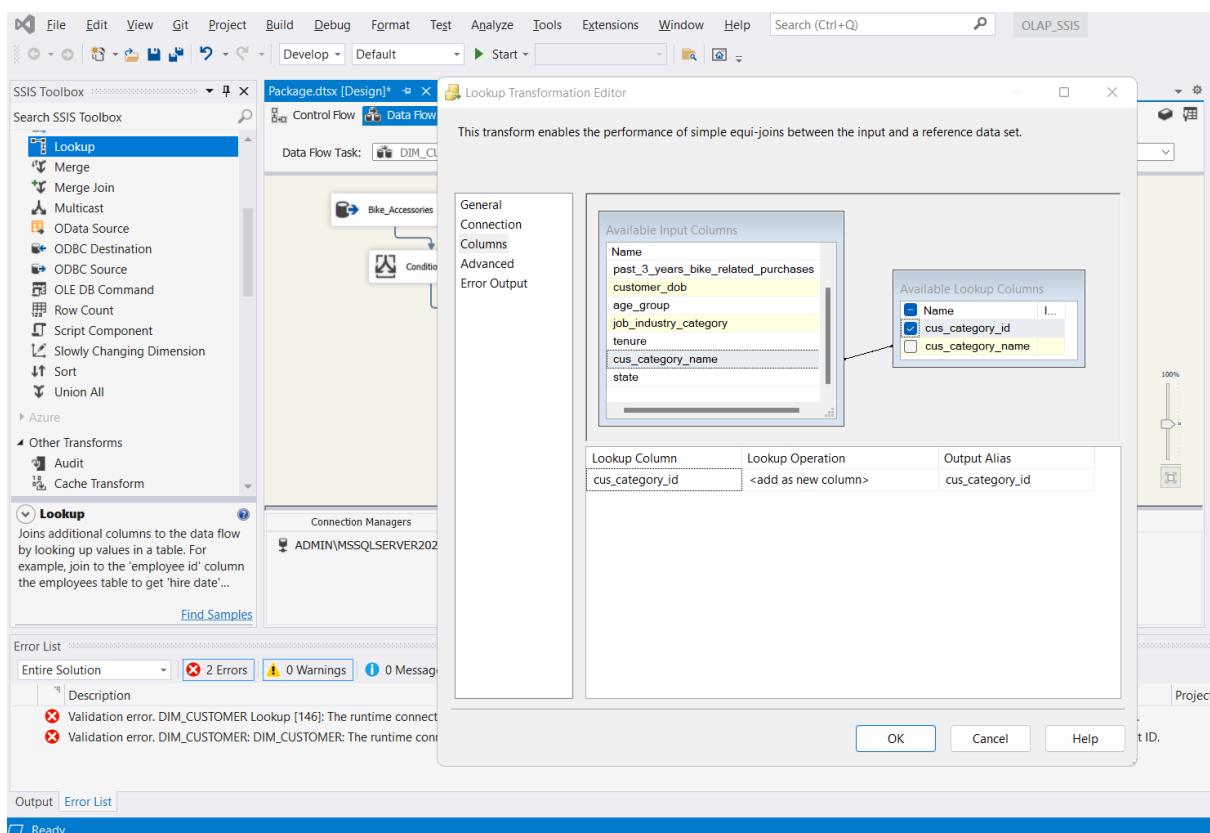
Tại “General”, tại mục “Specify how to handle rows with no matching entries” chọn “Redirect rows no to match no output”.



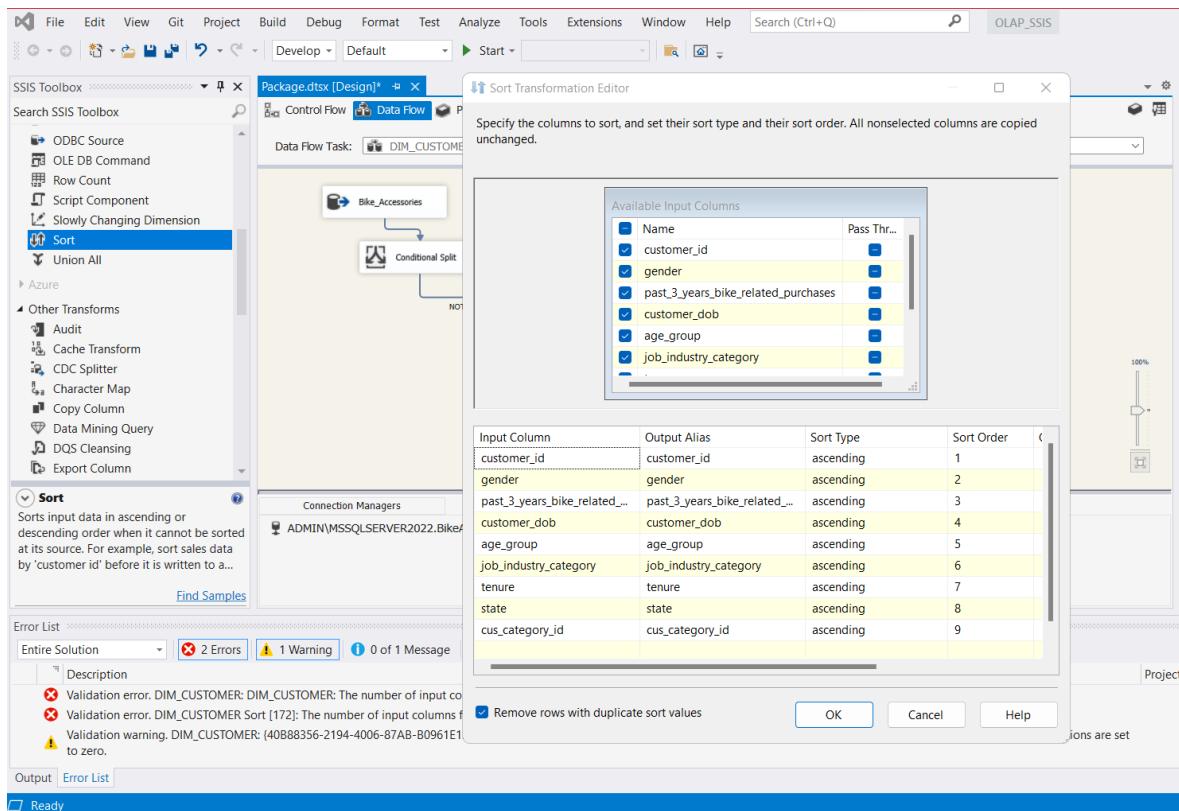
Tạo Connection, Chọn bảng DIM_CUS_CATEGORY.



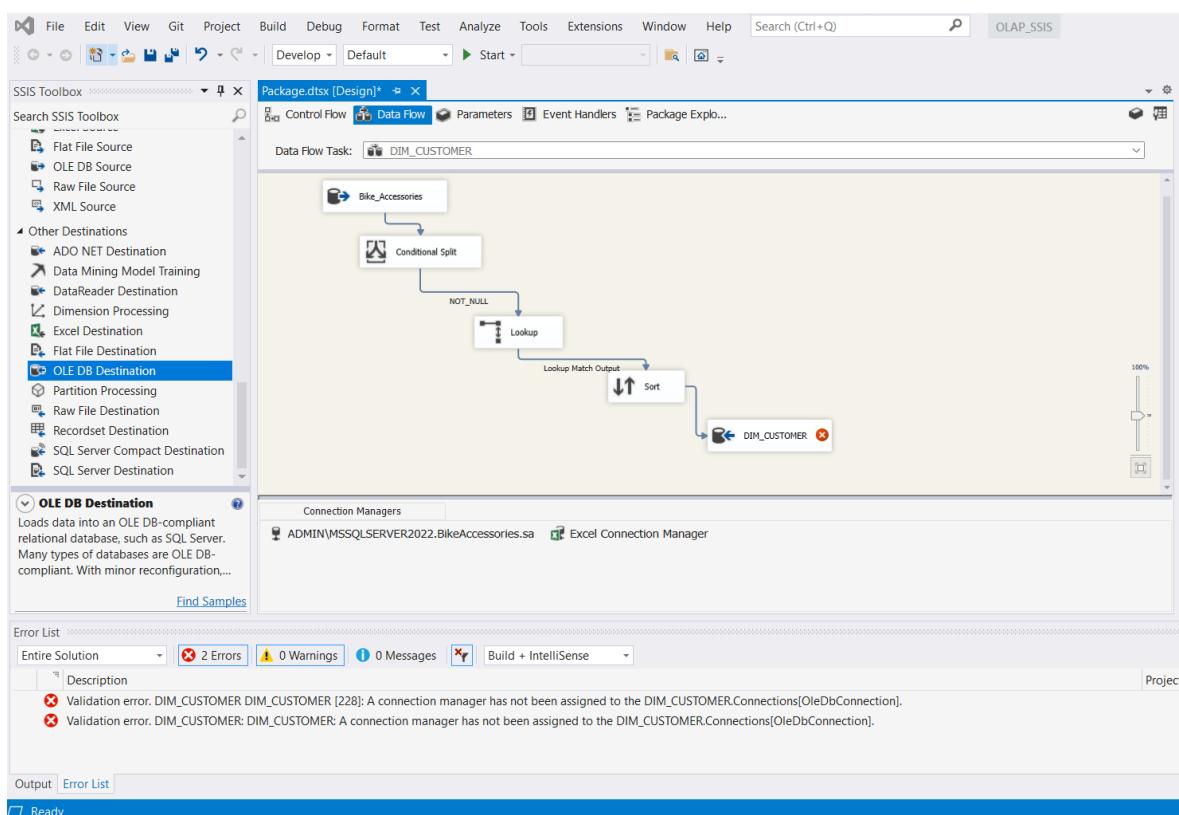
Kết nối Source và bảng DIM_CUS_CATEGORY.

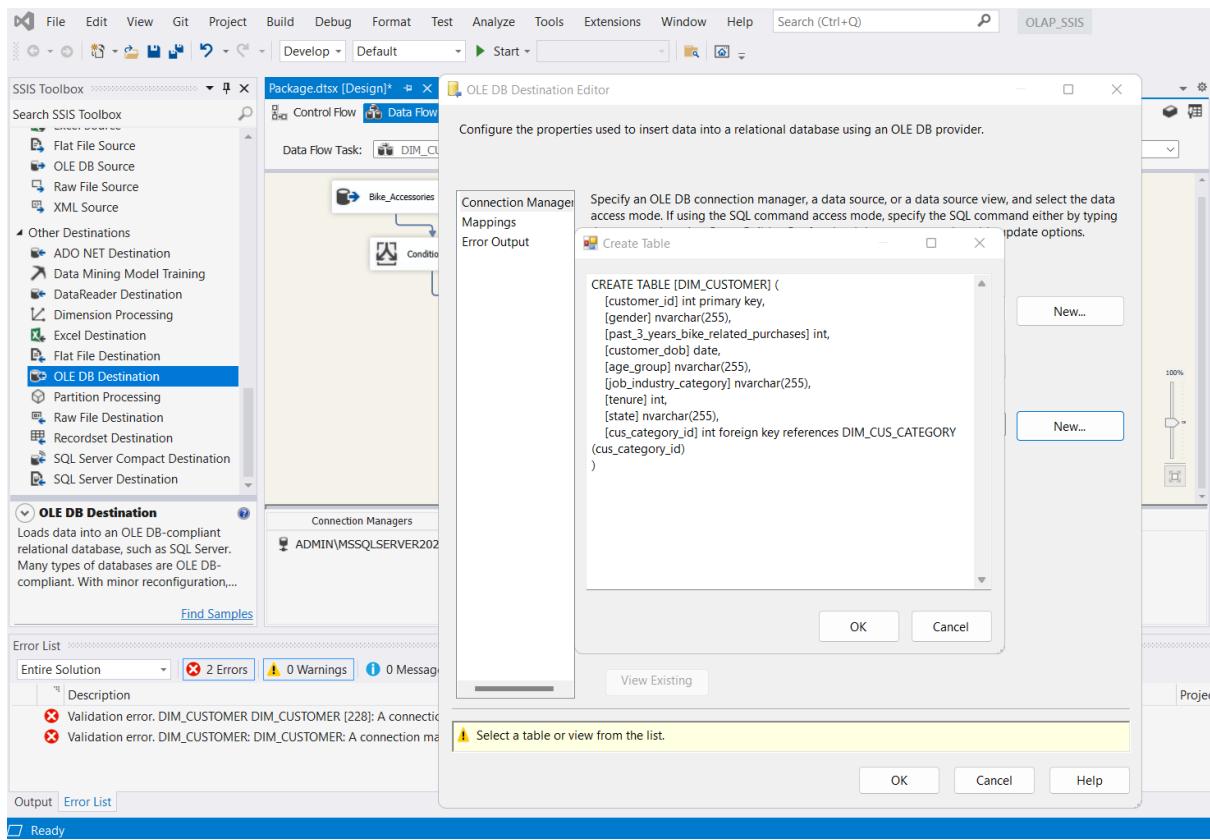


Bước 5: Dùng “Sort” để sắp xếp lại dữ liệu và tick vào ô “Remove rows with duplicate sort values” để xóa những dữ liệu bị trùng.

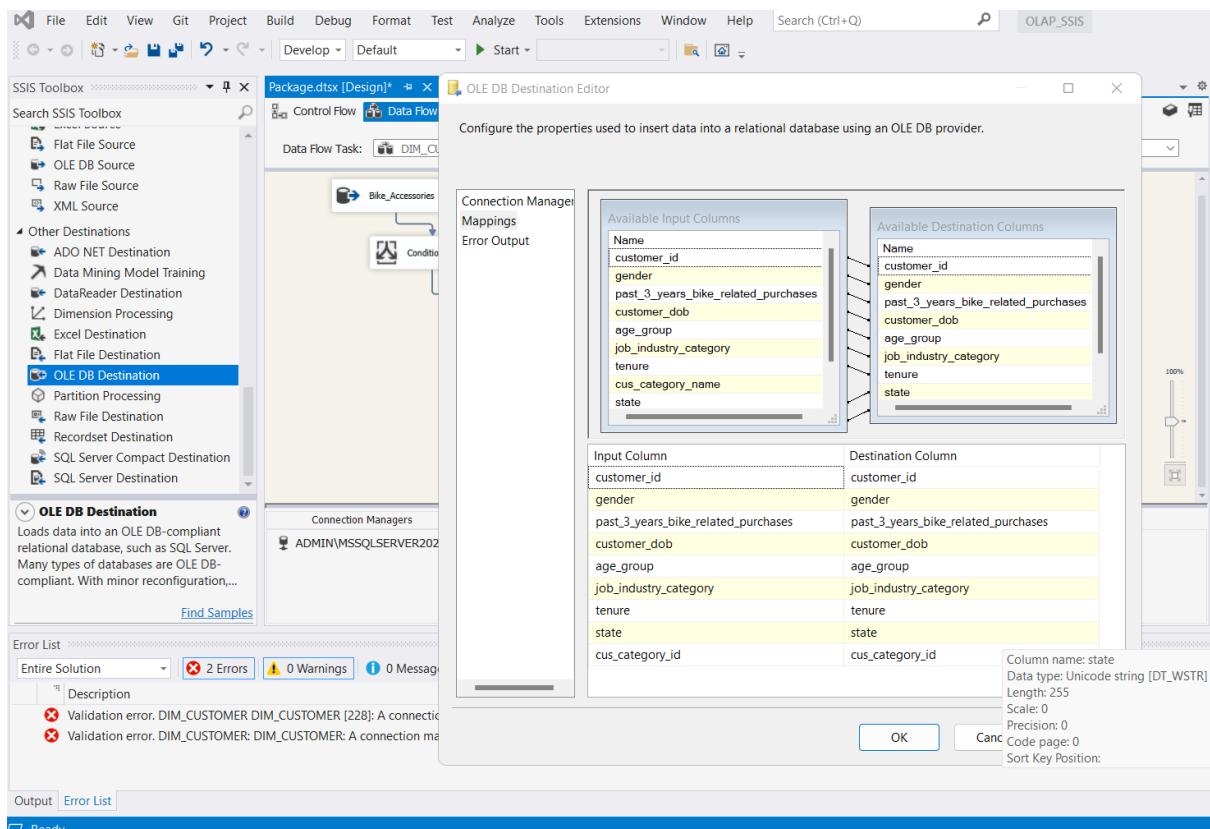


Bước 6: Tiếp theo khởi tạo “OLE DB Destination”, sau đó tạo bảng DIM_CUSTOMER.

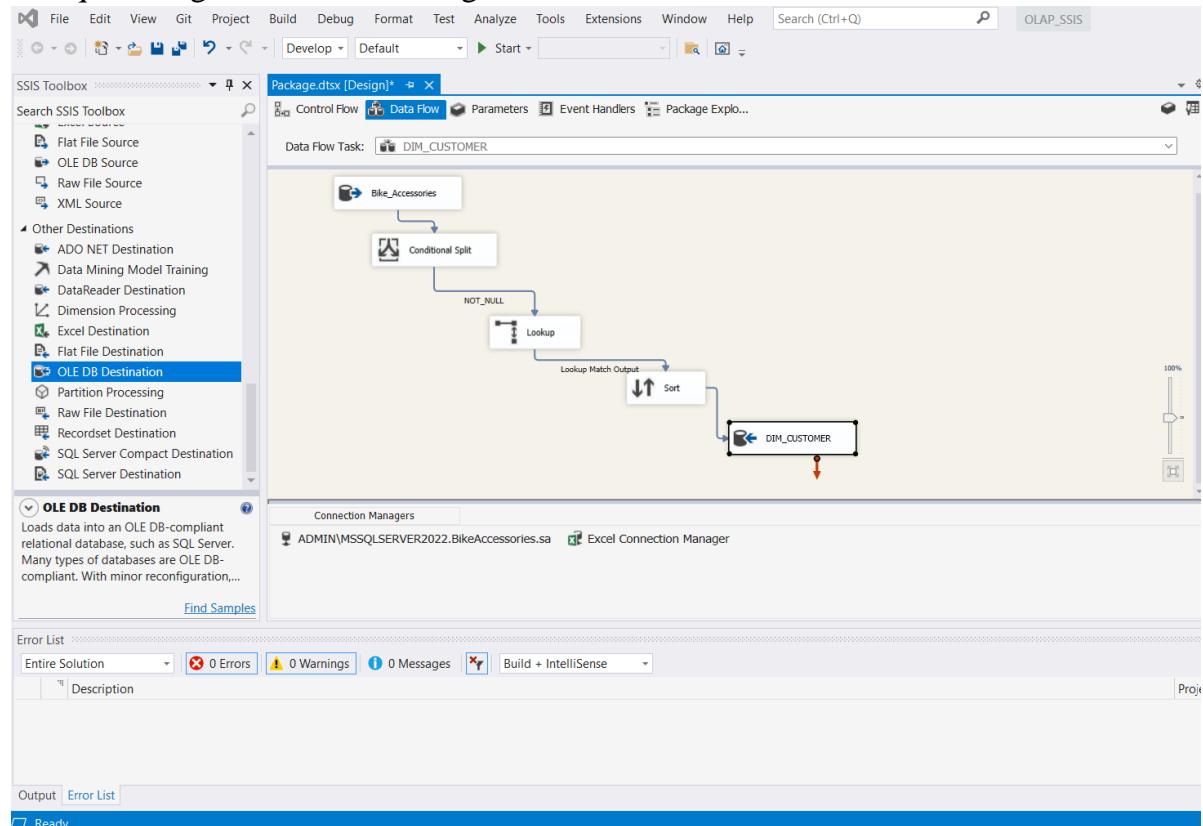




Bước 7: Qua tab “Mappings” để kiểm tra.

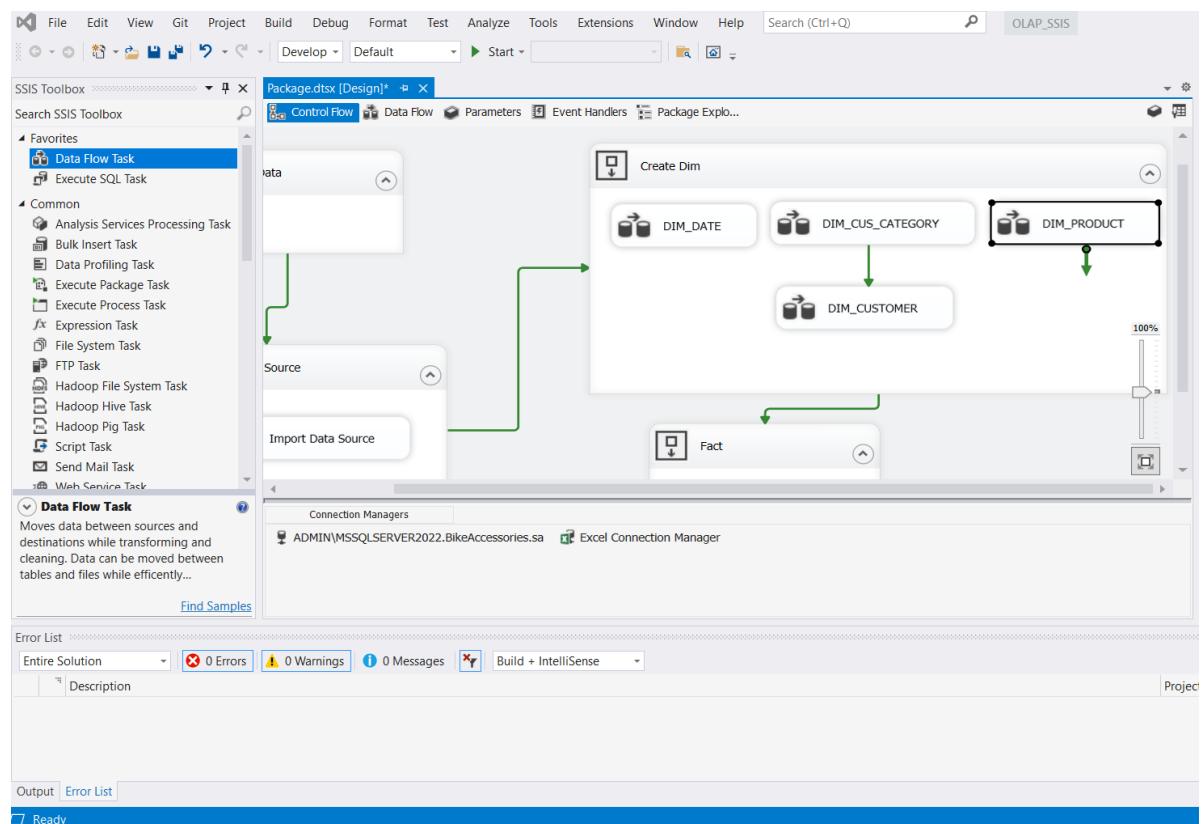


Kết quả luồng thực hiện của bảng DIM_CUSTOMER.

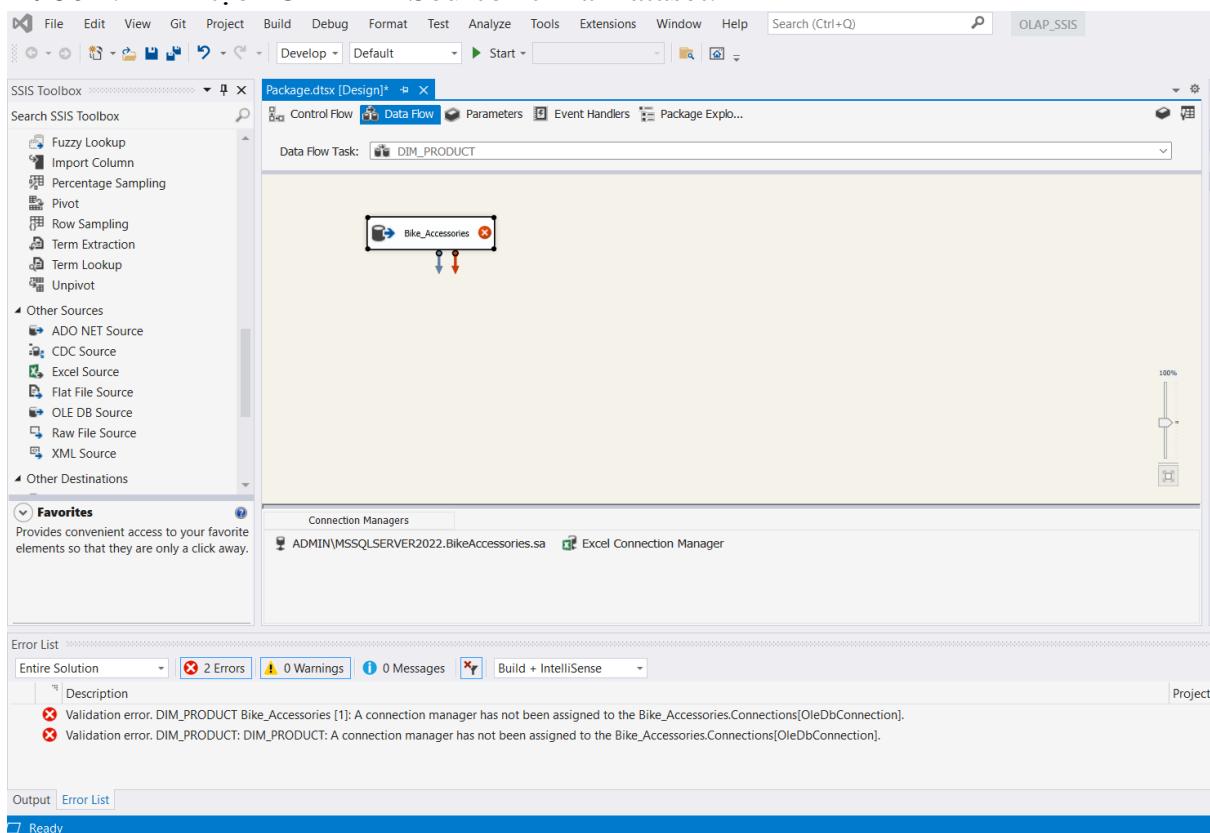


3.7 Tạo bảng DIM_PRODUCT

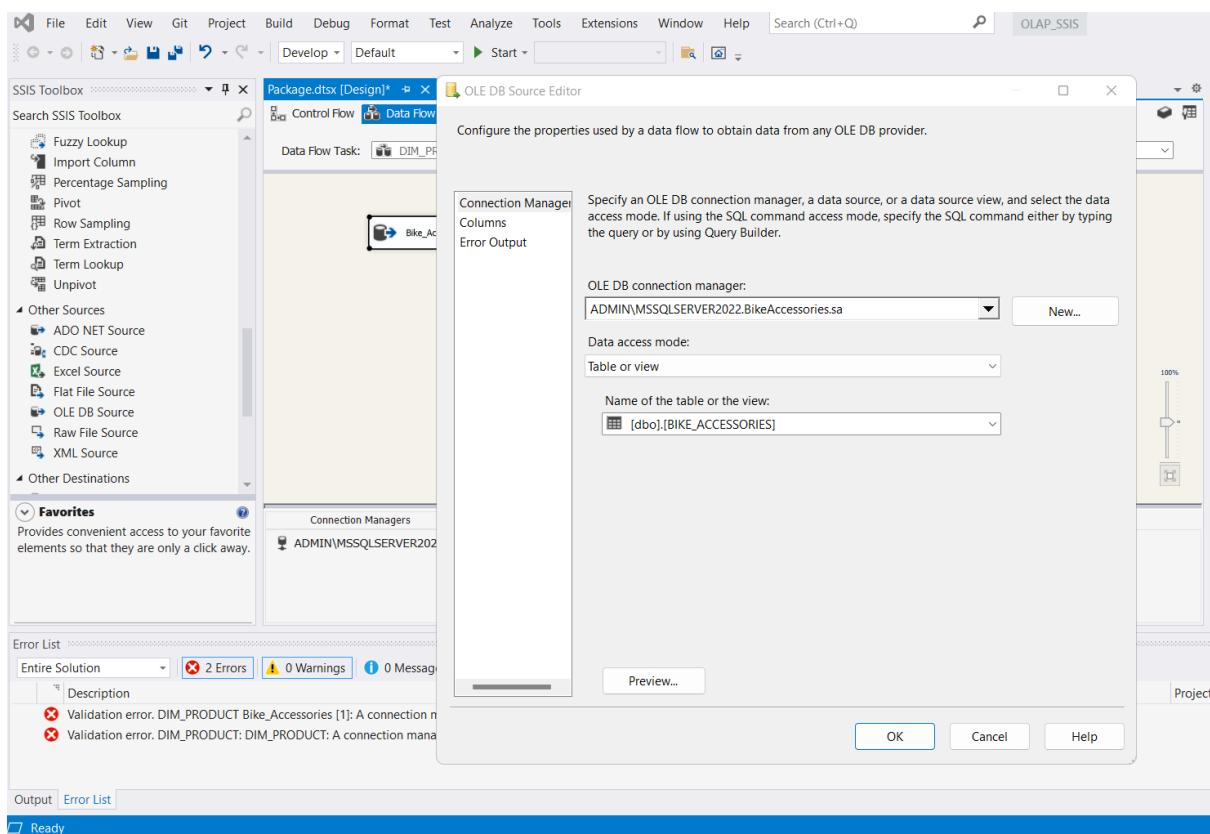
Bước 1: Kéo “Data Flow Task” vào Container đặt tên là DIM_PRODUCT.



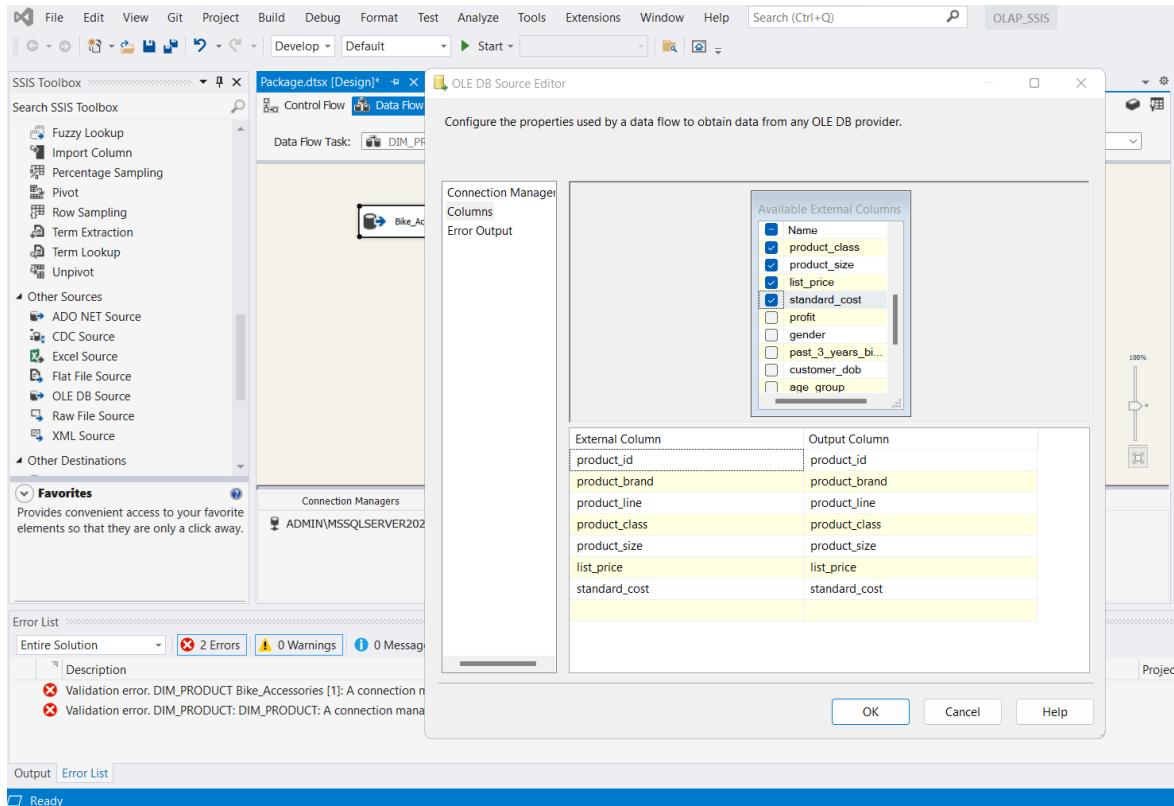
Bước 2: Khởi tạo “OLE DB Source” chứa Dataset.



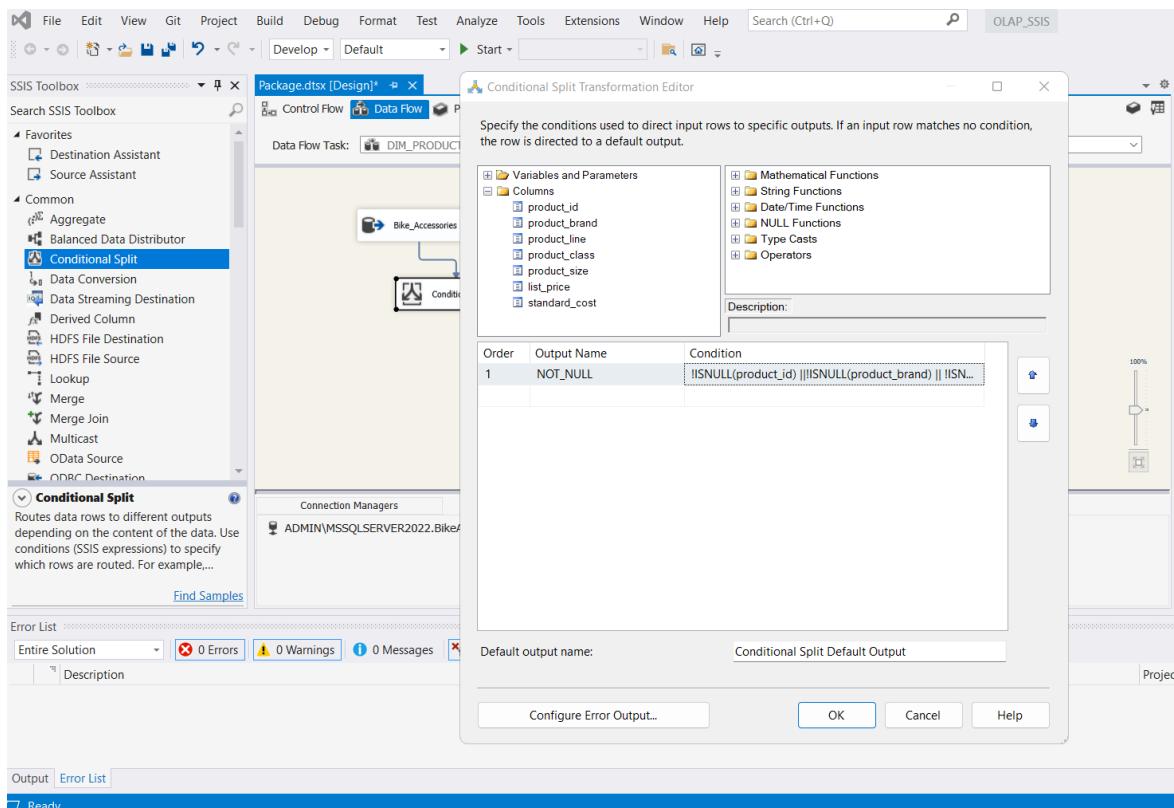
Bước 3: Chọn “Connection BikeAccessories” và chọn bảng BIKE_ACCESSORIES.



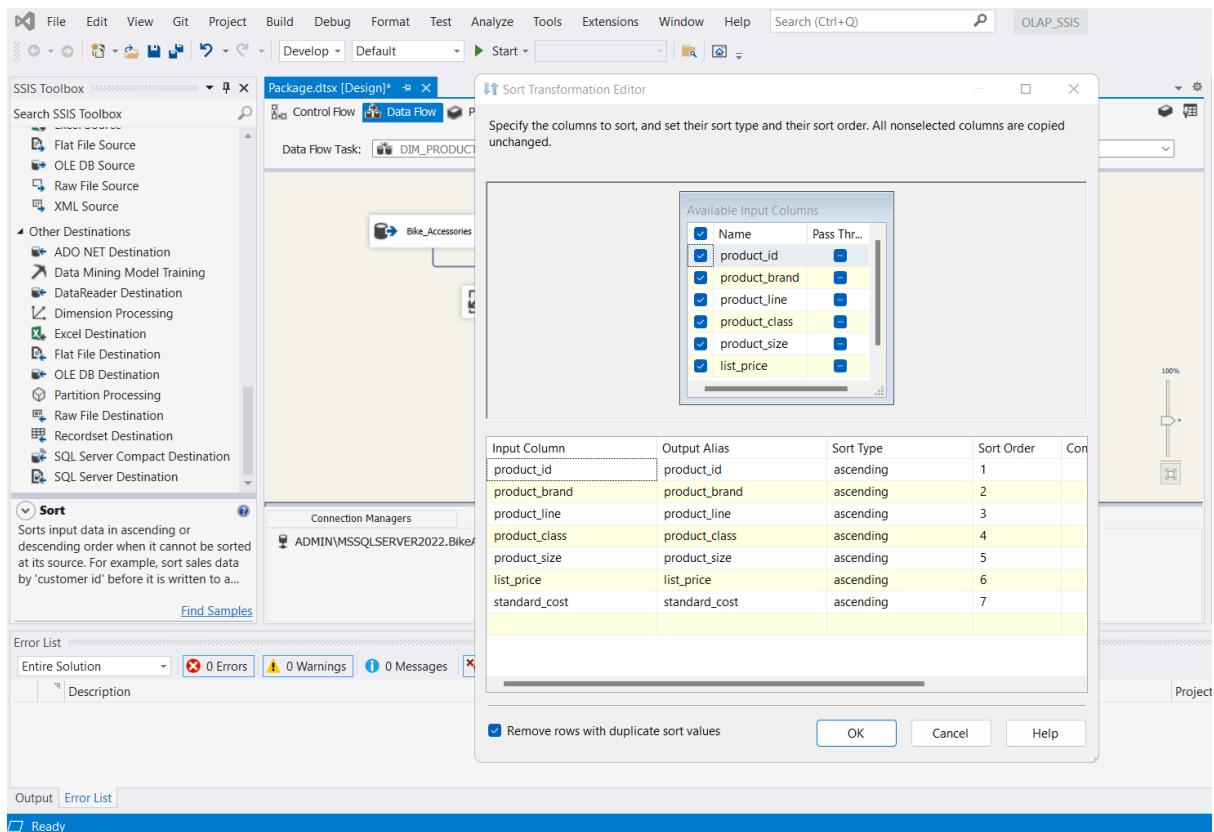
Bước 4: Preview bảng. Ở mục tab Column chọn thuộc tính thuộc bảng DIM_PRODUCT sau đó chọn OK.



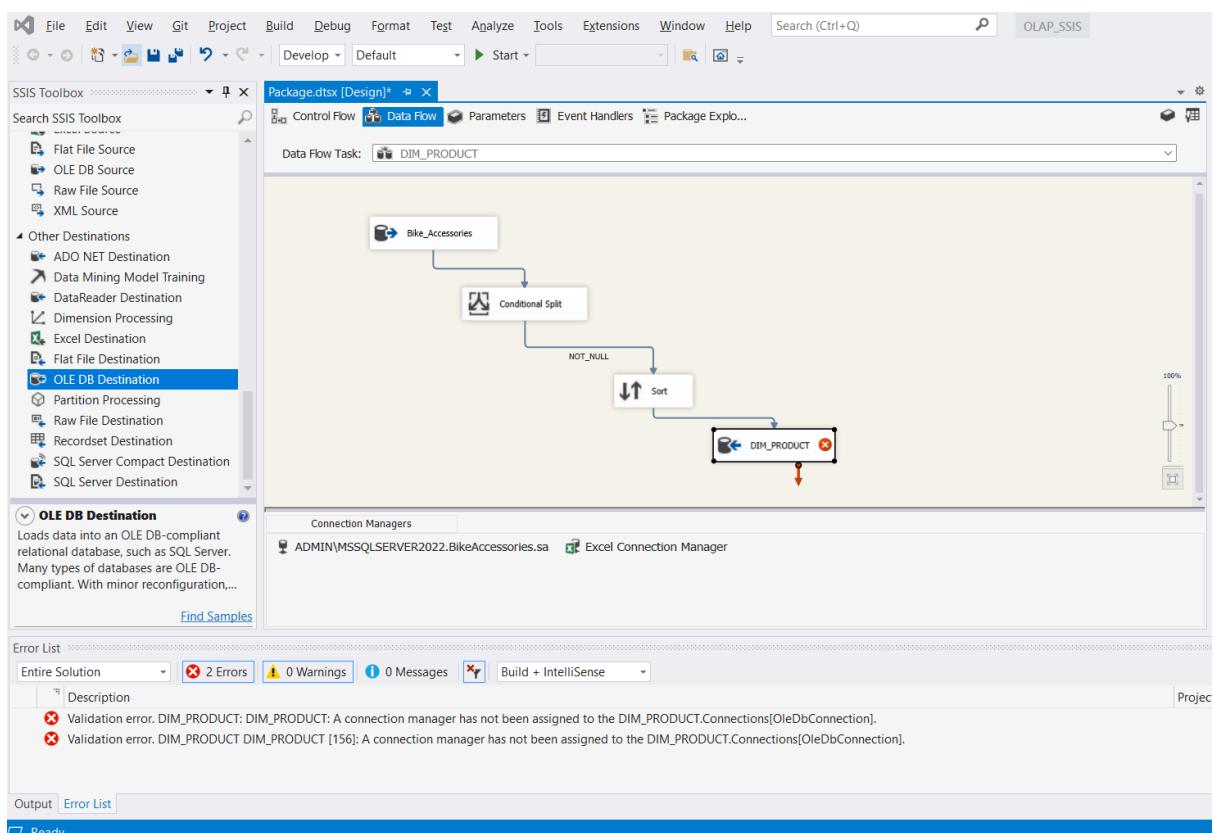
Bước 5: Dùng công cụ “Conditional Split” và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.

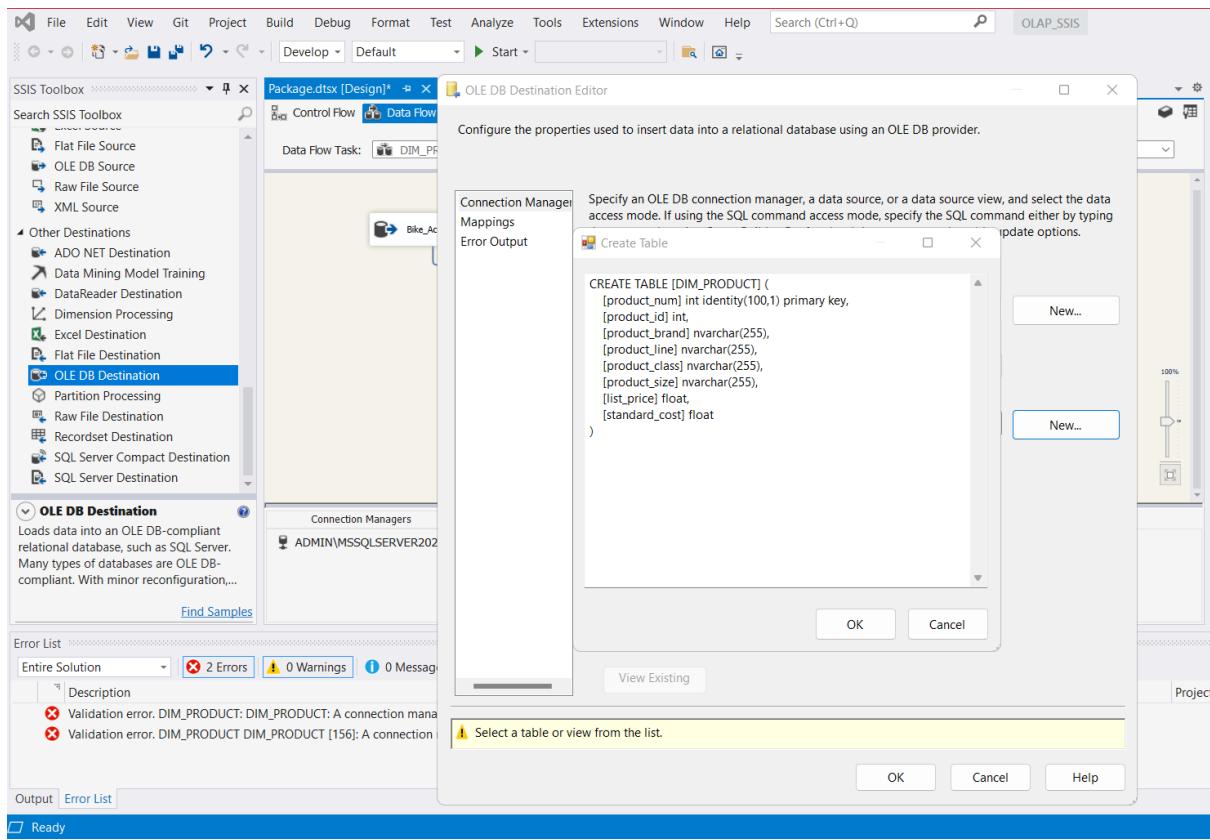


Bước 6: Dùng “Sort” để sắp xếp lại dữ liệu và tick vào ô “Remove rows with duplicate sort values” để xóa những dữ liệu bị trùng.

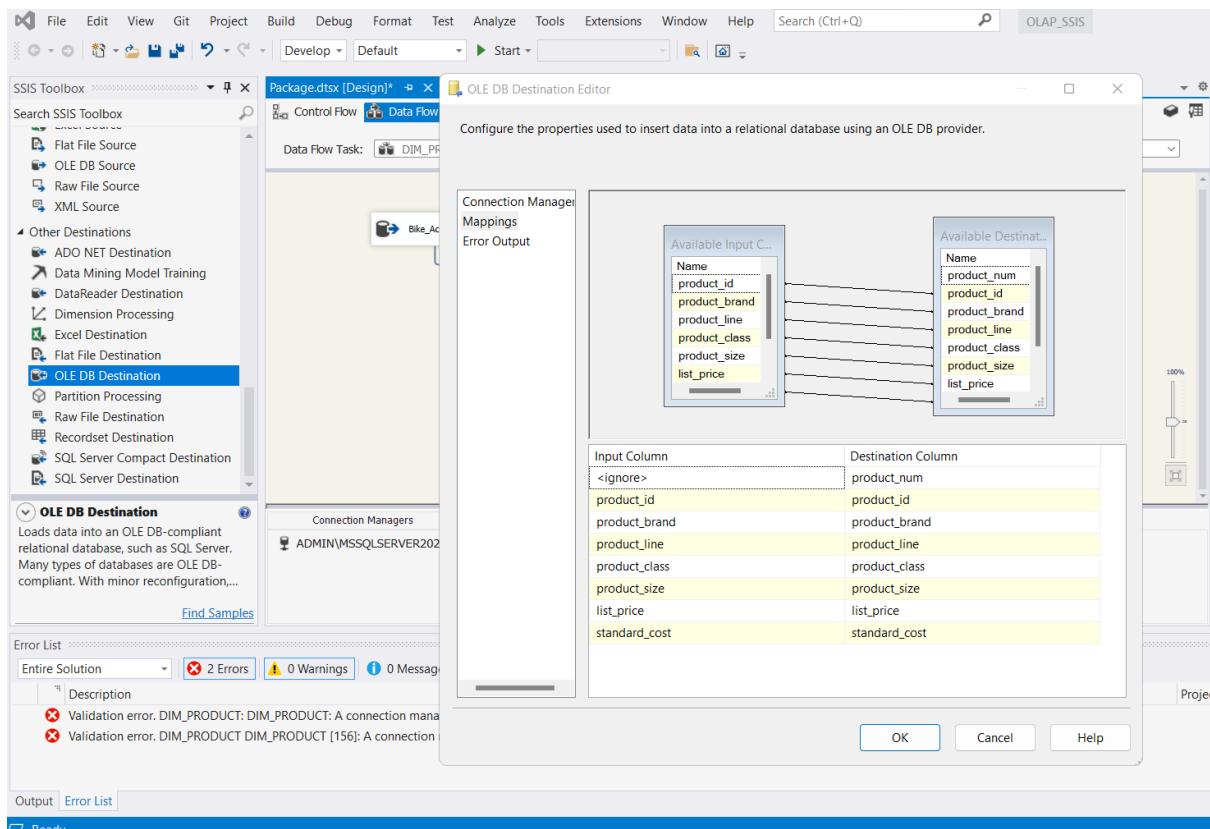


Bước 7: Tạo một “OLE Destination” sau đó tạo bảng DIM_PRODUCT.

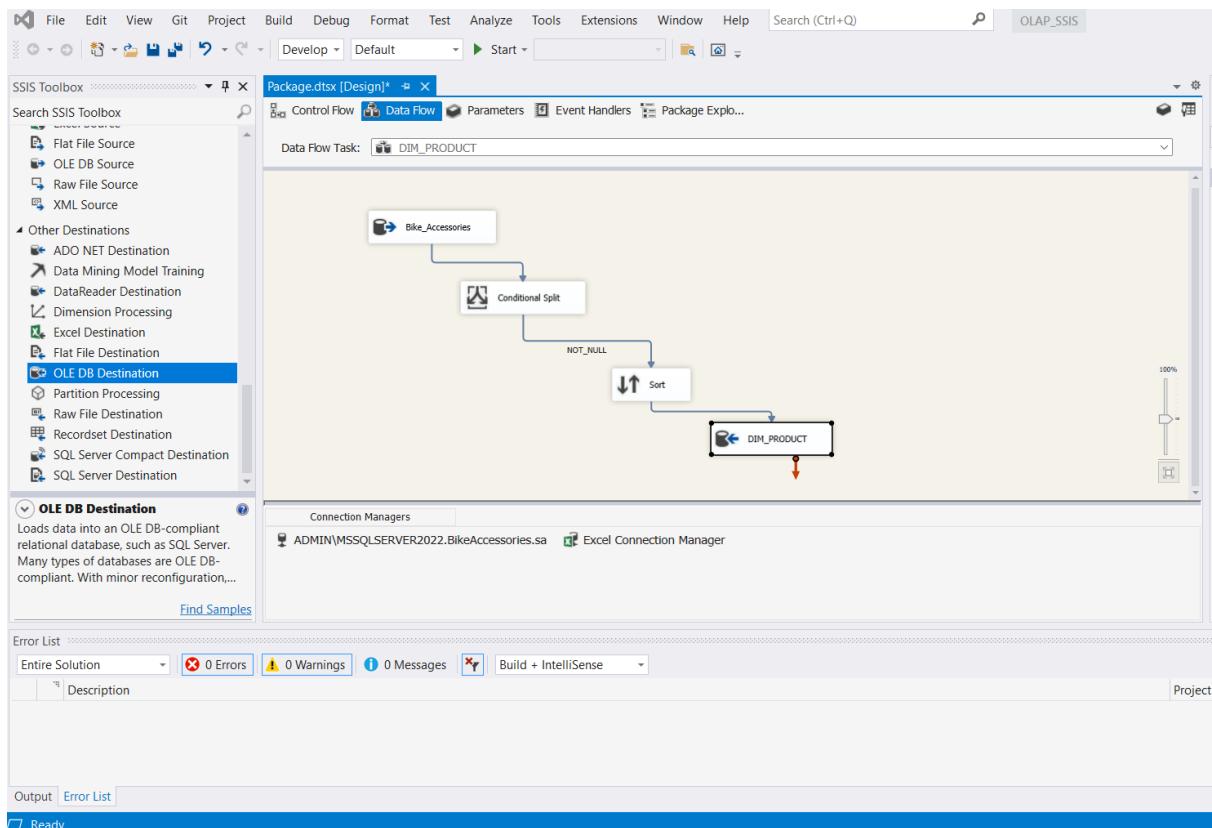




Bước 8: Qua tab “Mappings” để kiểm tra.

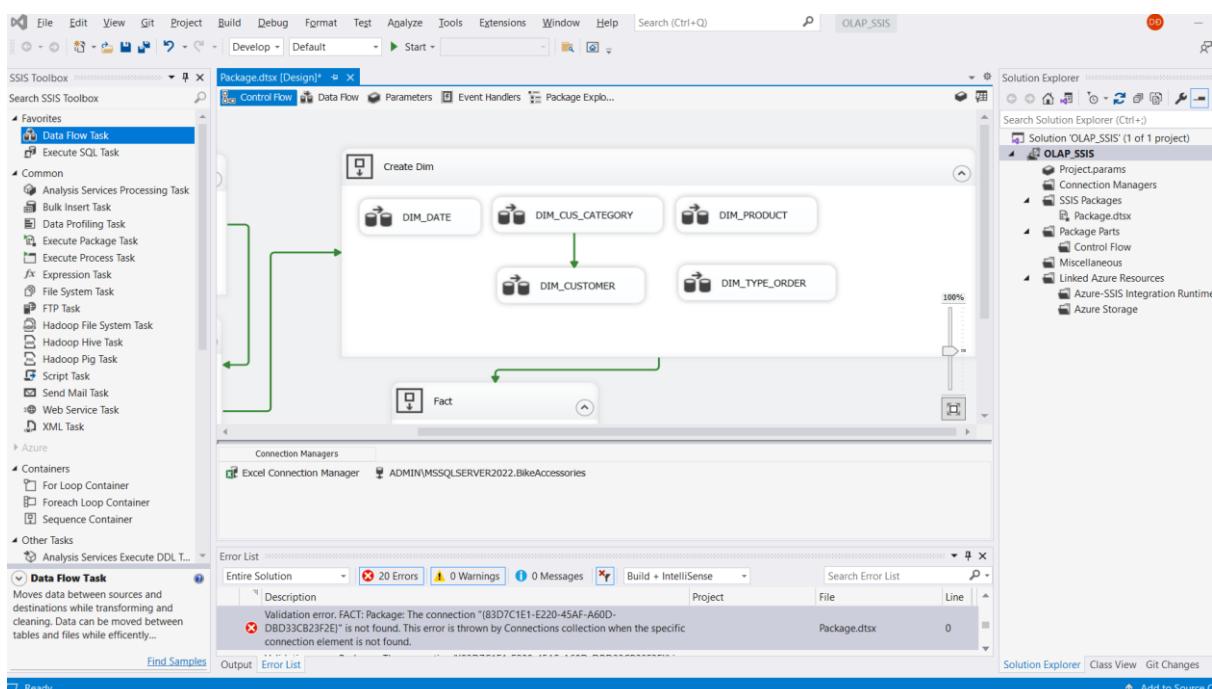


Kết quả luồng thực hiện của bảng DIM_PRODUCT.

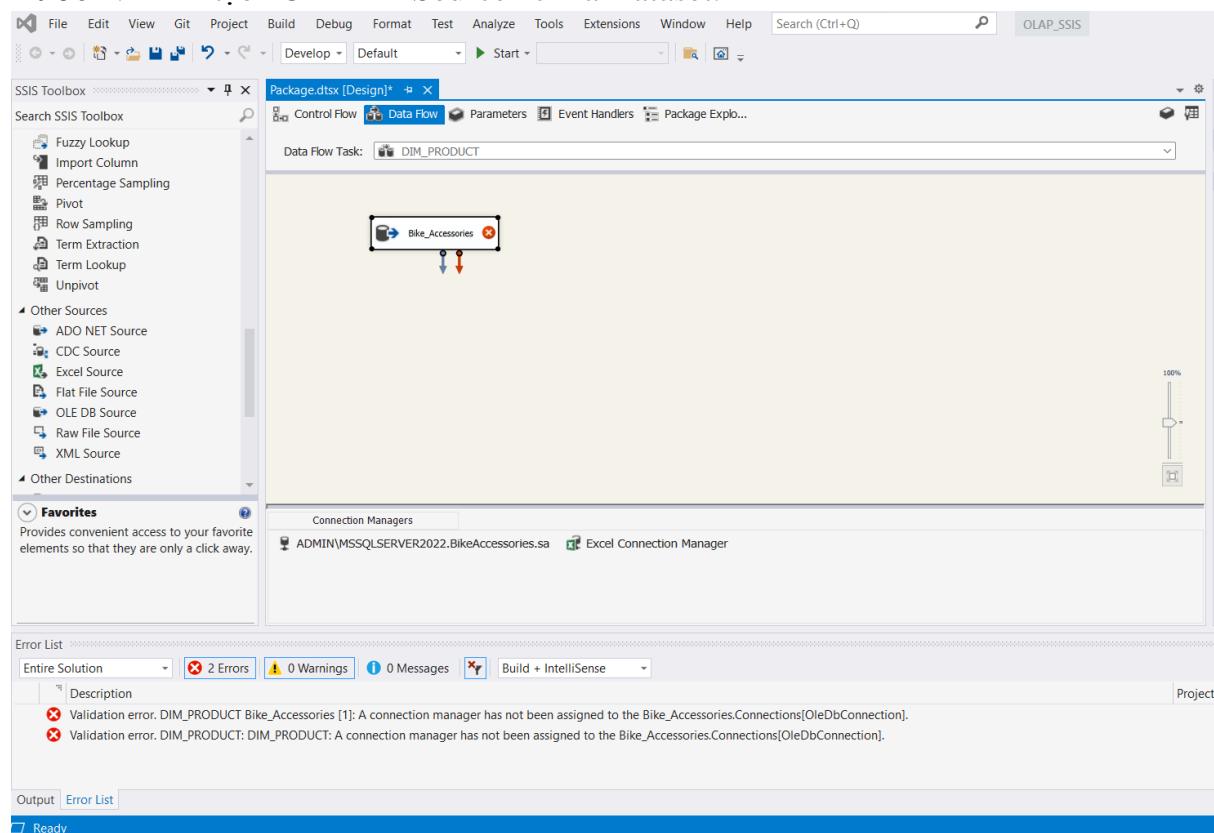


3.8 Tạo bảng DIM_TYPE_ORDER

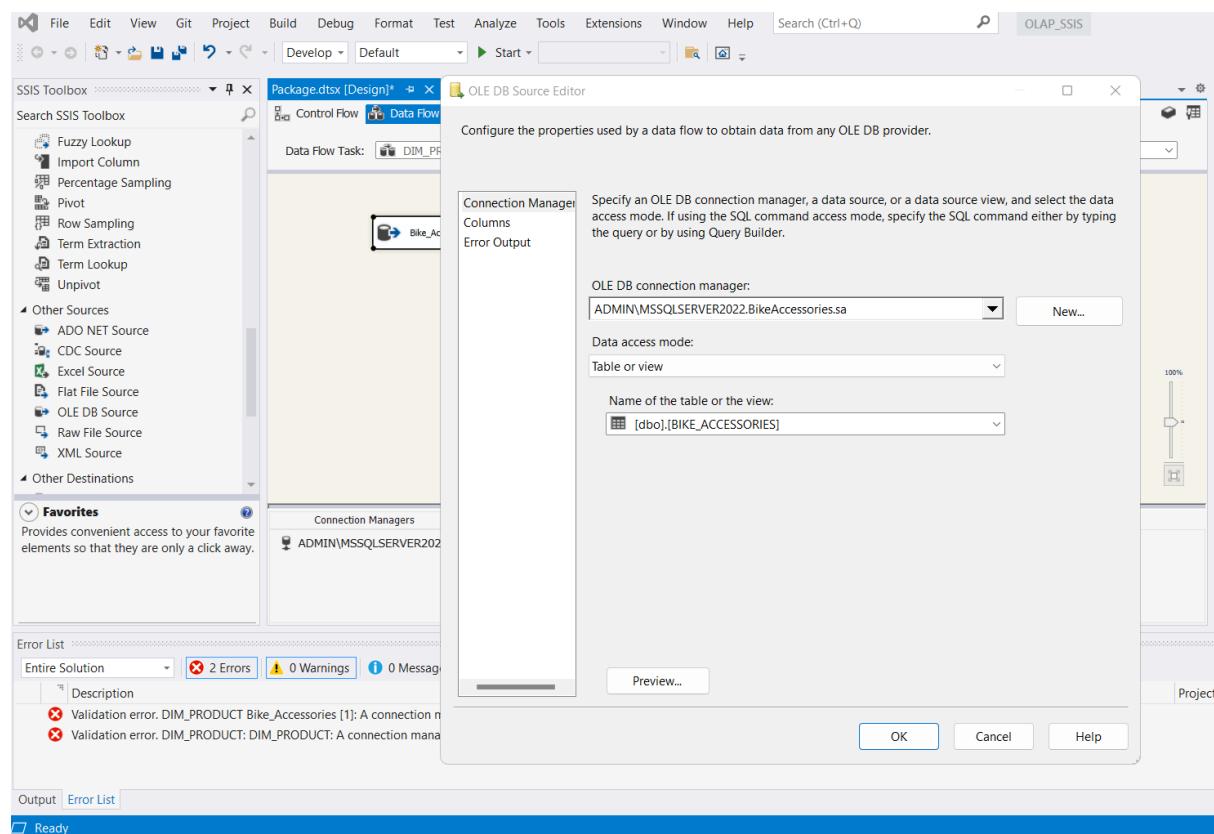
Bước 1: Kéo “Data Flow Task” vào Container đặt tên là DIM_TYPE_ORDER.



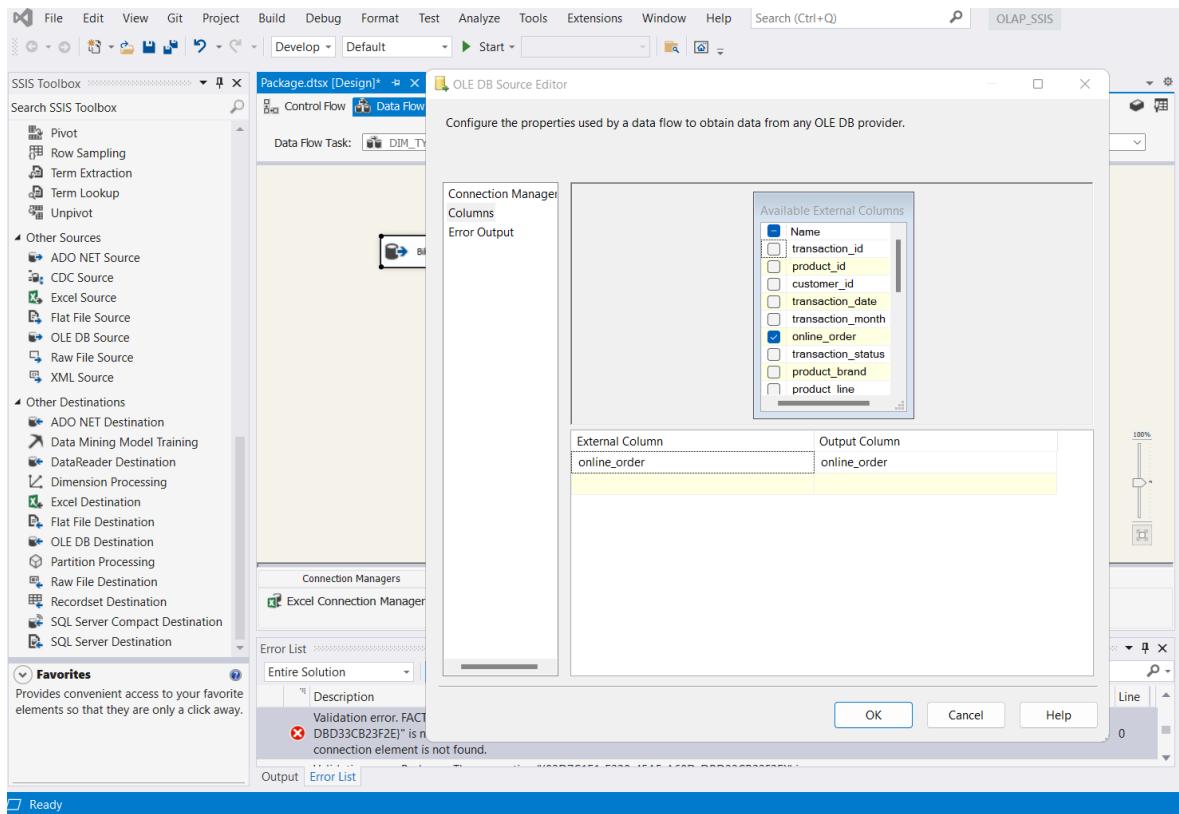
Bước 2: Khởi tạo “OLE DB Source” chứa Dataset.



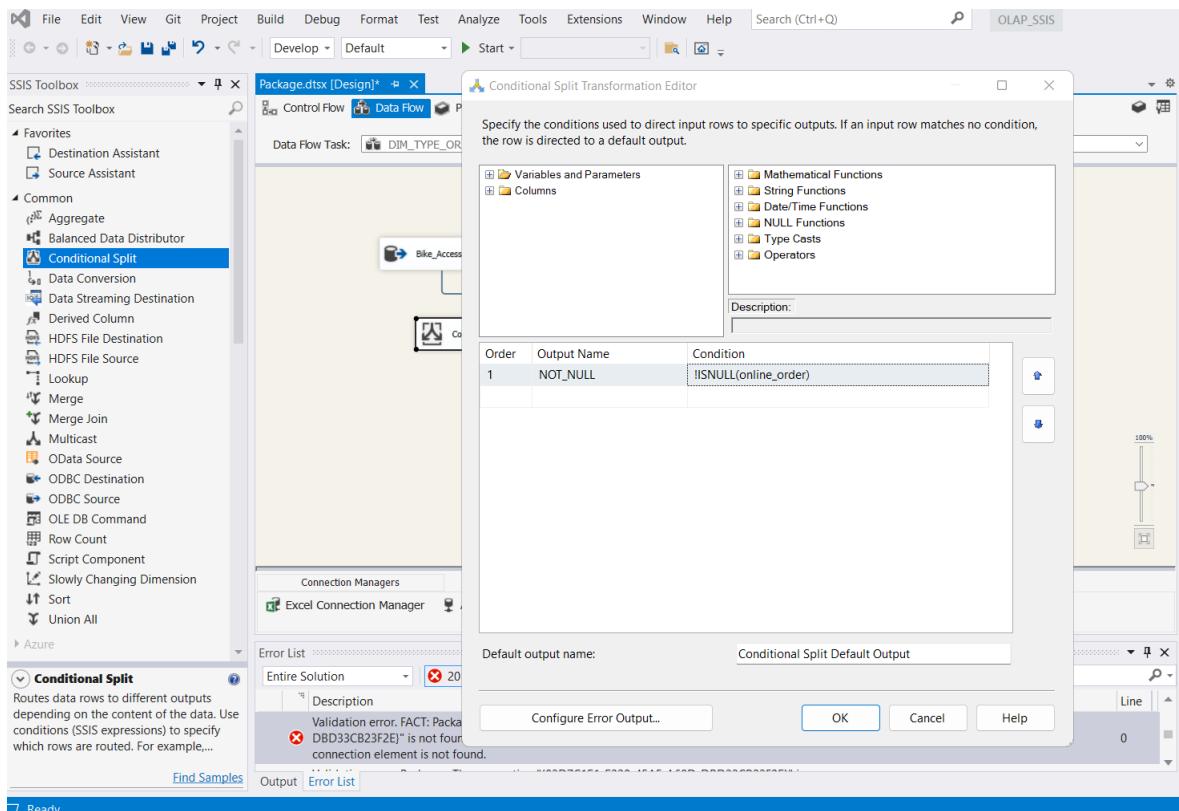
Bước 3: Chọn “Connection BikeAccessories” và chọn bảng BIKE_ACCESSORIES.



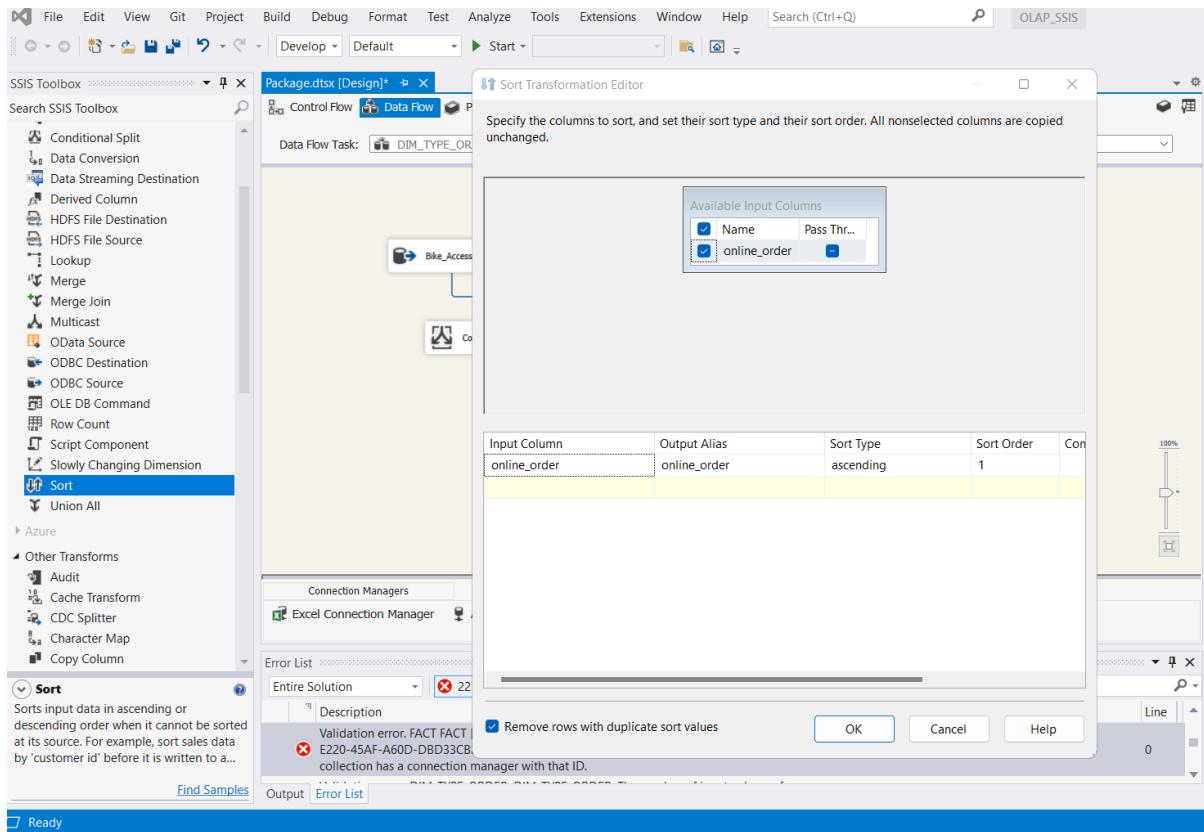
Bước 4: Preview bảng. Ở mục tab Column chọn thuộc tính thuộc bảng DIM_TYPE_ORDER sau đó chọn OK.



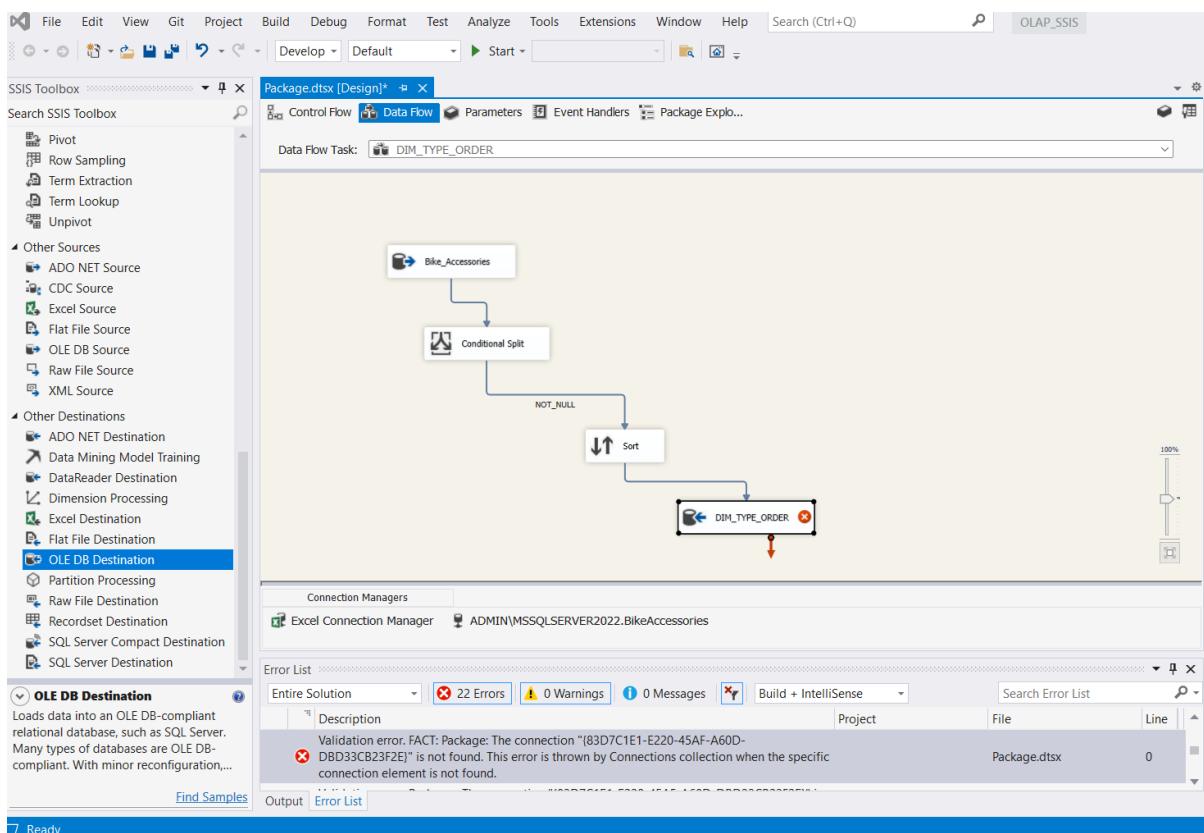
Bước 5: Dùng công cụ “Conditional Split” và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.

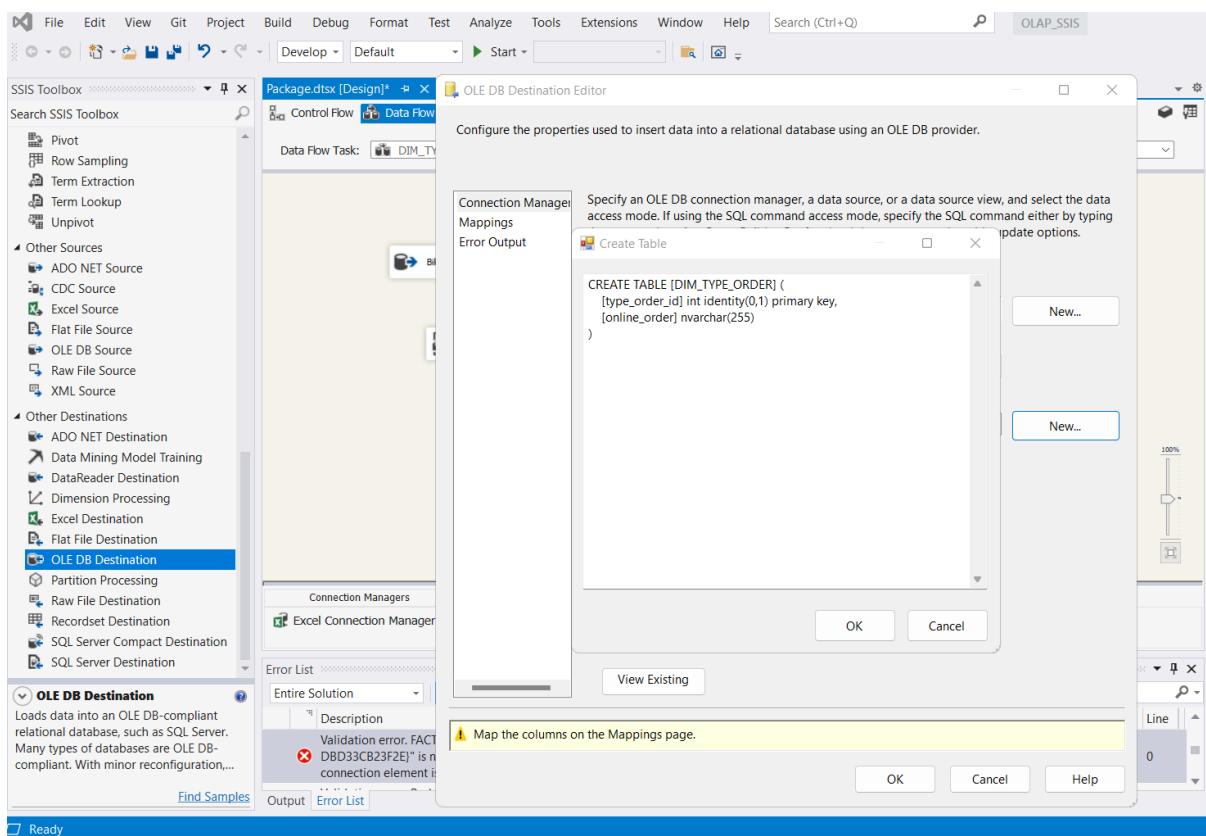


Bước 6: Dùng “Sort” để sắp xếp lại dữ liệu và tick vào ô “Remove rows with duplicate sort values” để xóa những dữ liệu bị trùng.

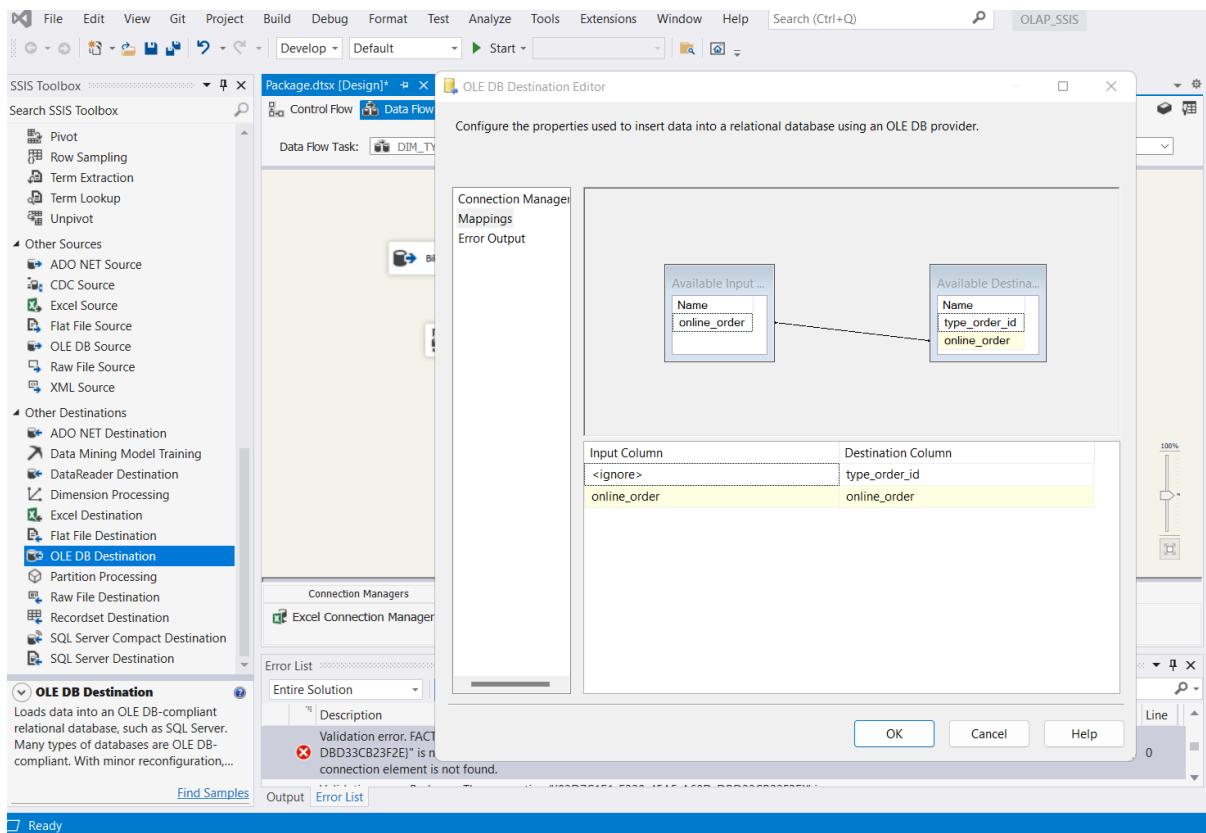


Bước 7: Tạo một “OLE Destination” sau đó tạo bảng DIM_TYPE_ORDER.

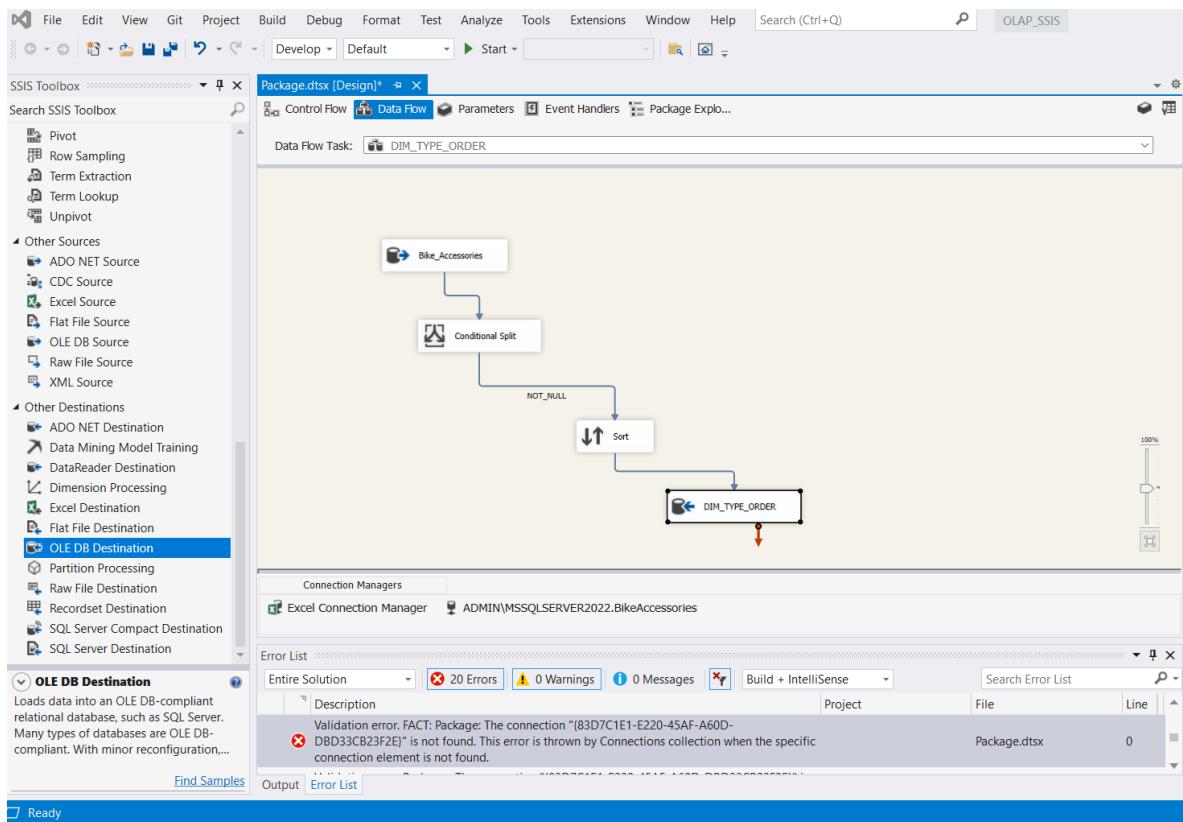




Bước 8: Qua tab “Mappings” để kiểm tra.

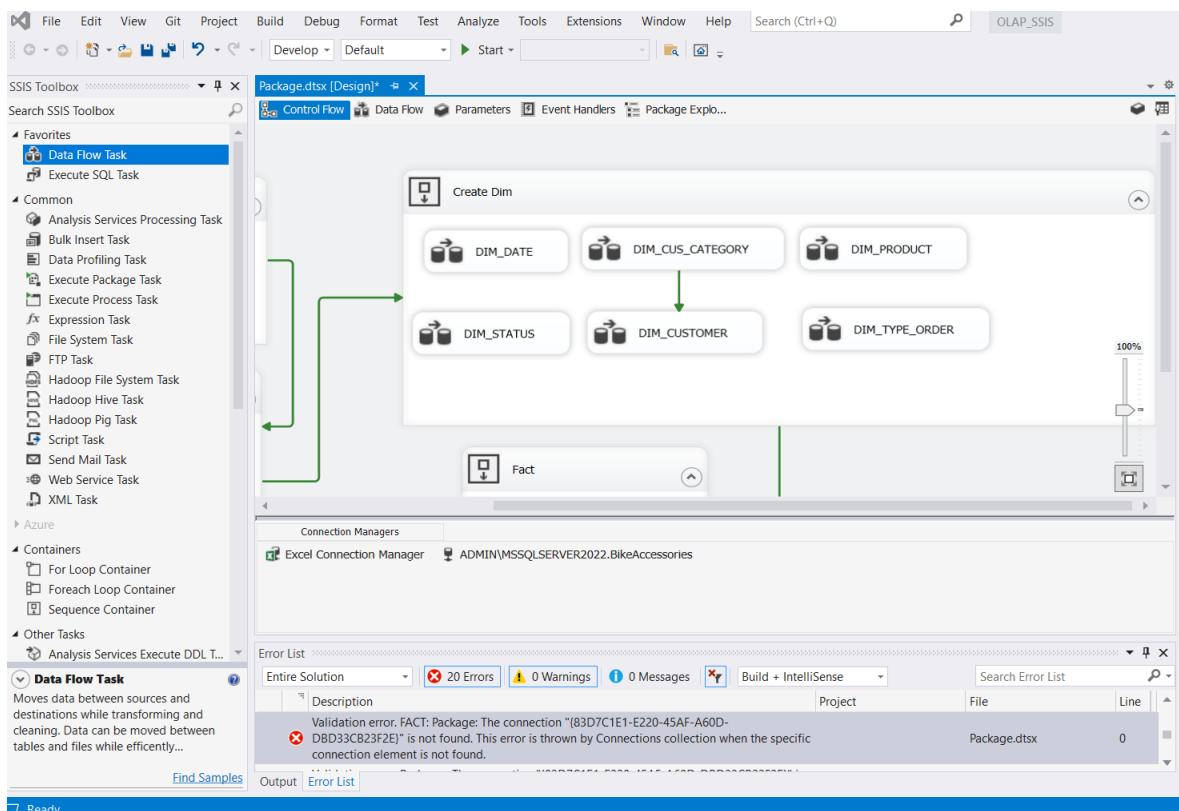


Kết quả luồng thực hiện của bảng DIM_TYPE_ORDER.

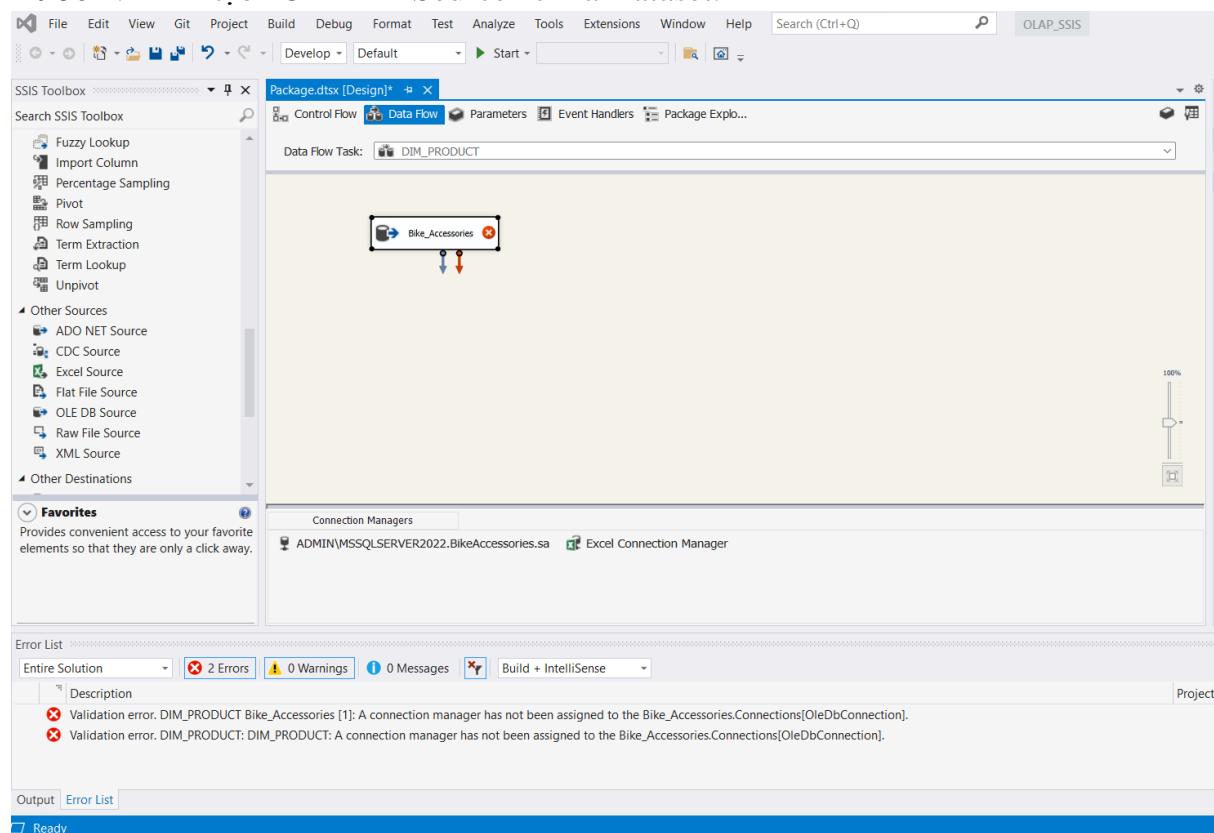


3.9 Tạo bảng DIM_STATUS

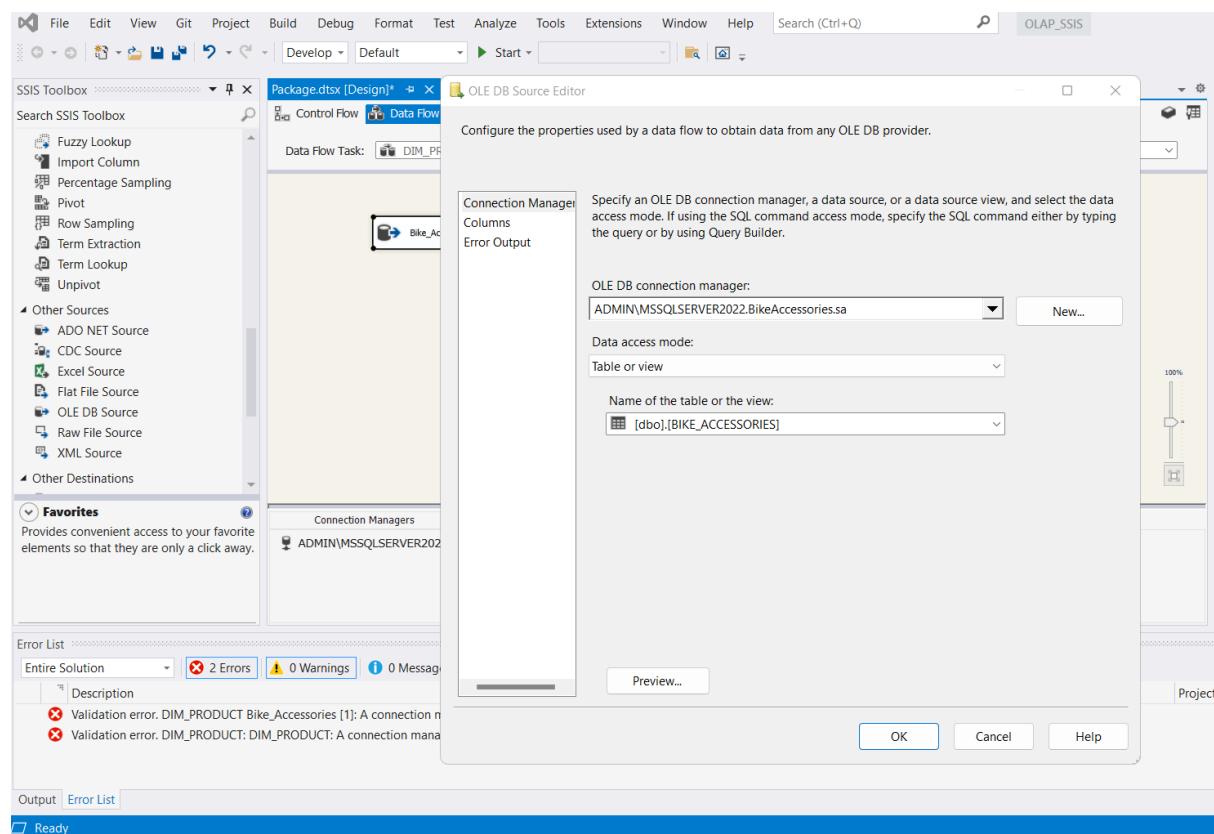
Bước 1: Kéo “Data Flow Task” vào Container đặt tên là DIM_STATUS.



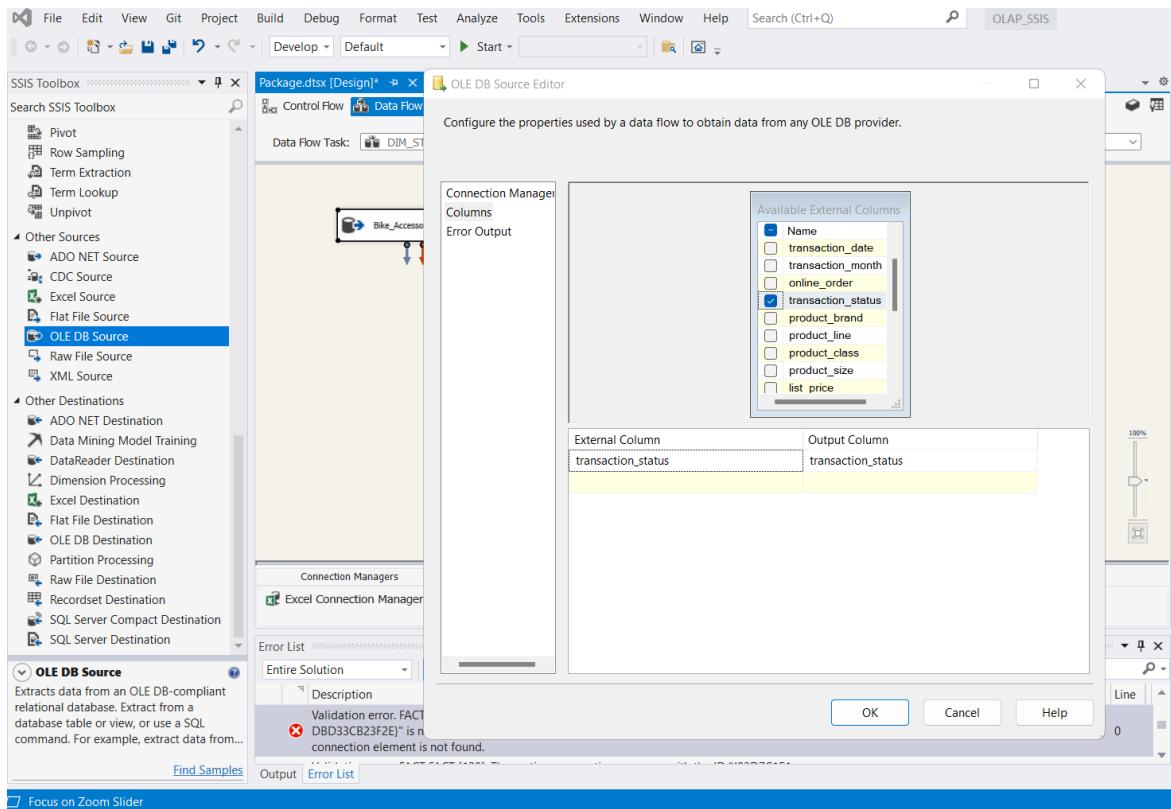
Bước 2: Khởi tạo “OLE DB Source” chứa Dataset.



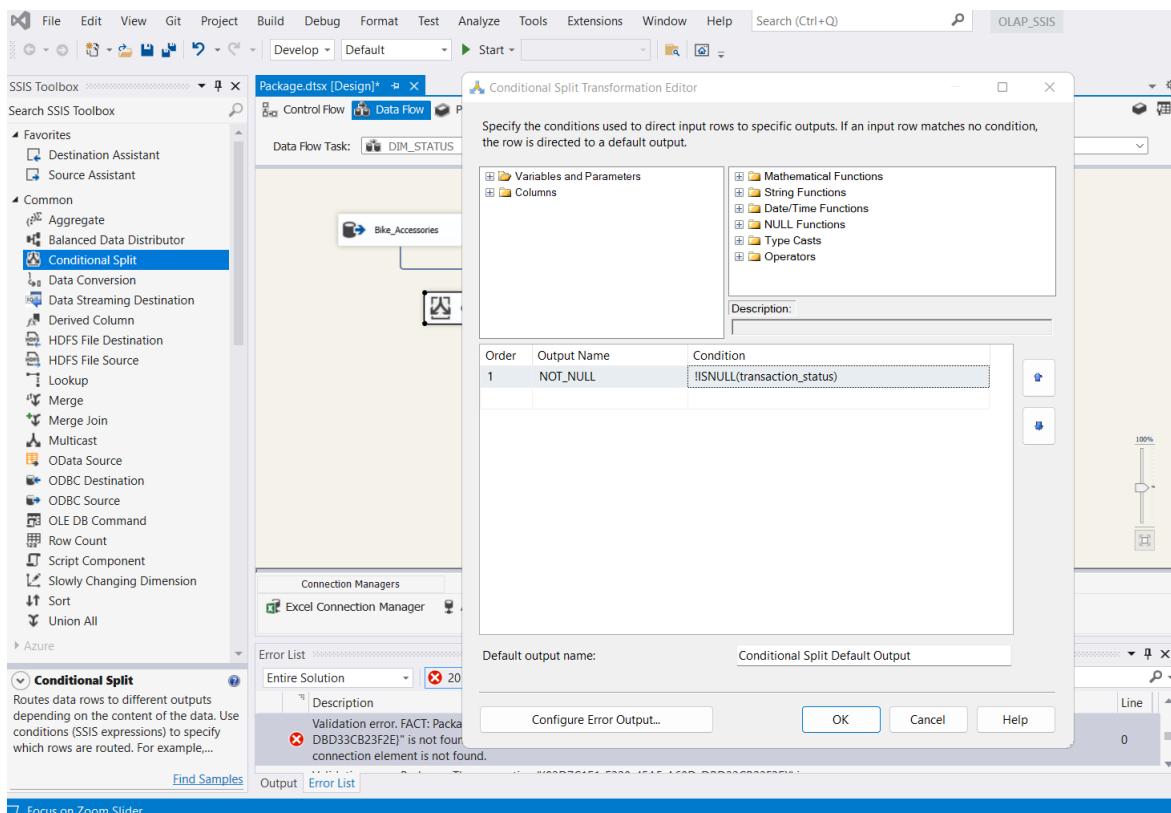
Bước 3: Chọn “Connection BikeAccessories” và chọn bảng BIKE_ACCESSORIES.



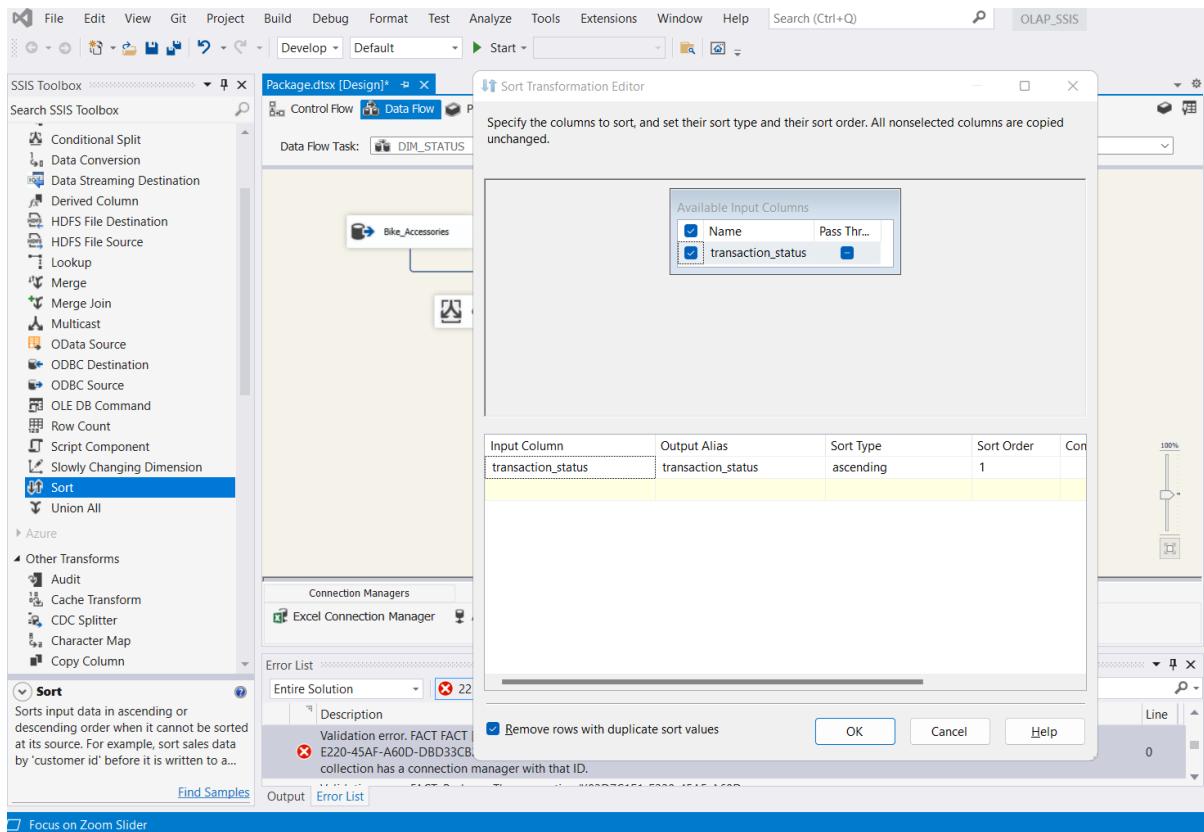
Bước 4: Preview bảng. Ở mục tab Column chọn thuộc tính thuộc bảng DIM_STATUS sau đó chọn OK.



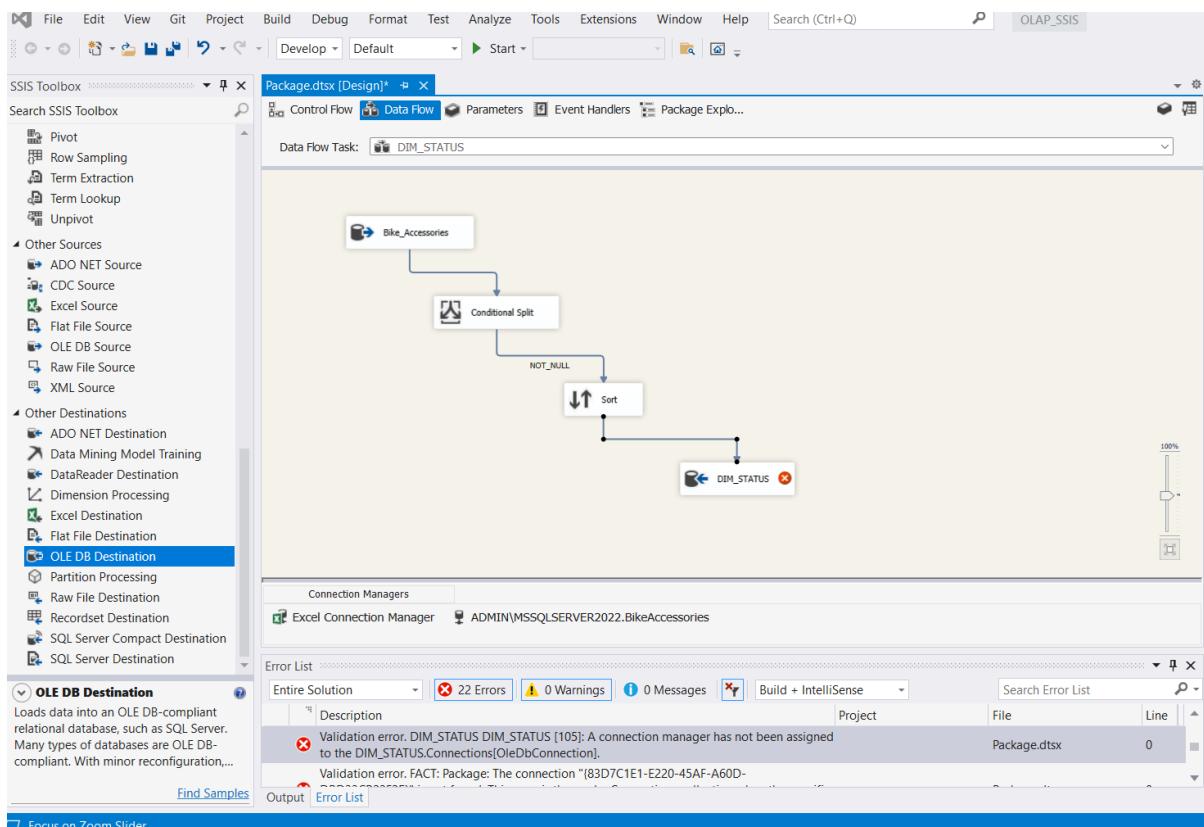
Bước 5: Dùng công cụ “Conditional Split” và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.

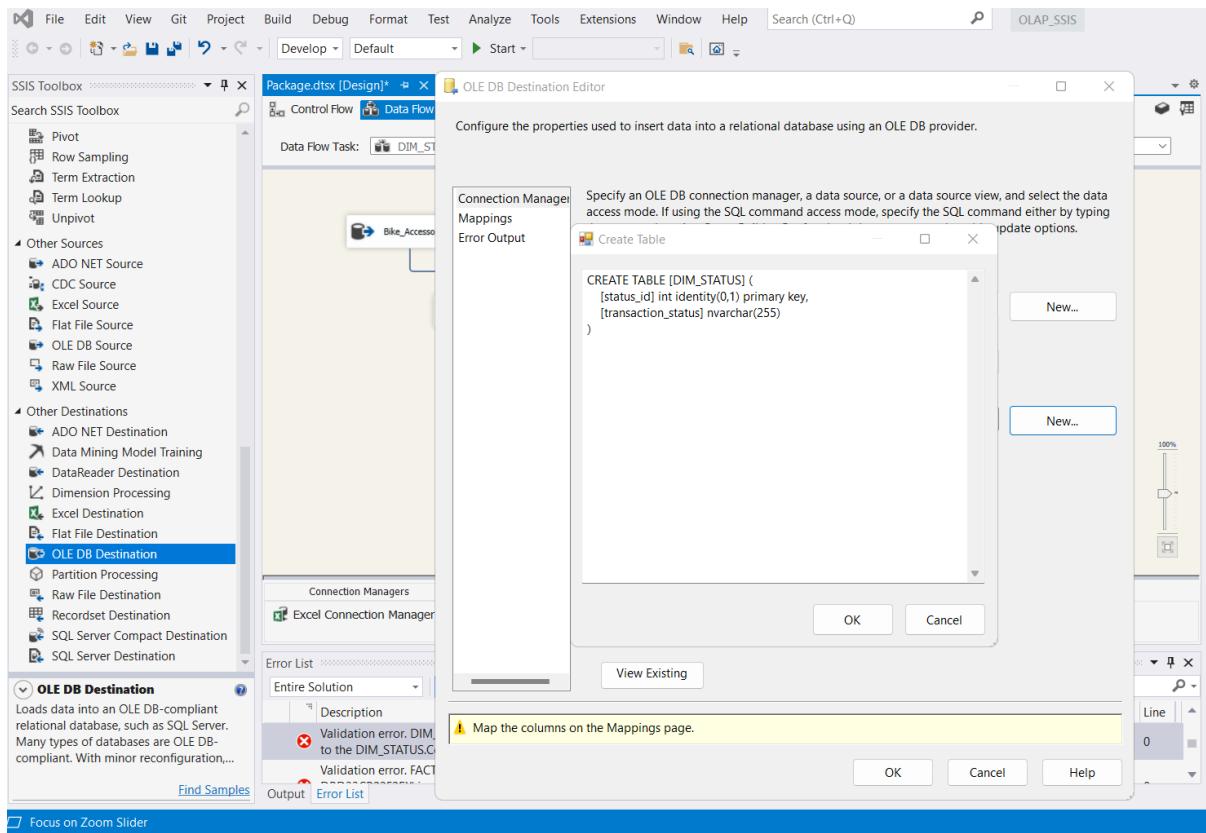


Bước 6: Dùng “Sort” để sắp xếp lại dữ liệu và tick vào ô “Remove rows with duplicate sort values” để xóa những dữ liệu bị trùng.

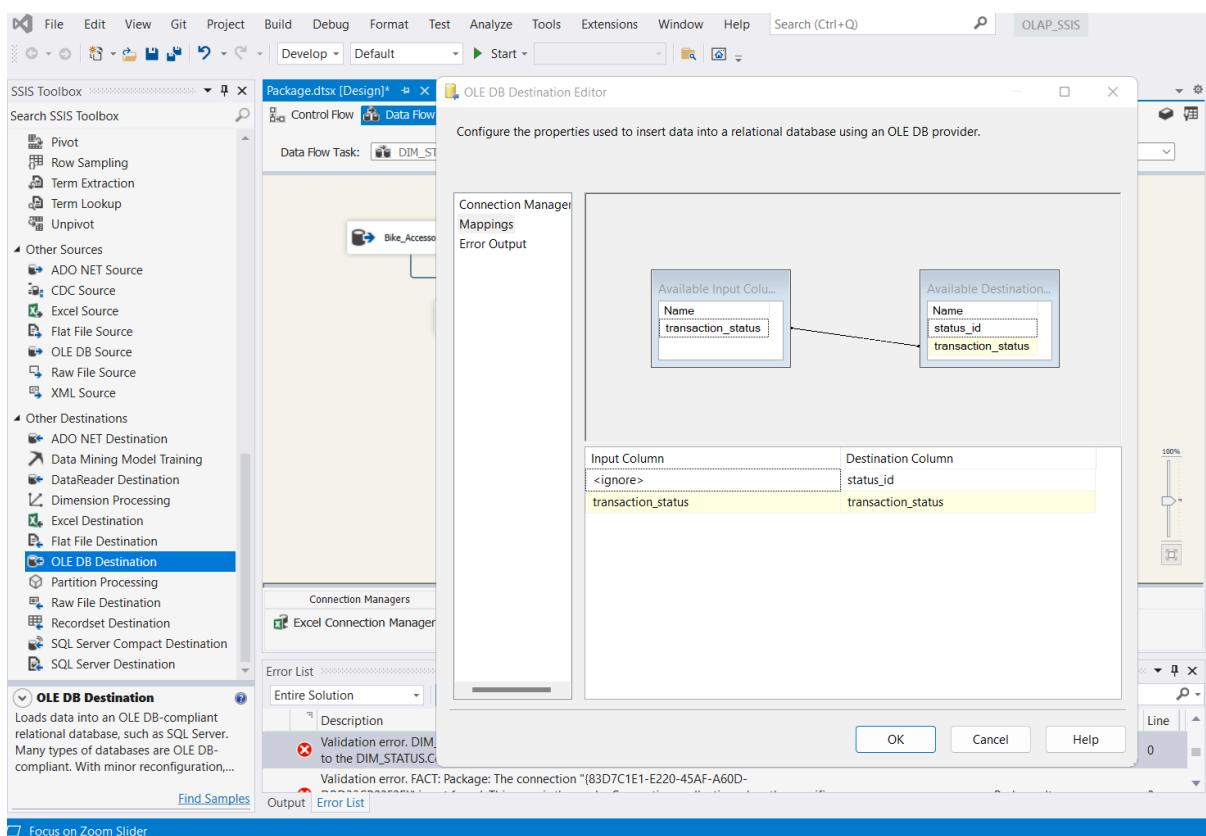


Bước 7: Tạo một “OLE Destination” sau đó tạo bảng DIM_STATUS.

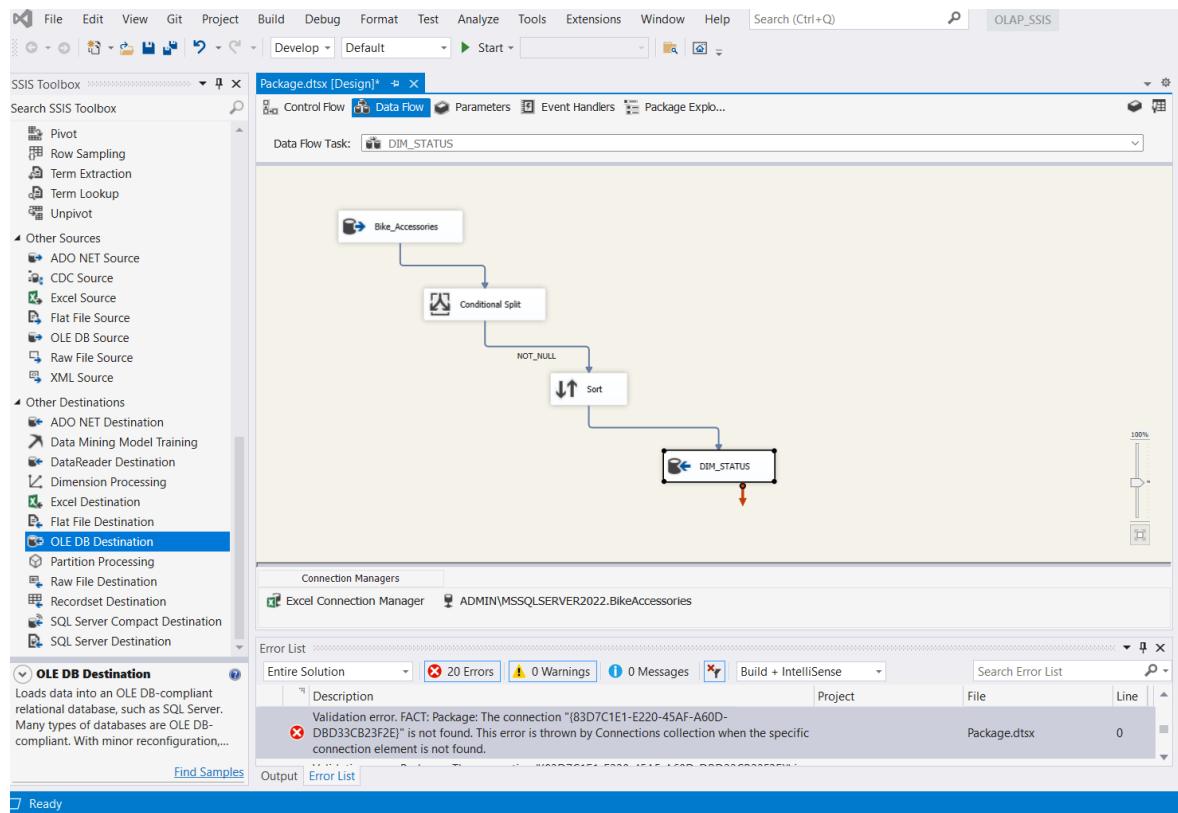




Bước 8: Qua tab “Mappings” để kiểm tra.

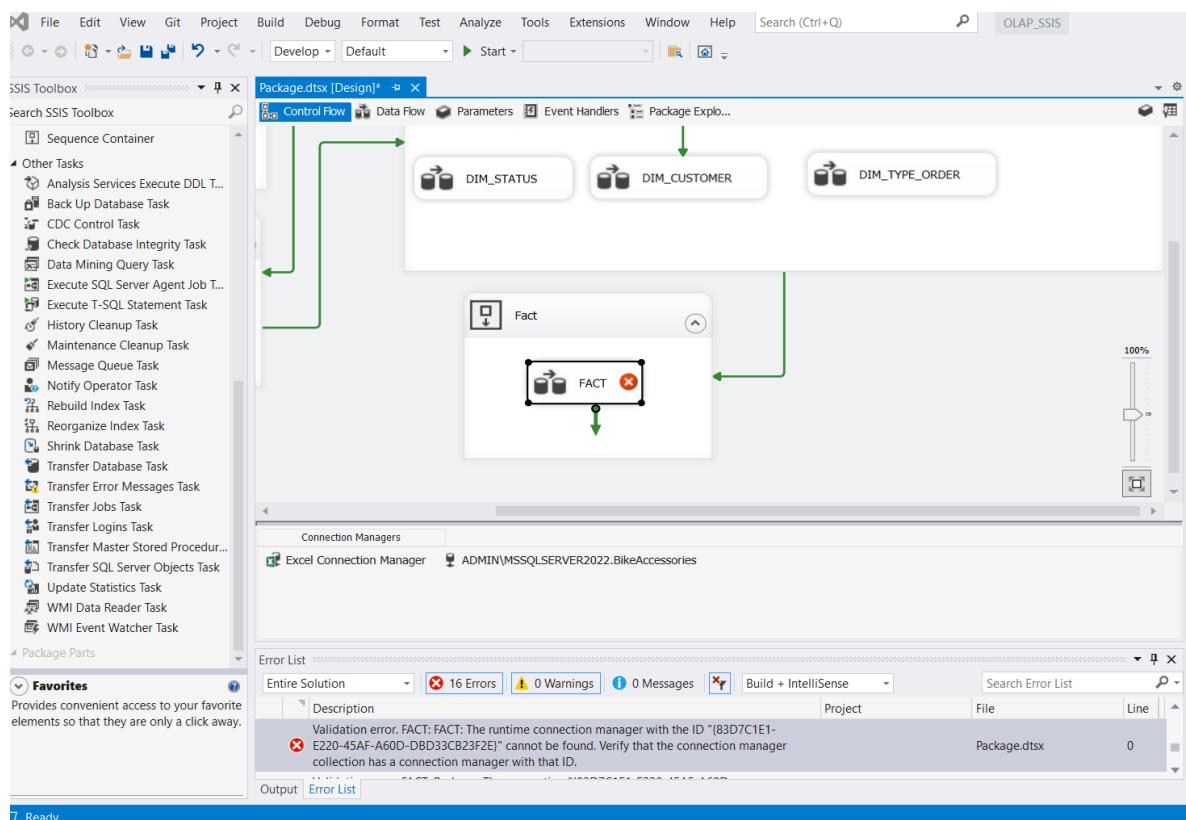


Kết quả luồng thực hiện của bảng DIM_STATUS.

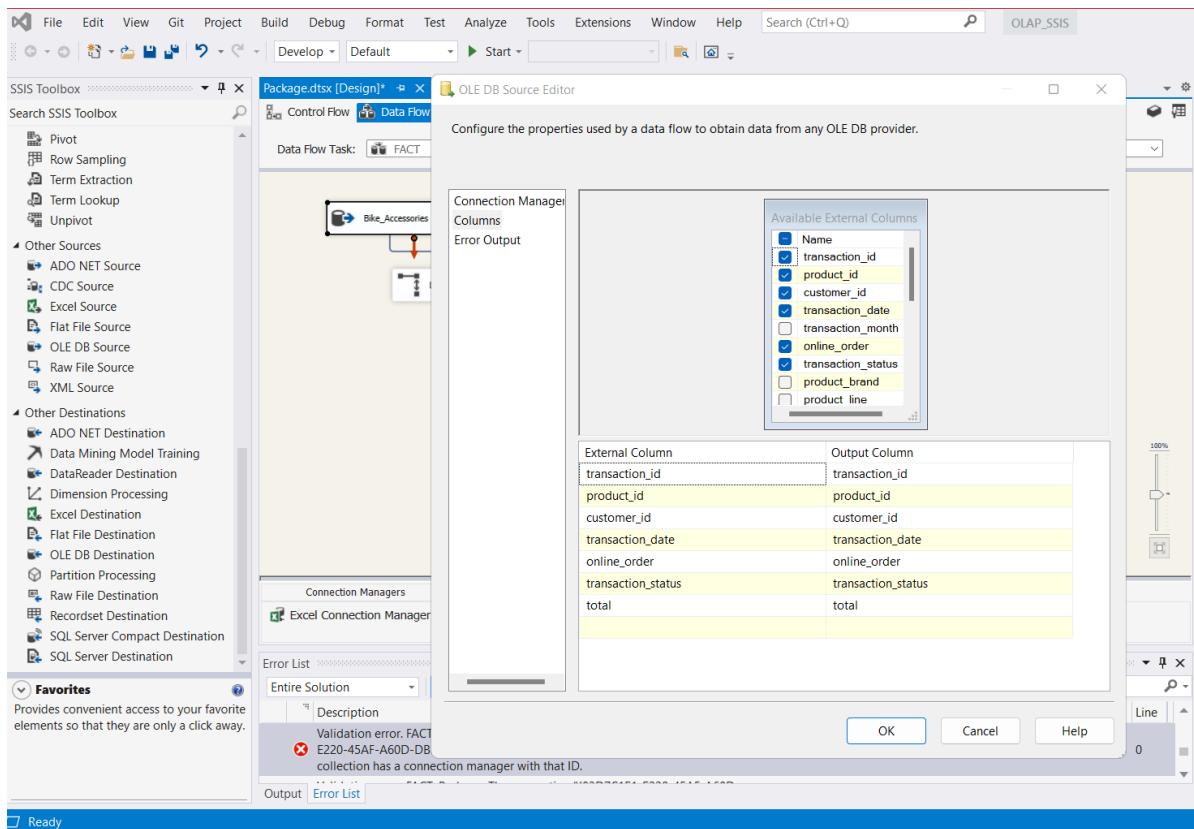


3.10 Tạo bảng FACT

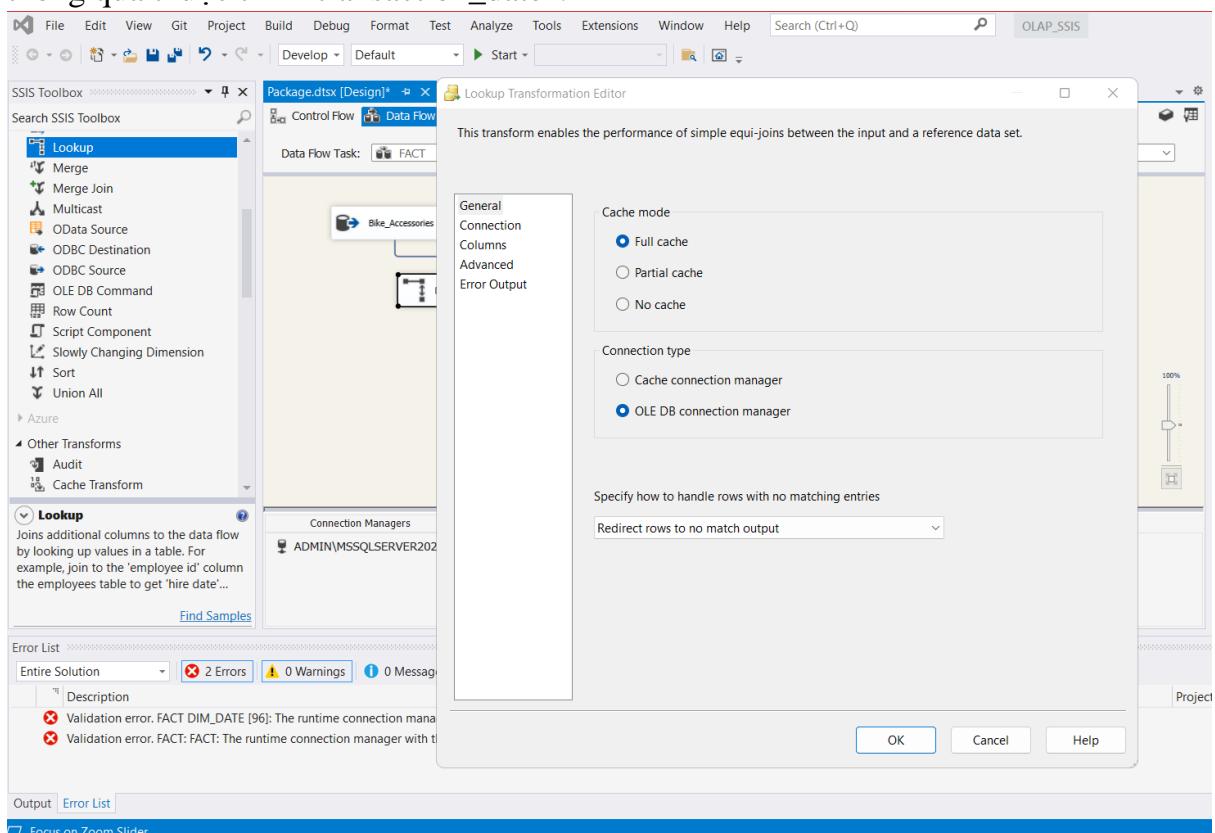
Bước 1: Kéo “Data Flow Task” vào Container đặt tên là FACT.

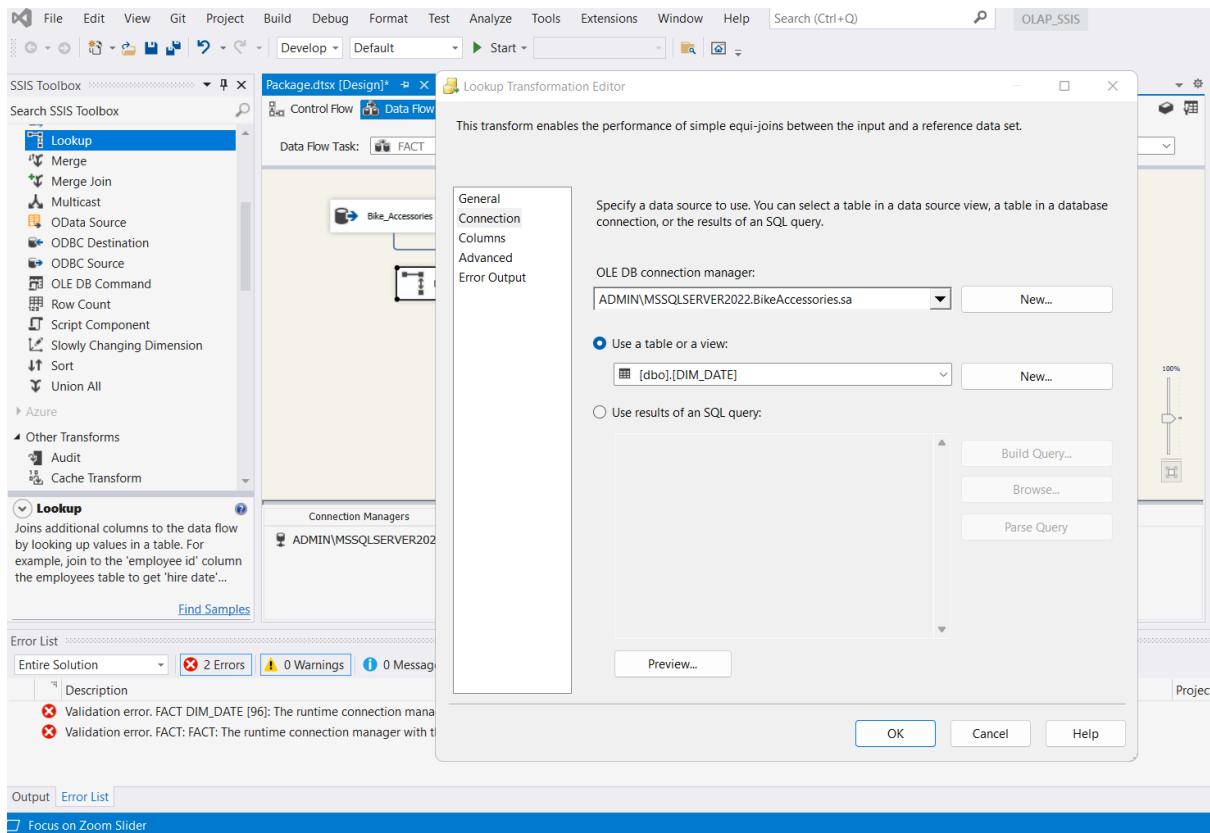


Bước 2: Khởi tạo “OLE DB Source” chứa Dataset và chọn thuộc tính cho bảng FACT.

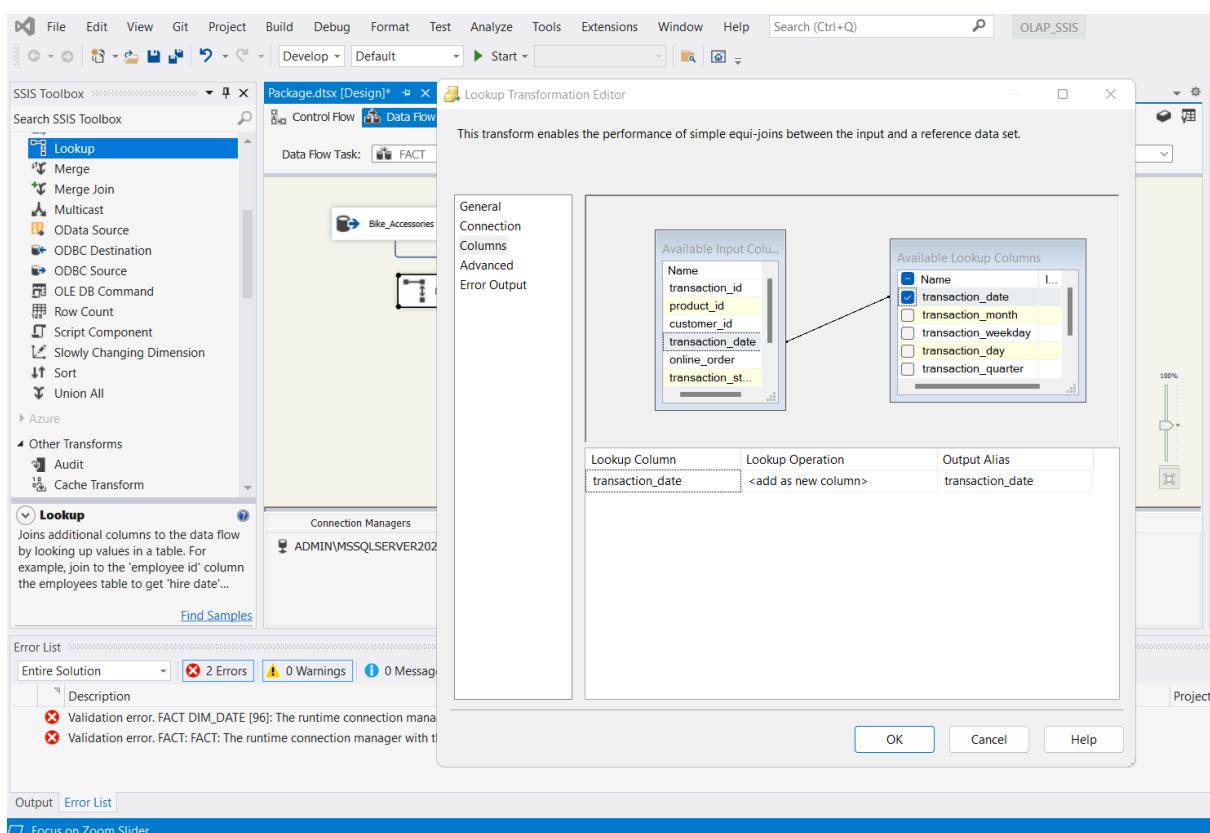


Bước 3: Dùng “Lookup” để kết bảng BIKE_ACCESSORIES và bảng DIM_DATE thông qua thuộc tính “transaction_date”.

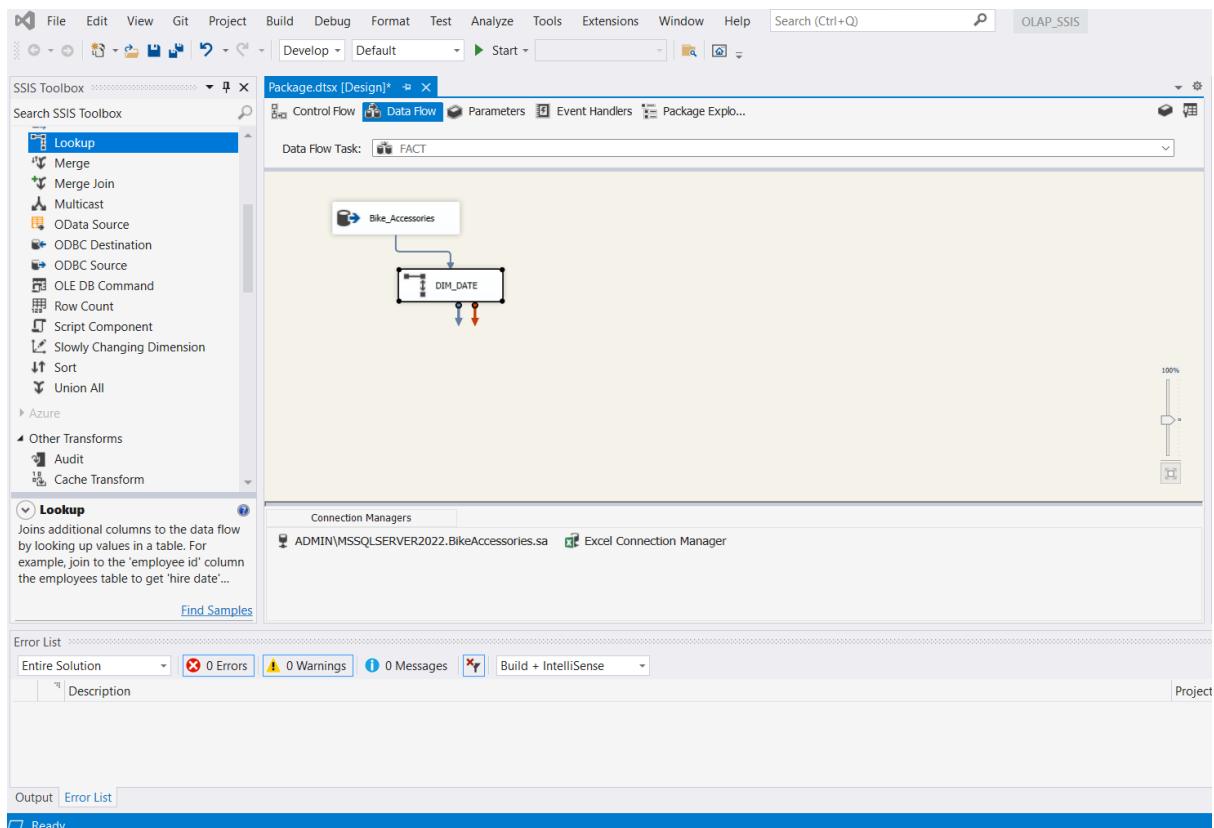




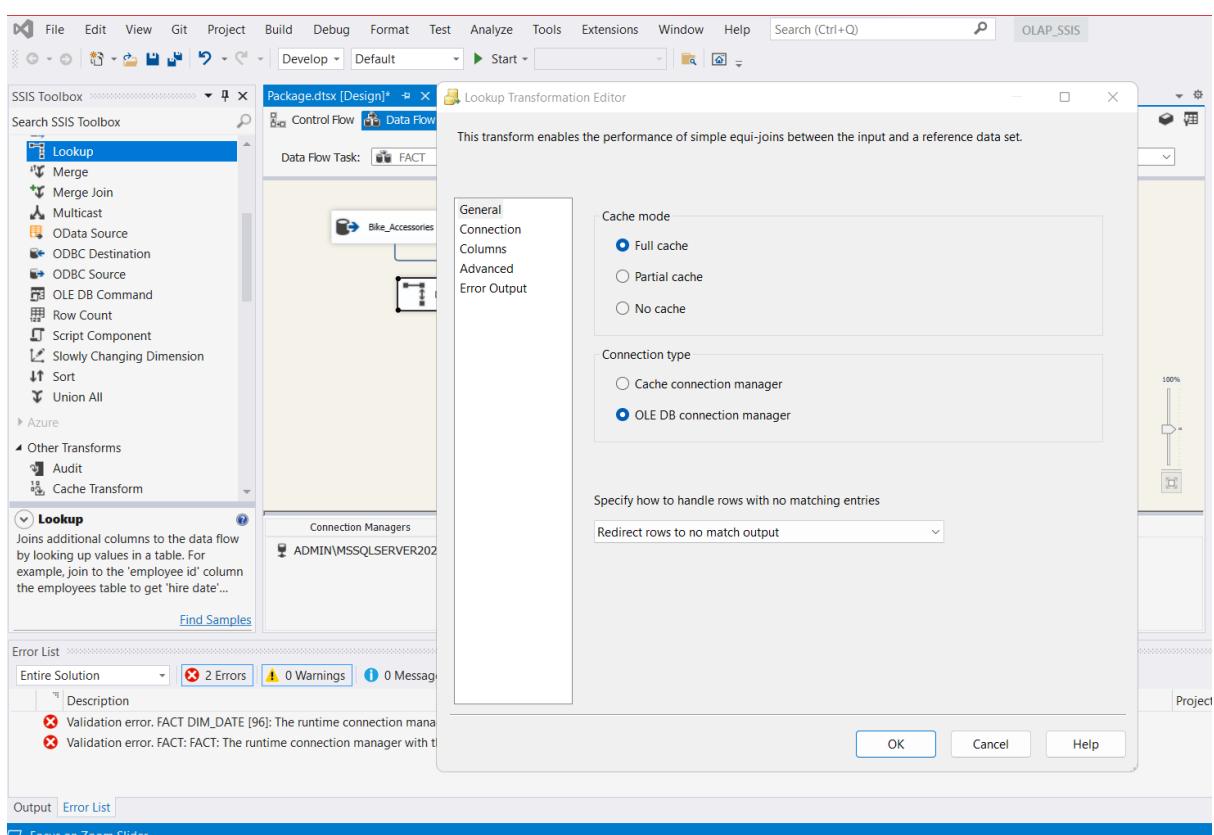
Kết nối Source và bảng DIM_DATE.

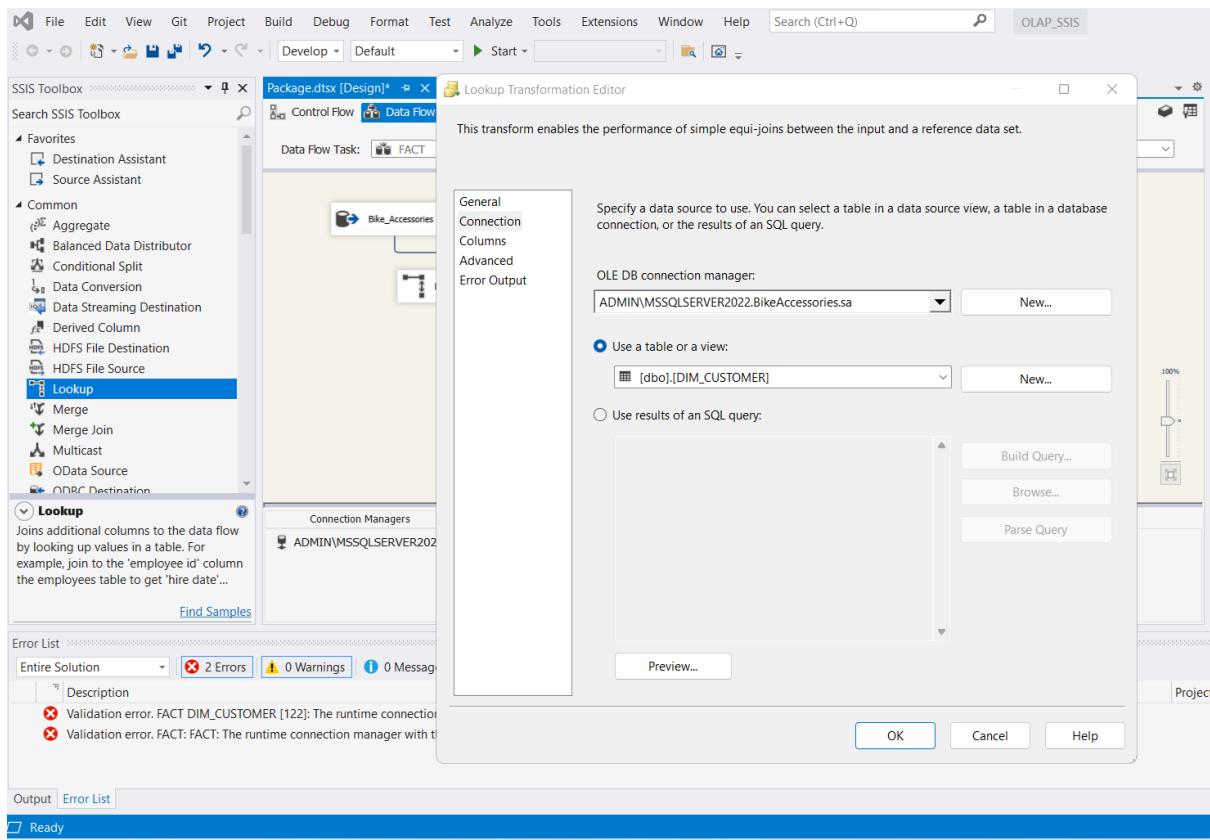


Kết quả sau khi thực hiện Lookup với bảng DIM_DATE.

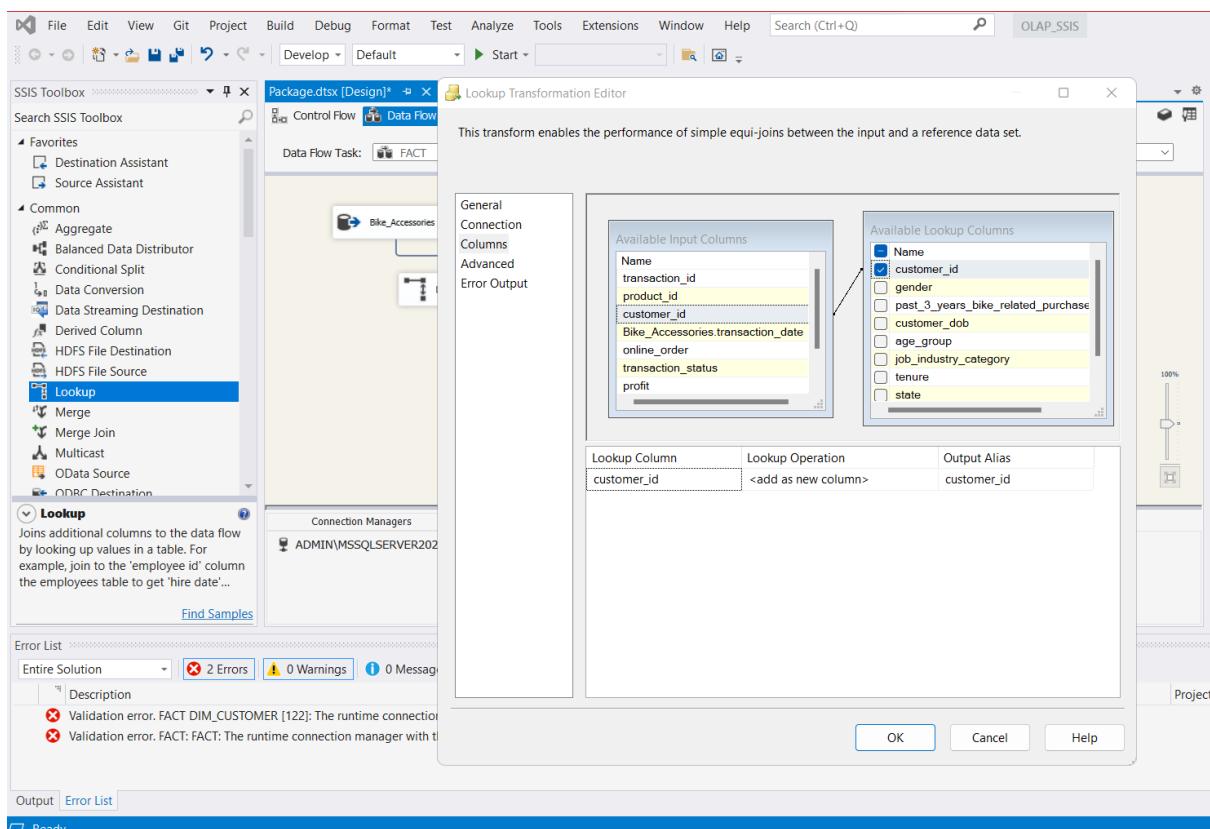


Bước 4: Dùng “Lookup” để kết bảng BIKE_ACCESSORIES và bảng DIM_CUSTOMER thông qua thuộc tính “customer_id”.

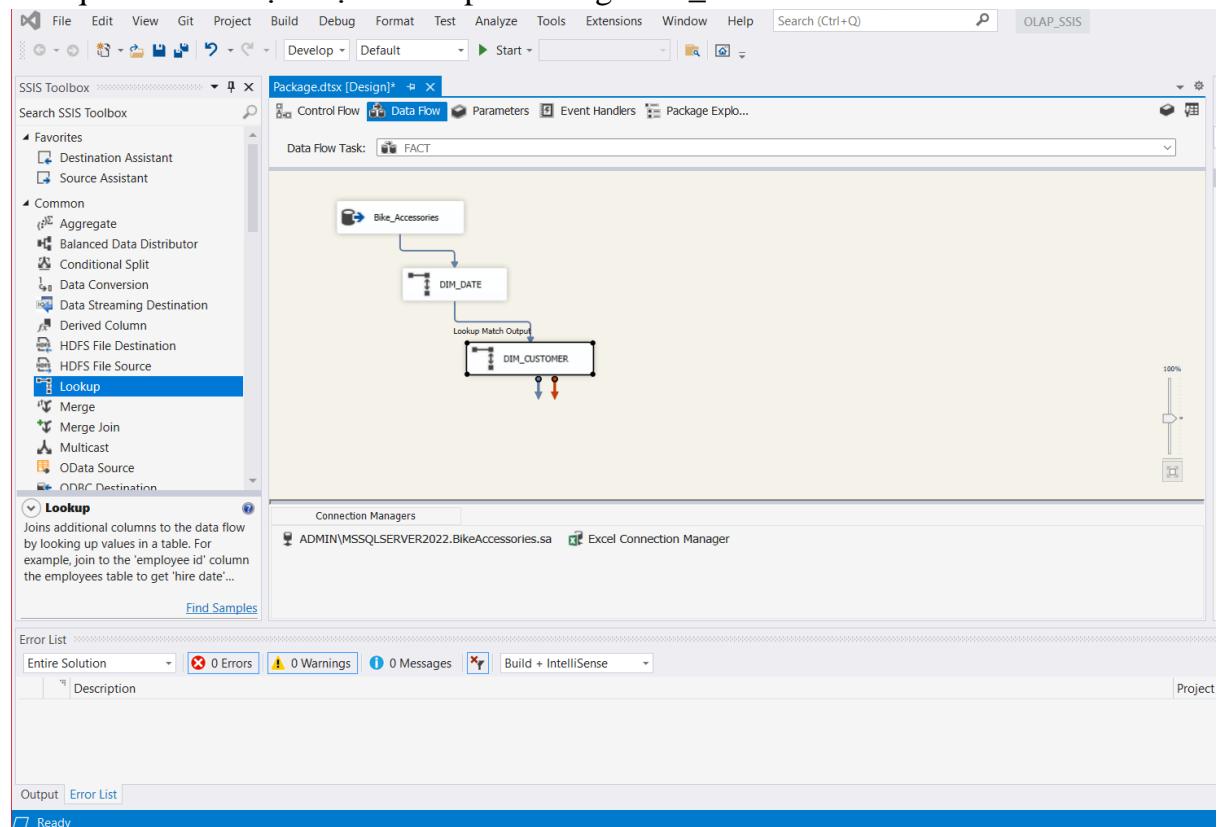




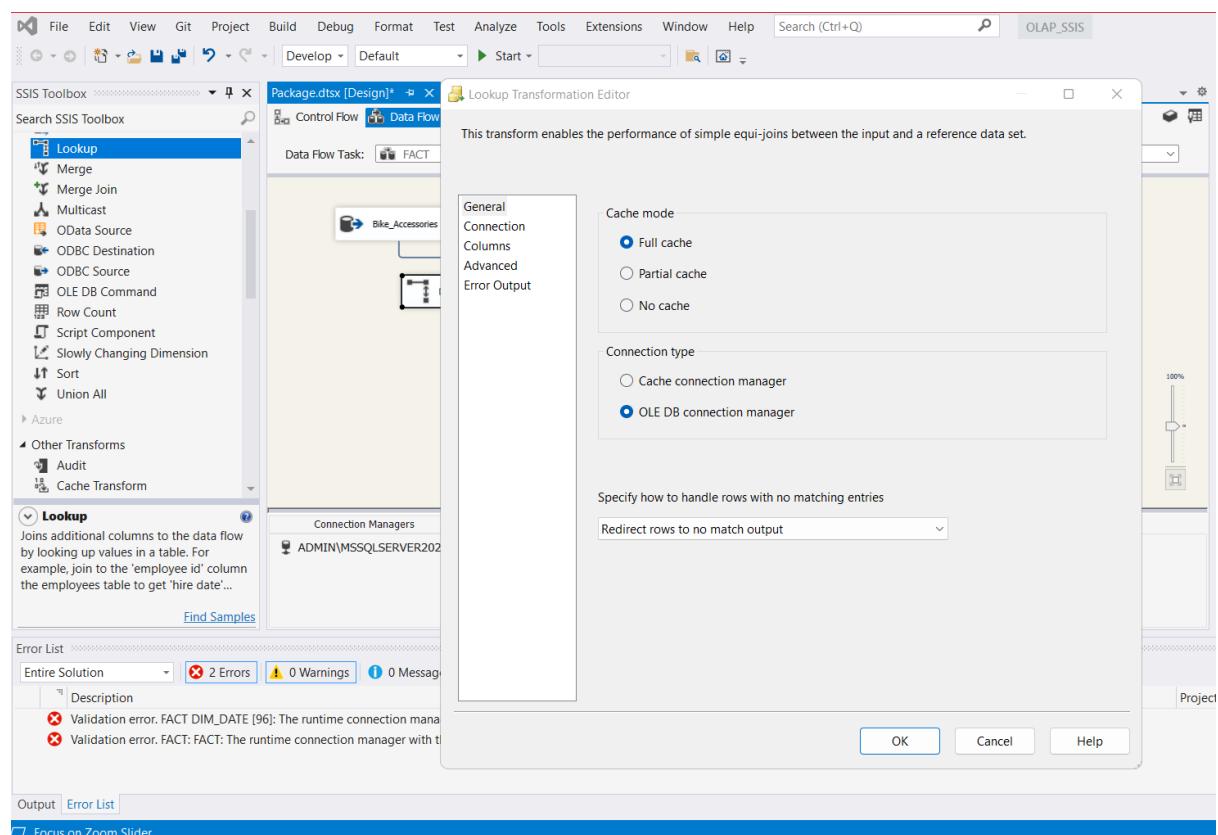
Kết nối Source và bảng DIM_CUSTOMER.

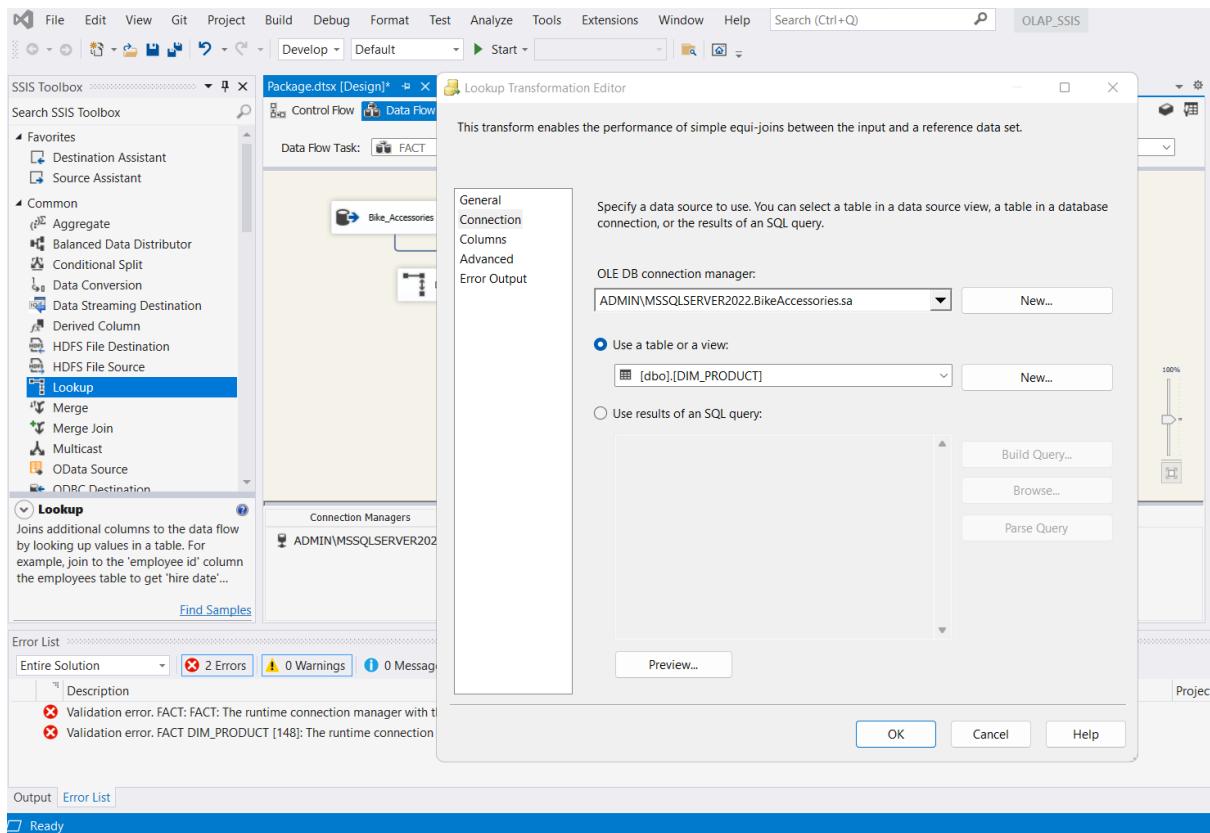


Kết quả sau khi thực hiện Lookup với bảng DIM_CUSTOMER.

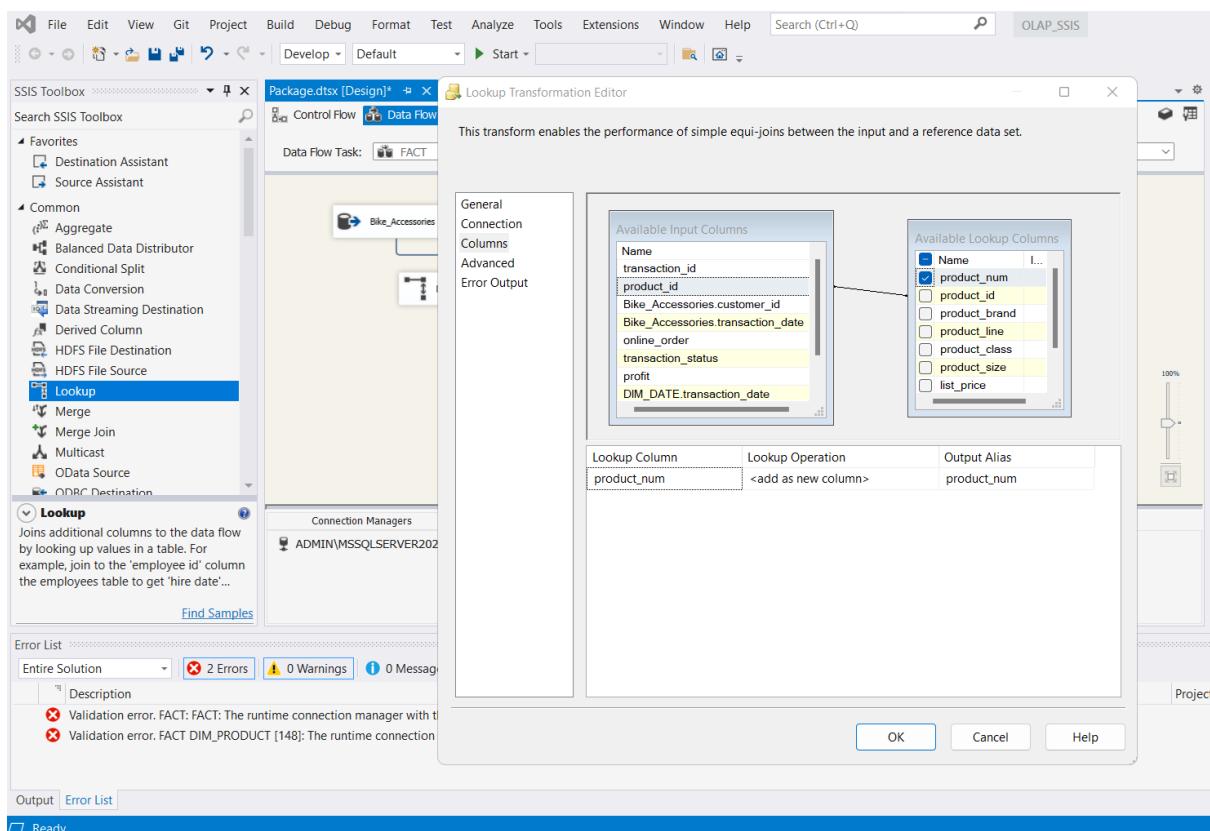


Bước 5: Dùng “Lookup” để kết bảng BIKE_ACCESSORIES và bảng DIM_PRODUCT thông qua thuộc tính “product_num”.

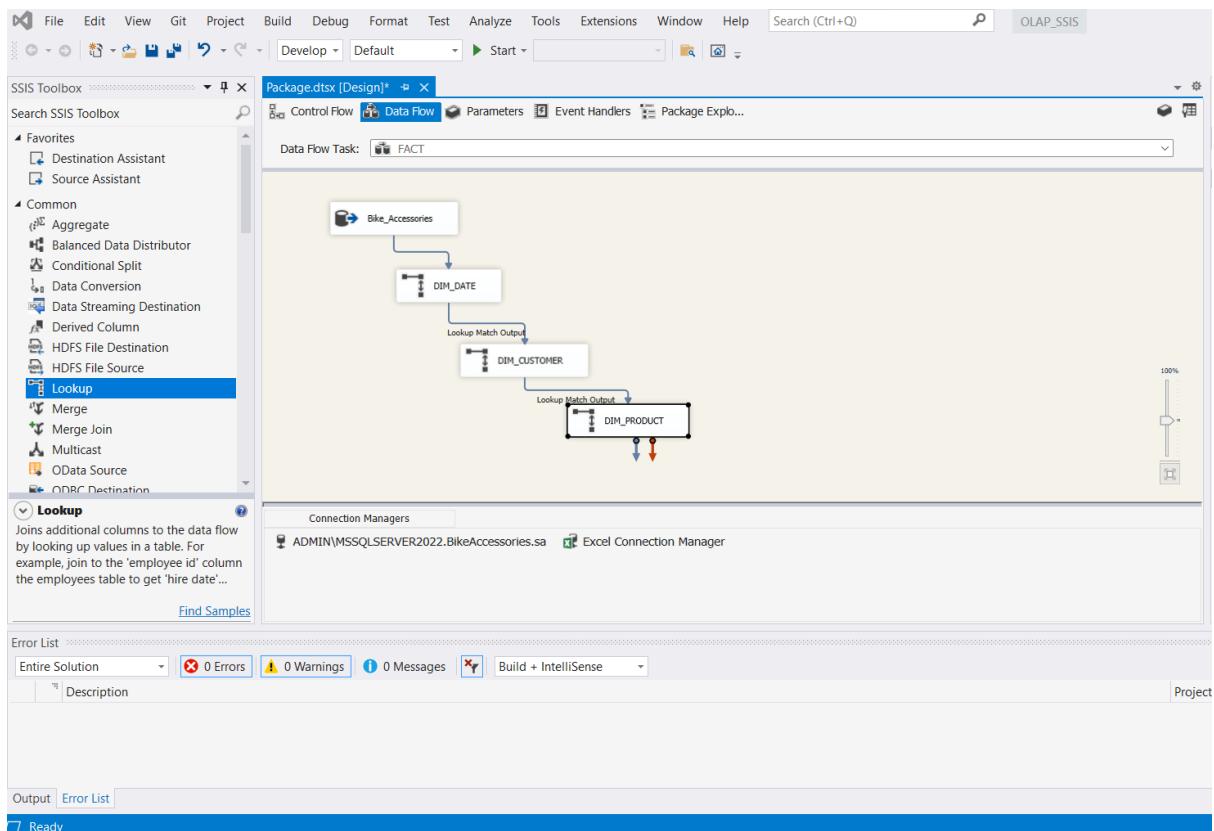




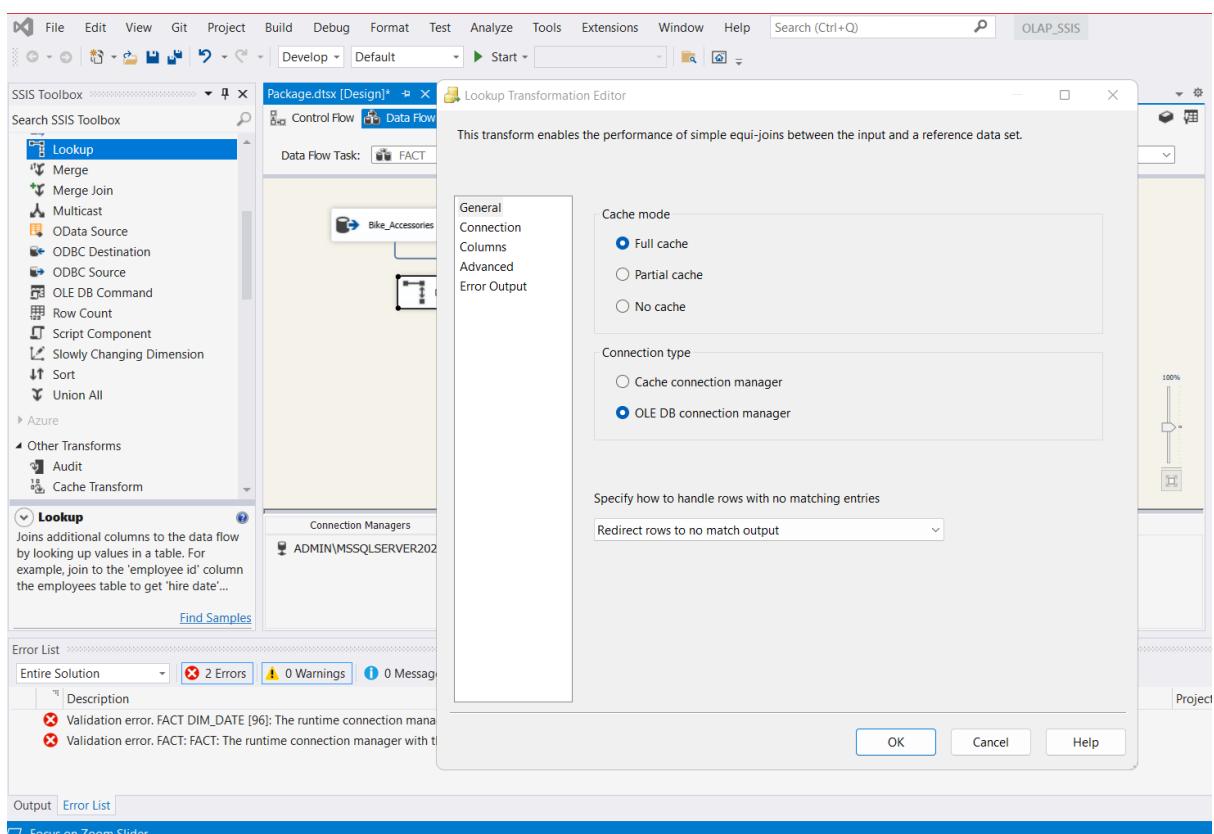
Kết nối Source và bảng DIM_PRODUCT.

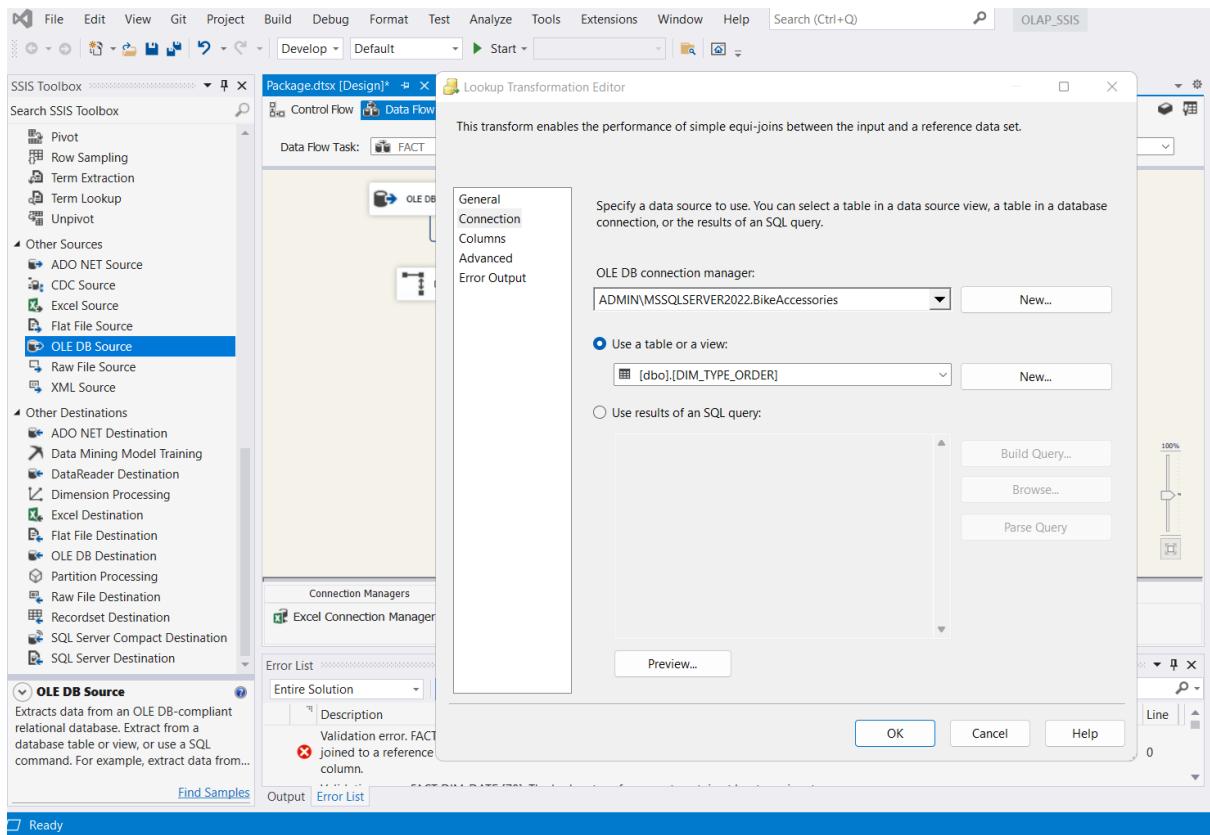


Kết quả sau khi thực hiện Lookup với bảng DIM_PRODUCT.

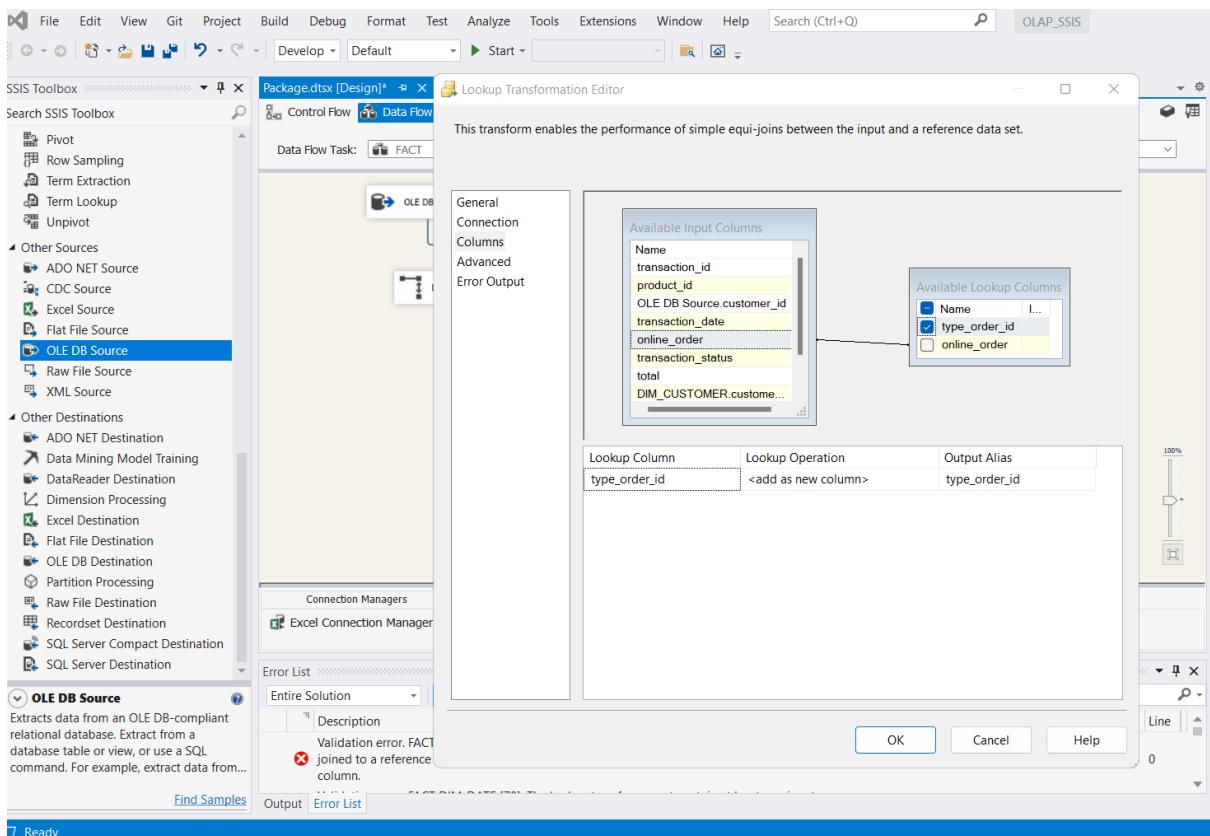


Bước 6: Dùng “Lookup” để kết bảng BIKE_ACCESSORIES và bảng DIM_TYPE_ORDER thông qua thuộc tính “online_order”.

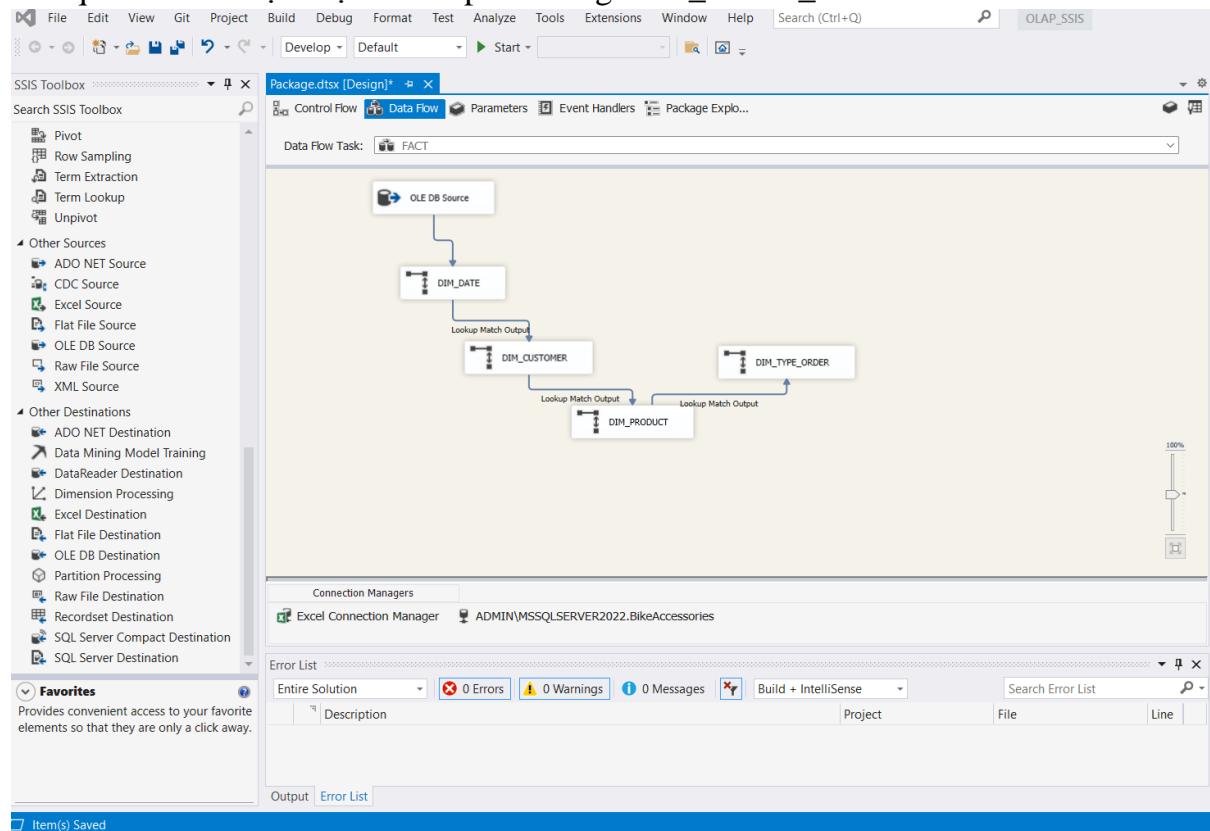




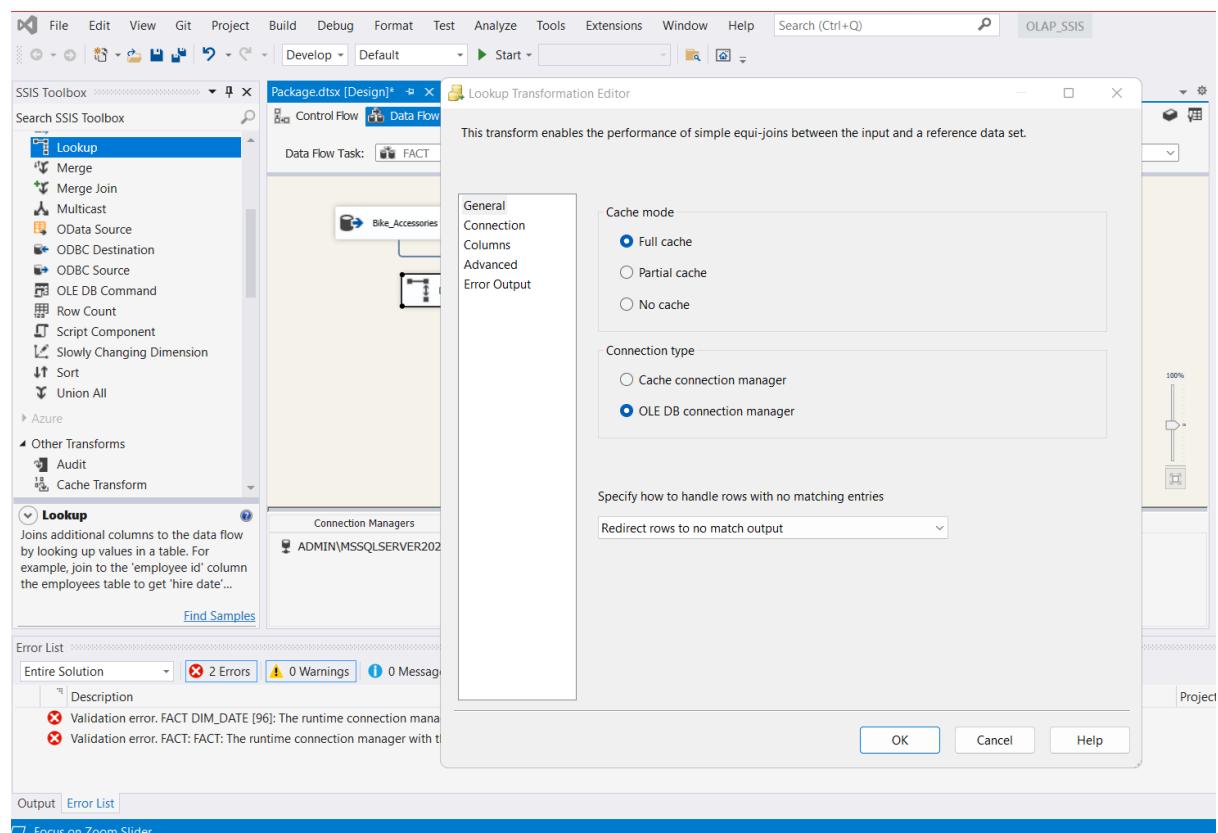
Kết nối Source và bảng DIM_TYPE_ORDER.

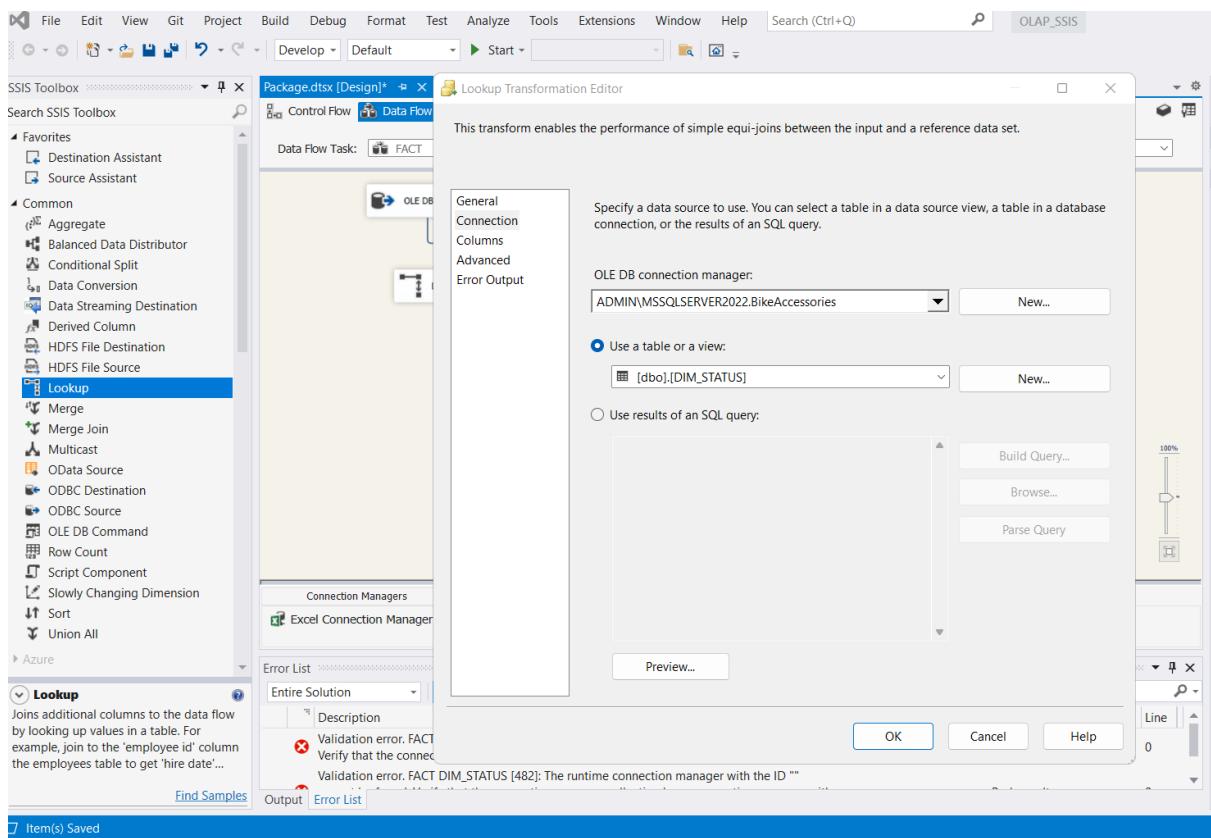


Kết quả sau khi thực hiện Lookup với bảng DIM_TYOE_ORDER.

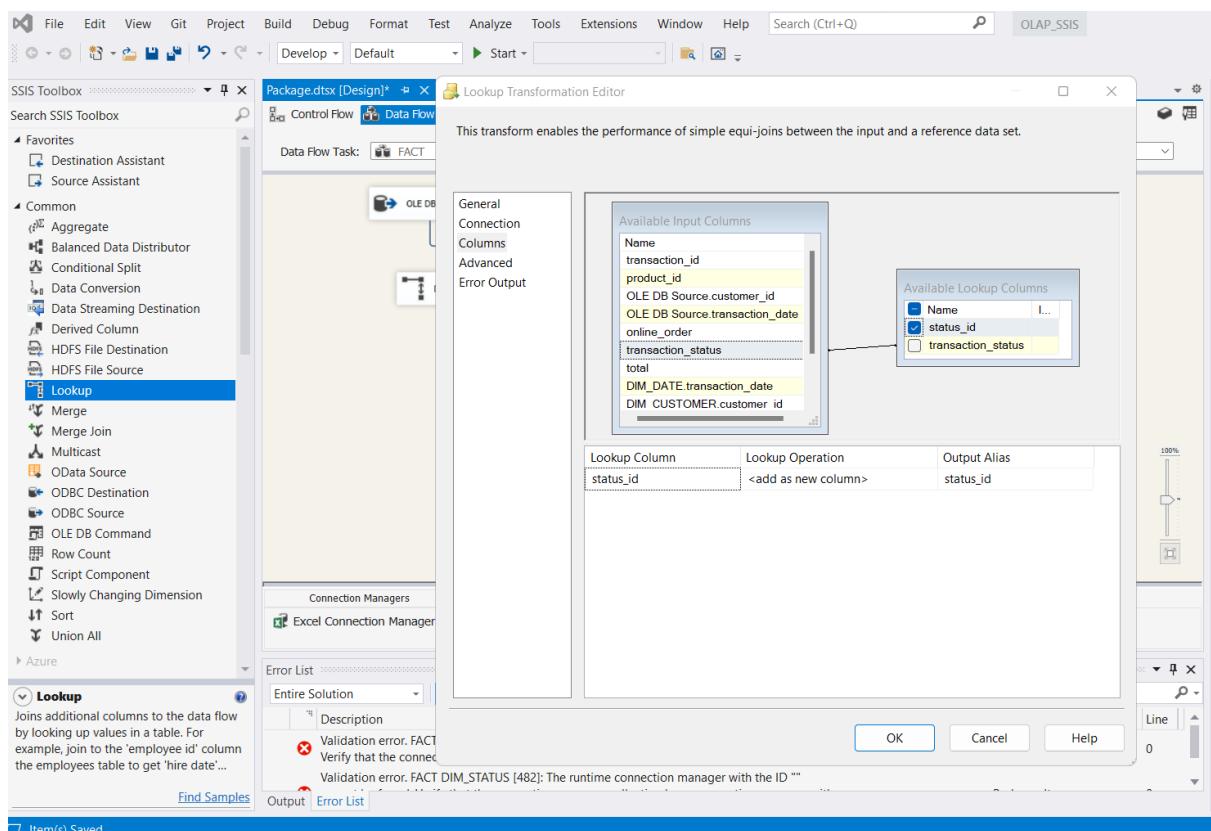


Bước 7: Dùng “Lookup” để kết bảng BIKE_ACCESSORIES và bảng DIM_STATUS thông qua thuộc tính “online_order”.

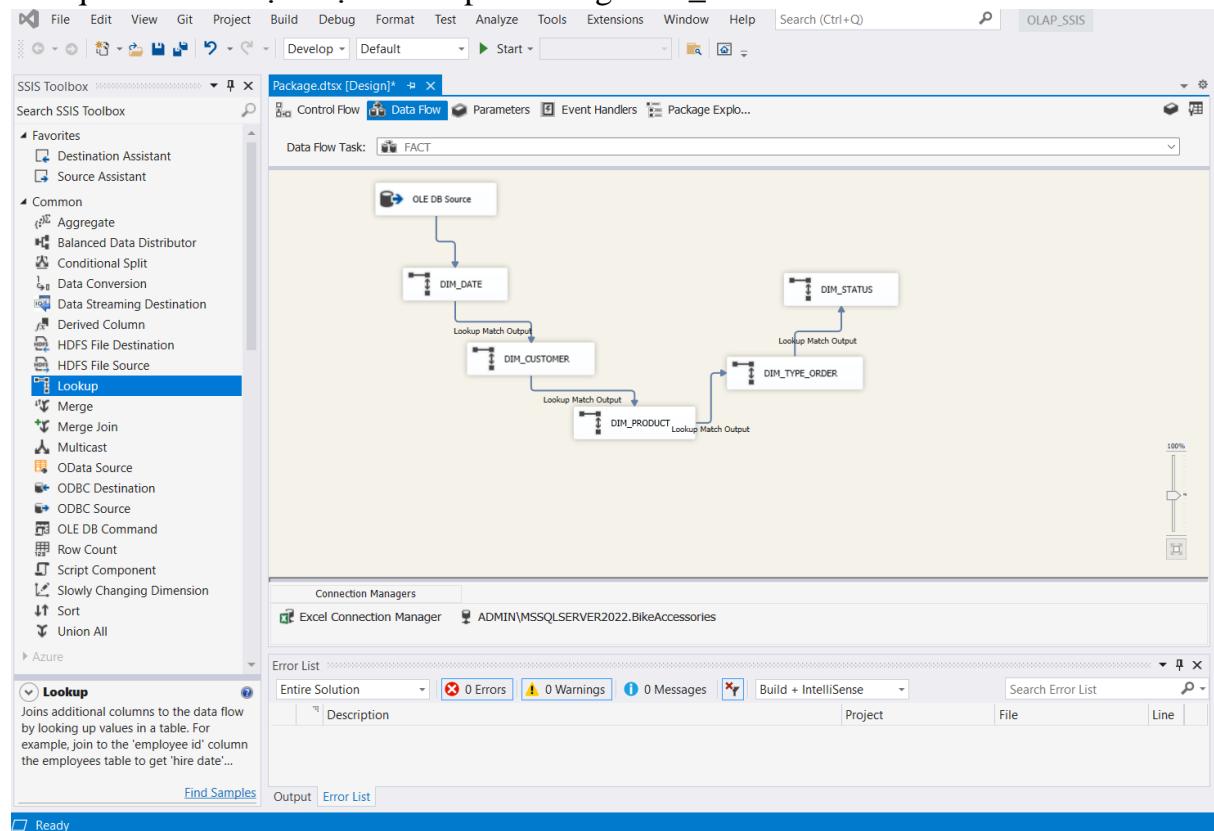




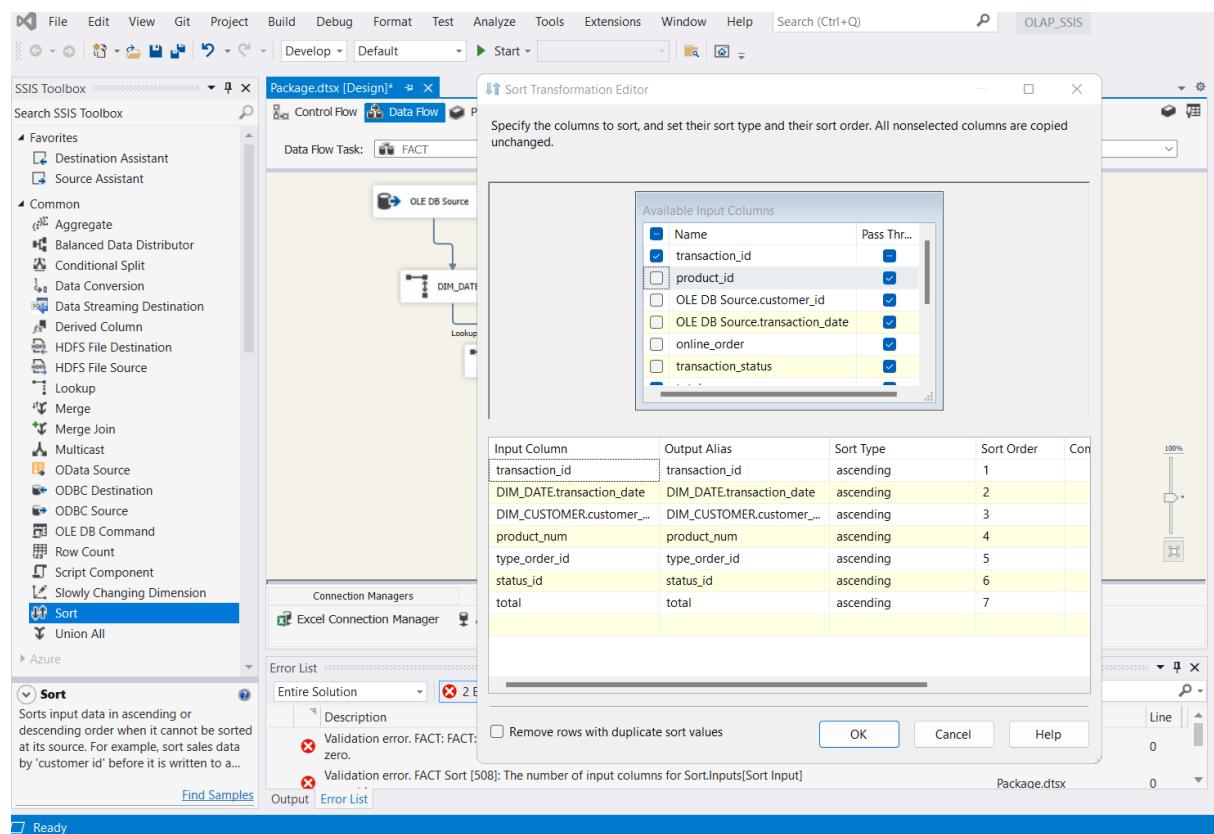
Kết nối Source và bảng DIM_STATUS.



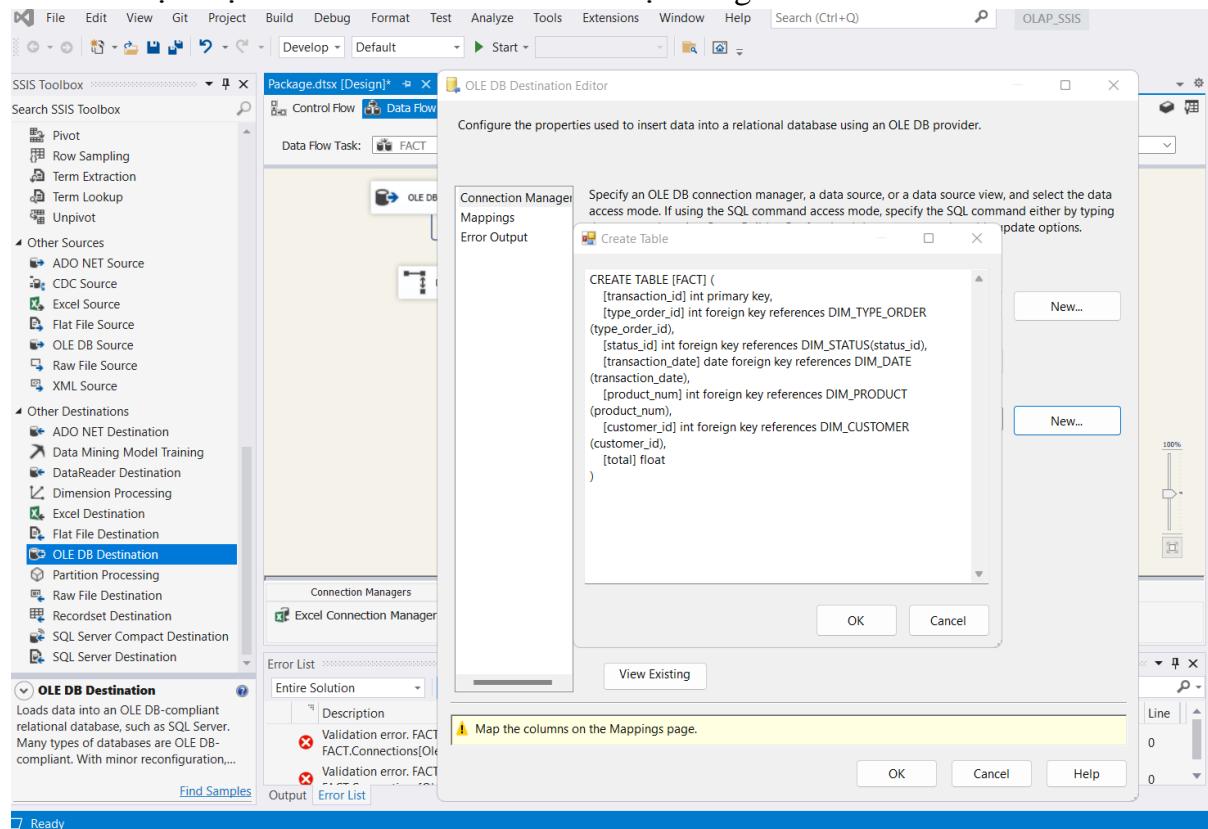
Kết quả sau khi thực hiện Lookup với bảng DIM_STATUS.



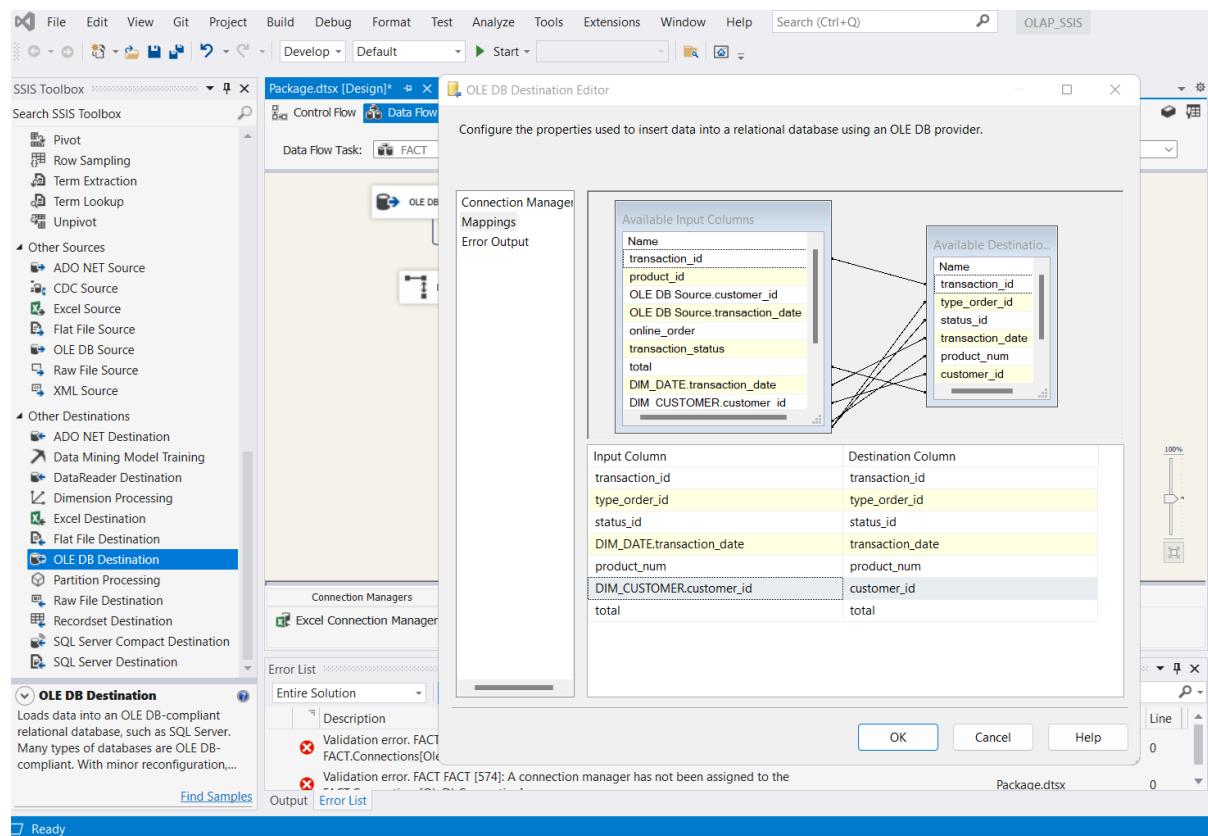
Bước 8: Dùng “Sort” để sắp xếp lại dữ liệu.



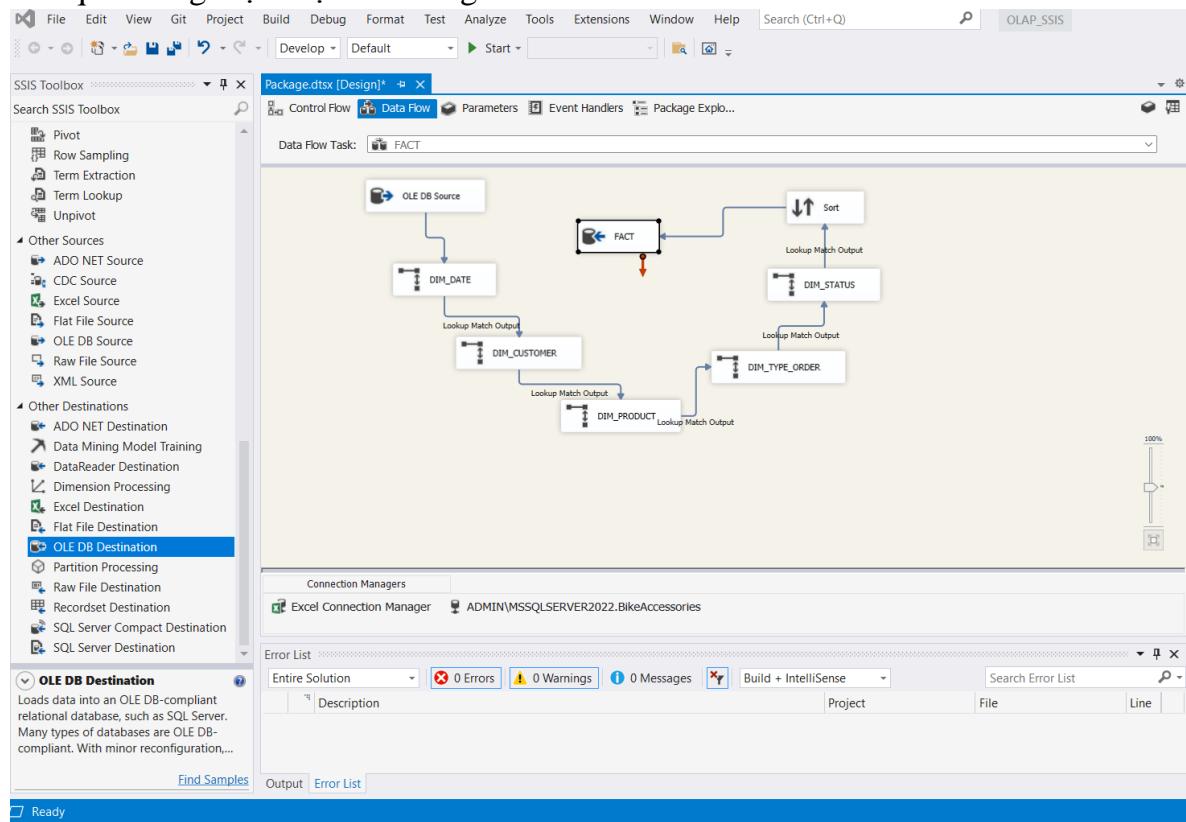
Bước 9: Tạo một “OLE Destination” sau đó tạo bảng FACT.



Bước 10: Qua tab “Mappings” để kiểm tra.

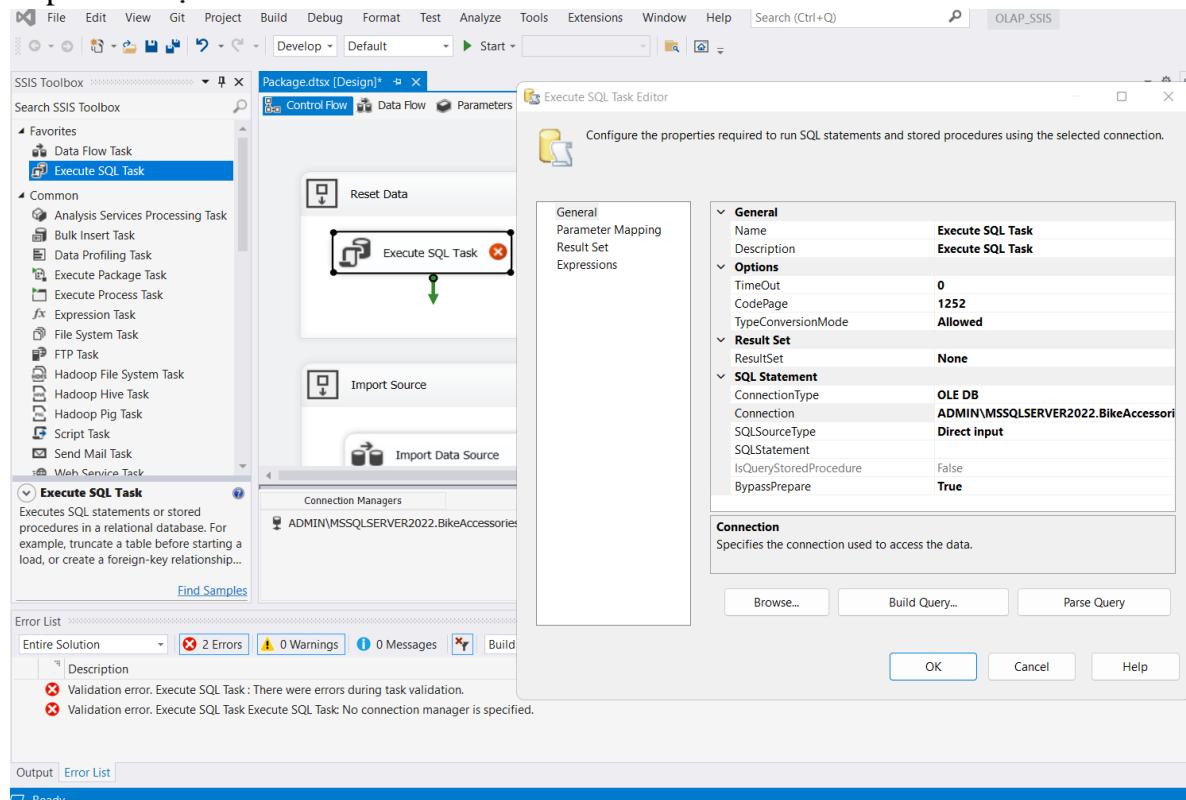


Kết quả luồng thực hiện của bảng FACT.

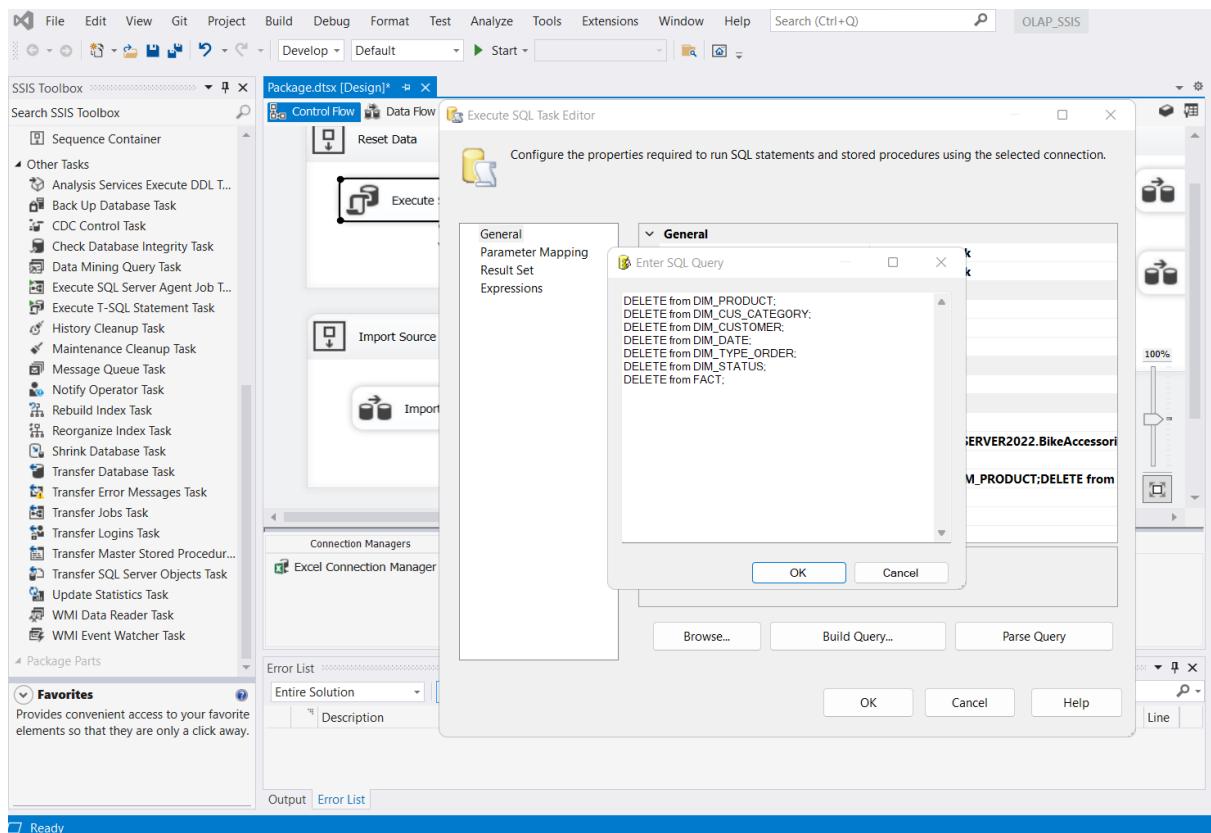


3.11 Import dữ liệu vào Database

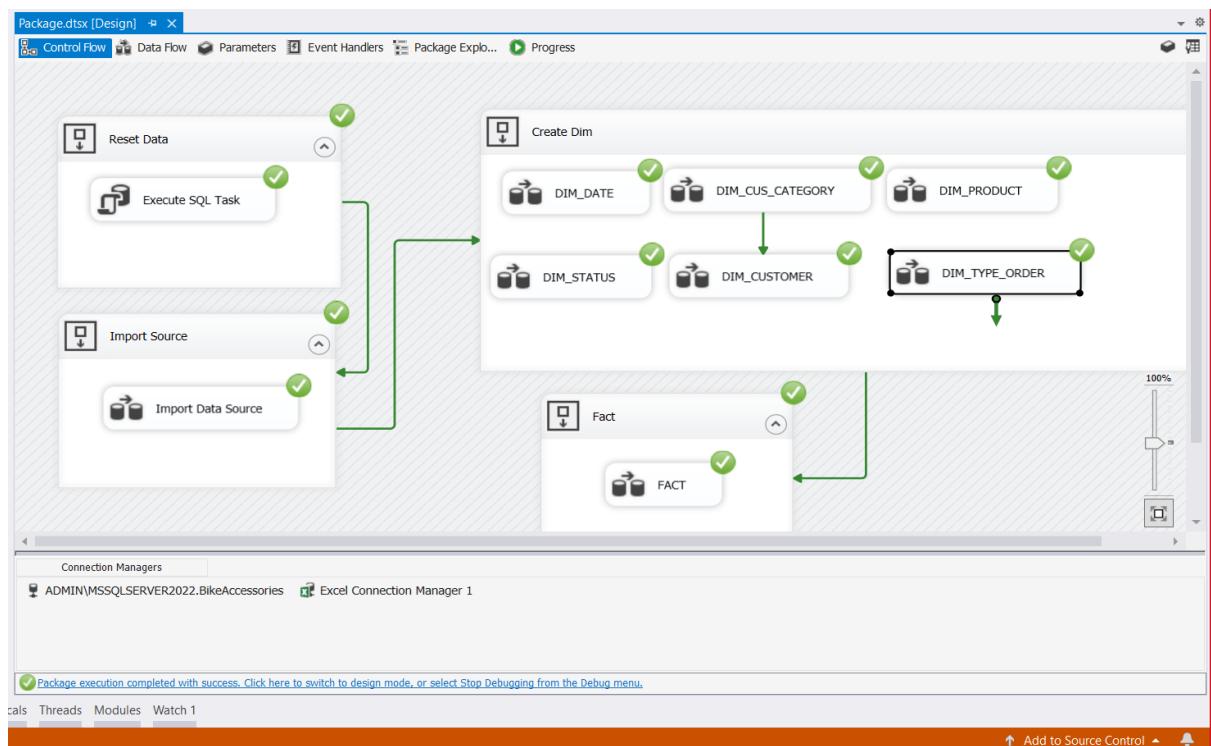
Bước 1: Ở SSIS Toolbox kéo thả “Execute SQL Task” vào Control Flow để thực thi import dữ liệu vào database.



Bước 2: Thực hiện câu truy vấn để xóa hết dữ liệu ban đầu (nếu có).



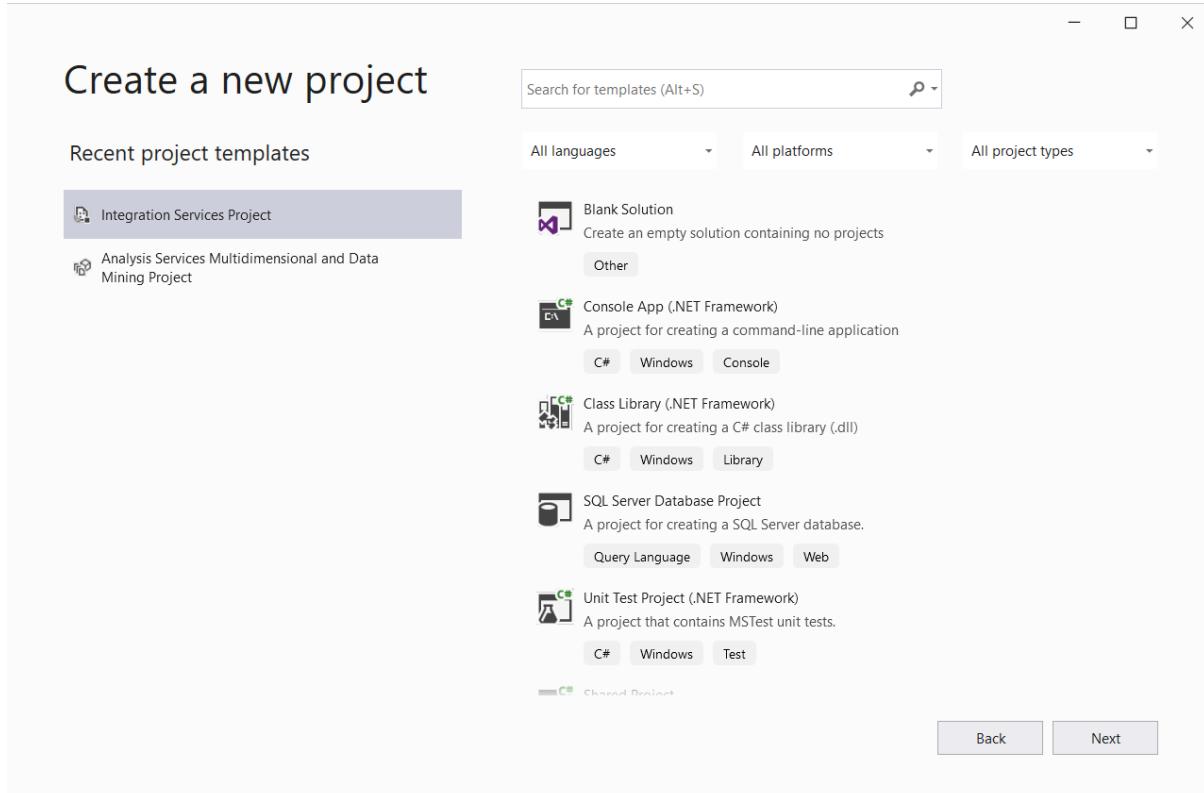
Kết quả luồng thực hiện của quá trình SSIS



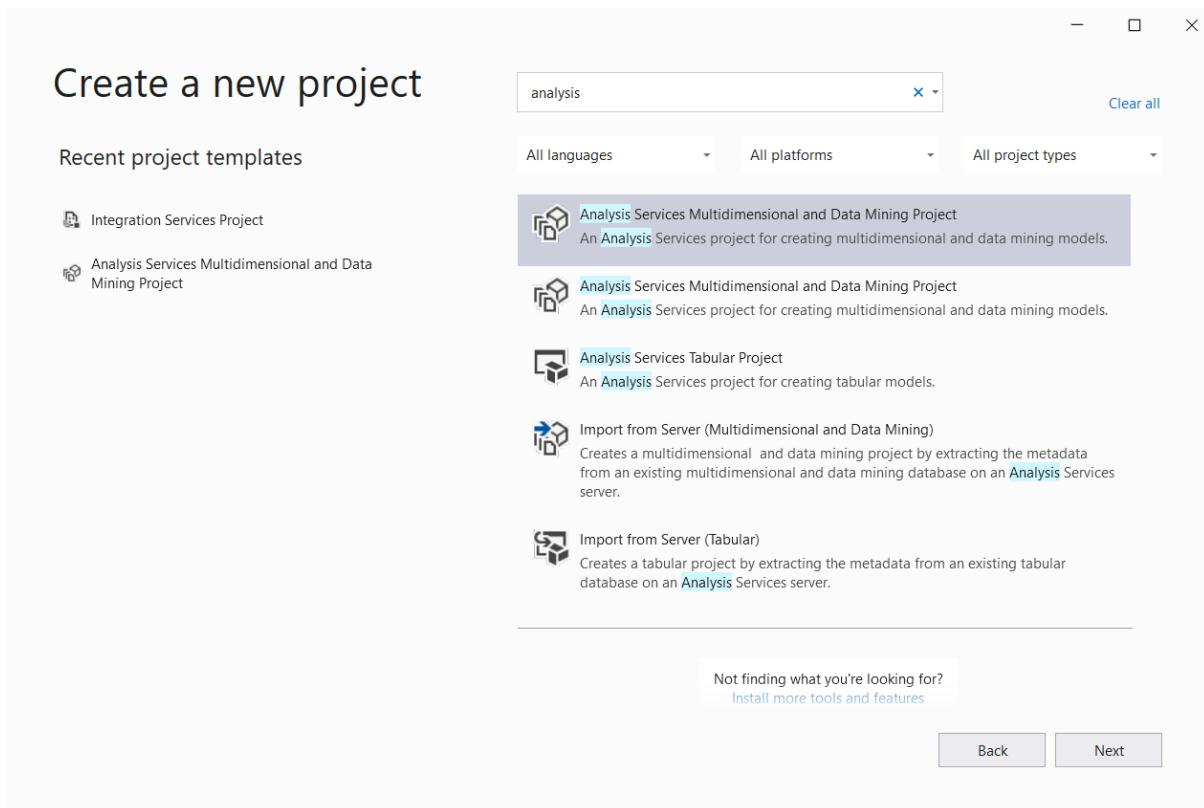
CHƯƠNG 3. QUÁ TRÌNH SSAS

1. Tạo Project SSAS trong Visual Studio 2019

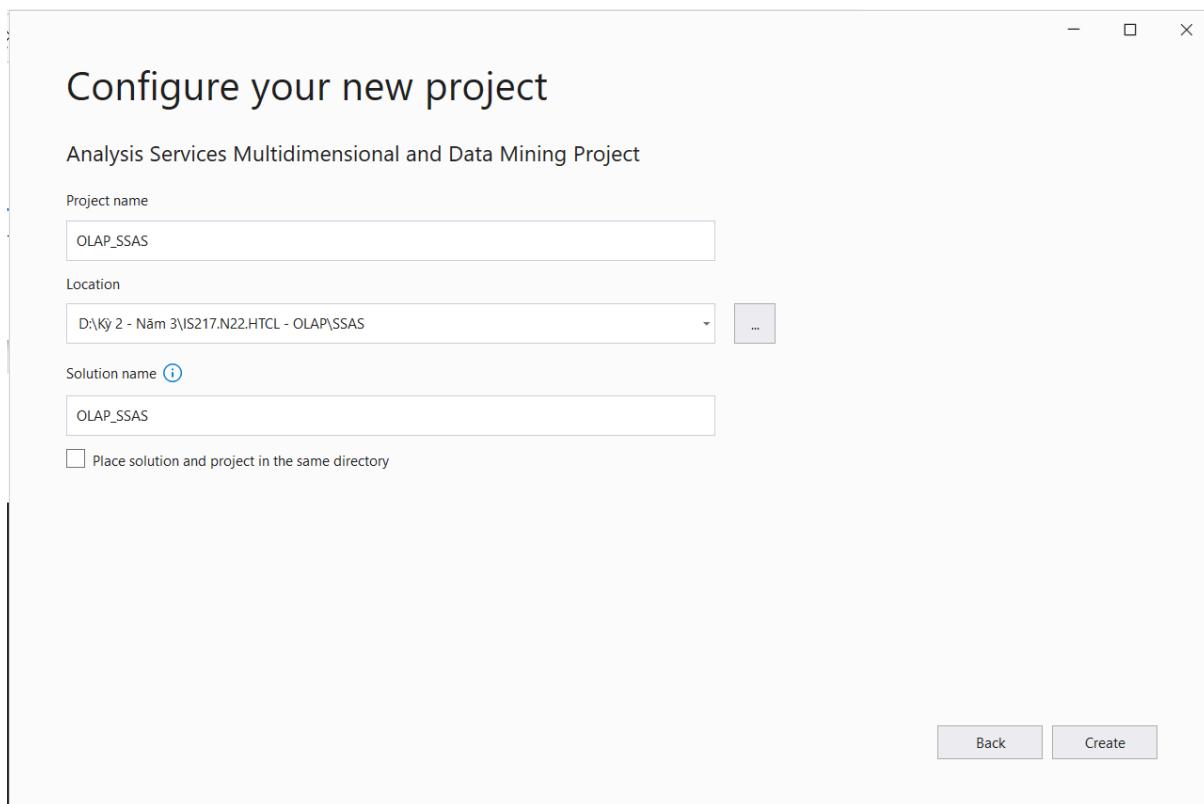
Bước 1: Khởi động Visual Studio 2019.



Bước 2: Tìm từ khóa Analysis, sau đó chọn “Analysis Services Multidimensional and Data Mining Project”, sau đó nhấn “Next”.

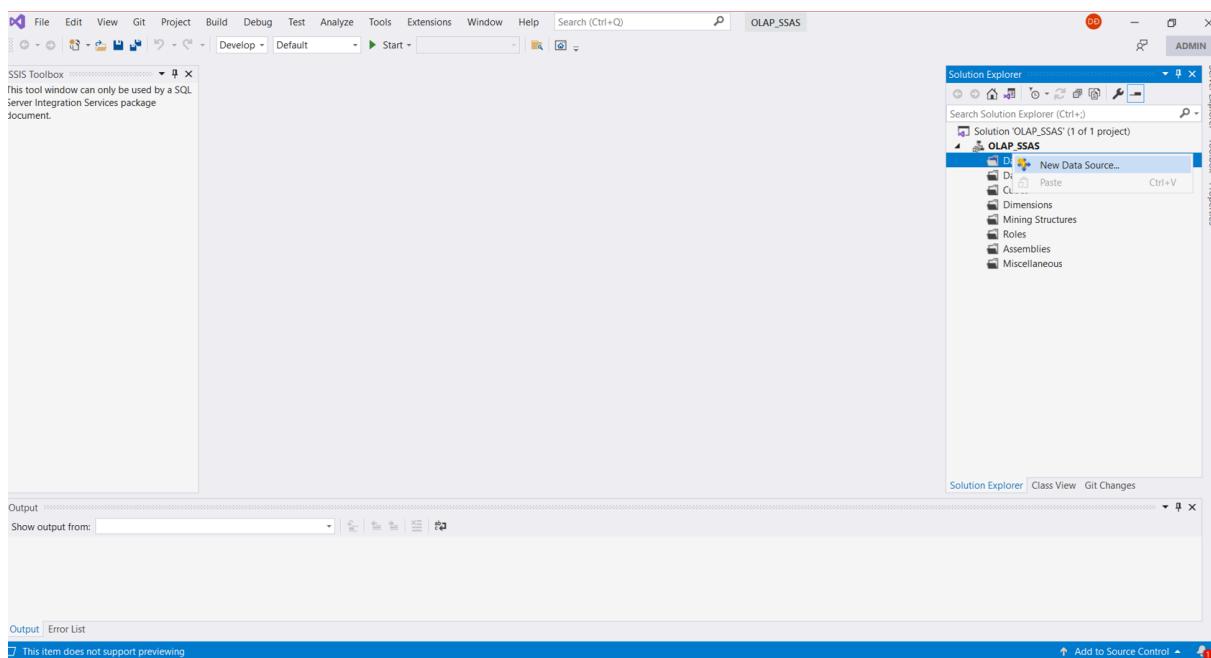


Bước 3: Đặt tên cho project và nhấn “Create”, giao diện quá trình SSAS hiện ra.

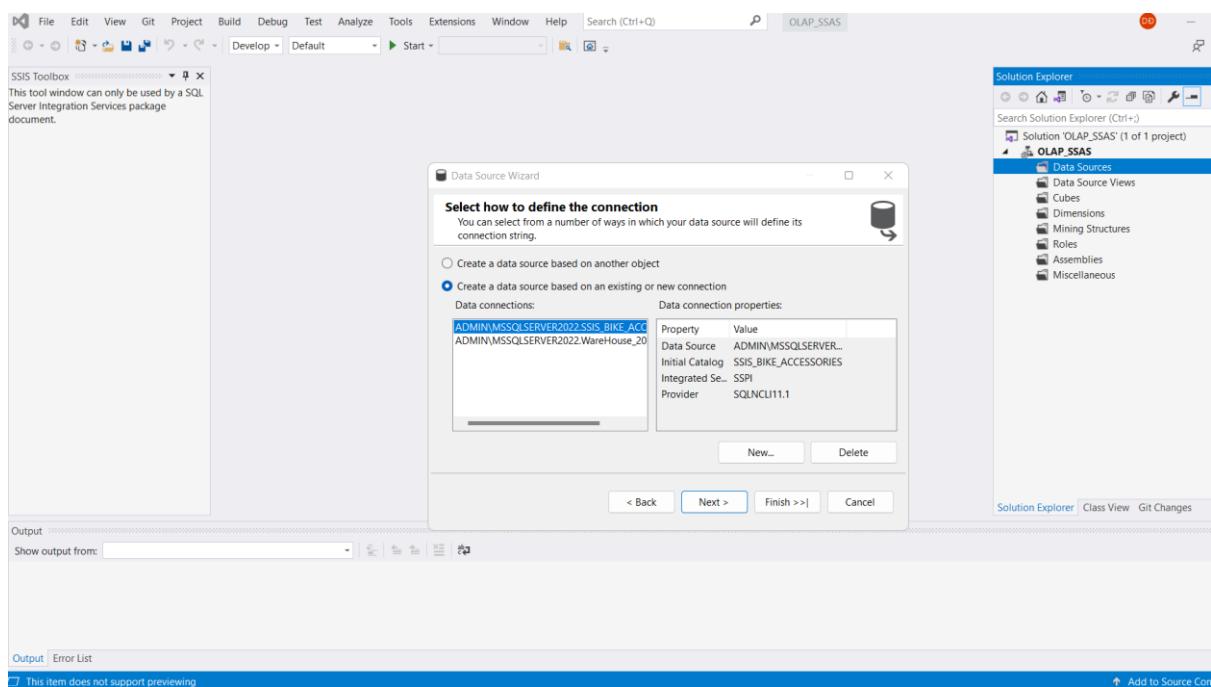


2. Connect đến data source

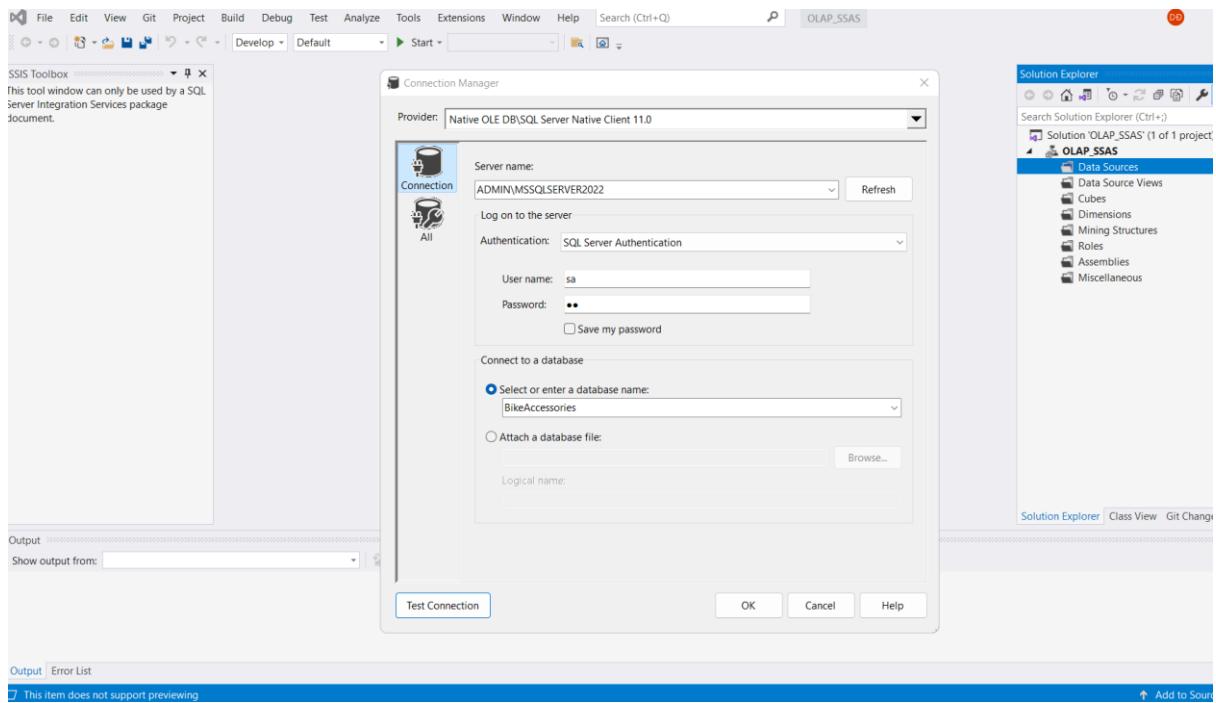
Bước 1: Nhảy phải Data Source → New Data Source.



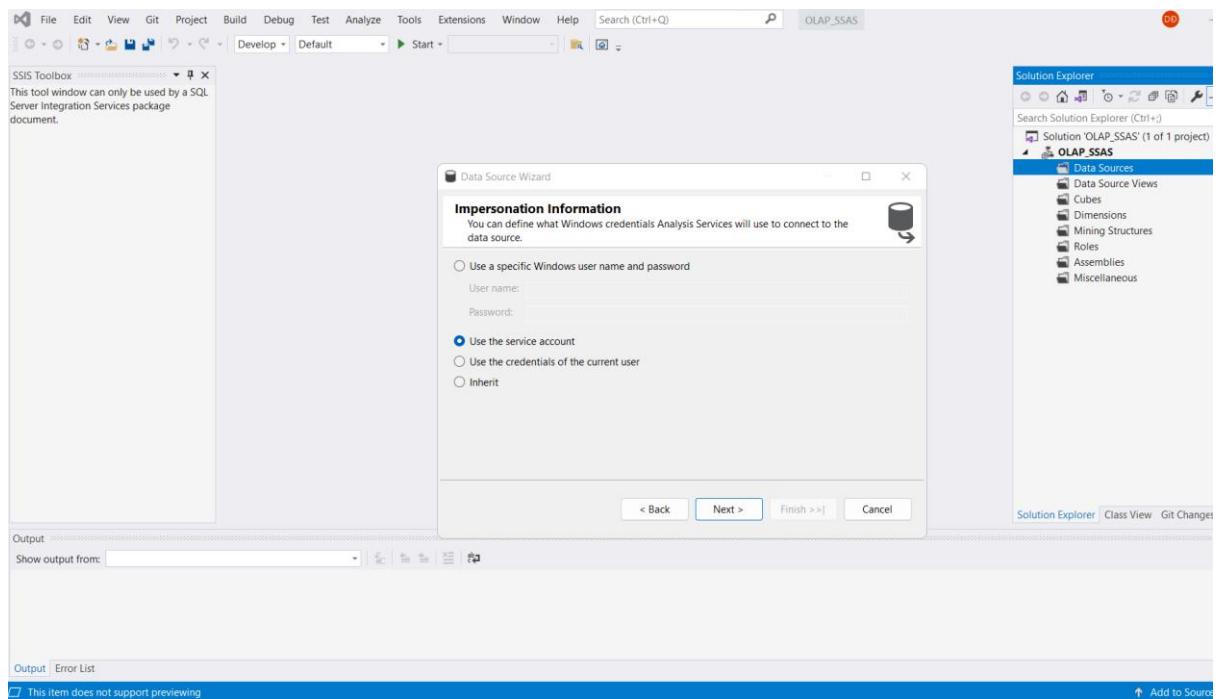
Bước 2: Chọn “Create a data source based on existing or new connection”.



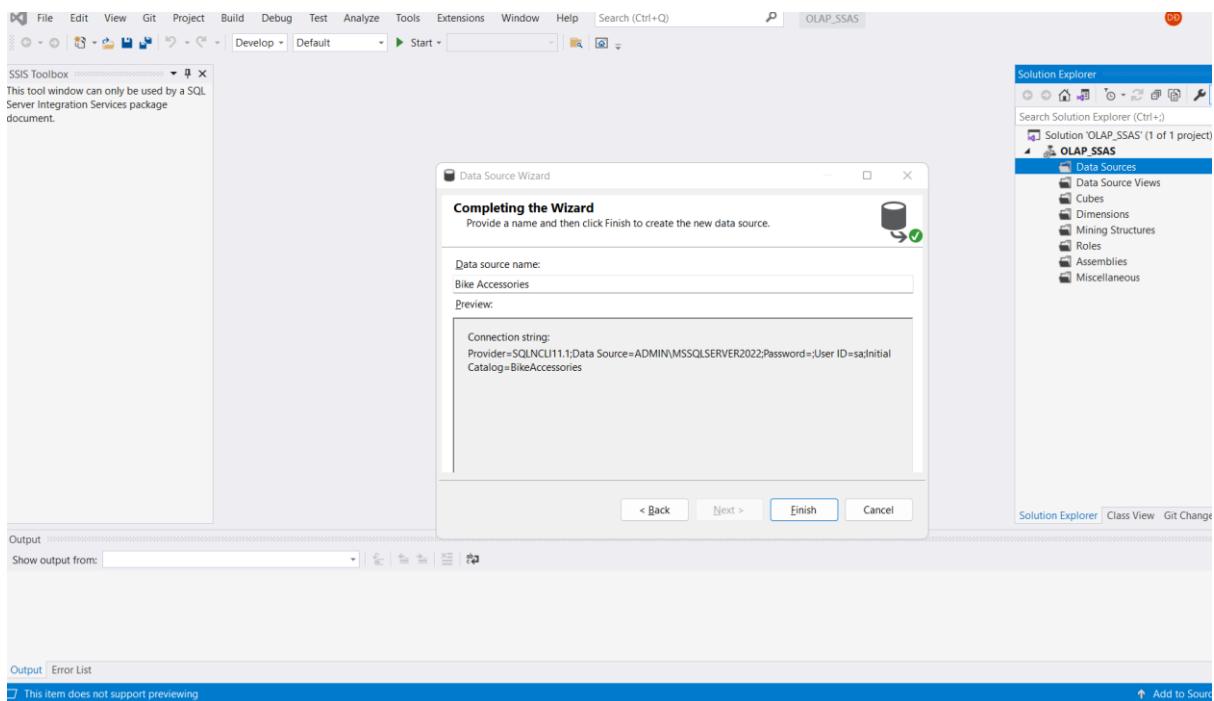
Bước 3: Click “New”, sau đó điền thông tin của Connection Manager.



Bước 4: Chọn “Use the service account”.

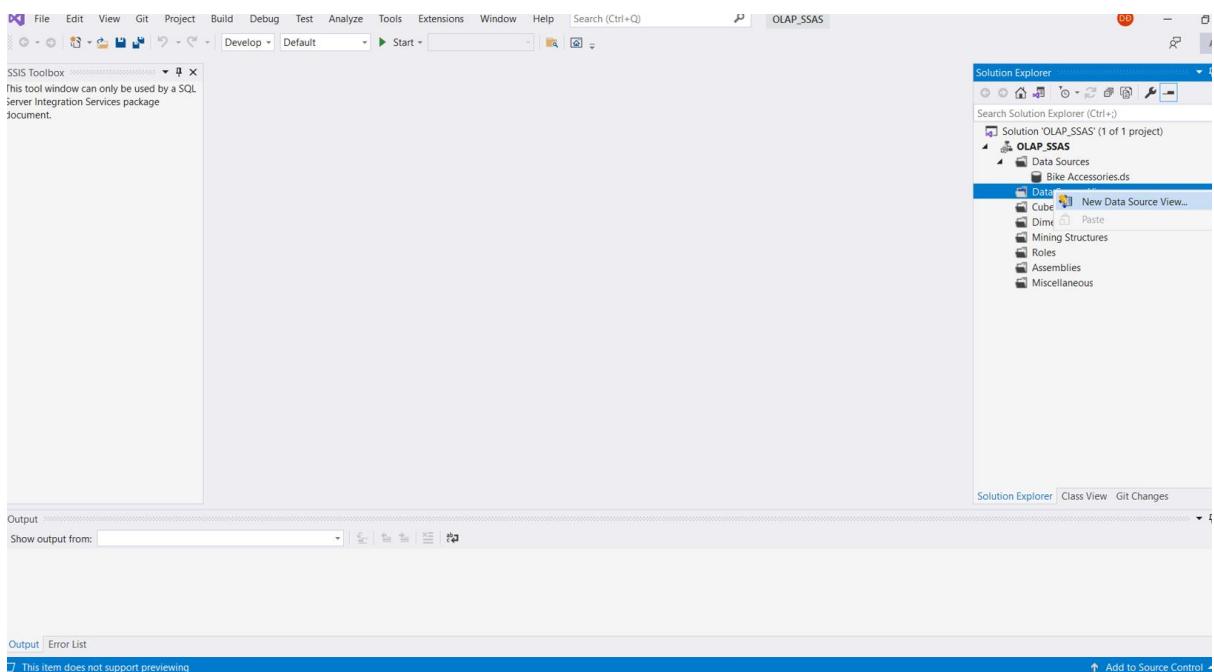


Bước 5: Chọn Next → Data source → Finish để hoàn tất.

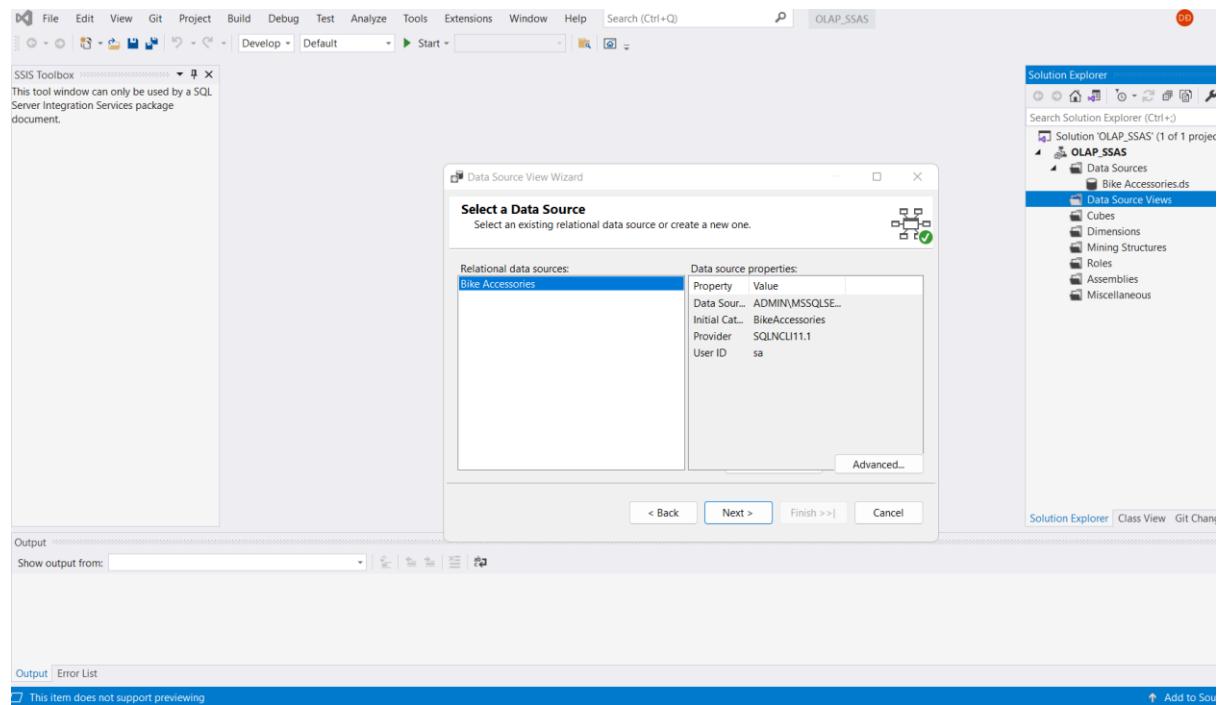


3. Tạo Data Source View

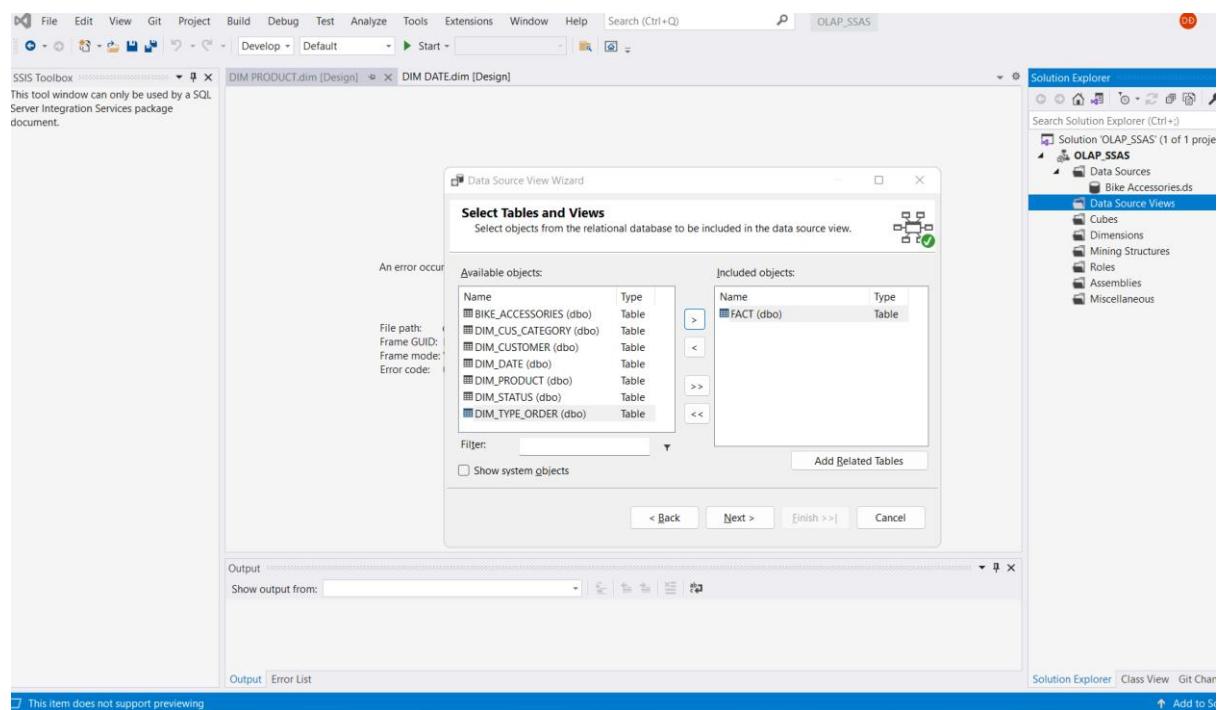
Bước 1: Bên góc phải phần “Solution Explorer”, chuột phải Data Source View → New Data Source View.



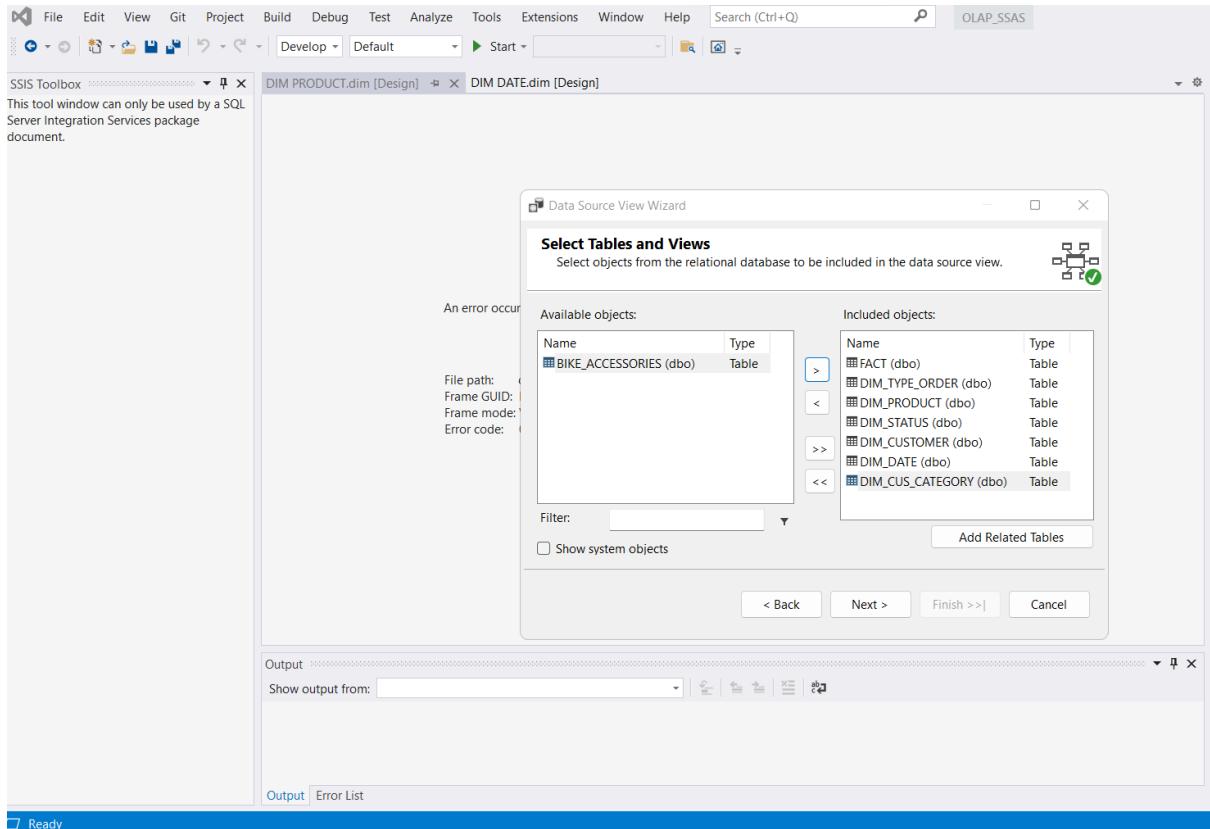
Bước 2: Kết nối với Database đã kết nối trước rồi nhấn “Next” để tiếp tục.



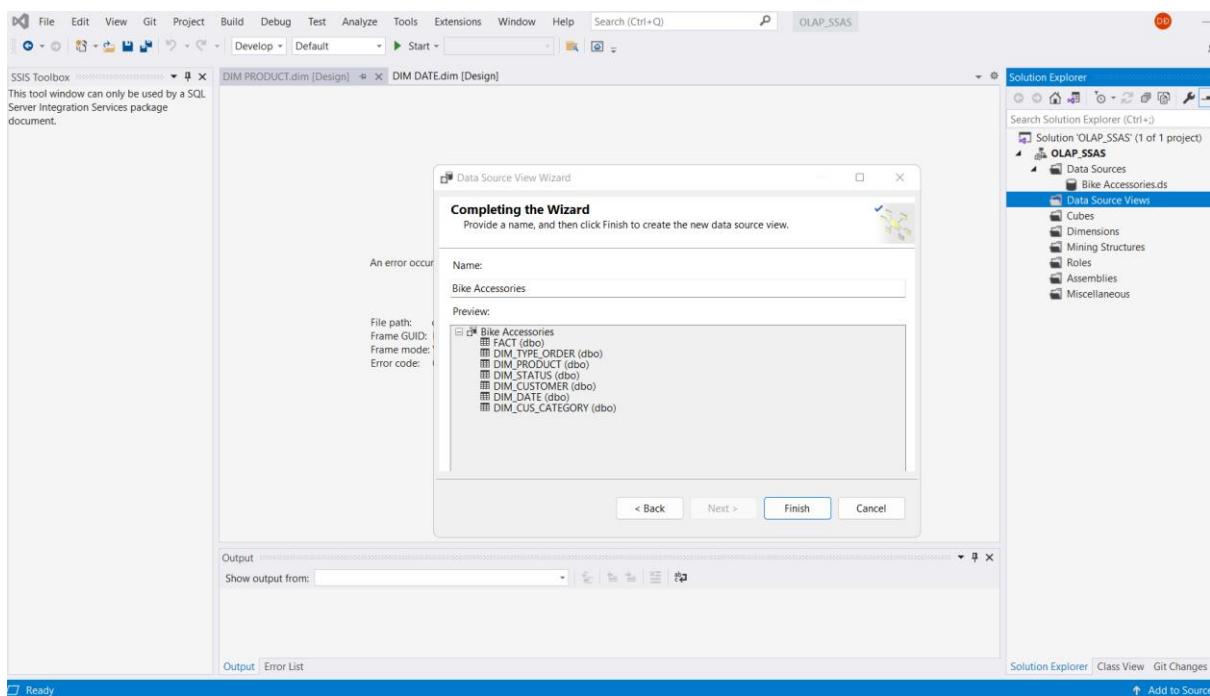
Bước 3: Chọn bảng Fact sau đó click “>”.



Bước 4: Tại phần “Included objects” Chọn “Add Related Tables”, sau đó chọn “Next”.

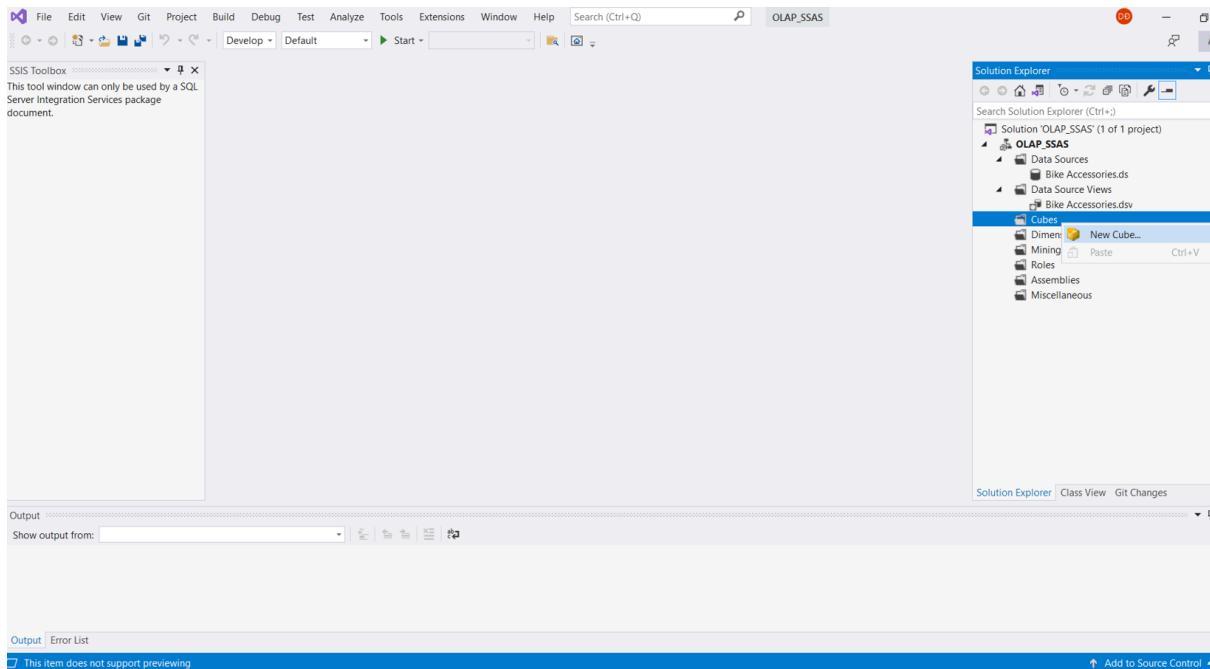


Bước 5: Kiểm tra các bảng, sau đó nhấn ‘Finish’.

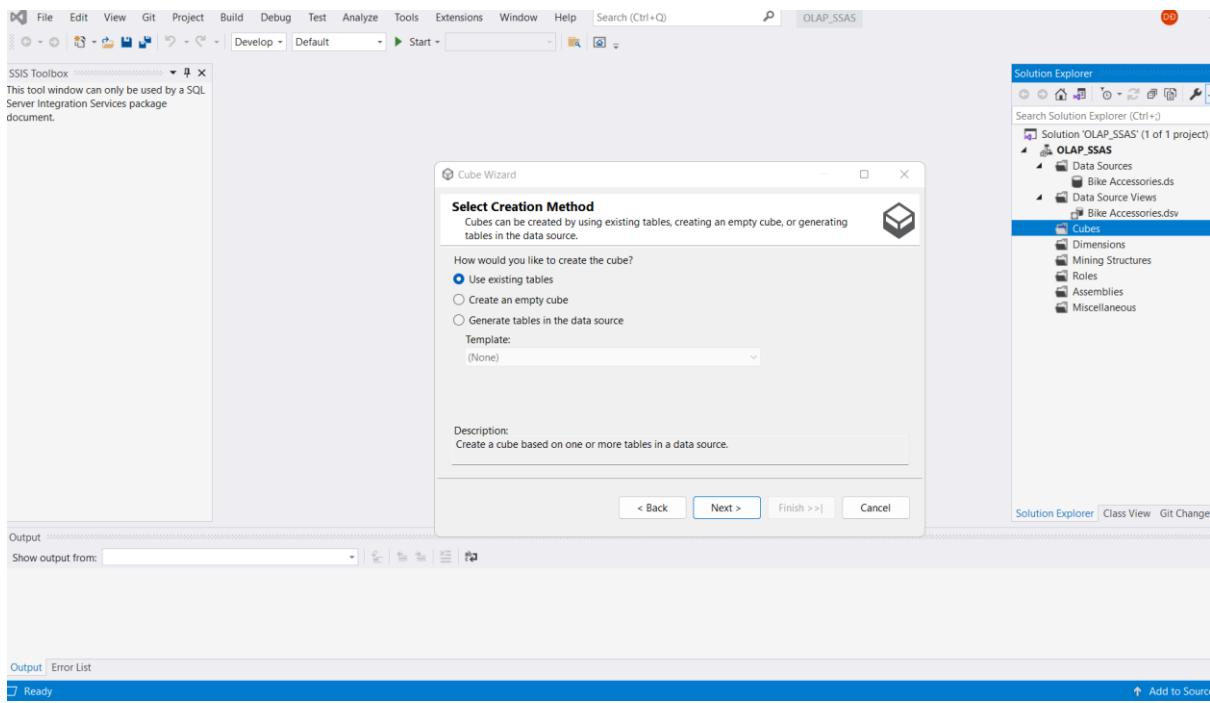


4. Tạo Cube và Dimensions

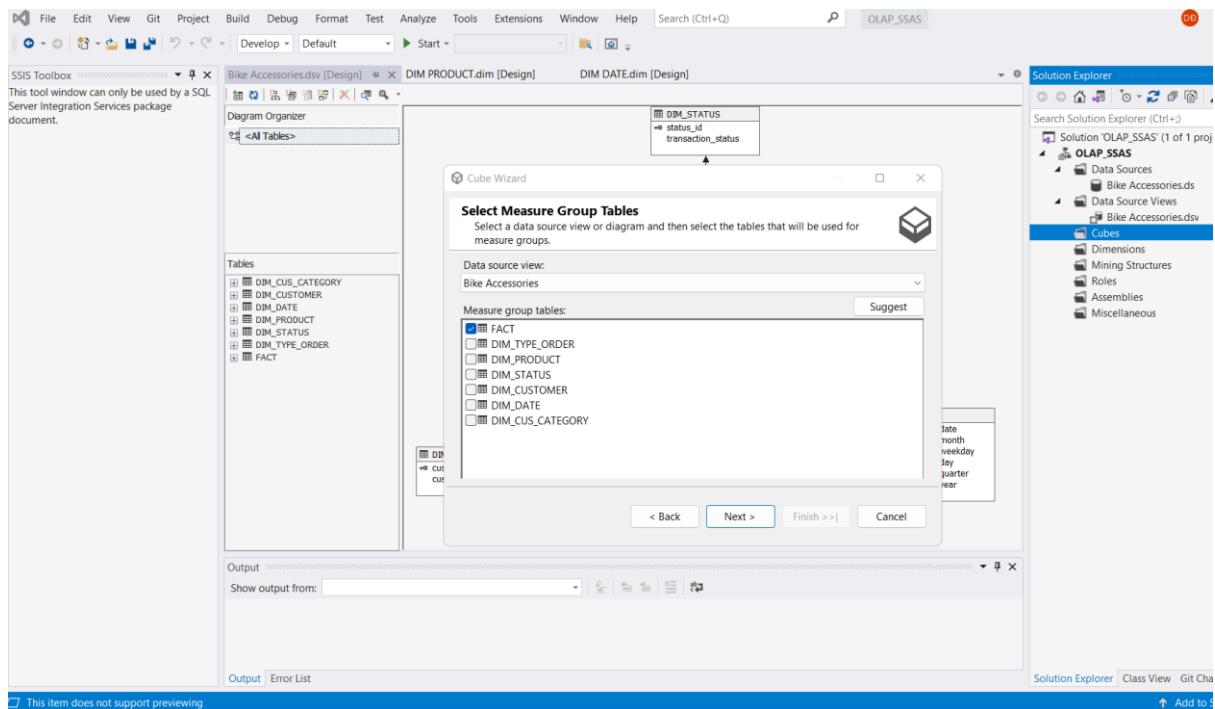
Bước 1: Chọn Cube → New Cubes.



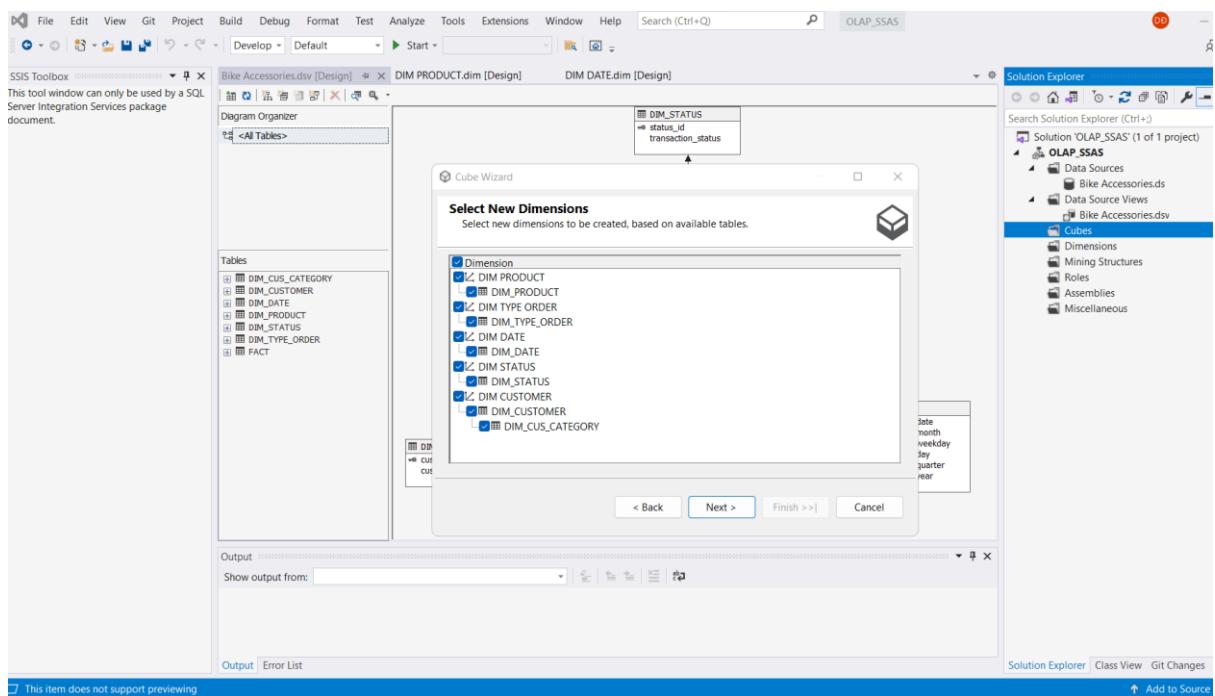
Bước 2: Chọn Use existing table → Next.

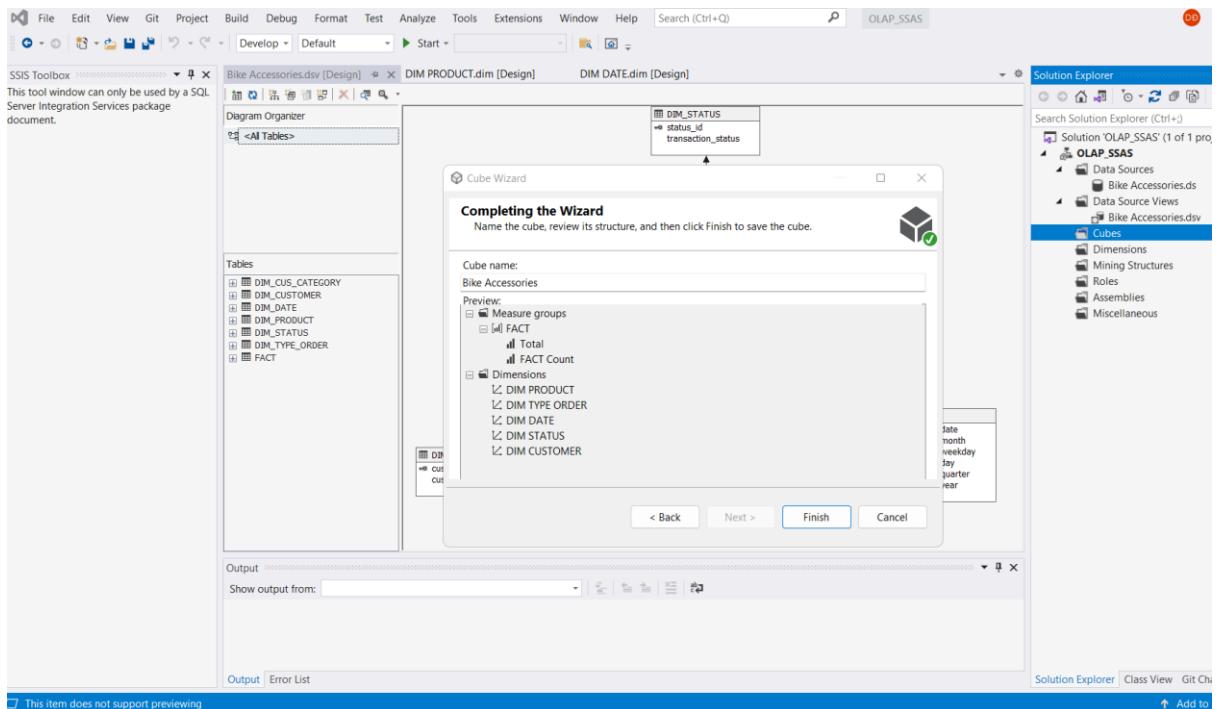


Bước 3: Chọn “Measure group table” là bảng “FACT”.

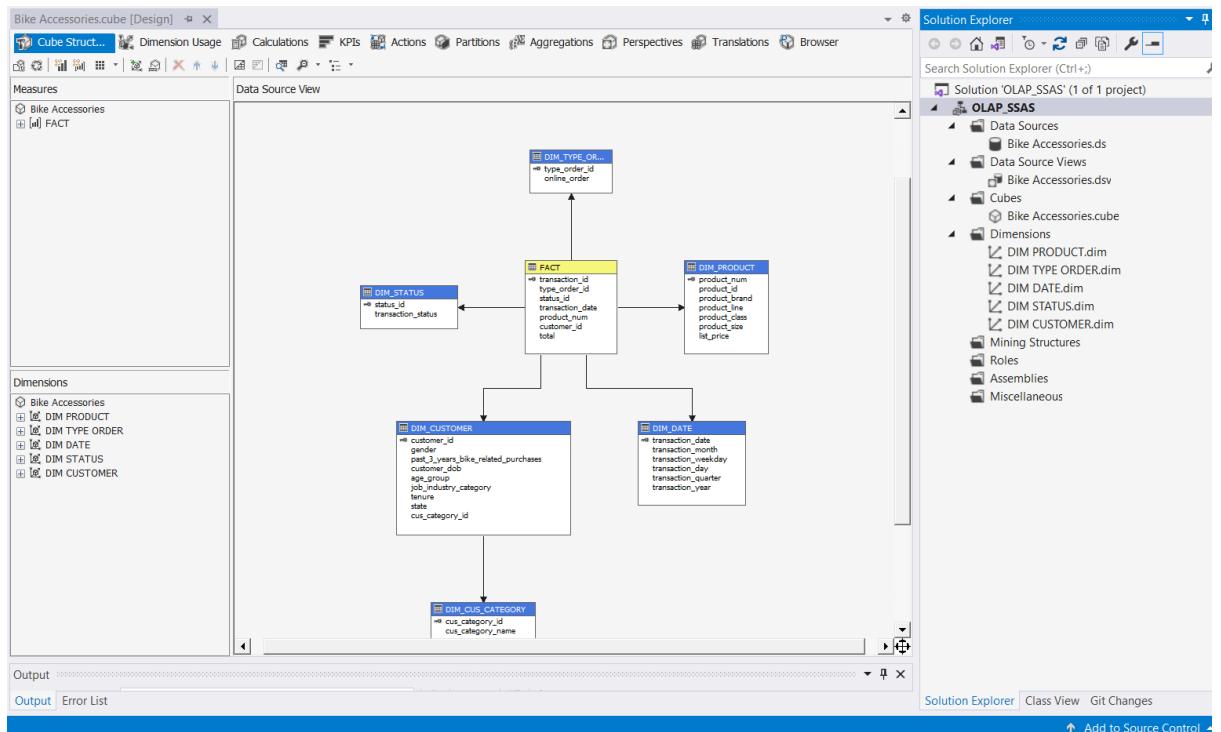


Bước 4: Chọn những Dimension cần add vào sau đó chọn “Next”, kiểm tra và nhấn “Finish”.





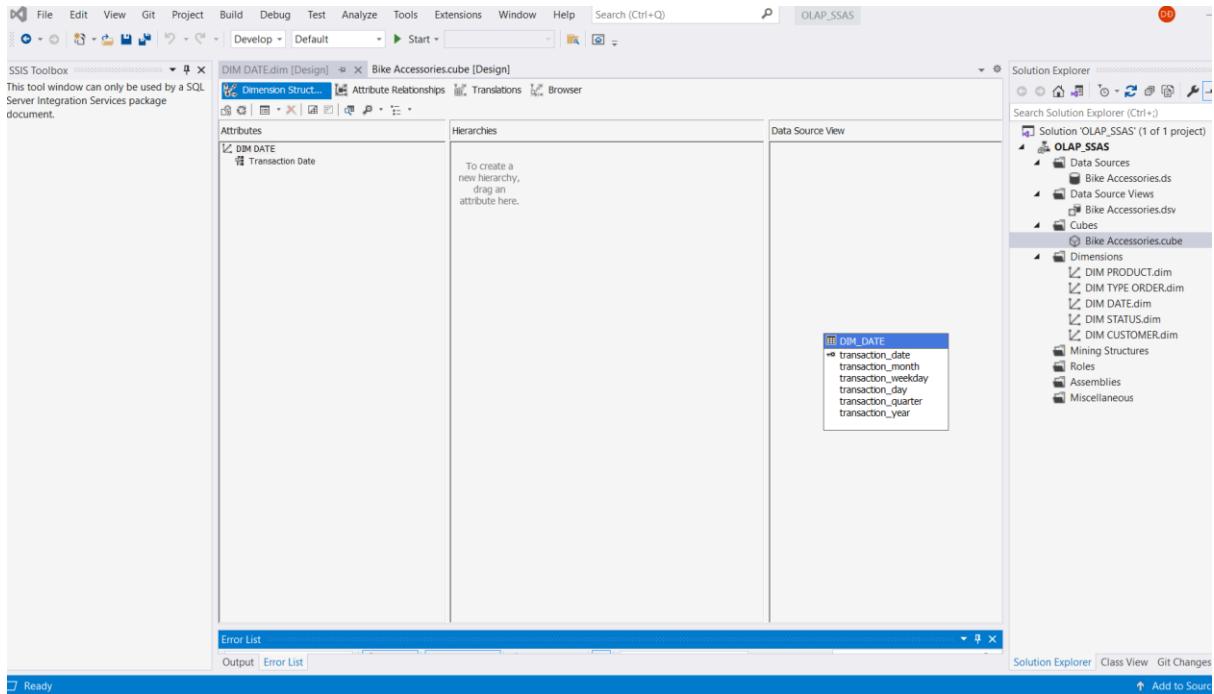
Bước 5: Sau đó các bảng Dim sẽ được tạo tự động.



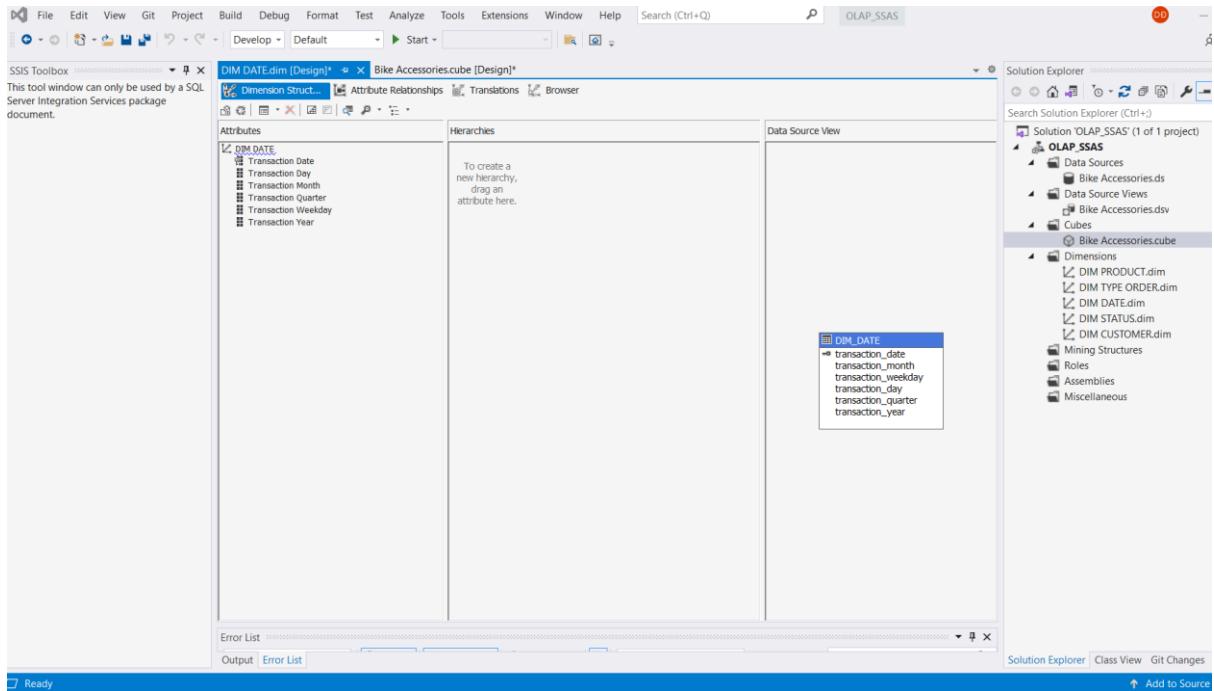
5. Thao tác trên các bảng Dim

5.1 Bảng DIM_DATE

Bước 1: Chọn bảng Dim Date.dim từ khung “Search Solution Explorer”.

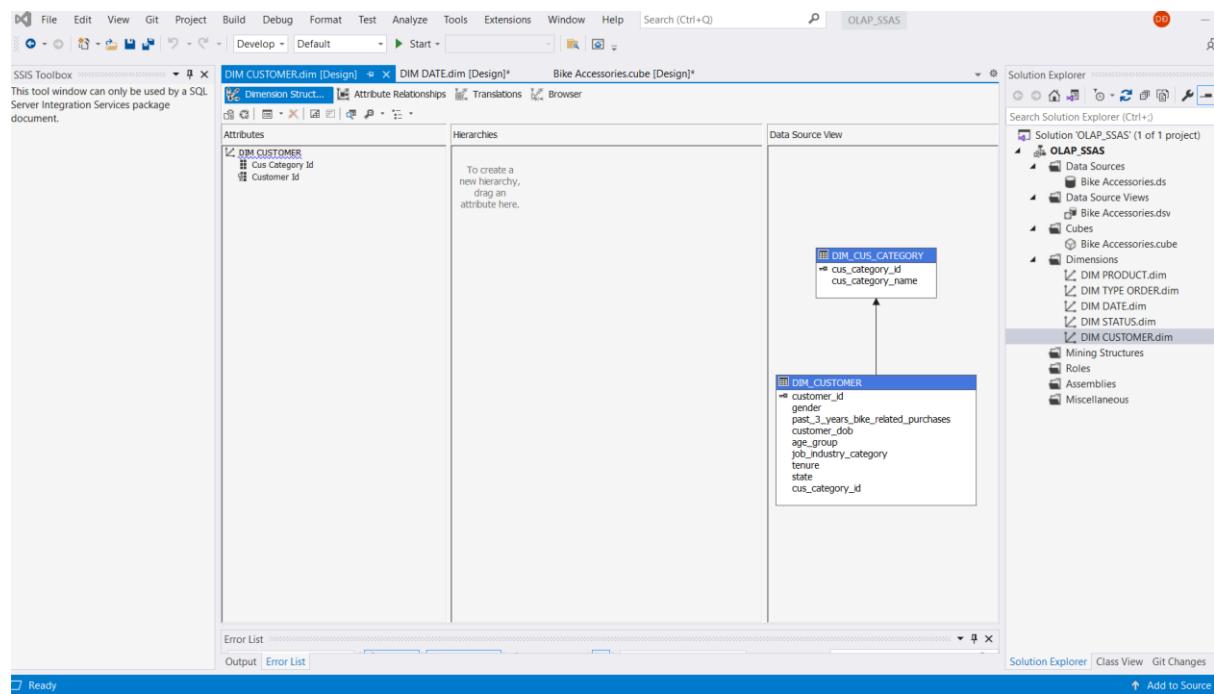


Bước 2: Chọn những thuộc tính của bảng Dim Date chưa xuất hiện ở cột “Attributes”. Nhấn giữ chuột trái và kéo từ cột “Data Source View” sang cột “Attributes”.

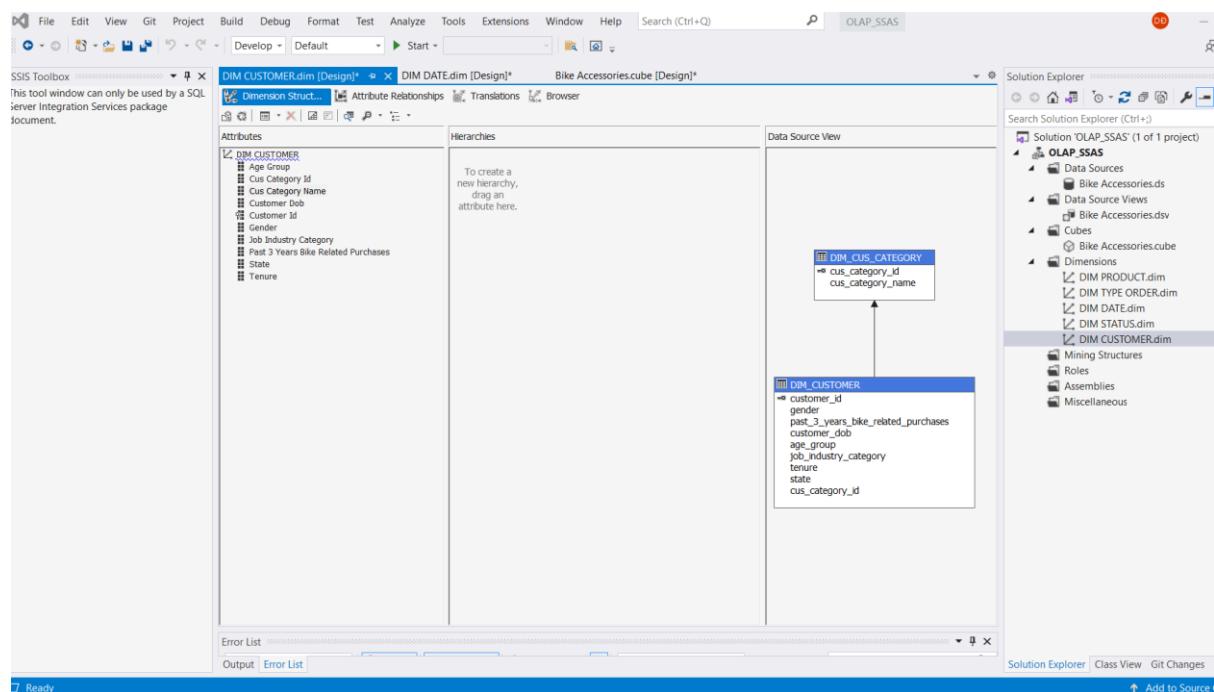


5.2 Bảng DIM_CUSTOMER

Bước 1: Chọn bảng Dim Customer.dim từ khung “Search Solution Explorer”.

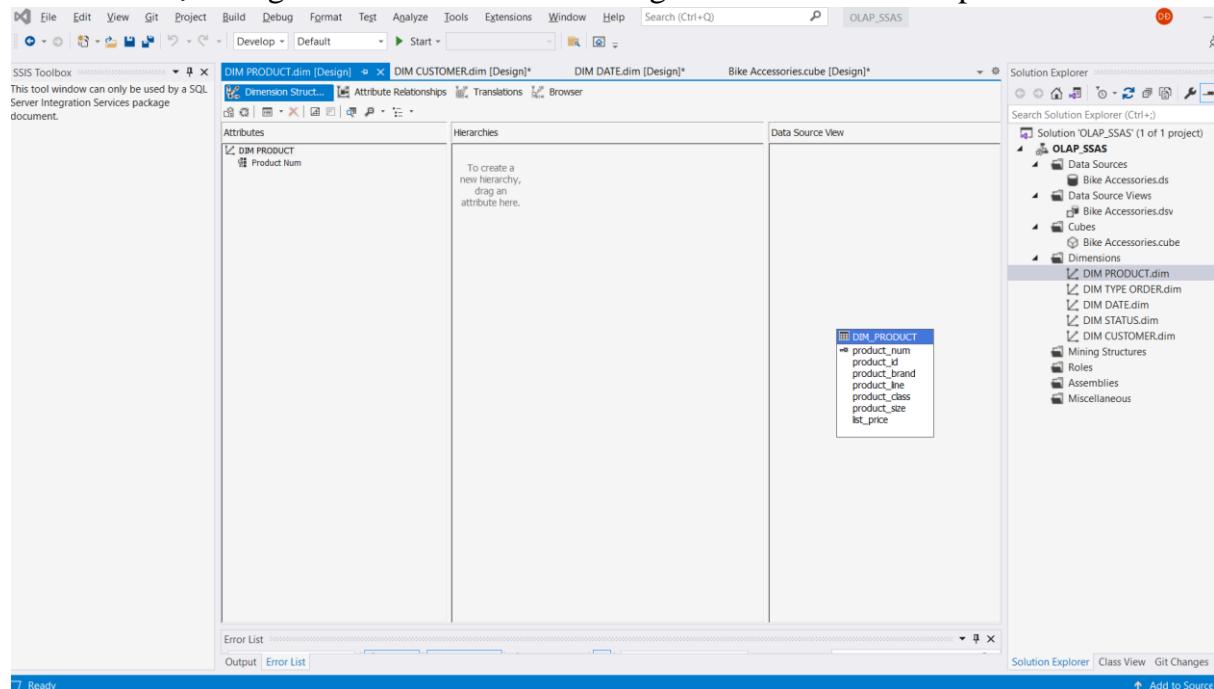


Bước 2: Chọn những thuộc tính của bảng Dim Customer chưa xuất hiện ở cột “Attributes”. Nhấn giữ chuột trái và kéo từ cột “Data Source View” sang cột “Attributes”.

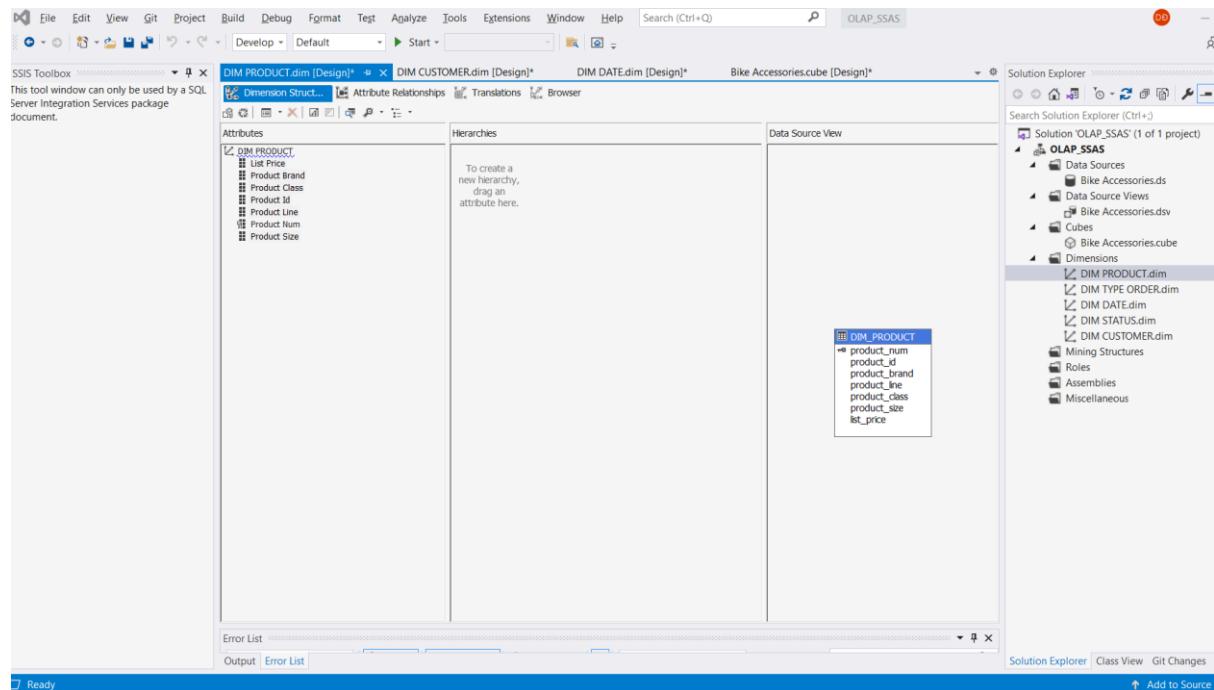


5.3 Bảng DIM_PRODUCT

Bước 1: Chọn bảng Dim Product.dim từ khung “Search Solution Explorer”.

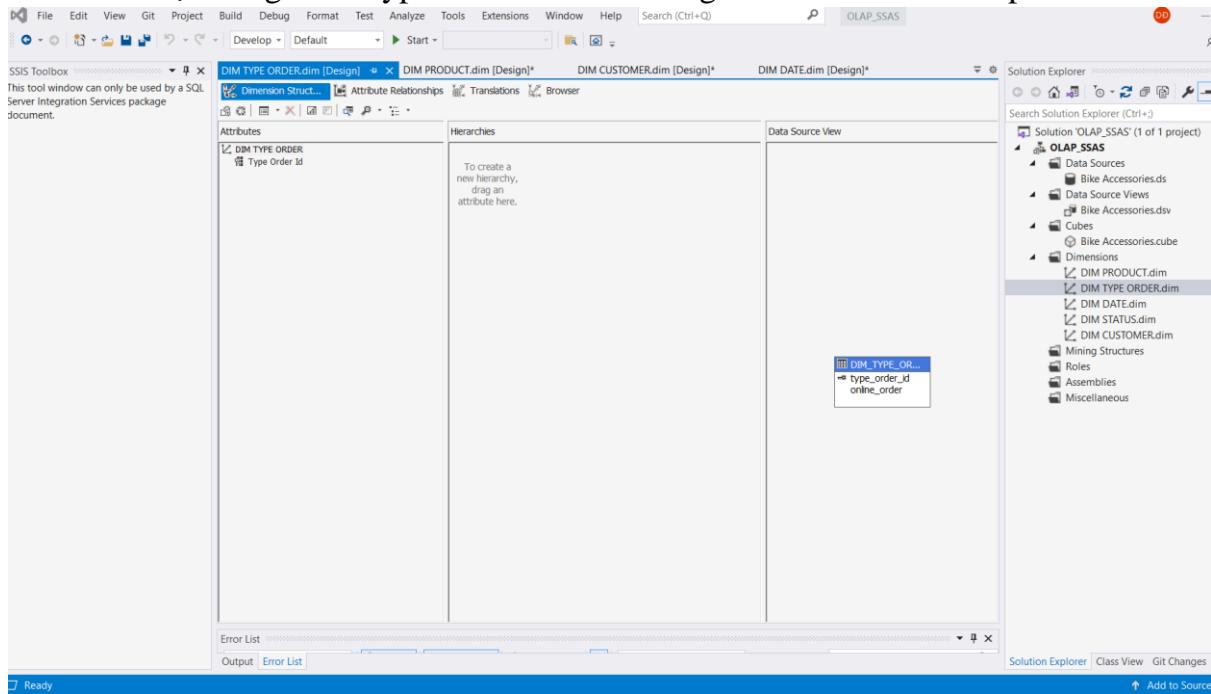


Bước 2: Chọn những thuộc tính của bảng Dim Product chưa xuất hiện ở cột “Attributes”. Nhấn giữ chuột trái và kéo từ cột “Data Source View” sang cột “Attributes”.

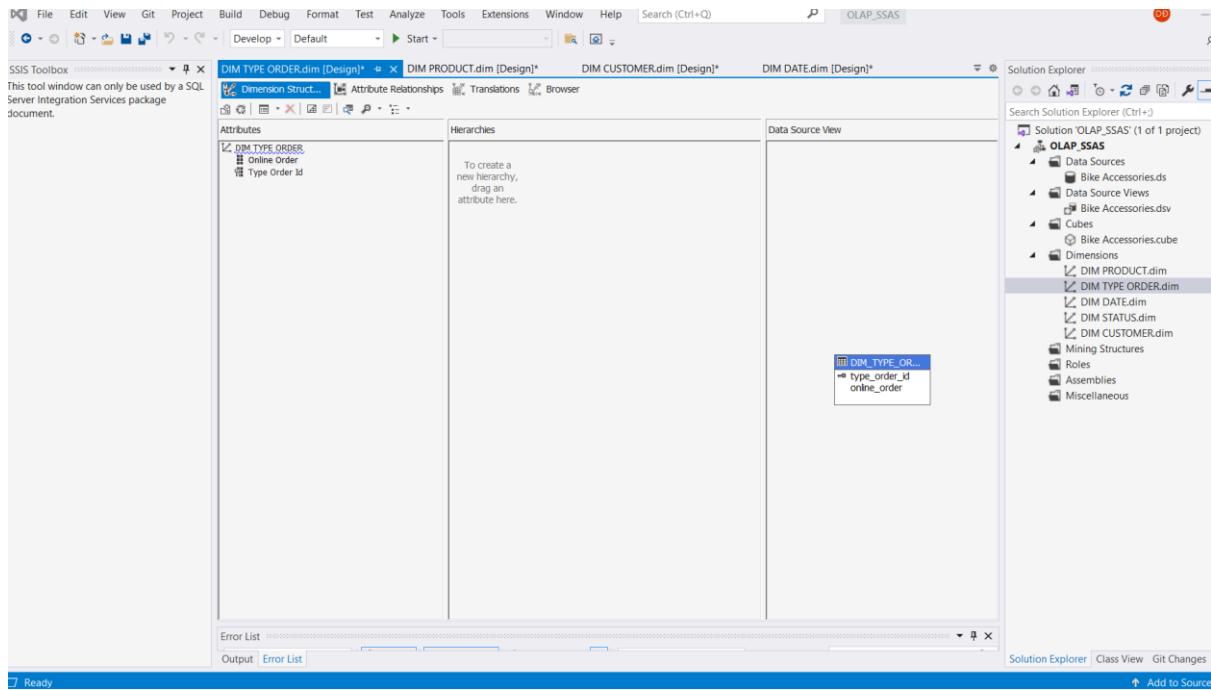


5.4 Bảng DIM_TYPE_ORDER

Bước 1: Chọn bảng Dim Type Order.dim từ khung “Search Solution Explorer”.

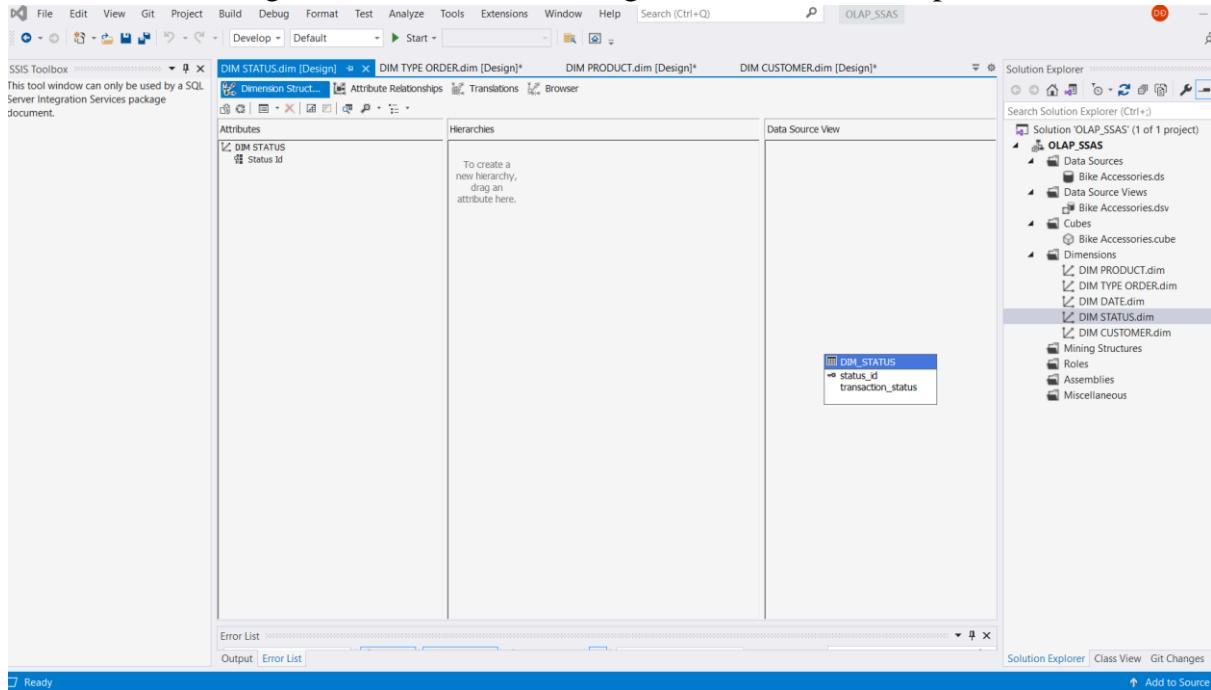


Bước 2: Chọn những thuộc tính của bảng Dim Type Order chưa xuất hiện ở cột “Attributes”. Nhấn giữ chuột trái và kéo từ cột “Data Source View” sang cột “Attributes”.

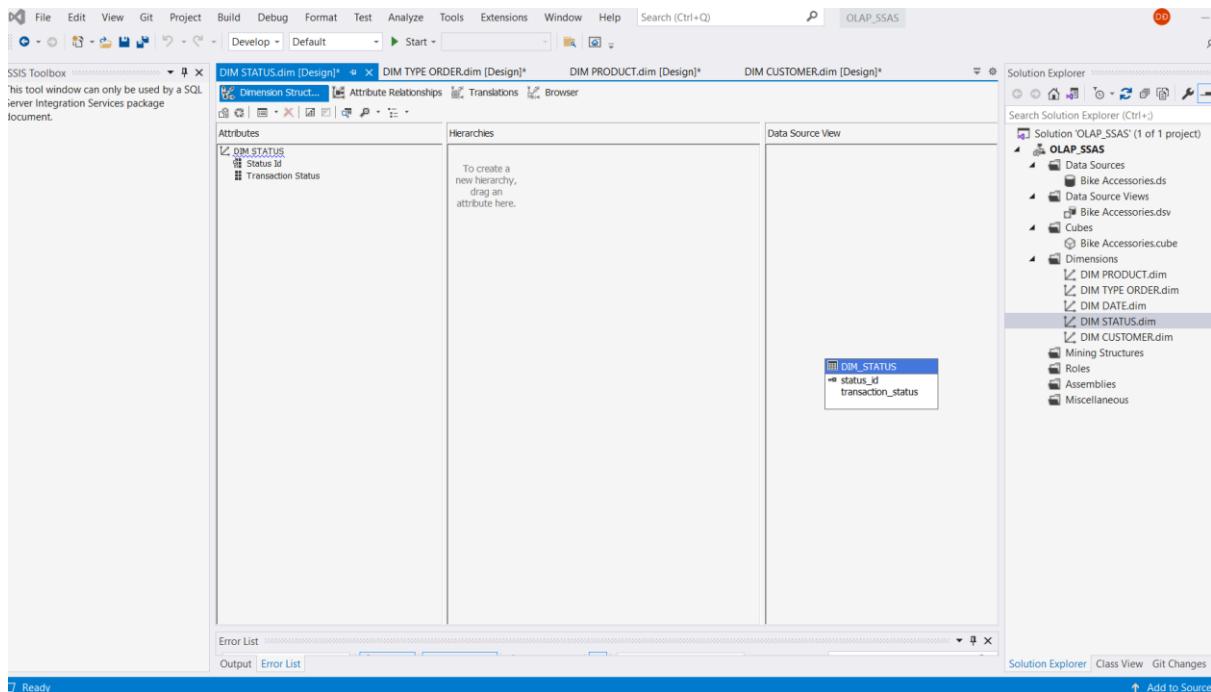


5.5 Bảng DIM_STATUS

Bước 1: Chọn bảng Dim Status.dim từ khung “Search Solution Explorer”.

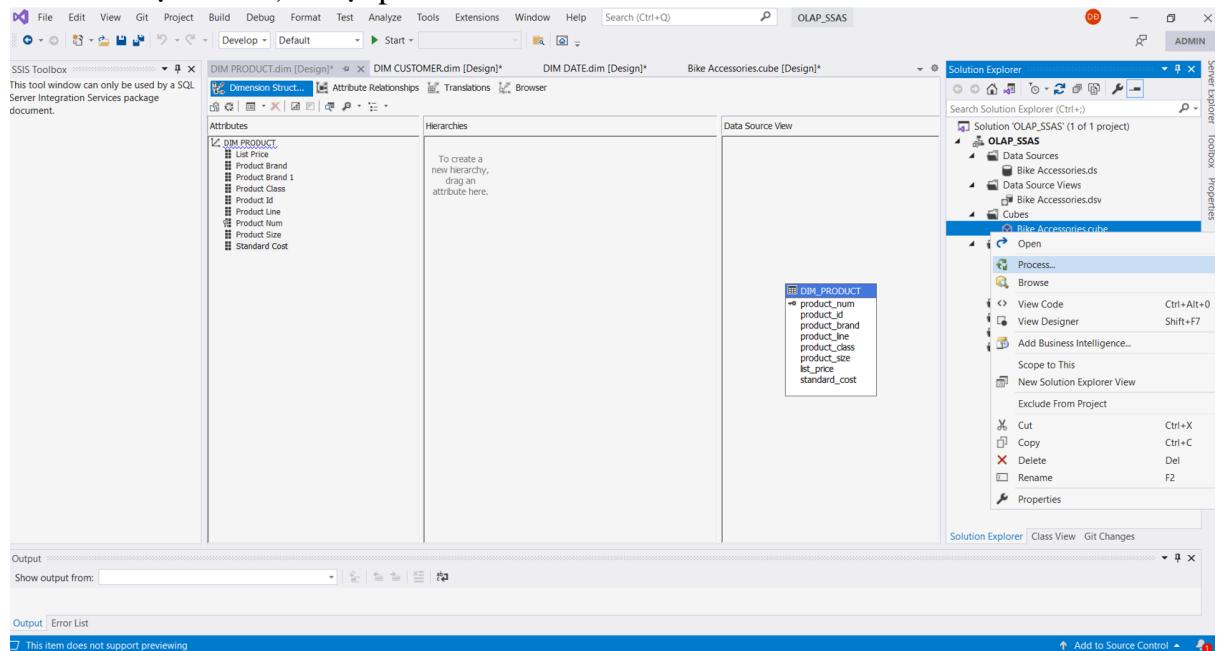


Bước 2: Chọn những thuộc tính của bảng Dim Status chưa xuất hiện ở cột “Attributes”. Nhấn giữ chuột trái và kéo từ cột “Data Source View” sang cột “Attributes”.

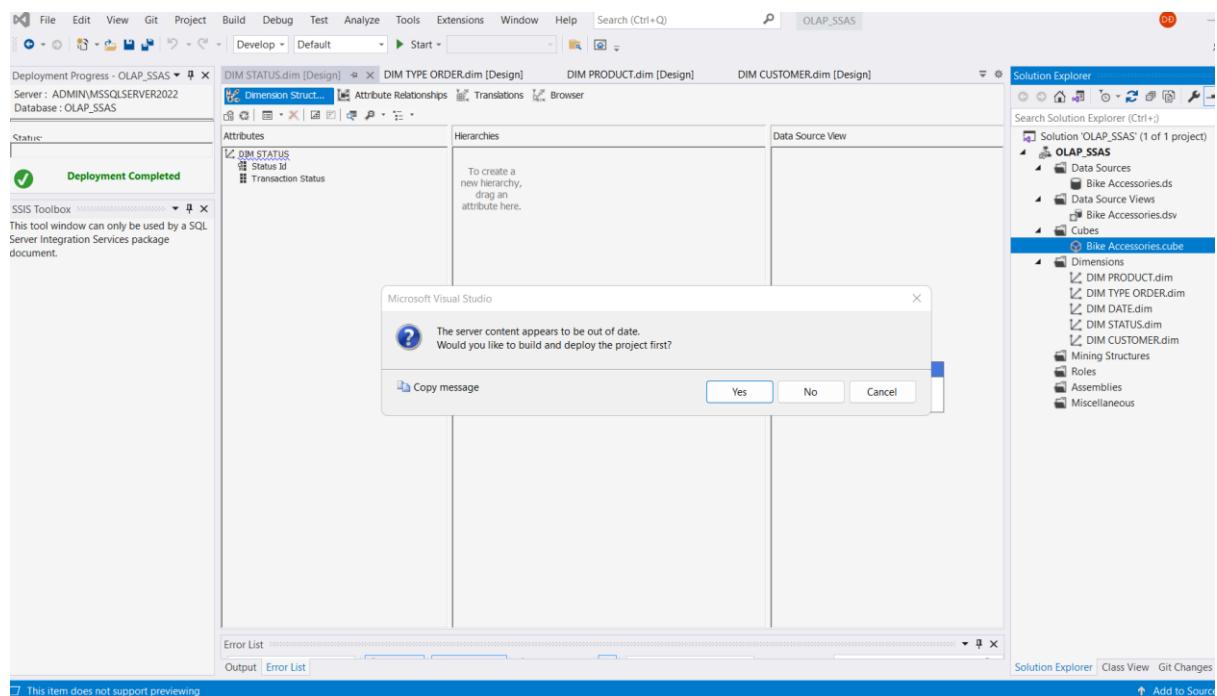


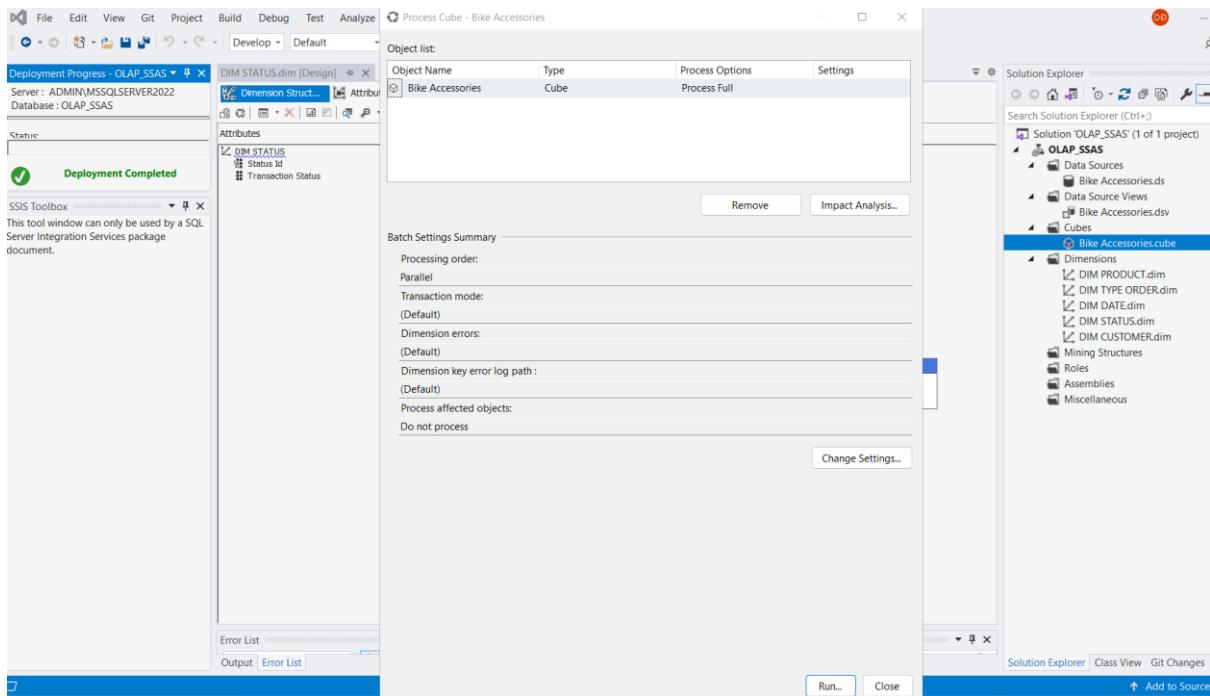
6. Deploy và process project

Bước 1: Tại Cubes, chuột phải vào Health Ware House → Process.

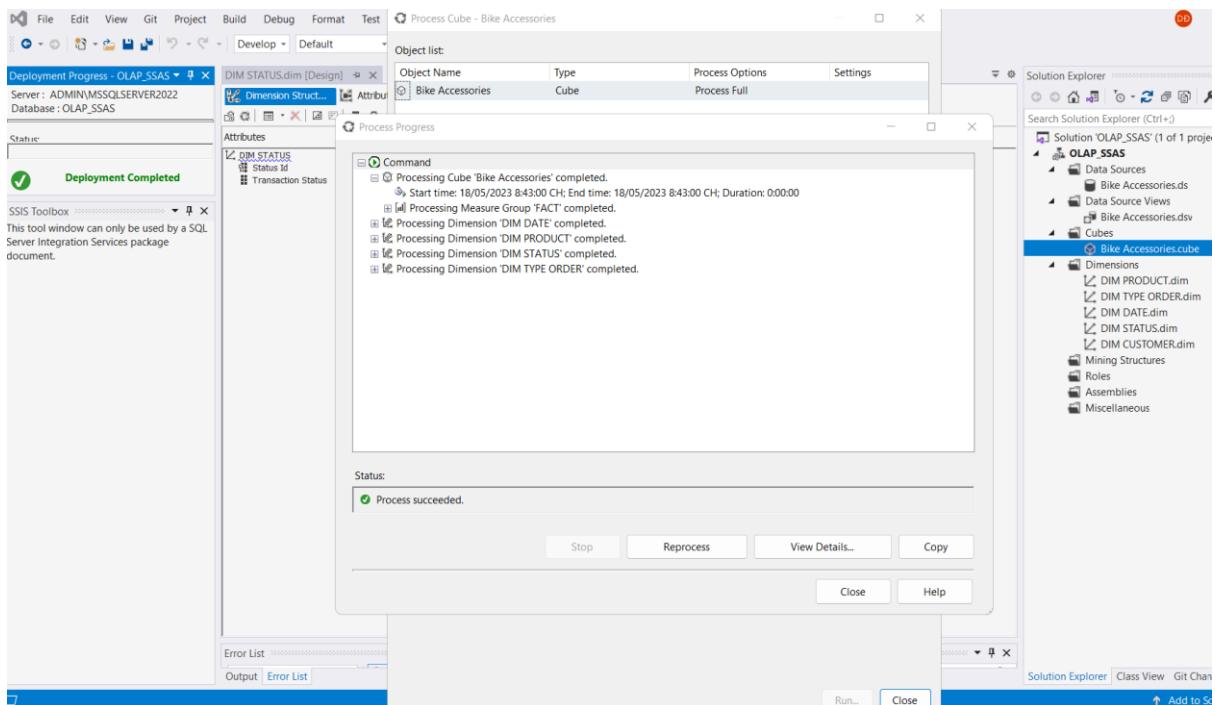


Bước 2: Chọn “Yes”. Sau đó chọn “Run” để bắt đầu thực hiện.





Bước 3: Sau khi chạy thành công ấn “Close” để hoàn tất.



7. Thực hiện 10 câu truy vấn

7.1 Thông kê tổng số giao dịch của tiểu bang NWS

❖ SSAS

Bước 1: Kéo thuộc tính “State” từ “DIM CUSTOMER” và “FACT Count” từ “FACT” sang khung truy vấn.

Dimension	Hierarchy	Operator	Filter Expression	Param...
DIM CUSTOMER	State	Equal	{NSW}	<input type="checkbox"/> <input checked="" type="checkbox"/>

Bước 2: Ở khung Filter chọn “DIM_CUSTOMER” cho “Dimension”, “State” cho “Hierarchy”, “Equal” cho “Operator” và “NSW” cho “Filter Expression”.

Bước 3: Chọn “Click to execute the query”.

State	FACT Count
NSW	6459

❖ Truy vấn bằng MDX

Query:

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON COLUMNS,
{ [DIM CUSTOMER].[State].&[NSW] } ON ROWS
FROM [Bike Accessories]
```

Kết quả:

```
--Câu 1. Thống kê tổng số giao dịch của tất cả những khách hàng ở tiểu bang NSW
SELECT NON EMPTY { [Measures].[FACT Count] } ON COLUMNS,
{ [DIM CUSTOMER].[State].[NSW] } ON ROWS
FROM [Bike Accessories]
```

	FACT Count
NSW	6459

Query executed successfully.

❖ Truy vấn bằng Pivot table trên Excel

Bước 1: Chọn tab “PivotTable Analyze”.

Bước 2: Ở mục Trường PivotTable kéo chọn thuộc tính “FACT Count” vào cột “Values” và “State” vào cột “Filters” và chọn điều kiện lọc của State là “NSW”.

State	FACT Count
NSW	6459

7.2 Thống kê doanh số của từng thương hiệu qua các quý

❖ SSAS

Bước 1: Kéo thuộc tính “Product Brand” từ DIM PRODUCT, “Transaction Quarter” từ DIM DATE và “Total” từ “FACT” sang khung truy vấn.

The screenshot shows the Microsoft Analysis Services (SSAS) Cube Design interface. The top menu bar includes options like Cube Struct..., Dimension Usage, Calculations, KPIs, Actions, Partitions, Aggregations, Perspectives, Translations, and Browser. The Language is set to Default. The main workspace is titled "Bike Accessories.cube [Design]". On the left, there's a tree view of the cube structure under "Bike Accessories" and "Metadata". Under "Measure Group", there are several items listed. In the center, a query grid is displayed with columns "Product Brand", "Transaction Quarter", and "Total". A tooltip "Click to execute the query." is shown near the grid. At the bottom, there's an "Error List" pane with tabs for Output and Error List.

Bước 2: Chọn “Click to execute the query”.

Product Brand	Transaction Quarter	Total
Giant Bicycles	1	1417346,99
Giant Bicycles	2	1454357,43
Giant Bicycles	3	1480068,6
Giant Bicycles	4	1570107,02
Norco Bicycles	1	788452,96000006
Norco Bicycles	2	746614,15000005
Norco Bicycles	3	722697,55000005
Norco Bicycles	4	757092,75000005
OHM Cycles	1	518744,25000001
OHM Cycles	2	539314,24000001
OHM Cycles	3	508365,46000001
OHM Cycles	4	505821,44000001
WeareA2B	1	874835,46000008
WeareA2B	2	845858,90000008
WeareA2B	3	894693,21000009
WeareA2B	4	890060,78000009

❖ Truy vấn bằng MDX

Query:

```
SELECT NON EMPTY { [Measures].[Total] } ON 0,
NON EMPTY CrossJoin ([DIM PRODUCT].[Product Brand].Children,
[DIM DATE].[Transaction Quarter].Children ) ON 1
FROM [Bike Accessories]
```

Kết quả:

The screenshot shows the SQL Server Management Studio interface with several tabs open:

- MDXQuery2.mdx - A...in\Administrator***: Shows a cube definition for "Bike Accessories" with measures like Total, Measures, KPIs, and various dimensions.
- SQLQuery12.sql**: Contains a query titled "Câu 2. Thống kê doanh số của từng thương hiệu qua các quý". The query uses SELECT NON EMPTY { [Measures].[Total] } ON 0, NON EMPTY CrossJoin ([DIM PRODUCT].[Product Brand].Children, [DIM DATE].[Transaction Quarter].Children) ON 1 FROM [Bike Accessories]
- SQLQuery11.sql - AD...cessories (sa (80))***
- SQLQuery10.sql - AD...cessories (sa (69))***

The results pane displays a table of sales data:

		Total
Giant Bicycles	1	1417346.99
Giant Bicycles	2	1454357.43
Giant Bicycles	3	1480068.6
Giant Bicycles	4	1570107.02
Norco Bicycles	1	78452.960000006
Norco Bicycles	2	746614.150000005
Norco Bicycles	3	72697.550000005
Norco Bicycles	4	757092.750000005
OHM Cycles	1	518744.250000001
OHM Cycles	2	539314.240000001
OHM Cycles	3	508365.460000001
OHM Cycles	4	505821.440000001
WeareA2B	1	874835.460000008
WeareA2B	2	845858.900000008
WeareA2B	3	894693.210000009
WeareA2B	4	890060.780000009

Message bar at the bottom: "Query executed successfully." and "ADMIN|MSSQLSERVER2022 Admin\Administrator OLAP_SSAS 00:00:01"

❖ Truy vấn bằng Pivot table trên Excel

Bước 1: Chọn tab “PivotTable Analyze”.

Bước 2: Ở mục Trường PivotTable kéo chọn các thuộc tính “Product Brand”, “Transaction Quarter” vào cột “Rows” và “Total” vào cột “Values”.

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable Fields pane on the right is configured as follows:

- Rows:** Transaction Quarter (selected)
- Values:** Total (selected)
- Fields:** DIM PRODUCT (List Price), Product Brand (selected)

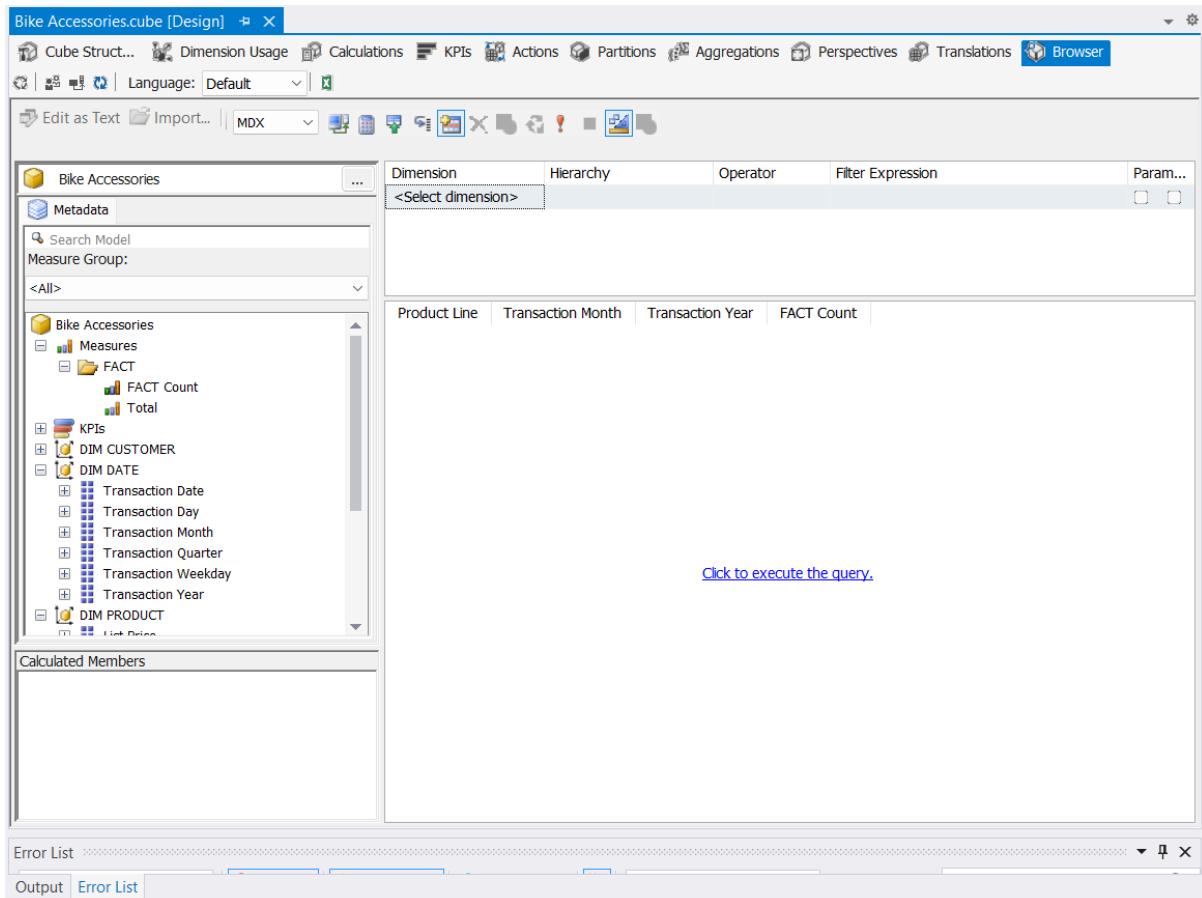
The main table area shows sales data by brand and quarter:

		Total
Row Labels	1	
	Giant Bicycles	1417346.99
	Norco Bicycles	78452.96
	OHM Cycles	518744.25
	WeareA2B	874835.46
Row Labels	2	
	Giant Bicycles	1454357.43
	Norco Bicycles	746614.15
	OHM Cycles	539314.24
	WeareA2B	845858.90
Row Labels	3	
	Giant Bicycles	1480068.6
	Norco Bicycles	72697.55
	OHM Cycles	508365.46
	WeareA2B	894693.21
Row Labels	4	
	Giant Bicycles	1570107.02
	Norco Bicycles	757092.75
	OHM Cycles	505821.44
	WeareA2B	890060.78
Grand Total		14514431.19

7.3 Thống kê tổng đơn đặt hàng online của từng dòng sản phẩm trong tháng 4 năm 2017

❖ SSAS

Bước 1: Kéo thuộc tính “Transaction Month” và “Transaction Year” từ “DIM DATE”, “Product Line” từ “PRODUCT” và “FACT Count” từ “FACT” sang khung truy vấn.



Bước 2: Ở khung Filter chọn “DIM DATE”, “DIM TYPE ORDER” cho “Dimension”, “Transaction Month”, “Transaction Year” cho “Hierarchy”, “Equal” cho “Operator” và “4”, “2017”, “True” cho “Filter Expression”.

Bước 3: Chọn “Click to execute the query”.

Product Line	Transaction Month	Transaction Year	FACT Count
Mountain	4	2017	19
Road	4	2017	119
Standard	4	2017	348
Touring	4	2017	35

❖ Truy vấn bằng MDX

Query:

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON 0,
NON EMPTY CrossJoin ([DIM PRODUCT].[Product Line].Children, [DIM
DATE].[Transaction Month].Children, [DIM DATE].[Transaction
Year].Children ) ON 1
FROM (SELECT ({[DIM TYPE ORDER].[Online Order].&[True]}) ON 0
FROM (SELECT ({[DIM DATE].[Transaction Year].&[2017]}) ON 0
FROM ( SELECT ({[DIM DATE].[Transaction Month].&[4]}) ON 0
FROM [Bike Accessories]))
WHERE ([DIM TYPE ORDER].[Online Order].&[True])
```

Kết quả:

-- Câu 3.Thống kê tổng đơn đặt hàng online của từng dòng sản phẩm trong tháng 4 năm 2017

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON 0,
NON EMPTY CrossJoin ([DIM PRODUCT].[Product Line].Children, [DIM DATE].[Transaction Month].Children,
[DIM DATE].[Transaction Year].Children ) ON 1
FROM ( SELECT ( {[DIM TYPE ORDER].[Online Order].&[True]} ) ON 0
      FROM ( SELECT ( {[DIM DATE].[Transaction Year].&[2017]} ) ON 0
              FROM ( SELECT ( {[DIM DATE].[Transaction Month].&[4]} ) ON 0
                      FROM [Bike Accessories])) )
WHERE ( [DIM TYPE ORDER].[Online Order].&[True] )
```

Product Line	Transaction Month	FACT Count
Mountain	4. 2017	19
Road	4. 2017	119
Standard	4. 2017	348
Touring	4. 2017	35

❖ Truy vấn bằng Pivot table trên Excel

Bước 1: Chọn tab “PivotTable Analyze”.

Bước 2: Ở mục Trường PivotTable kéo thuộc tính “FACT Count” vào cột “Values”, thuộc tính “Transaction Month”, “Transaction Year”, “Online Order” vào cột “Filters”, thuộc tính “Product Line” vào cột “Rows” và chọn điều kiện lọc của “Transaction Month”, “Transaction Year”, “Online Order” lần lượt là “4”, “2017”, “True”.

7.4 Thông kê tổng hóa đơn của từng nhóm khách hàng theo tháng trong năm

❖ SSAS

Bước 1: Kéo thuộc tính “Cus Category Name” từ DIM CUSTOMER, “Transaction Month”, “Transaction Year” từ DIM DATE và “Fact Count” từ “FACT” sang khung truy vấn.

Bước 2: Chọn “Click to execute the query”.

Transaction Month	Transaction Year	Cus Category Name	FACT Count
1	2017	Almost Lost Cust...	107
1	2017	Becoming Loyal	80
1	2017	Evasive Customer	112
1	2017	High Risk Customer	78
1	2017	Late Bloomer	102
1	2017	Losing Customer	111
1	2017	Lost Customer	55
1	2017	Platinum Customer	61
1	2017	Potential Customer	109
1	2017	Recent Customer	111
1	2017	Very Loyal	98
10	2017	Almost Lost Cust...	70
10	2017	Becoming Loyal	71
10	2017	Evasive Customer	11
10	2017	High Risk Customer	208
10	2017	Late Bloomer	114
10	2017	Losing Customer	204
10	2017	Lost Customer	5
10	2017	Platinum Customer	84
10	2017	Potential Customer	89
10	2017	Recent Customer	129
10	2017	Very Loyal	86

❖ Truy vấn bằng MDX

Query:

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON 0,
NON EMPTY CrossJoin ([DIM DATE].[Transaction Month].Children,
[DIM DATE].[Transaction Year].Children,
[DIM CUSTOMER].[Cus Category Name].Children ) ON 1
FROM [Bike Accessories]
```

Kết quả:

--Câu 4. Thống kê tổng hóa đơn theo từng nhóm khách hàng theo tháng trong năm

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON 0,
NON EMPTY CrossJoin ([DIM DATE].[Transaction Month].Children,
[DIM DATE].[Transaction Year].Children, [DIM CUSTOMER].[Cus Category Name].Children ) ON 1
FROM [Bike Accessories]
```

	FACT Count	
1 2017	Almost Lost Customer	107
1 2017	Becoming Loyal	80
1 2017	Evasive Customer	112
1 2017	High Risk Customer	78
1 2017	Late Bloomer	102
1 2017	Losing Customer	111
1 2017	Lost Customer	55
1 2017	Platinum Customer	61
1 2017	Potential Customer	109
1 2017	Recent Customer	111
1 2017	Very Loyal	98
10 2017	Almost Lost Customer	70
10 2017	Becoming Loyal	71
10 2017	Evasive Customer	11
10 2017	High Risk Customer	208
10 2017	Late Bloomer	114

Query executed successfully.

❖ Truy vấn bằng Pivot table trên Excel

Bước 1: Chọn tab “PivotTable Analyze”.

Bước 2: Ở mục Trường PivotTable kéo chọn các thuộc tính “Transaction Month”, “Transaction Year” và “Cus Category Name” vào cột “Rows” và “FACT Count” vào cột “Values”.

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable Fields pane on the right is open, showing the following fields:

- Rows: Cus Category Name, Transaction Month, Transaction Year, Cus Category Name (duplicated)
- Values: FACT Count

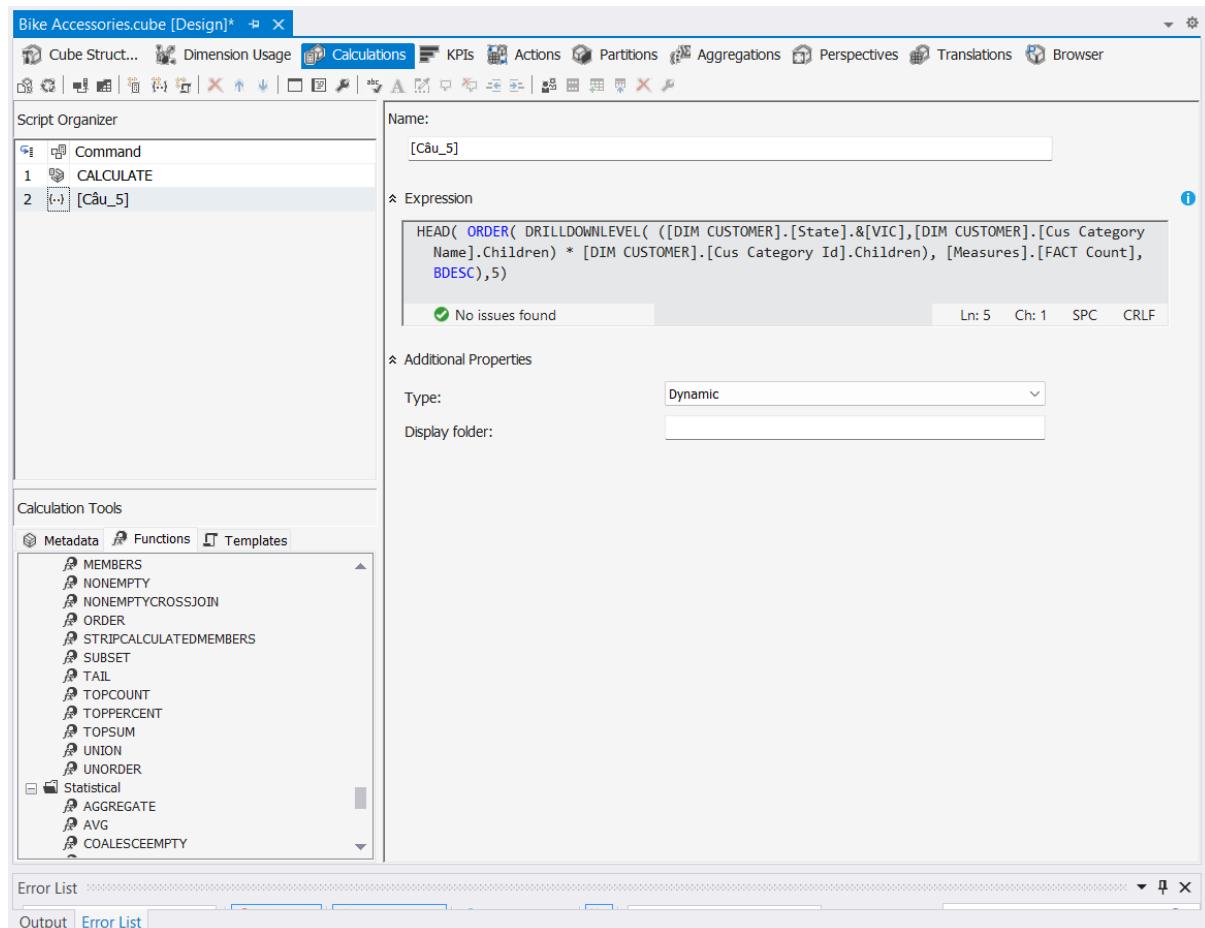
The main table area displays the following data:

Category	Month	Value
Almost Lost Customer	2017	107
Becoming Loyal	2017	80
Evasive Customer	2017	112
High Risk Customer	2017	78
Late Bloomer	2017	102
Losing Customer	2017	111
Lost Customer	2017	55
Platinum Customer	2017	61
Potential Customer	2017	109
Recent Customer	2017	111
Very Loyal	2017	98
Almost Lost Customer	2017	70
Becoming Loyal	2017	71
Evasive Customer	2017	11
High Risk Customer	2017	208
Late Bloomer	2017	114
Losing Customer	2017	204

7.5 Thống kê top 5 nhóm khách hàng thuộc bang VIC mua hàng nhiều nhất

❖ SSAS

Bước 1: Ở tab “Calculations” thêm 1 “Script Organizer” có tên là [Câu_5].



Script:

```
HEAD( ORDER( DRILLDOWNLEVEL( ([DIM  
CUSTOMER].[State].&[VIC], [DIM CUSTOMER].[Cus Category  
Name].Children) * [DIM CUSTOMER].[Cus Category Id].Children),  
[Measures].[FACT Count], BDESC), 5)
```

Bước 2: Kéo biến “Cus Category Name” và “FACT Count” sang khung truy vấn.

The screenshot shows the Microsoft Analysis Services (SSMS) cube design interface. The title bar reads "Bike Accessories.cube [Design]". The main area is the MDX query editor with the following content:

```
Dimension Hierarchy Operator Filter Expression Param...
<Select dimension>
```

Cus Category Name	FACT Count
-------------------	------------

In the left pane, the cube structure is displayed under "Bike Accessories". Under "Measures", there is a folder "FACT" containing "FACT Count" and "Total". Under "DIM CUSTOMER", there are various dimensions: Age Group, Cus Category Id, Cus Category Name, Customer Dob, Customer Id, Gender, Job Industry Category, Past 3 Years Bike Related Purchases, State, and Tenure. The "Cus Category Name" dimension is selected, indicated by a blue selection bar.

At the bottom right of the query editor, there is a link: "Click to execute the query."

At the bottom of the interface, there is an "Error List" tab.

Bước 3: Tại “Dimensions” chọn “DIM CUSTOMER”, ô “Hierarchy” chọn “Cus Category Name”, ô “Operator” chọn “Custom” và ô “Filter Expression” điền câu lệnh ở bước 1.

The screenshot shows the Analysis Services Designer application window titled "Bike Accessories.cube [Design]". The top menu bar includes options like Cube Struct..., Dimension Usage, Calculations, KPIs, Actions, Partitions, Aggregations, Perspectives, Translations, and Browser. The Language is set to Default. The main area has tabs for Edit as Text, Import..., and MDX. Below these are various toolbar icons. On the left, there's a tree view of the cube structure under "Bike Accessories" (Measures, FACT, KPIs, DIM CUSTOMER), and a section for Calculated Members. The right side contains a table with columns: Dimension, Hierarchy, Operator, Filter Expression, and Parameters. A dropdown says "<Select dimension>". The table row shows "DIM CUSTOMER" with "Cus Category Name" as the hierarchy, "Custom" as the operator, and a filter expression: "HEAD(ORDER(DRILLDOWNLEVEL(([DIM CUSTOMER].[Cus Category Name].ALL)), 1))". The table also has columns for Cus Category Name and FACT Count. A message "Click to execute the query." is displayed. At the bottom, there are Error List and Output tabs.

Bước 4: Chọn “Click to execute the query”.

This screenshot is identical to the one above, but it shows the results of the executed MDX query. The table in the main pane now displays data for the "FACT" measure, specifically the "FACT Count" for different customer categories. The data is as follows:

Cus Category Name	FACT Count
High Risk Customer	297
Losing Customer	284
Potential Customer	315
Recent Customer	394
Very Loyal	297

❖ Truy vấn bằng MDX

Query:

```
SELECT [Measures].[FACT Count] ON 0,
[DIM CUSTOMER].[Cus Category Name].Children ON 1
FROM ( SELECT ( TAIL ( ORDER (
DRILLDOWNLEVEL(([DIM CUSTOMER].[State].&[VIC],
[DIM CUSTOMER].[Cus Category Name].Children)
* [DIM CUSTOMER].[Cus Category Id].Children),
[Measures].[FACT Count], BASC),5 ) ) ON 0
FROM [Bike Accessories]);
```

Kết quả:

-- Câu 5. Thống kê top 5 nhóm khách hàng thuộc bang VIC mua hàng nhiều nhất

```
SELECT [Measures].[FACT Count] ON 0,
[DIM CUSTOMER].[Cus Category Name].Children ON 1
FROM ( SELECT ( TAIL ( ORDER (
DRILLDOWNLEVEL ( ([DIM CUSTOMER].[State].&[VIC],
[DIM CUSTOMER].[Cus Category Name].Children) *
[DIM CUSTOMER].[Cus Category Id].Children),
[Measures].[FACT Count], BASC),5 ) ) ON 0
FROM [Bike Accessories]);
```

	FACT Count
High Risk Customer	297
Losing Customer	284
Potential Customer	315
Recent Customer	394
Very Loyal	297

Query executed successfully.

❖ Truy vấn bằng Pivot table trên Excel

Bước 1: Chọn tab “PivotTable Analyze”.

Bước 2: Ở mục Trường PivotTable kéo chọn thuộc tính “Cus Category Name” vào cột “Rows”, “FACT Count” vào cột “Values”, “State” vào cột “Filters”.

PivotTable Fields

Choose fields to add to report:

Search

- Customer Dob
- Customer Id
- Gender
- Job Industry Category
- Past 3 Years Bike Related Purchases
- State

Drag fields between areas below:

Filters	Columns
Rows	Values
Cus Category Name	FACT Count

Bước 3: Row Labels → Values Filters → Top 10...

Value Filters

Search Cus Category Name

- (Select All)
- Almost Lost Customer
- Becoming Loyal
- Evasive Customer
- High Risk Customer
- Late Bloomer
- Losing Customer
- Lost Customer
- Platinum Customer
- Potential Customer

OK Cancel

Top 10...

Bước 4: Tại cửa sổ “Show” sửa số lượng hiển thị lại thành 5 và nhấn “OK”.

The screenshot shows a Microsoft Excel spreadsheet titled "Pivot_Table_OLAP.xlsx". A PivotTable is displayed in the range B5:P11. The rows are labeled "Row Labels" and "FACT Count". The columns are labeled "State" and "VIC". The data includes various customer categories and their counts, such as "Almost Lost Customer" (275), "Becoming Loyal" (233), "Evasive Customer" (276), etc., ending with a "Grand Total" of 2978. A "Top 10 Filter (Cus Category Name)" dialog box is open over the table, showing "Top" selected, "5" items, and "by FACT Count". The "PivotTable Fields" pane on the right lists fields like Customer Dob, Customer Id, Gender, Job Industry Category, Past 3 Years Bike Related Purchases, and State, with State checked. The "Filters" section shows State selected. The "Rows" section shows Cus Category Name selected, and the "Values" section shows FACT Count selected.

Kết quả:

The screenshot shows the same Microsoft Excel spreadsheet after applying the filter. The PivotTable now displays only the top 5 categories: High Risk Customer (297), Losing Customer (284), Potential Customer (315), Recent Customer (394), and Very Loyal (297), with a Grand Total of 1587. The "Top 10 Filter" dialog box is no longer visible. The "PivotTable Fields" pane remains the same as in the previous screenshot.

7.6 Thống kê những khách hàng mua hàng hơn 9 lần và tổng số tiền mua hơn 10000\$

❖ SSAS

Bước 1: Ở tab “Calculations” thêm 1 “Script Organizer” có tên là [Câu_6].

Name: [Câu_6]

Expression

```
INTERSECT(
    FILTER([DIM CUSTOMER].[Customer Id].[Customer Id], [Measures].[FACT Count] > 9),
    FILTER([DIM CUSTOMER].[Customer Id].[Customer Id], [Measures].[Total]>10000)
)
```

No issues found

Ln: 8 Ch: 1 SPC CRLF

Additional Properties

Type: Dynamic

Display folder:

Calculation Tools

Measure Group: <All>

Bike Accessories

- Measures
 - FACT
 - FACT Count
 - Total
- DIM CUSTOMER
 - Age Group
 - Cus Category Id
 - Members
 - Cus Category Id
 - Cus Category Name

Error List

Output Error List

Script:

```
INTERSECT(
    FILTER([DIM CUSTOMER].[Customer Id].[Customer Id],
    [Measures].[FACT Count] > 9),
    FILTER([DIM CUSTOMER].[Customer Id].[Customer Id],
    [Measures].[Total]>10000)
)
```

Bước 2: Kéo biến “Customer Id”, “Total” và “FACT Count” sang khung truy vấn.

The screenshot shows the Microsoft Analysis Services (SSMS) cube design interface. The left pane displays the cube structure with dimensions like 'Bike Accessories' and 'DIM CUSTOMER', and measures like 'FACT Count', 'Total', and 'Customer Id'. The right pane shows a query builder with columns 'Dimension', 'Hierarchy', 'Operator', and 'Filter Expression'. The 'Dimension' column has a dropdown labeled '<Select dimension>'. Below the query builder is a table with three columns: 'Customer Id', 'Total', and 'FACT Count'. A message 'Click to execute the query.' is visible. At the bottom, there is an 'Error List' window.

Bước 3: Tại “Dimensions” chọn “DIM CUSTOMER”, ô “Hierarchy” chọn “Customer Id”, ô “Operator” chọn “Custom” và ô “Filter Expression” điền câu lệnh ở bước 1.

The screenshot shows the Analysis Services Designer application window titled "Bike Accessories.cube [Design]*". The top menu bar includes options like Cube Struct..., Dimension Usage, Calculations, KPIs, Actions, Partitions, Aggregations, Perspectives, Translations, and Browser. The MDX tab is selected in the toolbar. The main area contains a query editor with the following content:

```

SELECT
    INTERSECT(
        FILTER([DIM CUSTOMER].[Customer Id], Customer Id = 637),
        FILTER([FACT].[Total], Total > 1000)
    )
    ON COLUMNS
    FROM [Bike Accessories]

```

The results grid displays the following data:

Customer Id	Total	FACT Count
637	11494,76	10
936	14747,19	10
1129	14466,38	10
1913	14652,7	12
1991	15263,75	10
2183	14307,58	10
2476	11885,89	11
2874	13730,94	10

A message "Click to execute the query." is visible in the results grid area.

Bước 4: Chọn “Click to execute the query”.

The screenshot shows the Analysis Services Designer application window titled "Bike Accessories.cube [Design]*". The top menu bar includes options like Cube Struct..., Dimension Usage, Calculations, KPIs, Actions, Partitions, Aggregations, Perspectives, Translations, and Browser. The MDX tab is selected in the toolbar. The main area contains a query editor with the same query as the previous screenshot:

```

SELECT
    INTERSECT(
        FILTER([DIM CUSTOMER].[Customer Id], Customer Id = 637),
        FILTER([FACT].[Total], Total > 1000)
    )
    ON COLUMNS
    FROM [Bike Accessories]

```

The results grid displays the following data, identical to the previous screenshot:

Customer Id	Total	FACT Count
637	11494,76	10
936	14747,19	10
1129	14466,38	10
1913	14652,7	12
1991	15263,75	10
2183	14307,58	10
2476	11885,89	11
2874	13730,94	10

❖ Truy vấn bằng MDX

Query:

```
SELECT { [Measures].[Total], [Measures].[FACT Count] } ON 0,
{ ([DIM CUSTOMER].[Customer Id].Children ) } ON 1
FROM ( SELECT
( INTERSECT(
    FILTER([DIM CUSTOMER].[Customer Id].Children,
[Measures].[FACT Count] > 9),
    FILTER([DIM CUSTOMER].[Customer Id].Children,
[Measures].[Total]>10000))) ON 0
FROM [Bike Accessories])
```

Kết quả:

The screenshot shows the SQL Server Management Studio interface with multiple tabs open. The main query window contains the MDX code provided above. The results pane displays a table with three columns: Customer ID, Total, and FACT Count. The data is as follows:

	Total	FACT Count
637	11494.76	10
936	14747.19	10
1129	14466.38	10
1913	14652.7	12
1991	15263.75	10
2183	14307.58	10
2476	11885.89	11
2874	13730.94	10

7.7 Thống kê danh sách và số lượng đơn hàng khách hàng nữ đã mua và đã thanh toán tại cửa hàng ngày 8/3/2017 không thuộc bang VIC

❖ SSAS

Bước 1: Kéo thuộc tính “Customer Id”, “State” từ “DIM CUSTOMER” và “FACT Count” từ “FACT” sang khung truy vấn.

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface for the 'Bike Accessories.cube' in design mode. The left pane displays the cube structure with dimensions like 'DIM CUSTOMER' and measures like 'FACT Count'. The right pane shows a query grid with columns 'Customer Id', 'State', and 'FACT Count'. A message 'Click to execute the query.' is displayed in the grid area. The bottom pane shows the error list.

Bước 2: Tại “Dimensions” chọn lần lượt “DIM CUSTOMER”, “DIM TYPE ORDER”, “DIM STATUS”, “DIM DATE” và “DIM CUSTOMER”, ô “Hierarchy” chọn lần lượt tương ứng với “Dimensions” là “Gender”, “Online Order”, “Transaction Status”, “Transaction Date” và “State”, ô “Operator” chọn “Equal” ngoại trừ ô cuối chọn “Not Equal” và ô “Filter Expression” chọn lần lượt “Female”, “False”, “Approved”, “2017-03-08” và “VIC”.

Bike Accessories.cube [Design]*

Cube Struct... Dimension Usage Calculations KPIs Actions Partitions Aggregations Perspectives Translations Browser

Language: Default

Edit as Text Import... MDX

Dimension Hierarchy Operator Filter Expression Param...

DIM CUSTOMER	Gender	Equal	{ Female }	<input type="checkbox"/>	<input type="checkbox"/>
DIM TYPE ORDER	Online Order	Equal	{ False }	<input type="checkbox"/>	<input type="checkbox"/>
DIM STATUS	Transaction Status	Equal	{ Approved }	<input type="checkbox"/>	<input type="checkbox"/>
DIM DATE	Transaction Date	Equal	{ 2017-03-08 }	<input type="checkbox"/>	<input type="checkbox"/>
DIM CUSTOMER	State	Not Equal	{ VIC }	<input type="checkbox"/>	<input type="checkbox"/>

<Select dimension>

Customer Id State FACT Count

Click to execute the query.

Bike Accessories

Metadata

Search Model

Measure Group:

All

Bike Accessories

Measures

FACT

FACT Count

Total

KPIs

DIM CUSTOMER

Câu_6

- Age Group
- Cus Category Id
- Cus Category Name
- Customer Dob
- Customer Id
- Gender
- Job Industry Category
- Past 3 Years Bike Related Purchases
- State

Calculated Members

Error List

Output Error List

Bước 3: Chọn “Click to execute the query”, ta có kết quả.

Bike Accessories.cube [Design]*

Cube Struct... Dimension Usage Calculations KPIs Actions Partitions Aggregations Perspectives Translations Browser

Language: Default

Edit as Text Import... MDX

Dimension Hierarchy Operator Filter Expression Param...

DIM CUSTOMER	Gender	Equal	{ Female }	<input type="checkbox"/>	<input type="checkbox"/>
DIM TYPE ORDER	Online Order	Equal	{ False }	<input type="checkbox"/>	<input type="checkbox"/>
DIM STATUS	Transaction Status	Equal	{ Approved }	<input type="checkbox"/>	<input type="checkbox"/>
DIM DATE	Transaction Date	Equal	{ 2017-03-08 }	<input type="checkbox"/>	<input type="checkbox"/>
DIM CUSTOMER	State	Not Equal	{ VIC }	<input type="checkbox"/>	<input type="checkbox"/>

<Select dimension>

Customer Id State FACT Count

292	QLD	1
681	NSW	1
871	NSW	1
1104	NSW	1
2361	QLD	1

Bike Accessories

Metadata

Search Model

Measure Group:

All

Bike Accessories

Measures

FACT

FACT Count

Total

KPIs

DIM CUSTOMER

Câu_6

- Age Group
- Cus Category Id
- Cus Category Name
- Customer Dob
- Customer Id
- Gender
- Job Industry Category
- Past 3 Years Bike Related Purchases
- State

Calculated Members

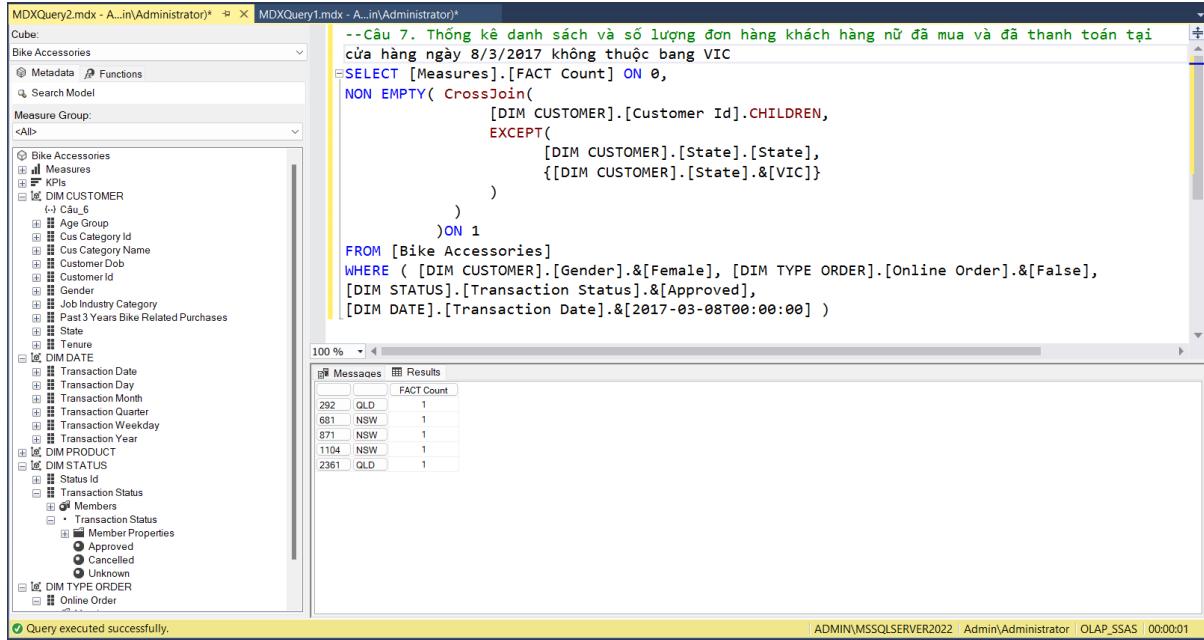
Error List

Output Error List

❖ Truy vấn bằng MDX

Query:

```
SELECT [Measures].[FACT Count] ON 0,
NON EMPTY( CrossJoin(
    [DIM CUSTOMER].[Customer Id].CHILDREN,
    EXCEPT(
        [DIM CUSTOMER].[State].[State],
        {[DIM CUSTOMER].[State].&[VIC]}
    )))ON 1
FROM [Bike Accessories]
WHERE ( [DIM CUSTOMER].[Gender].&[Female], [DIM TYPE
ORDER].[Online Order].&[False],
[DIM STATUS].[Transaction Status].&[Approved],
[DIM DATE].[Transaction Date].&[2017-03-08T00:00:00] )
```



❖ Truy vấn bằng Pivot table trên Excel

Bước 1: Chọn tab “PivotTable Analyze”.

Bước 2: Ở mục Trường PivotTable kéo chọn thuộc tính “Customer Id” vào cột “Rows”, “FACT Count” vào cột “Values”, các thuộc tính “Gender”, “Online Order”, “Transaction Status”, “Transaction Date” và “State” vào cột “Filters”.

A screenshot of Microsoft Excel showing a PivotTable setup. The PivotTable Fields pane on the right lists fields from two dimensions: DIM STATUS (Status Id, Transaction Status) and DIM TYPE ORDER (Type Order Id). The Filters section at the bottom of the pane includes Gender, Online Order, and Transaction Status. The Rows section has Customer Id, and the Values section has FACT Count.

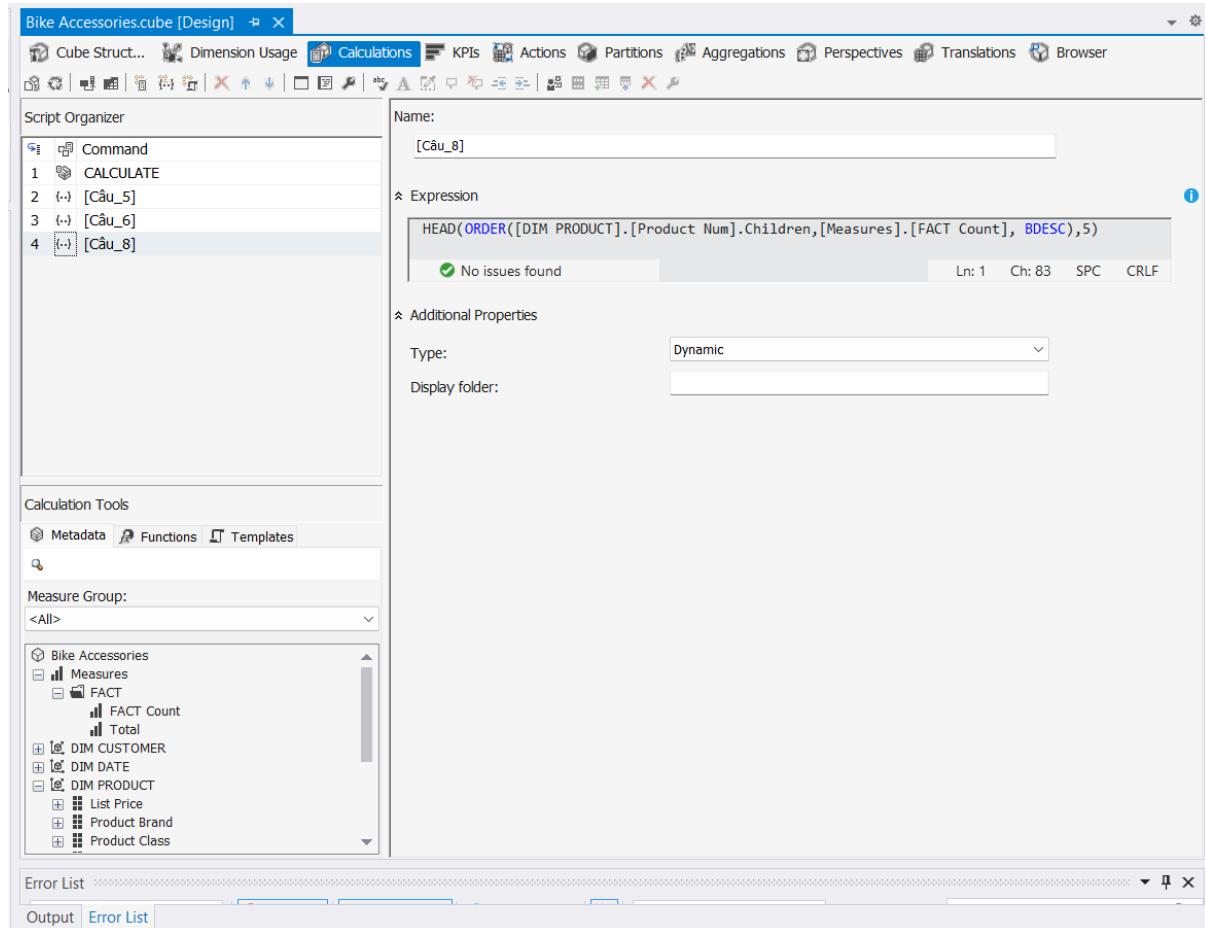
Bước 3: Chọn điều kiện lọc tương ứng với các cột “Filters” ở bước 2 lần lượt là “Female”, “False”, “Approved”, “2017-03-08” và “VIC”, sau đó kéo cột “State” từ “Filters” xuống “Rows” ta có kết quả như hình.

A screenshot of Microsoft Excel showing the PivotTable after applying filters. The Row Labels now show gender, online order status, transaction status, and transaction date. The Rows section now includes State, and the Values section still has FACT Count.

7.8 Thống kê top 5 sản phẩm bán chạy nhất trong từng bang

❖ SSAS

Bước 1: Ở tab “Calculations” thêm 1 “Script Organizer” có tên là [Câu_8].



Script:

```
HEAD(ORDER([DIM PRODUCT].[Product  
Num].Children,[Measures].[FACT Count], BDESC),5)
```

Bước 2: Kéo biến “Product Num”, “State” và “FACT Count” sang khung truy vấn

Bước 3: Tại “Dimensions” chọn “DIM PRODUCT”, ô “Hierarchy” chọn “Product Num”, ô “Operator” chọn “Custom” và ô “Filter Expression” điền câu lệnh ở bước 1.

Bước 4: Chọn “Click to execute the query”.

❖ Truy vấn bằng MDX

Query:

```
SELECT [Measures].[FACT Count] ON 0,
([DIM PRODUCT].[Product Num].Children,
[DIM CUSTOMER].[State].Children ) ON 1
FROM ( SELECT ( HEAD ( ORDER (
[DIM PRODUCT].[Product Num].Children,[Measures].[FACT
Count], BDESC),5) ) ON 0
FROM [Bike Accessories])
```

Kết quả:

MDXQuery2.mdx - A...in\Administrator* MDXQuery1.mdx - A...in\Administrator*

Cube: Bike Accessories

Metadata Functions Search Model Measure Group: <All>

Bike Accessories Measures KPIs DIM CUSTOMER DIM DATE DIM PRODUCT DIM STATUS DIM TYPE ORDER

```
--Câu 8. Thống kê top 5 sản phẩm bán chạy nhất trong từng bang
SELECT [Measures].[FACT Count] ON 0,
([DIM PRODUCT].[Product Num].Children,
[DIM CUSTOMER].[State].Children) ON 1
FROM ( SELECT ( HEAD ( ORDER (
[DIM PRODUCT].[Product Num].Children,[Measures].[FACT Count], BDESC),5)
) ON 0
) ON 0
FROM [Bike Accessories]
```

100 % Results

	FACT Count
100 NSW	366
100 QLD	144
100 VIC	146
112 NSW	156
112 QLD	62
112 VIC	82
123 NSW	114
123 QLD	46
123 VIC	56
164 NSW	109
164 QLD	59
164 VIC	38
191 NSW	100
191 QLD	50
191 VIC	60

Query executed successfully.

ADMIN\MSSQLSERVER2022 | Admin\Administrator | OLAP_SSAS | 00:00:01

❖ Truy vấn bằng Pivot table trên Excel

Bước 1: Chọn tab “PivotTable Analyze”.

Bước 2: Ở mục Trường PivotTable kéo chọn thuộc tính “Customer Category Name” vào cột “Rows”, “FACT Count” vào cột “Values”, “State” vào cột “Filters”.

PivotTable Fields

Choose fields to add to report:

Search

Drag fields between areas below:

Filters Columns

Rows Values

Product Num FACT Count

State

Defer Layout Update Update

Row Labels	FACT Count
100	366
100 QLD	144
100 VIC	146
112	156
112 QLD	62
112 VIC	82
114	114
114 NSW	22
114 QLD	14
114 VIC	8
115	89
115 NSW	89
115 QLD	37
115 VIC	47
116	28
116 NSW	28
116 QLD	10
116 VIC	13
117	13

Bước 3: Row Labels → Values Filters → Top 10...

The screenshot shows the Microsoft Excel interface with the PivotTable Fields ribbon tab selected. In the 'Rows' section, 'Product Num' and 'State' are listed. In the 'Values' section, 'FACT Count' is selected. A 'Value Filters' dialog box is open, with 'Top 10...' selected from the dropdown menu. The main worksheet displays a list of products and their FACT Count values.

Product Num	FACT Count
100	366
QLD	144
VIC	146
NSW	156
QLD	62
VIC	82
NSW	22
QLD	14
VIC	8
NSW	89
QLD	37
VIC	47
NSW	28
QLD	10
VIC	13

Bước 4: Tại cửa sổ “Show” sửa số lượng hiển thị lại thành 5 và nhấn “OK”.

The screenshot shows the Microsoft Excel interface with the PivotTable Fields ribbon tab selected. The 'Rows' section shows 'Product Num' and 'State'. The 'Values' section shows 'FACT Count'. A 'Value Filters' dialog box is open, showing a dropdown menu with 'Top 10...' selected. The main worksheet displays a list of products and their FACT Count values, with the top 5 items shown.

Product Num	FACT Count
100	366
QLD	144
VIC	146
NSW	156
QLD	62
VIC	82
NSW	22
QLD	14
VIC	8
NSW	89
QLD	37
VIC	47
NSW	28
QLD	10
VIC	13

Kết quả:

A screenshot of Microsoft Excel showing a PivotTable. The PivotTable Fields pane on the right lists fields such as Customer Dob, Customer Id, Gender, Job Industry Category, Past 3 Years Bike Related Purchases, and State. The State field is checked. The main table shows FACT Count for various states and years.

		FACT Count
100	NSW	366
100	QLD	144
100	VIC	146
112	NSW	156
112	QLD	62
112	VIC	82
123	NSW	114
123	QLD	46
123	VIC	56
164	NSW	109
164	QLD	59
164	VIC	38
191	NSW	100
191	QLD	50
191	VIC	60
Grand Total		1588

7.9 Thống kê những khách hàng tiềm năng thuộc nhóm khách hàng từ 20 đến dưới 40 tuổi mua sản phẩm trong quý 1

❖ SSAS

Bước 1: Kéo biến “Customer Id”, “Total” và “FACT Count” sang khung truy vấn.

The screenshot shows the Microsoft Analysis Services (SSAS) MDX Editor. The left pane displays the cube structure, including dimensions like DIM CUSTOMER and measures like FACT Count, Total, and Age Group. The right pane shows a query grid with columns Customer Id, FACT Count, and Total. A message 'Click to execute the query.' is visible at the bottom of the grid area.

Bước 2: Tại “Dimensions” lần lượt chọn “DIM CUSTOMER”, “DIM CUSTOMER”, “DIM DATE”, ô “Hierarchy” lần lượt chọn “Cus Category Name”, “Age Group”, “Transaction Quarter”, ô “Operator” lần lượt chọn “Potential Customer”, “U40” và “1”.

Dimension	Hierarchy	Operator	Filter Expression	Param...
DIM CUSTOMER	Cus Category Name	Equal	{ Potential Customer }	<input type="checkbox"/> <input checked="" type="checkbox"/>
DIM CUSTOMER	Age Group	Equal	{ U40 }	<input type="checkbox"/> <input checked="" type="checkbox"/>
DIM DATE	Transaction Quarter	Equal	{ 1 }	<input type="checkbox"/> <input checked="" type="checkbox"/>
<Select dimension>				

Bước 3: Chọn “Click to execute the query”.

❖ Truy vấn bằng MDX

Query:

```

SELECT NON EMPTY { [Measures].[FACT Count],
[Measures].[Total] } ON 0,
NON EMPTY { ([DIM CUSTOMER].[Customer Id].Children) } ON 1
FROM(SELECT( { [DIM DATE].[Transaction Quarter].&[1] } ) ON 0
FROM(SELECT ( { [DIM CUSTOMER].[Age Group].&[U40] } ) ON 0
FROM(SELECT ( { [DIM CUSTOMER].[Cus Category
Name].&[Potential Customer] } ) ON 0
FROM [Bike Accessories]))
WHERE( [DIM CUSTOMER].[Cus Category Name].&[Potential
Customer],
[DIM CUSTOMER].[Age Group].&[U40], [DIM DATE].[Transaction
Quarter].&[1] )

```

--Câu 9. Thống kê những khách hàng tiềm năng thuộc nhóm khách hàng từ 20 đến dưới 40 tuổi
--mua sản phẩm trong quý 1

```

SELECT NON EMPTY { [Measures].[FACT Count], [Measures].[Total] } ON 0,
NON EMPTY { ([DIM CUSTOMER].[Customer Id].Children) } ON 1
FROM ( SELECT ( { [DIM DATE].[Transaction Quarter].&[1] } ) ON 0
FROM ( SELECT ( { [DIM CUSTOMER].[Age Group].&[U40] } ) ON 0
FROM ( SELECT ( { [DIM CUSTOMER].[Cus Category Name].&[Potential Customer] } ) ON 0
FROM [Bike Accessories])))
WHERE ( [DIM CUSTOMER].[Cus Category Name].&[Potential Customer],
[DIM CUSTOMER].[Age Group].&[U40], [DIM DATE].[Transaction Quarter].&[1] )

```

	FACT Count	Total
47	1	569,56
90	1	912,52
148	1	1458,17
149	1	1073,07
208	3	4084,45
221	1	1466,68
233	1	1458,17
250	2	3065,42
337	1	1280,28
345	1	1992,93
352	1	1992,93
367	2	3209,07
382	1	774,53
408	1	912,52
475	2	2905,45
488	2	1186,36

Query executed successfully.

❖ Truy vấn bằng Pivot table trên Excel

Bước 1: Chọn tab “PivotTable Analyze”.

Bước 2: Ở mục Trường PivotTable kéo chọn thuộc tính “Customer Id” vào cột “Rows”, “FACT Count”, “Total” vào cột “Values”, các thuộc tính “Cus Category Name”, “Age Group”, “Transaction Quarter” vào cột “Filters”.

PivotTable Fields

Choose fields to add to report:

- Transaction Day
- Transaction Month
- Transaction Quarter
- Transaction Weekday
- Transaction Year
- DIM PRODUCT
- List Price

Drag fields between areas below:

Filters

Cus Category Name	Age Group	Transaction Quarter
-------------------	-----------	---------------------

Columns

Values

Rows

Customer Id	FACT Count	Total
-------------	------------	-------

Bước 3: Chọn điều kiện lọc tương ứng với các cột “Filters” ở bước 2 lần lượt là “Potential Customer”, “U40”, “1”, ta có kết quả như hình.

A screenshot of Microsoft Excel showing a PivotTable. The PivotTable Fields pane on the right lists fields: Transaction Day, Transaction Month, Transaction Quarter (checked), Transaction Weekday, Transaction Year, DIM PRODUCT, and List Price. The main table shows data for Customer Id, FACT Count, and Total across various categories.

7.10 Thống kê top 3 tháng có đơn hàng bị hủy nhiều nhất

❖ SSAS

Bước 1: Ở tab “Calculations” thêm 1 “Script Organizer” có tên là [Câu_10].

The screenshot shows the Microsoft Analysis Services (SSAS) Management Studio interface. The 'Script Organizer' window is active, displaying a script named [Câu_10] with the following expression:

```
TopCount(
    [DIM DATE].[Transaction Month].Children, 3,
    [Measures].[FACT Count]
)
```

The 'Measure Group' pane on the left shows dimensions like Status, Transaction Status, and Member Properties, along with measures such as Approved, Cancelled, and Unknown. The 'Error List' pane at the bottom is empty.

Script:

```
TopCount(  
    [DIM DATE].[Transaction Month].Children, 3,  
    [Measures].[FACT Count]  
)
```

Bước 2: Kéo biến “Transaction Month” và “FACT Count” sang khung truy vấn.

The screenshot shows the Microsoft Analysis Services (SSAS) MDX Editor. The title bar says "Bike Accessories.cube [Design]". The menu bar includes "Cube Struct...", "Dimension Usage", "Calculations", "KPIs", "Actions", "Partitions", "Aggregations", "Perspectives", "Translations", and "Browser". The toolbar has icons for "Edit as Text", "Import...", "MDX", and other tools. The main workspace has a table with columns: Dimension, Hierarchy, Operator, Filter Expression, and Param... . A dropdown menu under "Dimension" says "<Select dimension>". Below the table, there are two buttons: "Transaction Month" and "FACT Count". To the left is a tree view of dimensions: DIM DATE (with Transaction Date, Transaction Day, Transaction Month, Transaction Quarter, Transaction Weekday, Transaction Year), DIM PRODUCT (with List Price, Product Brand, Product Class, Product Id, Product Line, Product Num, Product Size), and DIM STATUS (with Status Id). At the bottom, there are tabs for "Error List" and "Output", with "Error List" selected.

Bước 3: Tại “Dimensions” chọn “DIM STATUS”, “DIM DATE”, ô “Hierarchy” chọn tương ứng “Transaction Status”, “Transaction Month”, ô “Operator” chọn “Equal” và “Custom”, ô “Filter Expression” lần lượt chọn “Cancelled” và điền câu lệnh ở bước 1.

The screenshot shows the Analysis Services Designer application window titled "Bike Accessories.cube [Design]". The top menu bar includes options like Cube Struct..., Dimension Usage, Calculations, KPIs, Actions, Partitions, Aggregations, Perspectives, Translations, and Browser. The MDX query editor at the top has tabs for "Edit as Text" and "Import...". The main area contains a table for defining filters:

Dimension	Hierarchy	Operator	Filter Expression	Param...
DIM STATUS	Transaction Status	Equal	{ Cancelled }	<input type="checkbox"/>
DIM DATE	Transaction Month	Custom	TopCount([DIM DATE].[Transaction Mo...]	<input type="checkbox"/>
<Select dimension>				

Below the table is a results grid titled "Transaction Month | FACT Count":

Transaction Month	FACT Count
4	12
5	12
9	11

A message "Click to execute the query." is displayed next to the results grid. On the left side, there is a navigation pane with sections for Bike Accessories, Measures, FACT, KPIs, and DIM CUSTOMER. The DIM CUSTOMER section is expanded, showing various dimensions like Age Group, Cus Category Id, etc. At the bottom, there are tabs for "Output" and "Error List".

Bước 4: Chọn “Click to execute the query”.

This screenshot is identical to the previous one, except the results grid now displays the data from the executed query:

Transaction Month	FACT Count
4	12
5	12
9	11

- ❖ Truy vấn bằng MDX

Query:

```
SELECT [Measures].[FACT Count] ON 0,
       TopCount(
           [DIM DATE].[Transaction Month].Children, 3, [DIM
DATE].[Transaction Date] ) ON 1
FROM [Bike Accessories]
WHERE [DIM STATUS].[Transaction Status].&[Cancelled];
```

-- Câu 10. Thống kê top 3 tháng có đơn hàng bị hủy nhiều nhất

```
SELECT [Measures].[FACT Count] ON 0,
       TopCount(
           [DIM DATE].[Transaction Month].Children, 3, [DIM
DATE].[Transaction Date] ) ON 1
FROM [Bike Accessories]
WHERE [DIM STATUS].[Transaction Status].&[Cancelled];
```

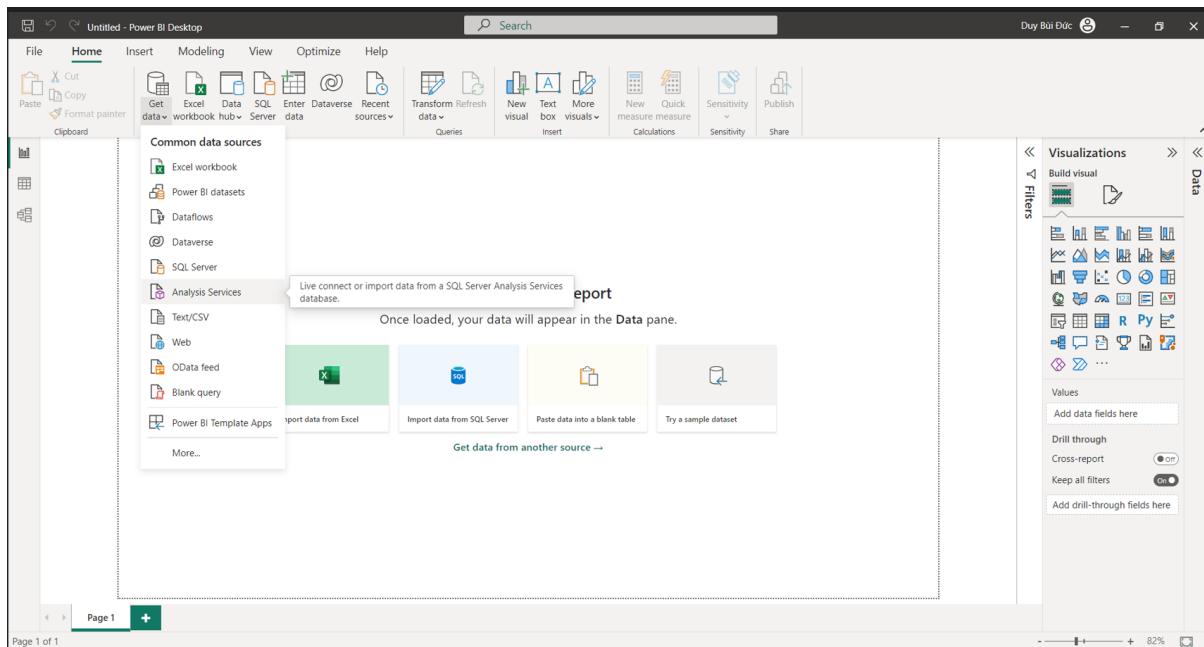
FACT Count	12
5	12
4	12
9	11

Query executed successfully.

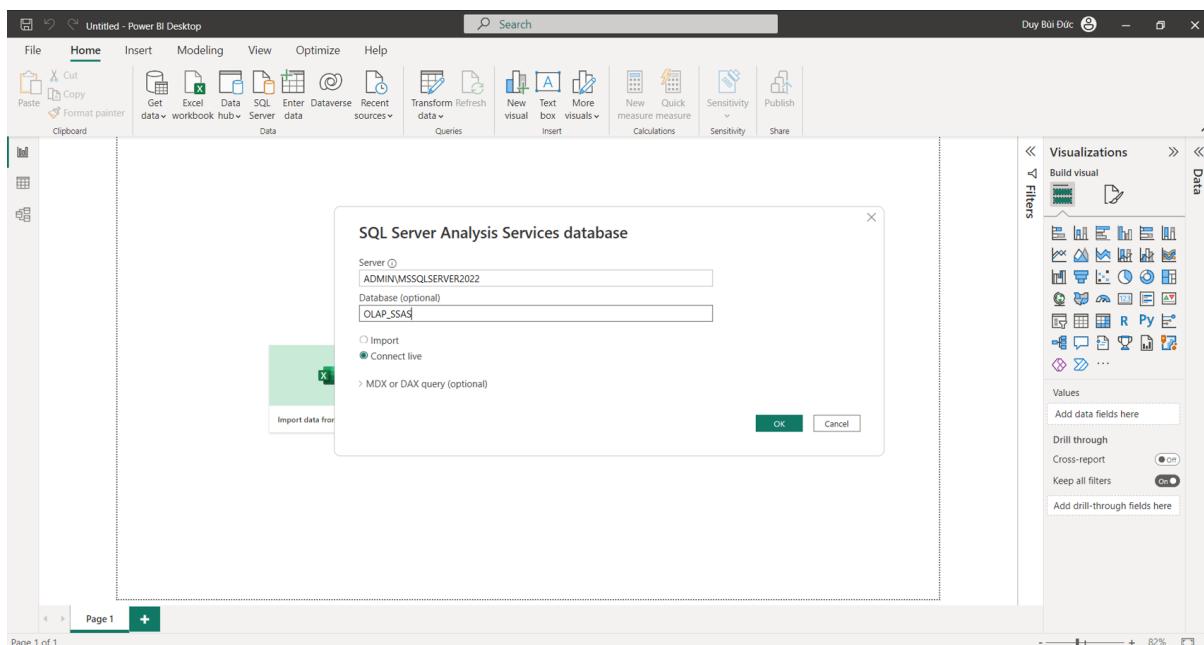
CHƯƠNG 4. REPORT BẰNG CÔNG CỤ POWER BI

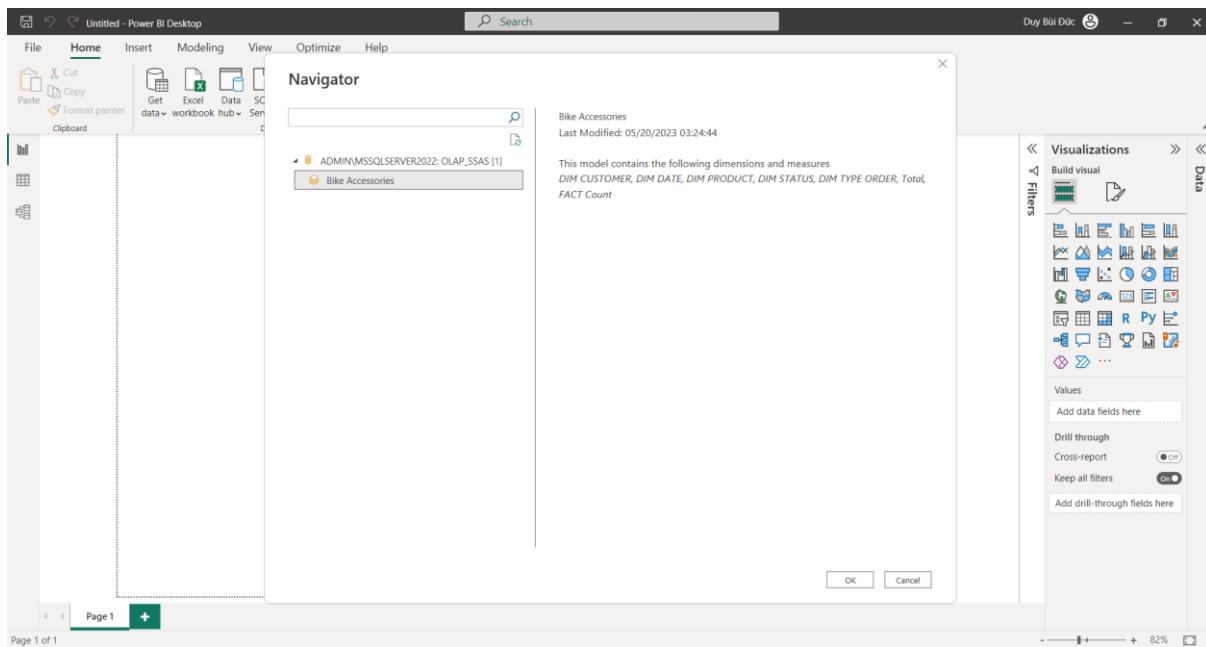
1. Đỗ dữ liệu vào project Power BI

Bước 1: Chọn Get Data → Analysis Server.

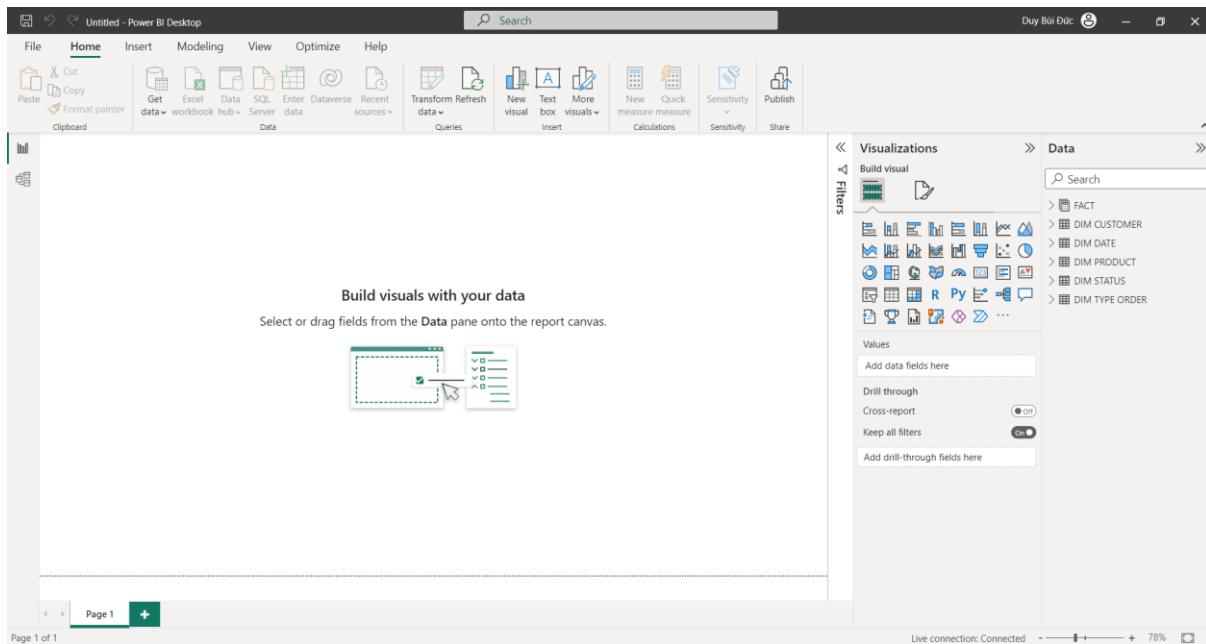


Bước 2: Điền tên Server và Database muốn kết nối để lấy dữ liệu.





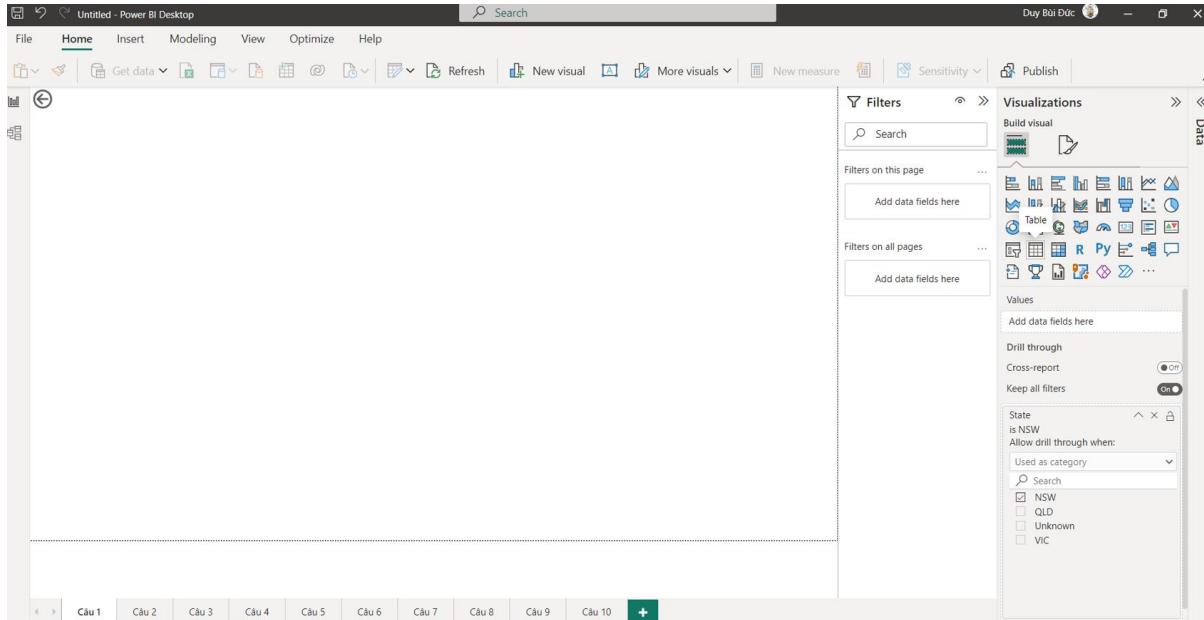
Kết quả: Dữ liệu từ Database OLAP_SSAS được đổ vào cột Fields trong PowerBI.



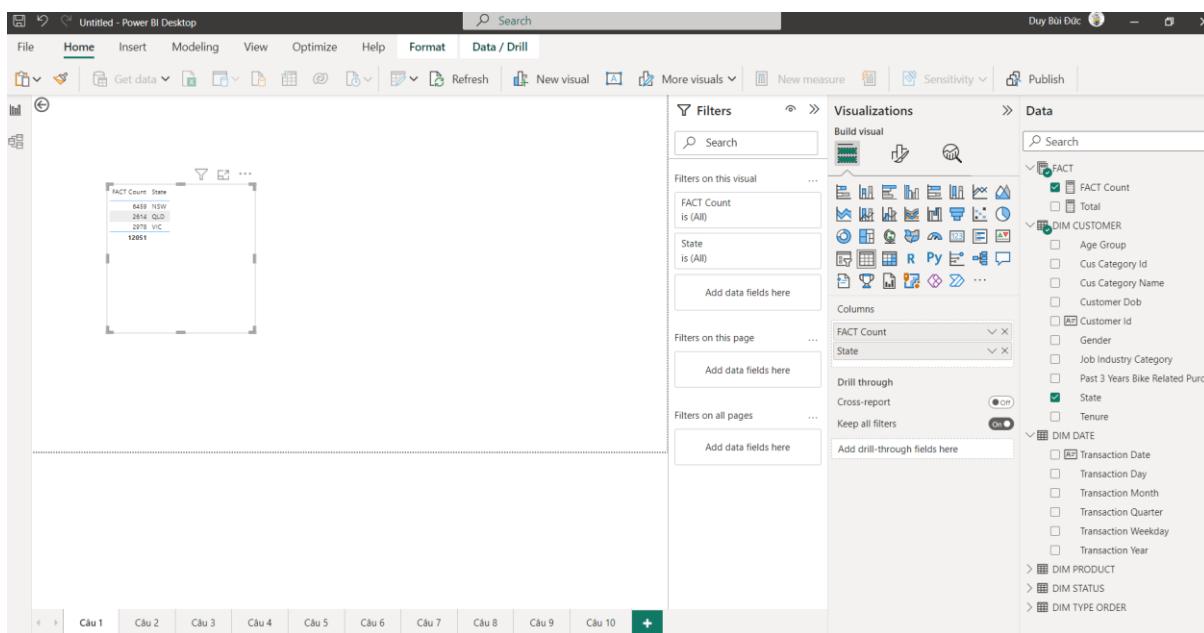
2. Thực hiện Report

2.1 Thống kê tổng số giao dịch ở tiểu bang NSW

Bước 1: Chọn loại Report Table.



Bước 2: Kéo thuộc tính “State”, “FACT Count” vào “Columns”.



Bước 3: Kéo thuộc tính “State” vào “Drill Through” tick chọn “NSW”.

The screenshot shows the Power BI Desktop interface. A single card visual is displayed on the canvas, showing the value '6459 NSW'. The 'Data / Drill' pane on the right is open, specifically the 'Drill through' section under 'State'. A dropdown menu is open, showing options: 'Search' (selected), 'NSW' (checked), 'QLD' (unchecked), 'Unknown' (unchecked), and 'VIC' (unchecked). Other filters like 'FACT Count' and 'State' are also visible in the pane.

Bước 4: Điều chỉnh Title và Style, ta có kết quả.

Tổng số giao dịch ở tiểu bang NWS	
FACT Count	State
6459	NSW

Table

Nhận xét: Có tổng 6459 giao dịch ở tiểu bang NWS.

2.2 Thống kê doanh số của từng thương hiệu qua các quý

Bước 1: Chọn loại Report Matrix.

The screenshot shows the Power BI Desktop interface. The top navigation bar includes File, Home, Insert, Modeling, View, Optimize, Help, and various ribbon tabs like Get data, Refresh, New visual, Sensitivity, and Publish. The left sidebar has sections for Filters, Visualizations, and Data. The 'Visualizations' section contains a grid of icons, with the 'Matrix' icon being highlighted. The 'Data' section on the right lists various dimensions and measures such as Cus Category Name, Customer Id, Gender, Job Industry Category, etc. The bottom navigation bar shows tabs Câu 1 through Câu 10, with 'Câu 2' currently selected.

Bước 2: Kéo thuộc tính “Product Brand” vào “Rows”, “Transaction Quarter” vào “Columns”, thuộc tính “Total” vào “Values”.

The screenshot shows the Power BI Desktop interface after creating a matrix visualization. The main area displays a table with four columns labeled 1, 2, 3, and 4, and a 'Total' row at the bottom. The table data is as follows:

	1	2	3	4	Total
Grand Bicycles	1,417,346.99	1,454,357.43	1,480,064.80	1,570,107.02	5,821,860.04
Norco Bicycles	788,452.98	740,514.15	722,073.75	737,292.75	3,00,487.41
Cannondale	874,831.48	845,838.90	894,693.21	900,060.78	3,505,448.35
Total	3,599,379.66	3,586,144.72	3,605,824.82	3,723,081.99	14,514,401.19

The 'Visualizations' pane on the right shows the matrix configuration: 'Rows' is set to 'Product Brand', 'Columns' is set to 'Transaction Quarter', and 'Values' is set to 'Total'. The 'Filters' pane shows 'Product Brand is (All)' and 'Total is (All)'. The 'Data' pane on the right lists various dimensions and measures. The bottom navigation bar shows tabs Câu 1 through Câu 10, with 'Câu 2' currently selected.

Bước 3: Duplicate Table vừa tạo sau đó điều chỉnh về dạng “Clustered Column Chart”, sau đó điều chỉnh dữ liệu. “X-axis” chọn “Transaction Quarter”, “Legend” chọn “Product Brand”, “Y-axis” chọn “Total”.

The screenshot shows the Power BI Desktop interface. On the left, there is a data grid titled "Product Brand" with columns 1, 2, 3, 4, and Total. The data includes sales figures for four brands: Giant Bicycles, Norco Bicycles, OHM Cycles, and WeareA2B across four quarters. A clustered column chart titled "Total by Transaction Quarter and Product Brand" is displayed below the grid, showing sales volume for each brand per quarter. The chart has "Transaction Quarter" on the X-axis and "Total" on the Y-axis. The Power BI visualizations pane is open on the right, showing various filters and data fields available for this report.

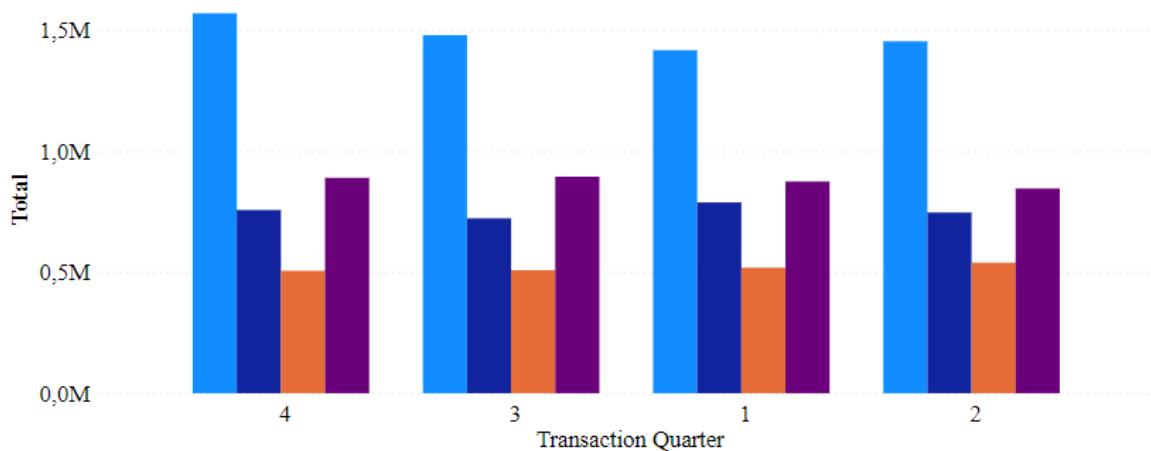
Bước 4: Điều chỉnh Title và Style, ta có kết quả

Product Brand	1	2	3	4	Total
Giant Bicycles	1.417.346,99	1.454.357,43	1.480.068,60	1.570.107,02	5.921.880,04
Norco Bicycles	788.452,96	746.614,15	722.697,55	757.092,75	3.014.857,41
OHM Cycles	518.744,25	539.314,24	508.365,46	505.821,44	2.072.245,39
WeareA2B	874.835,46	845.858,90	894.693,21	890.060,78	3.505.448,35
Total	3.599.379,66	3.586.144,72	3.605.824,82	3.723.081,99	14.514.431,19

Matrix

Doanh số của từng thương hiệu qua các quý

Product Brand • Giant Bicycles • Norco Bicycles • OHM Cycles • WeareA2B



Clustered Column Chart

Nhận xét: Doanh số của thương hiệu Giant Bicycles cao nhất trong tất cả các quý, ngược lại thì doanh số của thương hiệu OHM Cycles thấp nhất ở cả 4 quý.

2.3 Thống kê tổng đơn đặt hàng online của từng dòng sản phẩm trong tháng 4 năm 2017

Bước 1: Chọn loại Report Matrix.

The screenshot shows the Power BI Desktop interface with the 'Matrix' visual selected in the 'Visualizations' pane. The report canvas is empty, and the data pane on the right lists various dimensions and measures available for selection.

Bước 2: Kéo thuộc tính “Transaction Month”, “Transaction Year”, “Online Order” vào “Rows”, thuộc tính “Product Line” vào “Columns”, thuộc tính “FACT Count” vào “Values”.

The screenshot shows the Power BI Desktop interface with the Matrix visual now populated with data. The columns represent 'Product Line' categories (Mountain, Road, Standard, Touring) and the rows represent 'Transaction Month' categories (1 through 12). The values are 'FACT Count' for each combination, showing the number of orders per month and product line.

	1	2	3	4	5	6	7	8	9	10	11	12	Total
Mountain	24	39	24	25	39	29	35	39	30	28	30	20	360
Road	233	242	231	217	211	231	209	201	202	238	213	182	2668
Standard	71	72	72	73	72	73	65	65	65	65	65	65	850
Touring	55	59	52	50	55	46	59	60	57	54	55	52	694
Total	1024	1071	1009	1040	969	988	1000	1009	941	1011	1052	933	12051

Bước 3: Tại “Drill Through” ở “Transaction Month” tick chọn “4”, “Online Order” tick chọn “True”.

The screenshot shows the Power BI Desktop interface. On the left, there is a table visual titled "Product Line 4 Total" with the following data:

	19	119	348	35	521
Mountain	19	119	348	35	521
Road	119	119	348	35	521
Standard	348	348	348	35	521
Touring	35	35	35	35	521
Total	521	521	521	521	521

The "Data / Drill" pane on the right is open, showing various filters and drill-through settings. Under "Drill through", "Transaction Month" is set to "is 4". Under "Online Order", "is True" is checked.

Bước 4: Duplicate Table vừa tạo sau đó điều chỉnh về dạng “Pie Chart”, sau đó điều chỉnh dữ liệu: kéo thuộc tính “FACT Count” vào ô “Values”, thuộc tính “Product Line” vào ô “Details”.

The screenshot shows the Power BI Desktop interface with a pie chart visual titled "Tổng đơn đặt hàng online của từng dòng sản phẩm trong tháng 4 năm 2017". The chart displays the following proportions:

- Road: 35 (67%)
- Standard: 119 (22.84%)
- Touring: 35 (6.7%)
- Mountain: 19 (3.87%)

The "Data / Drill" pane on the right is open, showing the same filters and drill-through settings as the previous step, including "Transaction Month" set to "is 4" and "Online Order" set to "is True".

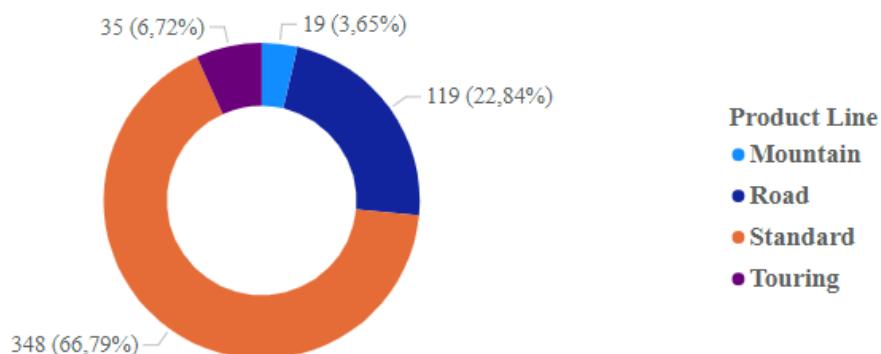
Bước 5: Điều chỉnh Title và Style, ta có kết quả

Tổng đơn đặt hàng online của từng dòng sản phẩm trong tháng 4 năm 2017

Product Line	4	Total
Mountain	19	19
Road	119	119
Standard	348	348
Touring	35	35
Total	521	521

Matrix

Tổng đơn đặt hàng online của từng dòng sản phẩm trong tháng 4 năm 2017



Pie Chart

Nhận xét: Dòng sản phẩm Standard được đặt hàng online chiếm tỷ lệ cao nhất chiếm tới 66,79%, tiếp đến là dòng Road chiếm 22,84% còn lại Touring và Mountain chiếm tỷ lệ thấp lần lượt là 6,72% và 3,65%.

2.4 Thống kê tổng hóa đơn của từng nhóm khách hàng theo tháng trong năm

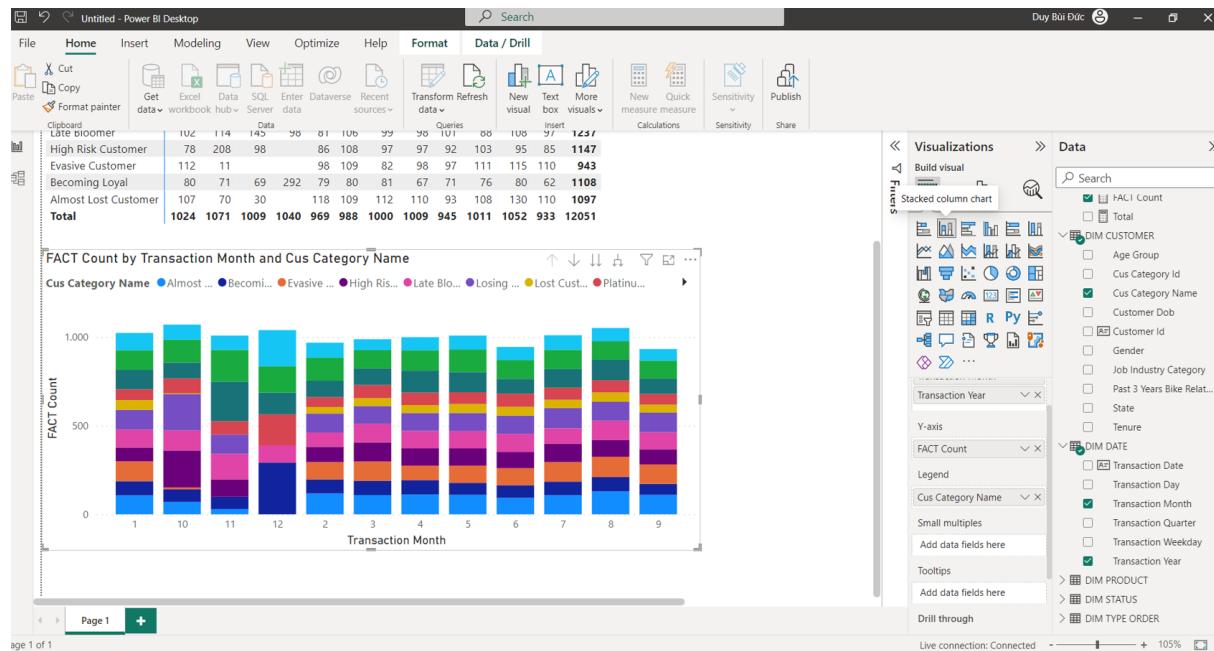
Bước 1: Chọn loại Report Matrix.

The screenshot shows the Power BI Desktop interface. The top menu bar has 'Home' selected. The ribbon includes 'Home', 'Insert', 'Modeling', 'View', 'Optimize', and 'Help'. Below the ribbon are icons for 'Cut', 'Copy', 'Format painter', 'Clipboard', 'Get data', 'Data', 'Transform data', 'Queries', 'New visual', 'Text box', 'More visuals', 'New measure', 'Quick measure', 'Calculations', 'Sensitivity', and 'Publish'. The 'Data' pane on the right lists fields: FACT, DIM CUSTOMER, DIM DATE, DIM PRODUCT, DIM STATUS, and DIM TYPE ORDER. A 'Visualizations' pane shows various chart and matrix icons. A 'Filters' pane is open, showing 'Cus Category Name is (All)', 'FACT Count is (All)', and 'Transaction Month is (All)'. A 'Values' pane shows 'Add data fields here'. The main canvas area shows a matrix visual being built with columns for 'Cus Category Name' and 'Transaction Month' and rows for 'Transaction Year'. The status bar at the bottom indicates 'Live connection: Connected' and '78%'. The page navigation bar shows 'Page 1'.

Bước 2: Kéo thuộc tính “Cus Category Name” vào “Rows”, thuộc tính “Transaction Month”, “Transaction Year” vào “Columns”, thuộc tính “FACT Count” vào “Values”.

The screenshot shows the Power BI Desktop interface with the matrix visual now fully configured. The 'Values' pane shows 'FACT Count' selected. The matrix visual on the canvas displays transaction counts for each category-month combination. The status bar at the bottom indicates 'Live connection: Connected' and '65%'. The page navigation bar shows 'Page 1'.

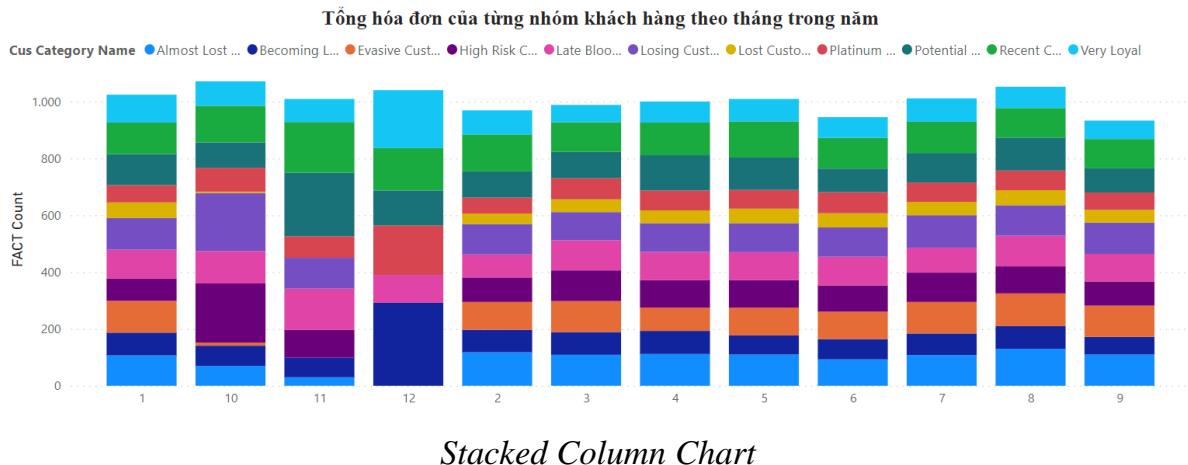
Bước 3: Duplicate Table vừa tạo sau đó điều chỉnh về dạng “Stacked Column Chart”, sau đó điều chỉnh dữ liệu. “X-axis” chọn “Transaction Month”, “Transaction Year”, “Legend” chọn “Cus Category Name”, “Y-axis” chọn “FACT Count”.



Bước 4: Điều chỉnh Title và Style, ta có kết quả

Cus Category Name	1	10	11	12	2	3	4	5	6	7	8	9	Total
Almost Lost Customer	107	70	30		118	109	112	110	93	108	130	110	1097
Becoming Loyal	80	71	69	292	79	80	81	67	71	76	80	62	1108
Evasive Customer	112	11			98	109	82	98	97	111	115	110	943
High Risk Customer	78	208	98		86	108	97	97	92	103	95	85	1147
Late Bloomer	102	114	145	98	81	106	99	98	101	88	108	97	1237
Losing Customer	111	204	108		107	99	101	102	103	114	107	110	1266
Lost Customer	55	5			37	45	45	51	50	47	53	45	433
Platinum Customer	61	84	76	174	56	74	70	66	74	68	69	61	933
Potential Customer	109	89	223	123	92	93	124	114	83	104	116	85	1355
Recent Customer	111	129	179	149	130	104	115	127	108	110	104	103	1469
Very Loyal	98	86	81	204	85	61	74	79	73	82	75	65	1063
Total	1024	1071	1009	1040	969	988	1000	1009	945	1011	1052	933	12051

Matrix



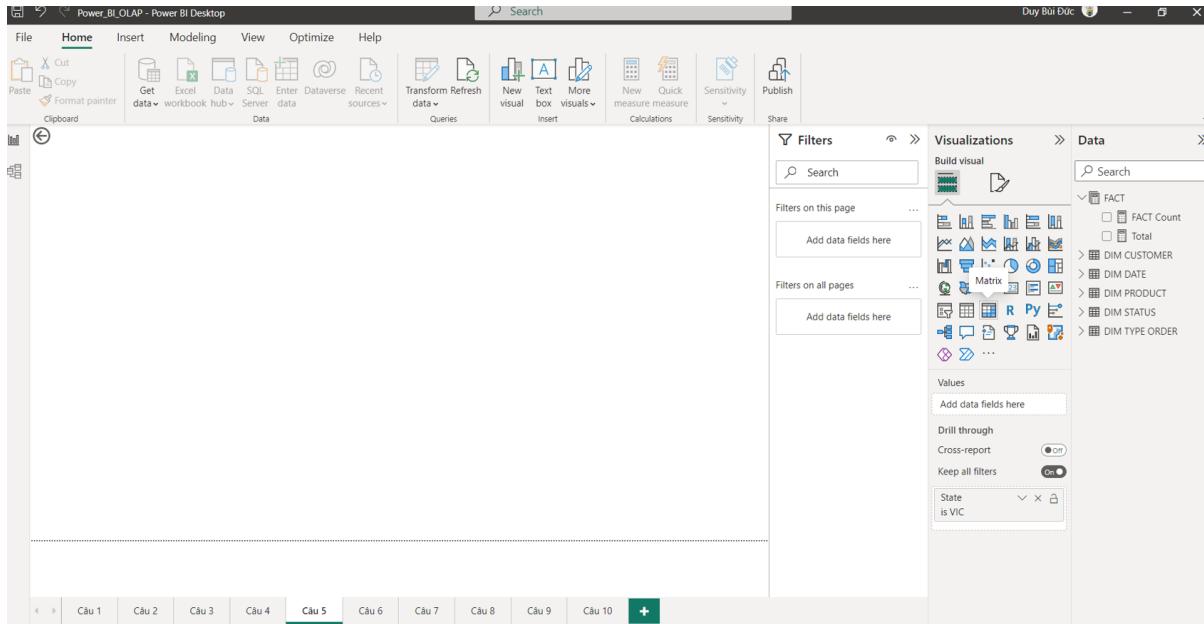
Nhận xét:

Hầu hết các tháng đều có tổng hóa đơn của các nhóm khách hàng khá đồng đều nhau trừ 3 tháng cuối năm tỷ lệ đặt đơn có sự chênh lệch giữa các nhóm khá rõ rệt:

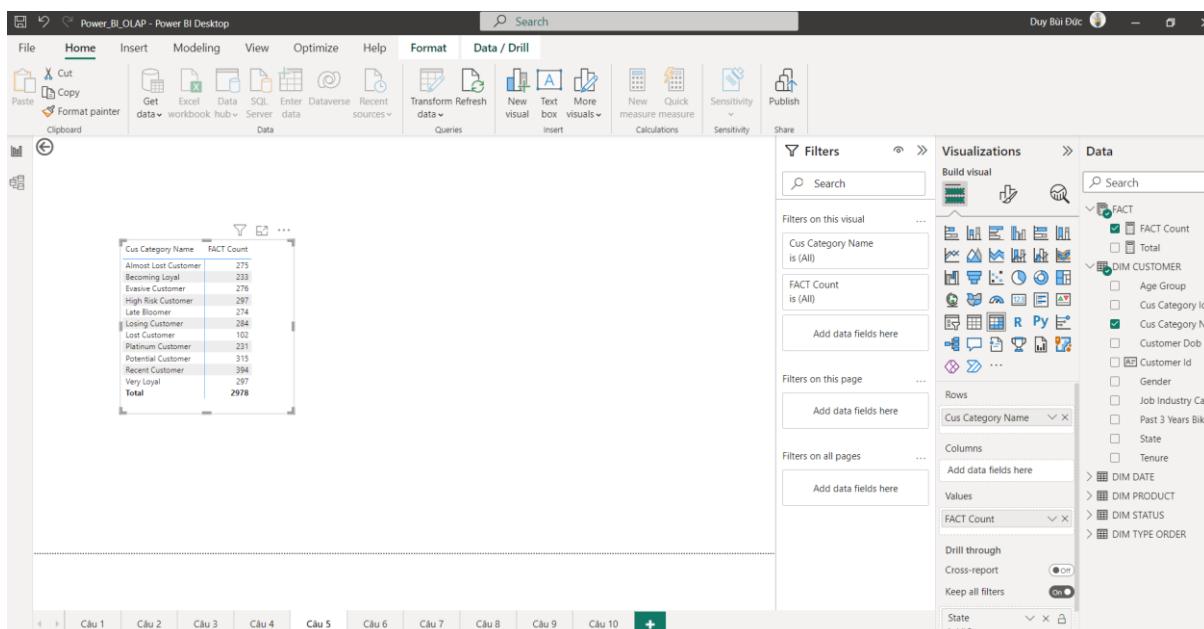
- Tháng 10 có 2 nhóm khách hàng đặt đơn nhiều nhất là *High Risk Customer* (208 đơn) và *Losing Customer* (204 đơn) ngược lại thì nhóm *Evasive Customer* và *Lost Customer* đặt rất ít đơn lần lượt là 5 đơn và 11 đơn.
- Tháng 11 có nhóm khách hàng đặt đơn nhiều nhất là *Potential Customer* (223 đơn) tiếp đến là các nhóm *Rencent Customer* (179 đơn) và *Late Bloomer* (145 đơn) ngược lại thì 2 nhóm không mua đơn nào là *Evasive Customer* và *Lost Customer*.
- Tương tự với tháng 12 có 2 nhóm *Becoming Loyal Customer* (292 đơn) và *Very Loyal* (204 đơn) các nhóm còn lại số lượng đặt đơn thì không chênh lệch lớn vẫn duy trì tương đương các tháng trước đó nhưng có tới 5 nhóm không có đơn nào trong tháng 12 là *Almost Lost Customer*, *Evasive Customer*, *High Risk Customer*, *Losing Customer* và *Lost Customer*.

2.5 Thống kê top 5 nhóm khách hàng thuộc bang VIC mua hàng nhiều nhất

Bước 1: Chọn loại Report Matrix.



Bước 2: Kéo thuộc tính “Cus Category Name” vào “Rows”, “FACT Count” vào “Values”, kéo “State” vào “Drill through” chọn “VIC”.



Bước 3: Ở “Filters” tại ô “Cus Category Name” chọn “Filter type” là “TopN”, “Show items” chọn “Top” và điền “5”, kéo thuộc tính “FACT Count” vào “By value” và nhấn “Apply filter”.

Cus Category Name	FACT Count
High Risk Customer	297
Losing Customer	284
Potential Customer	315
Recent Customer	394
Very Loyal	297
Total	1.587

Bước 4: Điều chỉnh Title và Style, ta có kết quả.

Top 5 nhóm khách hàng thuộc bang VIC mua hàng nhiều nhất

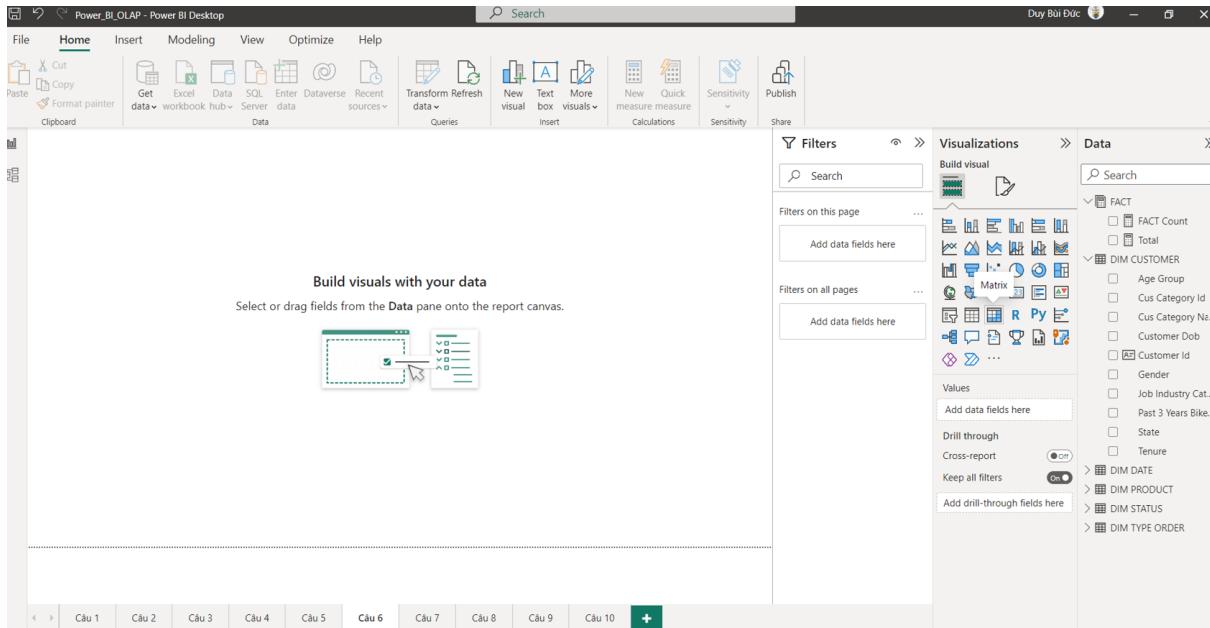
Cus Category Name	FACT Count
High Risk Customer	297
Losing Customer	284
Potential Customer	315
Recent Customer	394
Very Loyal	297
Total	1.587

Matrix

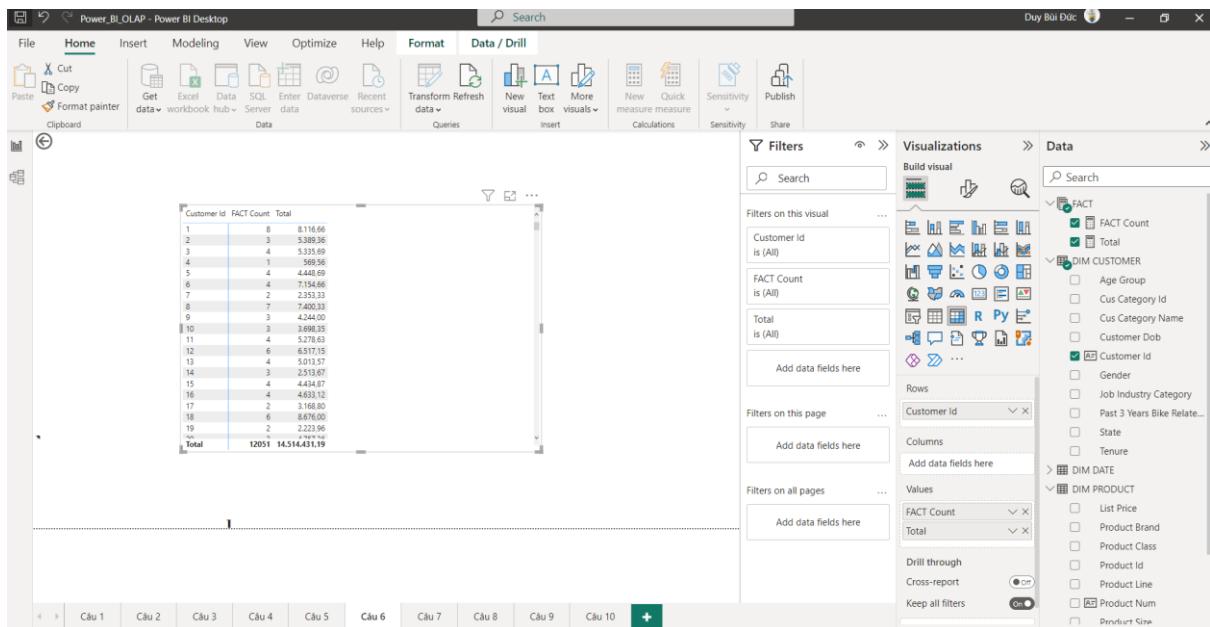
Nhận xét: Trong top 5 nhóm khách hàng mua nhiều nhất tại bang VIC thì khá tương đương nhau chỉ có nhóm *Recent Customer* là có số đơn cao vượt trội so với các nhóm còn lại.

2.6 Thống kê những khách hàng mua hàng hơn 9 lần và tổng số tiền mua hơn 10000\$

Bước 1: Chọn loại Report Matrix.



Bước 2: Kéo thuộc tính “Customer Id” vào “Rows”, “FACT Count”, “Total” vào “Values”.



Bước 3: Ở “Filters” tại ô “FACT Count” mục “Show items when the value” chọn “is greater than” điền “9” và nhấn “Apply filter”, tại ô “Total” mục “Show items when the value” chọn “is greater than” điền “10000” và nhấn “Apply filter”.

The screenshot shows the Power BI Desktop interface. On the left, there is a table visualization titled "Customer Id FACT Count Total". The table contains the following data:

Customer Id	FACT Count	Total
637	10	11.494,76
936	10	14.747,19
1129	10	14.466,38
1913	12	14.652,70
1991	10	15.263,75
2183	10	14.307,58
2476	11	11.885,89
2874	10	13.730,94
Total	83	110.549,19

The "Filters" pane on the right shows two filters applied:

- For "FACT Count": "is greater than" 9, selected "And", "Apply filter".
- For "Total": "is greater than" 10000, selected "And", "Apply filter".

Bước 4: Duplicate Table vừa tạo sau đó điều chỉnh về dạng “Clustered Column Chart”, sau đó điều chỉnh dữ liệu: kéo thuộc tính “Total” vào ô “Y-axis”, thuộc tính “Customer Id” vào ô “X-axis”.

The screenshot shows the Power BI Desktop interface with a clustered bar chart titled "Những khách hàng mua hàng hơn 9 lần và tổng số tiền mua sắm hơn 10000\$". The chart displays the total amount spent by customers who have made more than 9 purchases. The Y-axis is labeled "Total" and the X-axis is labeled "Customer Id". The chart has three bars representing different customers.

The "Properties" pane on the right shows the following settings for the chart:

- Title:** Nhng kha>ch hngmua hng hngn 9 lnvàtngso tnnmua
- Font:** Times New Roman, size 20
- Text color:** Black
- Background color:** Red
- Horizontal alignment:** Center

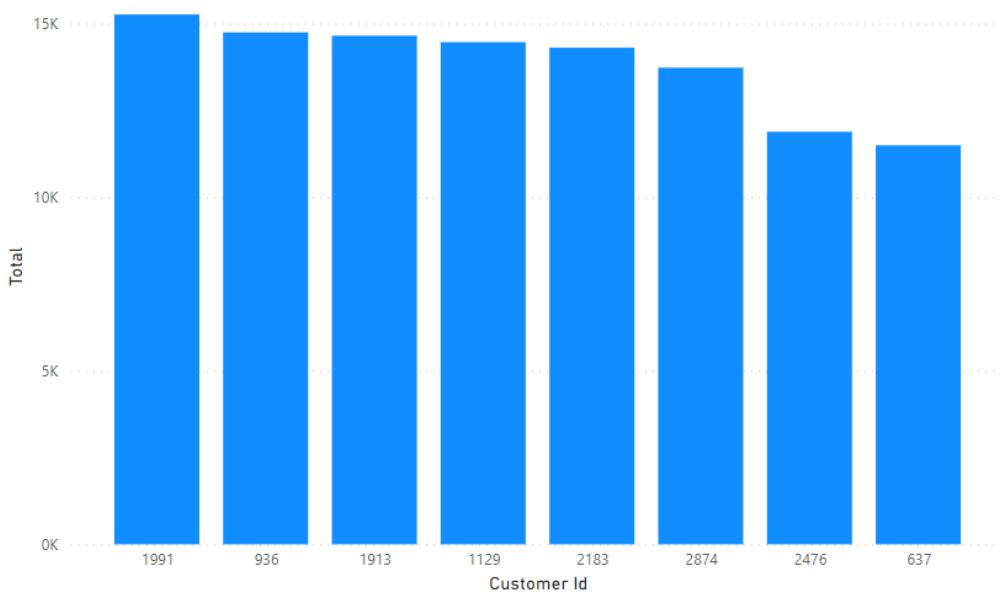
Bước 5: Điều chỉnh Title và Style, ta có kết quả.

Những khách hàng mua hàng hơn 9 lần và tổng số tiền mua hơn 10000\$

Customer Id	FACT Count	Total
637	10	11.494,76
936	10	14.747,19
1129	10	14.466,38
1913	12	14.652,70
1991	10	15.263,75
2183	10	14.307,58
Total	83	110.549,19

Matrix

Những khách hàng mua hàng hơn 9 lần và tổng số tiền mua hơn 10000\$



Cluttered Column Chart

Nhận xét: Đa số khách hàng đều mua nhiều nhất là 10 đơn, trừ khách hàng có id 1913 là mua 12 đơn, khách hàng mua đơn thanh toán nhiều nhất là khách hàng có id 1991 tổng tiền đã thanh toán là 15263,75\$.

2.7 Thống kê danh sách và số lượng đơn hàng khách hàng nữ đã mua và đã thanh toán tại cửa hàng ngày 8/3/2017 không thuộc bang VIC

Bước 1: Chọn loại Report Matrix.

Bước 2: Kéo thuộc tính “Customer Id”, “State” vào “Rows”, “FACT Count” vào “Values”.

Bước 3: Ở “Filters” tại ô “State” mục “Filter type” tick chọn “NSW”, “QLD”.

The screenshot shows the Power BI Desktop interface. On the left, a matrix visualization displays data with columns 'Customer Id' and 'FACT Count'. The data includes rows for Customer Ids 292, 681, 871, 1104, and 2361, each with a FACT Count of 1. There are also rows for QLD and NSW states, each with a FACT Count of 1. A total row at the bottom has a FACT Count of 5. On the right, the 'Filters' pane is open, showing a 'Basic filtering' section where 'NSW' and 'QLD' are selected under the 'State' dropdown. The 'Data' pane on the far right lists various dimensions and their fields, with 'State' checked under 'DIM CUSTOMER'.

Bước 4: Điều chỉnh Title và Style, ta có kết quả.

Danh sách và số lượng đơn hàng khách hàng nữ đã mua và đã thanh toán tại cửa hàng ngày 8/3/2017 không thuộc bang VIC

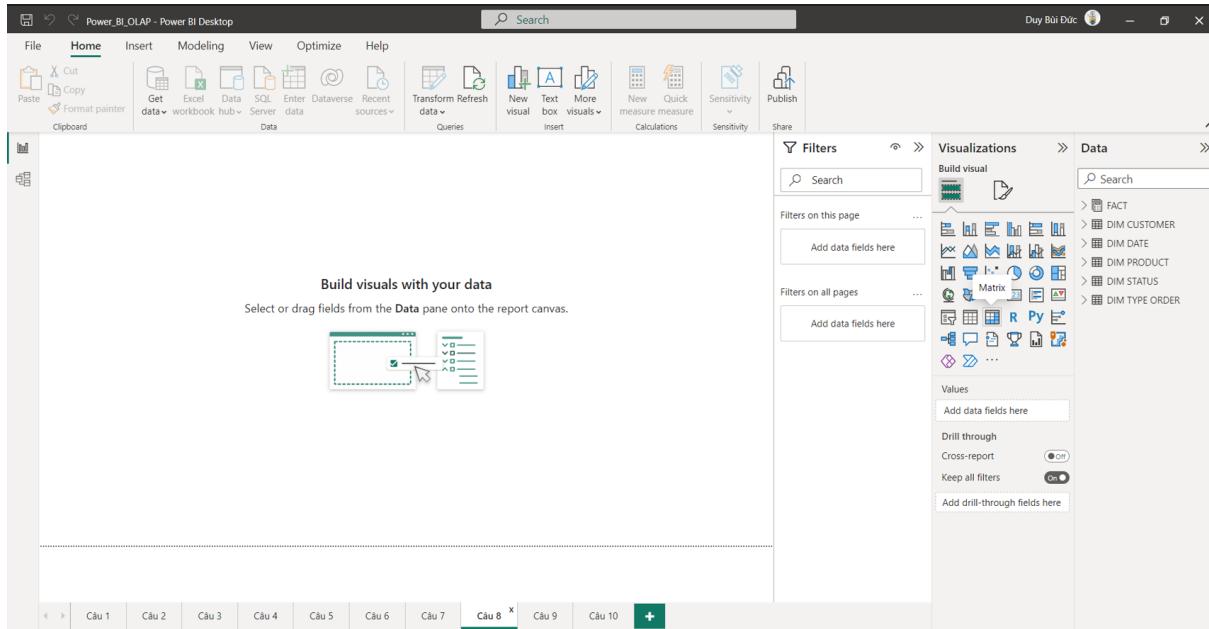
Customer Id	FACT Count
292	1
QLD	1
681	1
NSW	1
871	1
NSW	1
1104	1
NSW	1
2361	1
QLD	1
Total	5

Matrix

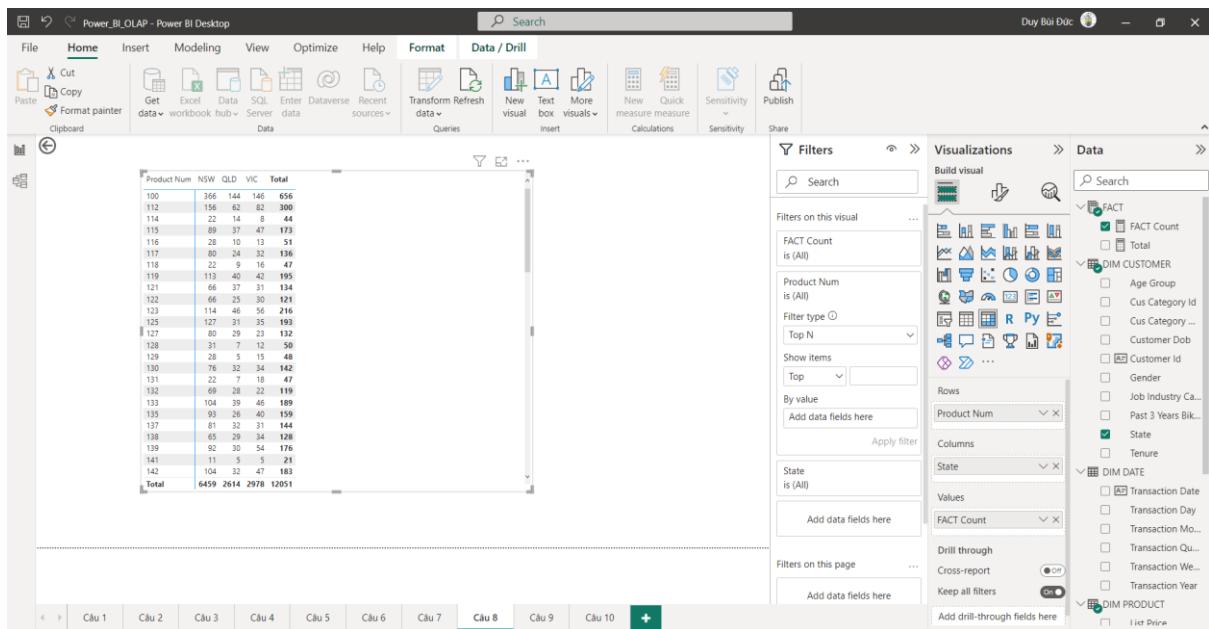
Nhận xét: Số lượng khách hàng nữ mua hàng vào ngày 8/3/2017 tại các bang là như nhau.

2.8 Thống kê top 5 sản phẩm bán chạy nhất trong từng bang

Bước 1: Chọn loại Report Matrix.



Bước 2: Kéo thuộc tính “Product Num” vào “Rows”, “State” vào “Columns”, “FACT Count” vào “Values”.



Bước 3: Ở “Filters” tại ô “Product Num” chọn “Filter type” là “TopN”, “Show items” chọn “Top” và điền “5”, kéo thuộc tính “FACT Count” vào “By value” và nhấn “Apply filter”.

The screenshot shows the Power BI Desktop interface. On the left, there is a table visual with columns: Product Num, NSW, QLD, VIC, Total. The data is as follows:

Product Num	NSW	QLD	VIC	Total
100	366	144	146	656
112	156	62	82	300
123	114	46	56	216
164	109	59	38	206
191	100	50	60	210
Total	845	361	382	1588

The “Filters” pane on the right is open, showing the following settings for the “Product Num” visual:

- Filter type: Top N
- Show items: Top 5
- By value: FACT Count
- Apply filter

The “Visualizations” pane shows a clustered column chart selected. The “Data” pane on the right lists various dimensions and facts, with “FACT Count” checked under the “FACT” section.

Bước 4: Duplicate Table vừa tạo sau đó điều chỉnh về dạng “Clustered Column Chart”, sau đó điều chỉnh dữ liệu. “X-axis” chọn “State”, “Legend” chọn “Product Num”, “Y-axis” chọn “FACT Count”.

The screenshot shows the Power BI Desktop interface with a clustered column chart visual. The chart displays “FACT Count by State and Product Num”. The X-axis is labeled “State” and shows categories NSW, VIC, and QLD. The Y-axis is labeled “FACT Count” and ranges from 0 to 400. The legend indicates five product numbers: 100 (blue), 112 (orange), 123 (purple), 164 (pink), and 191 (yellow). The chart data is as follows:

State	Product Num	FACT Count
NSW	100	366
NSW	112	156
NSW	123	114
NSW	164	109
NSW	191	100
VIC	100	144
VIC	112	62
VIC	123	46
VIC	164	59
VIC	191	50
QLD	100	146
QLD	112	82
QLD	123	56
QLD	164	38
QLD	191	60

The “Filters” pane on the right is open, showing the following settings for the “FACT Count by State and Product Num” visual:

- Filter type: Top N
- Show items: Top 5
- By value: FACT Count
- X-axis: State
- Y-axis: FACT Count
- Legend: Product Num

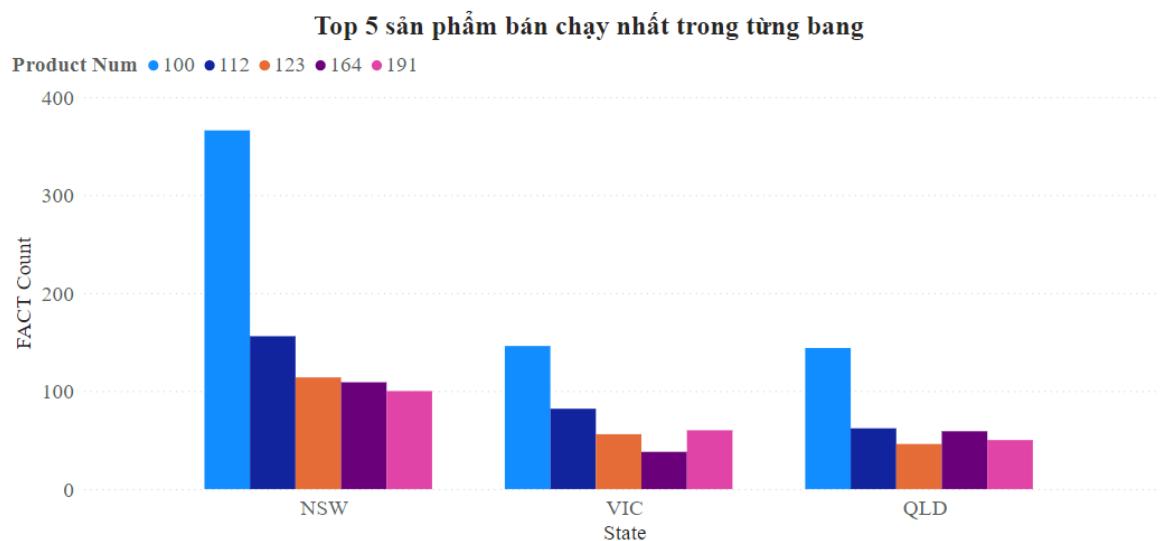
The “Visualizations” pane shows a clustered column chart selected. The “Data” pane on the right lists various dimensions and facts, with “FACT Count” checked under the “FACT” section.

Bước 5: Điều chỉnh Title và Style, ta có kết quả.

Top 5 sản phẩm bán chạy nhất trong từng bang

Product Num	NSW	QLD	VIC	Total
100	366	144	146	656
112	156	62	82	300
123	114	46	56	216
164	109	59	38	206
191	100	50	60	210
Total	845	361	382	1588

Matrix



Clustered Column Chart

Nhận xét: Sản phẩm với mã 100 bán chạy nhất trong cả 3 bang nhưng tại bang NSW thì bán nhiều nhất với tổng đơn là 336 gần như gấp 3 so với các sản phẩm còn lại trong top 5.

2.9 Thống kê những khách hàng tiềm năng thuộc nhóm khách hàng từ 20 đến dưới 40 tuổi mua sản phẩm trong quý 1

Bước 1: Chọn loại Report Matrix.

The screenshot shows the Power BI Desktop interface. The ribbon is at the top with tabs: File, Home (selected), Insert, Modeling, View, Optimize, Help. The Home tab has icons for Paste, Cut, Copy, Format painter, Get data (with sub-options: Excel workbook hub, Data, SQL Server, Enter data, Dataverse, Recent sources), Transform Refresh data (with sub-options: New visual, Text box, More visuals), Insert (with sub-options: New quick measure, Calculations, Sensitivity), Publish, and Share. The Data pane on the left lists fields: Customer Id, FACT Count, Total, Cus Category Name, Transaction Quarter, Age Group, Customer Dob, Gender, Job Industry Ca..., Past 3 Years Bic..., State, Tenure, DIM DATE, DIM CUSTOMER, DIM PRODUCT, and DIM STATUS. The Visualizations pane shows icons for various charts like Matrix, Line, Bar, Clustered Bar, Stacked Bar, Line with Area, Scatter, Radar, Map, Gauge, Card, Timeline, and R PY. The Data pane on the right shows filters for Customer Id, FACT Count, and Total. A matrix visual is being built on the canvas.

Bước 2: Kéo thuộc tính “Customer Id” vào “Rows”, “FACT Count”, “Total” vào “Values”, kéo các thuộc tính “Cus Category Name”, “Transaction Quarter”, “Age Group” vào “Drill through”.

The screenshot shows the Power BI Desktop interface with the Home tab selected. The ribbon has 'File', 'Home' (selected), 'Insert', 'Modeling', 'View', 'Optimize', 'Help', 'Format', and 'Data / Drill'. The Data pane on the left shows a matrix visual with columns 'Customer Id', 'FACT Count', and 'Total'. The matrix has 22 rows of data. The Visualizations pane on the right shows the matrix icon selected. The Data pane on the right shows filters for Customer Id, FACT Count, and Total. The ribbon tabs are: File, Home (selected), Insert, Modeling, View, Optimize, Help, Format, Data / Drill.

Bước 3: Ở “Drill through” của “Cus Category Name”, “Transaction Quarter”, “Age Group”, tick chọn lần lượt “Potential Customer”, “1”, “U40”.

The screenshot shows the Power BI Desktop interface. In the center, there is a matrix visualization with columns labeled 'Customer Id', 'FACT Count', and 'Total'. The data rows are numbered from 47 to 1001, with the total row at the bottom. The 'Total' column shows values such as 569,56, 912,52, etc., and the 'Total' row shows 118 and 143.195,44. To the right of the matrix is the 'Data' pane, which displays a hierarchical list of fields under 'FACT', 'DIM CUSTOMER', 'DIM DATE', and 'DIM PRODUCT'. Under 'Age Group', 'U40' is checked. The 'Used as category' dropdown also has 'U40' selected. The top ribbon shows tabs like File, Home, Insert, Modeling, View, Optimize, Help, and various icons for data import, transforms, and publishing.

Bước 4: Điều chỉnh Title và Style, ta có kết quả.

Những khách hàng tiềm năng thuộc nhóm khách hàng từ 20 đến dưới 40 tuổi mua sản phẩm trong quý 1

Customer Id	FACT Count	Total
47	1	569,56
90	1	912,52
148	1	1.458,17
149	1	1.073,07
208	3	4.084,45
221	1	1.466,68
233	1	1.458,17
250	2	3.065,42
337	1	1.280,28
345	1	1.992,93
352	1	1.992,93
367	2	3.209,07
382	1	774,53
408	1	912,52
475	2	2.905,45
488	2	1.186,36
546	1	175,89
553	1	1.403,50
777	1	774,53
859	1	1.311,44
Total	118	143.195,44

Matrix

Nhận xét: Những khách hàng tiềm năng thuộc nhóm khách hàng từ 20 đến dưới 40 tuổi mua sản phẩm trong quý 1 thì khách hàng có id 208 là tiềm năng nhất với số đơn là 3 và tổng thanh toán là 4084,45\$, tiếp đến là khách hàng có id 367, 250 và 475 đều có tổng đơn là 2 và tổng thanh toán xấp xỉ 3000\$ trở lên.

2.10 Thống kê top 3 tháng có đơn hàng bị hủy nhiều nhất

Bước 1: Chọn loại Report Matrix.

Bước 2: Kéo thuộc tính “Transaction Month” vào “Rows”, “FACT Count” vào “Values”, “Transaction Status” vào “Drill through”.

Transaction Month	FACT Count
1	1024
10	1071
11	1009
12	1040
2	969
3	988
4	1000
5	1009
6	945
7	1011
8	1052
9	933
Total	12051

Bước 3: Ở “Drill through” của “Transaction Status” chọn “Cancelled”, ở “Filters” tại ô “Transaction Month” chọn “Filter type” là “TopN”, “Show items” chọn “Top” và điền “3”, kéo thuộc tính “FACT Count” vào “By value” và nhấn “Apply filter”.

Bước 4: Điều chỉnh Title và Style, ta có kết quả

Top 3 tháng có đơn hàng bị hủy nhiều nhất

Transaction Month	FACT Count
4	12
5	12
9	11
Total	35

Matrix

Nhận xét: Tháng 4,5 có tổng đơn hàng bị hủy nhiều nhất với 12 đơn tiếp đến là tháng 9 với 11 đơn.

CHƯƠNG 5. DATA MINING

1. Phần mềm và ngôn ngữ sử dụng

Phần mềm Visual Studio Code, ngôn ngữ Python phiên bản 3.11.0.

2. Các bước xử lý chung

- Import thư viện

```
● ✓ import numpy as np
    import pandas as pd
    import seaborn as sns
    sns.set(style="whitegrid")
    ✓ import matplotlib.pyplot as plt
    %matplotlib inline
    import warnings
    warnings.filterwarnings('ignore')

    ✓ import category_encoders as ce
    from sklearn.model_selection import train_test_split
    from sklearn.ensemble import RandomForestClassifier
    from sklearn import svm
    from sklearn.metrics import accuracy_score
    from sklearn.metrics import confusion_matrix
    from sklearn.metrics import classification_report
    from sklearn.inspection import permutation_importance
    import sklearn.metrics as metrics
```

✓ 10.4s Python

- Import và xem tổng quan về dataset
 - Import dataset

```
data = 'Bike_Accessories_Sales.xlsx'
df = pd.read_excel(data)
```

✓ 6.7s Python

- Kích thước dataset

```
# print the shape
print('The shape of the dataset : ', df.shape)
```

✓ 0.1s Python

The shape of the dataset : (12051, 21)

- Xem dataset

df
✓ 0.1s

	transaction_id	product_id	customer_id	transaction_date	transaction_month	online_order	transaction_status	product_brand	product_line	product_class	product_size	total	list_price	gender	past_3_years_bike_related_purchases	customer_dob	age_group	job_industry_category	tenure	cus_category_name	state
0	5323	11	3010	2017-01-01	January	False	Approved	Giant	Electronics	Electronics	Small	100.0	100.0	Male	0	1999-12-31	18-24	Manufacturing	10	Electronics	California
1	3525	50	3490	2017-01-01	January	True	Approved	Giant	Electronics	Electronics	Medium	150.0	150.0	Male	0	1999-12-31	18-24	Manufacturing	10	Electronics	California
2	3460	64	234	2017-01-01	January	True	Approved	Giant	Electronics	Electronics	Large	200.0	200.0	Male	0	1999-12-31	18-24	Manufacturing	10	Electronics	California
3	140	11	462	2017-01-01	January	False	Approved	Giant	Electronics	Electronics	Small	50.0	50.0	Male	0	1999-12-31	18-24	Manufacturing	10	Electronics	California
4	19158	32	671	2017-01-01	January	False	Approved	Giant	Electronics	Electronics	Medium	100.0	100.0	Male	0	1999-12-31	18-24	Manufacturing	10	Electronics	California
...	
12046	7518	65	3433	2017-12-30	December	False	Approved	Wet	Food	Food	Small	50.0	50.0	Male	0	1999-12-31	18-24	Manufacturing	10	Food	California
12047	9820	31	1425	2017-12-30	December	True	Approved	Wet	Food	Food	Medium	100.0	100.0	Male	0	1999-12-31	18-24	Manufacturing	10	Food	California
12048	12667	89	1594	2017-12-30	December	True	Approved	Wet	Food	Food	Large	200.0	200.0	Male	0	1999-12-31	18-24	Manufacturing	10	Food	California
12049	16166	89	2345	2017-12-30	December	False	Approved	Wet	Food	Food	Small	50.0	50.0	Male	0	1999-12-31	18-24	Manufacturing	10	Food	California
12050	17182	26	3169	2017-12-30	December	True	Approved	Wet	Food	Food	Medium	100.0	100.0	Male	0	1999-12-31	18-24	Manufacturing	10	Food	California

12051 rows × 21 columns

➤ Thông tin tổng quan dataset

df.info()
✓ 0.1s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12051 entries, 0 to 12050
Data columns (total 21 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   transaction_id  12051 non-null   int64  
 1   product_id      12051 non-null   int64  
 2   customer_id     12051 non-null   int64  
 3   transaction_date 12051 non-null   datetime64[ns]
 4   transaction_month 12051 non-null   object  
 5   online_order    12051 non-null   bool   
 6   transaction_status 12051 non-null   object  
 7   product_brand   12051 non-null   object  
 8   product_line    12051 non-null   object  
 9   product_class   12051 non-null   object  
 10  product_size    12051 non-null   object  
 11  total           12051 non-null   float64 
 12  list_price      12051 non-null   float64 
 13  gender          12051 non-null   object  
 14  past_3_years_bike_related_purchases 12051 non-null   int64  
 15  customer_dob    12051 non-null   datetime64[ns]
 16  age_group       12051 non-null   object  
 17  job_industry_category 12051 non-null   object  
 18  tenure          12051 non-null   int64  
 19  cus_category_name 12051 non-null   object  
 20  state           12051 non-null   object  
dtypes: bool(1), datetime64[ns](2), float64(2), int64(5), object(11)
```

- Xóa các thuộc tính không có ý nghĩa cho việc phân loại

Các cột kiểu id, datetime và một số biến ko có ý nghĩa cho việc phân loại nên tiến hành xóa những cột này

```
columns_to_drop = ['transaction_id', 'product_id', 'customer_id',
                    'transaction_date', 'customer_dob', 'job_industry_category',
                    'past_3_years_bike_related_purchases', 'tenure']
df = df.drop(columns_to_drop, axis=1)
```

✓ 0.7s

Python

- Kiểm tra các giá trị còn thiếu

```
df.isnull().sum()
```

✓ 0.1s

Python

transaction_month	0
online_order	0
transaction_status	0
product_brand	0
product_line	0
product_class	0
product_size	0
total	0
list_price	0
gender	0
age_group	0
cus_category_name	0
state	0

dtype: int64

- Chuyển cột online_order kiểu boolean về kiểu int

```
df['online_order'] = df['online_order'].astype(int)
```

✓ 0.9s

Python

- Kiểm tra lại

```
df.online_order.unique()
```

✓ 0.8s

Python

array([0, 1])

- Chuyển đổi các biến phân loại về dạng số để train

```

type_transaction_month = {'January':1, 'February':2, 'March':3, 'April':4, 'May':5, 'June':6,
|   'July':7, 'August':8, 'September':9, 'October':10, 'November':11, 'December':12}
df.transaction_month = df.transaction_month.map(type_transaction_month)

type_product_brand = {'Giant Bicycles':1, 'Norco Bicycles':2, 'OHM Cycles':3, 'WeareA2B':4}
df.product_brand = df.product_brand.map(type_product_brand)

type_product_line = {'Standard':1, 'Road':2, 'Touring':3, 'Mountain':4}
df.product_line = df.product_line.map(type_product_line)

type_product_class = {'high':1, 'medium':2, 'low':3}
df.product_class = df.product_class.map(type_product_class)

type_product_size = {'medium':1, 'small':2, 'large':3}
df.product_size = df.product_size.map(type_product_size)

type_gender = {'Female':1, 'Male':2, 'U':3}
df.gender = df.gender.map(type_gender)

type_age_group = {'U20':1, 'U40':2, 'U60':3, 'U100':4}
df.age_group = df.age_group.map(type_age_group)

type_cus_category_name = {'Evasive Customer':1, 'Platinum Customer':2, 'Almost Lost Customer':
|   'Losing Customer':4, 'Becoming Loyal':5, 'Potential Customer':6,
|   'Lost Customer':7, 'Very Loyal':8, 'Recent Customer':9, 'Late Bloomer':10,
|   'High Risk Customer':11}
df.cus_category_name = df.cus_category_name.map(type_cus_category_name)

type_state = {'NSW':1, 'QLD':2, 'VIC':3}
df.state = df.state.map(type_state)

```

- Phân tích biến mục tiêu “transaction_status”

Kiểm tra các giá trị thiếu trong biến mục tiêu

```

# check for missing values
df['transaction_status'].isnull().sum()

```

✓ 0.1s

Python

0

Xem các giá trị duy nhất trong biến mục tiêu

```

# view the unique values
df['transaction_status'].unique()

```

✓ 0.2s

Python

array(['Approved', 'Cancelled'], dtype=object)

Số lượng các giá trị biến mục tiêu trong dataset

```
# view the frequency distribution of values  
df['transaction_status'].value_counts()
```

✓ 0.2s

Python

```
Approved      9444  
Cancelled     2607  
Name: transaction_status, dtype: int64
```

3. Chuẩn bị dữ liệu

- Tách thuộc tính mục tiêu “transaction_status”

```
X = df.drop(['transaction_status'], axis=1)  
y = df['transaction_status']
```

✓ 0.1s

Python

- Chia dataset để train và test

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)
```

✓ 0.1s

Python

- Số lượng thuộc tính và kích thước dữ liệu để train, test

```
# Kiểm tra kích thước của X_train và X_test  
X_train.shape, X_test.shape
```

✓ 0.9s

Python

```
((8435, 12), (3616, 12))
```

4. Train model

4.1 Random Forest Classifier model

- Chạy model

```
# instantiate the classifier with n_estimators = 100  
rfc_100 = RandomForestClassifier(n_estimators=100, random_state=0)
```

```
# fit the model to the training set  
rfc_100.fit(X_train, y_train)
```

✓ 1.1s

Python

```
▼      RandomForestClassifier  
RandomForestClassifier(random_state=0)
```

- Dự đoán

```
# Dự đoán trên tập test  
y_pred_rf =rf.predict(x_test)
```

Python

- Tính độ chính xác

```
print(classification_report(y_test, y_pred_rf))
```

Python

	precision	recall	f1-score	support
Approved	0.98	0.99	0.99	2849
Cancelled	0.96	0.94	0.95	767
accuracy			0.98	3616
macro avg	0.97	0.97	0.97	3616
weighted avg	0.98	0.98	0.98	3616

- Vẽ ma trận nhầm lẫn (Confusion matrix)

```
# In Ma trận nhầm lẫn và chia thành 4 phần
cm = confusion_matrix(y_test, y_pred_rf)
print('Confusion matrix\n\n', cm)
# Trực quan hóa ma trận nhầm lẫn với bản đồ nhiệt seaborn
cm_matrix = pd.DataFrame(data=cm, columns=['Actual Positive:1', 'Actual Negative:0'],
                           index=['Predict Positive:1', 'Predict Negative:0'])

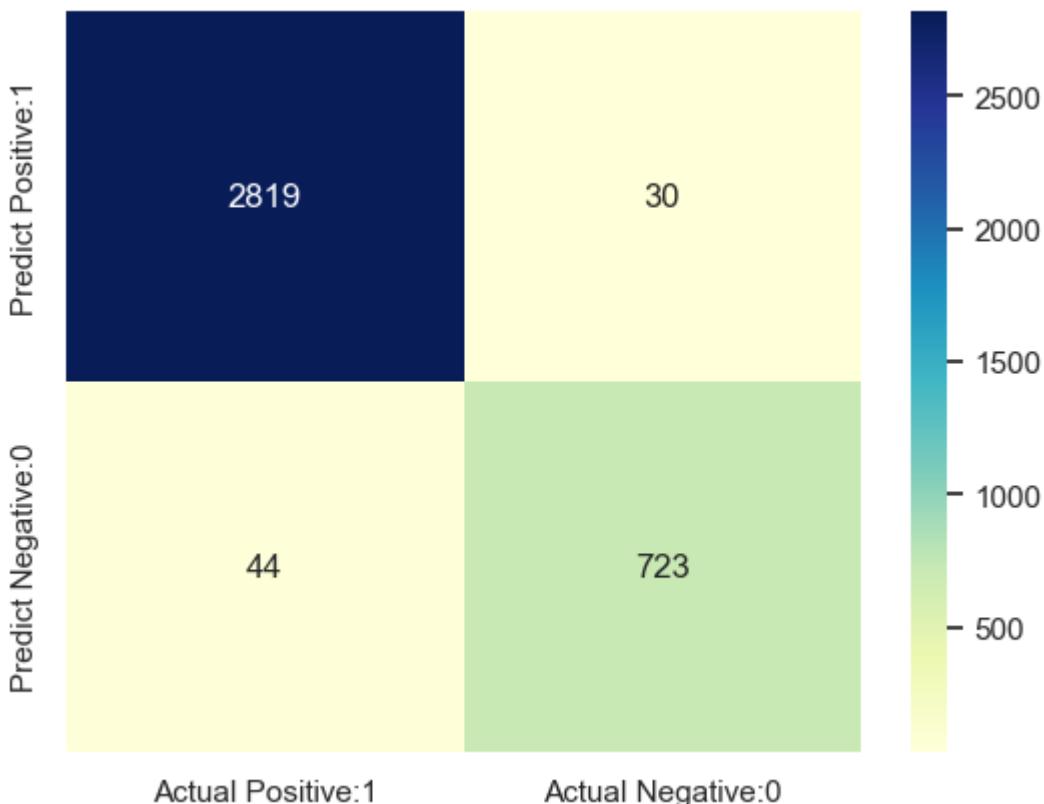
sns.heatmap(cm_matrix, annot=True, fmt='d', cmap='YlGnBu')
```

Python

Confusion matrix

```
[[2819    30]
 [ 44   723]]
```

<AxesSubplot: >



Nhận xét:

- ❖ Kết quả dự đoán đúng với thực tế Approved là 2819 dòng, với Cancelled là 30 dòng.
- ❖ Kết quả dự đoán sai với thực tế Approved là 44 dòng, với Cancelled là 723 dòng.
- Biểu diễn thuộc tính quan trọng
 - Tìm giá trị thuộc tính quan trọng

```
# Tìm giá trị thuộc tính quan trọng
rf.feature_importances_
```

✓ 0.1s

Python

```
array([0.56124771, 0.00788002, 0.22009439, 0.00987702, 0.00818148,
       0.00959001, 0.0481672 , 0.06882922, 0.00770916, 0.01371571,
       0.03371175, 0.01099634])
```

➤ Tìm tên thuộc tính quan trọng

```
# Tìm tên thuộc tính quan trọng
rf.feature_names_in_
```

✓ 0.8s

Python

```
array(['transaction_month', 'online_order', 'product_brand',
       'product_line', 'product_class', 'product_size', 'total',
       'list_price', 'gender', 'age_group', 'cus_category_name', 'state'],
      dtype=object)
```

➤ Tạo tệp dữ liệu mẫu

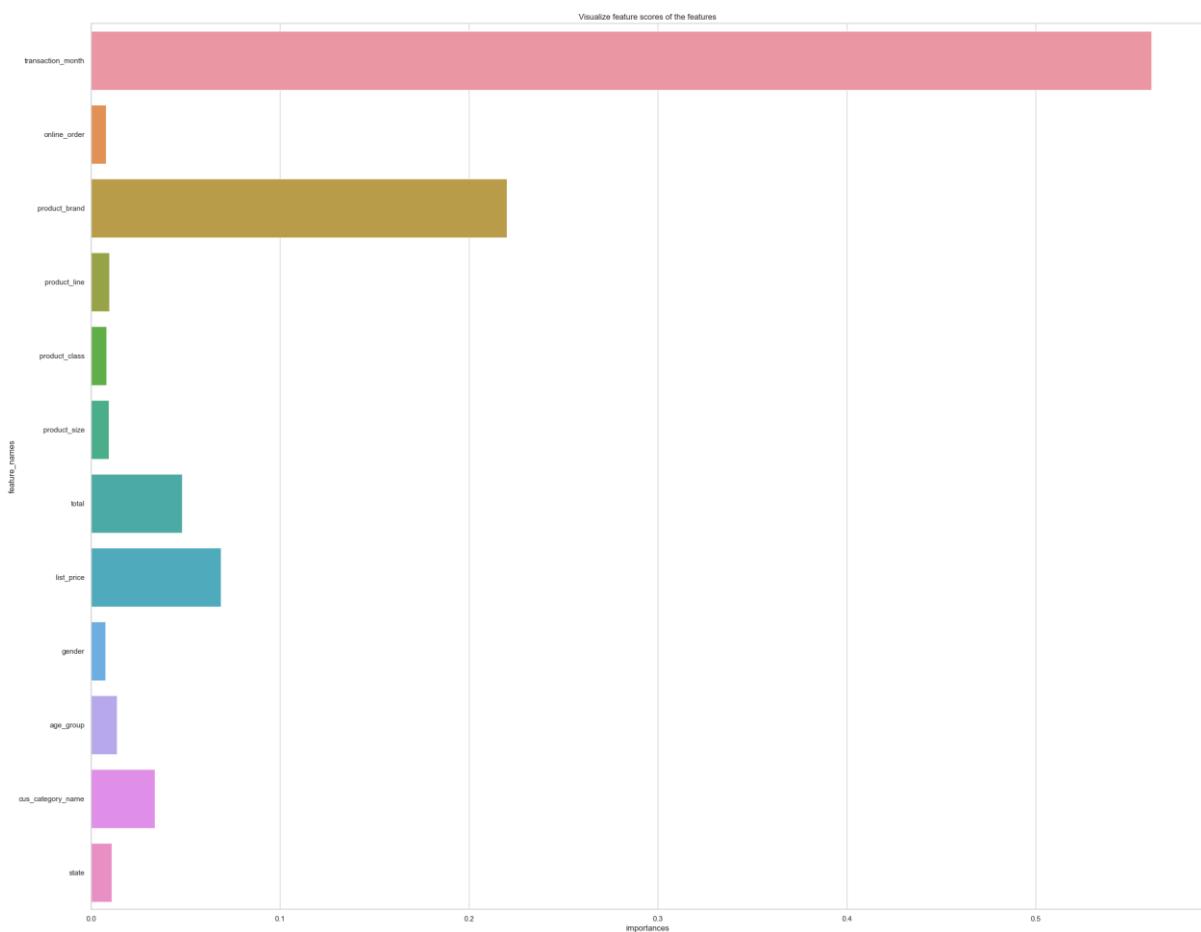
```
# Tạo tệp dữ liệu mẫu  
feature_names = rf.feature_names_in_  
importances = rf.feature_importances_  
✓ 0.7s
```

Python

➤ Tạo một biểu đồ barplot với seaborn

```
# Tạo một biểu đồ barplot với seaborn  
  
importances_df = pd.DataFrame({'feature_names':feature_names,  
                                "importances": importances})  
  
f, ax = plt.subplots(figsize=(30, 24))  
ax = sns.barplot(x=importances_df["importances"], y=importances_df["feature_names"]  
  
ax.set_title("Visualize feature scores of the features")  
# ax.set_yticklabels(rf.feature_names_in_)  
# ax.set_xlabel("Feature importance score")  
# ax.set_ylabel("Features")  
plt.show()  
✓ 0.9s
```

Python



Các thuộc tính ảnh hưởng nhất là: transaction_month, product_brand, total, list_price và cus_category_name.

- Vẽ cây rút ra tập luật:

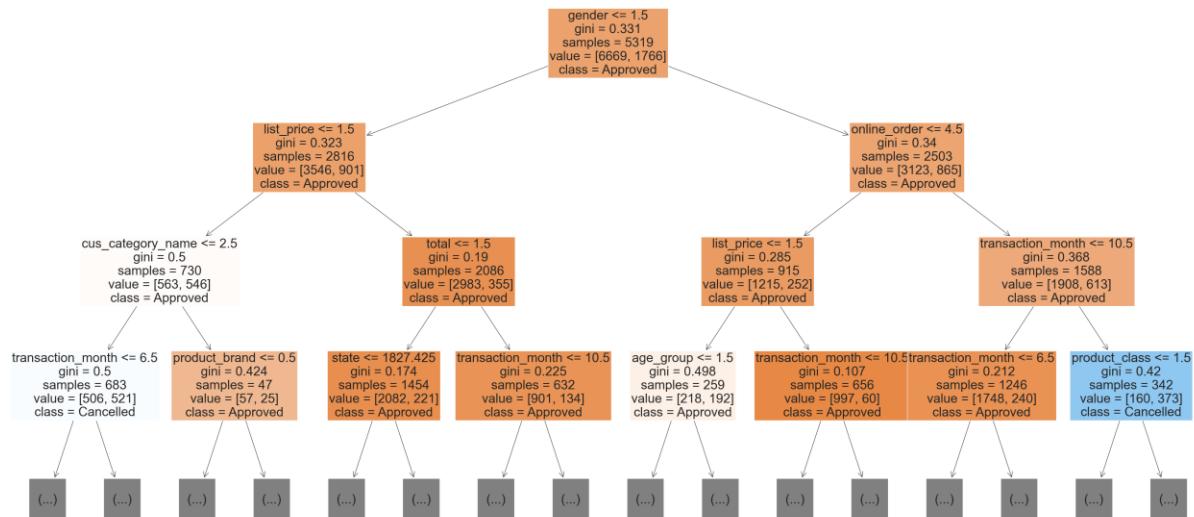
```

from sklearn import tree
# Truy cập vào cây quyết định đầu tiên trong Random Forest
decision_tree = rf.estimators_[0]

# Vẽ cây quyết định
fig, ax = plt.subplots(figsize=(50, 24))
tree.plot_tree(decision_tree, feature_names=feature_names,
               class_names=y_pred_rf, filled=True, fontsize=30, max_depth=3)
# Lưu hình ảnh của cây quyết định
plt.savefig('decision_tree_random_forest', dpi=100)
plt.show()

```

✓ 3.8s Python



Rút ra tập luật:

Tập luật 1: gender ≤ 1.5 , list_price ≤ 1.5 , cus_category_name ≤ 2.5 , transaction_month ≤ 6.5 .

Tập luật 2: gender ≤ 1.5 , list_price ≤ 1.5 , cus_category_name ≤ 2.5 , product_brand ≤ 0.5 .

Tập luật 3: gender ≤ 1.5 , list_price ≤ 1.5 , total ≤ 1.5 , state ≤ 1827.425 .

Tập luật 4: gender ≤ 1.5 , list_price ≤ 1.5 , total ≤ 1.5 , transaction_month ≤ 10.5 .

Tập luật 5: gender ≤ 1.5 , online_order ≤ 4.5 , list_price ≤ 1.5 , age_group ≤ 1.5 .

Tập luật 6: gender ≤ 1.5 , online_order ≤ 4.5 , list_price ≤ 1.5 , transaction_month ≤ 10.5 .

Tập luật 7: gender <= 1.5, online_order <= 4.5, transaction_month <= 10.5, transaction_month <= 6.5.

Tập luật 8: gender <= 1.5, online_order <= 4.5, transaction_month <= 10.5, product_class <= 1.5.

- Tạo dữ liệu để dự đoán

```
# Tạo dòng dữ liệu mới
new_data = {
    'transaction_month':7, 'online_order':0, 'product_brand':3,
    'product_line':1, 'product_class':1, 'product_size':1, 'total':1458.17,
    'list_price':583.27, 'gender':1, 'age_group':3, 'cus_category_name':10, 'state':1
}

✓ 0.7s
```

Python

```
# Chuyển từ data về DataFrame
new_data_df = pd.DataFrame([new_data])
```

✓ 0.7s

Python

- Dự đoán

```
# Dự đoán nhãn
predicted_label = rf.predict(new_data_df)

# In ra nhãn
print("transaction_status", predicted_label)
```

✓ 0.9s

Python

transaction_status ['Approved']

4.2 Support Vector Machine model

- Chạy model

```
# Tạo SupportVectorClassifier model
# SVM_model = svm.SVC(kernel='linear', probability=True)
SVM_model = svm.SVC(kernel='linear')

# fit the model to the training set
SVM_model.fit(X_train, y_train)
```

✓ 3m 16.2s

Python

```
SVC
SVC(kernel='linear')
```

- Kiểm tra thuộc tính phân lớp và tạo tập thuộc tính độc lập
 - Các thuộc tính phân lớp

```
# Thuộc tính phân lớp
SVM_model.classes_
✓ 0.1s
```

Python

```
array(['Approved', 'Cancelled'], dtype=object)
```

➤ Các thuộc tính độc lập

```
# Thuộc tính độc lập
feature_names_svm = ['transaction_month', 'online_order', 'product_brand',
                     'product_line', 'product_class', 'product_size', 'total',
                     'list_price', 'gender', 'age_group', 'cus_category_name', 'state']
features = np.array(feature_names_svm)
✓ 0.9s
```

- Dự đoán trên tập test

```
# Dự đoán trên tập test
y_pred_svm = SVM_model.predict(X_test)
✓ 0.1s
```

Python

- Tính độ chính xác

```
print(classification_report(y_test, y_pred_svm))
✓ 0.2s
```

Python

	precision	recall	f1-score	support
Approved	0.94	0.98	0.96	2849
Cancelled	0.91	0.77	0.83	767
accuracy			0.93	3616
macro avg	0.92	0.87	0.90	3616
weighted avg	0.93	0.93	0.93	3616

- Vẽ ra ma trận nhầm lẫn (Confusion Matrix)

```
# Print the Confusion Matrix and slice it into four pieces
cm = confusion_matrix(y_test, y_pred_svm)
print('Confusion matrix\n\n', cm)
# visualize confusion matrix with seaborn heatmap
cm_matrix = pd.DataFrame(data=cm, columns=['Actual Positive:1', 'Actual Negative:0'],
                           index=['Predict Positive:1', 'Predict Negative:0'])

sns.heatmap(cm_matrix, annot=True, fmt='d', cmap='YlGnBu')
```

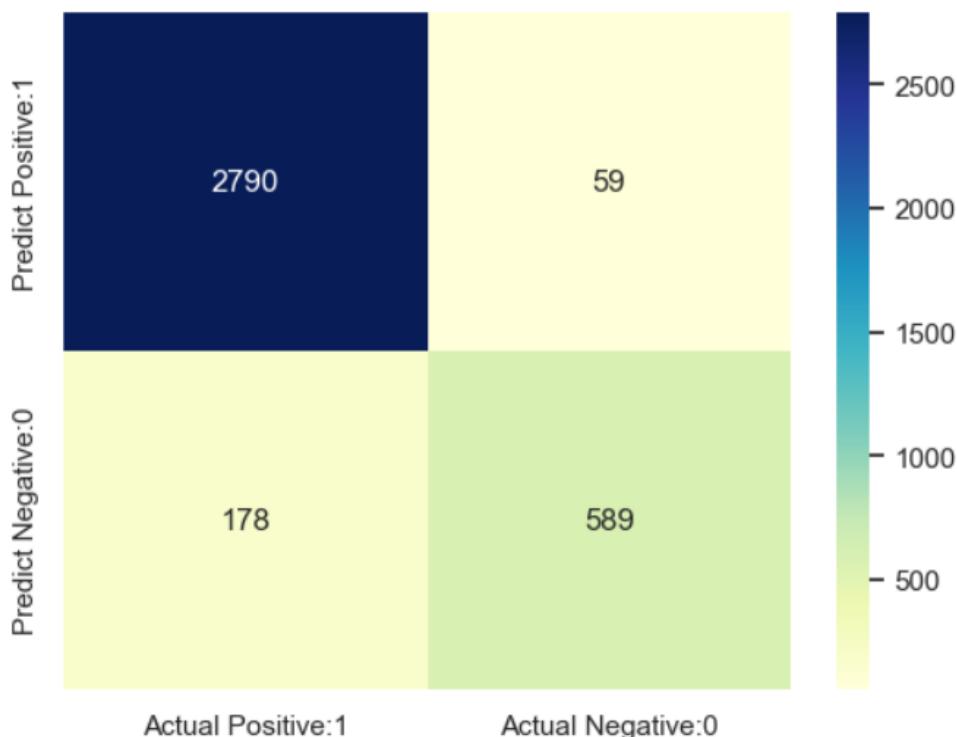
✓ 0.4s

Python

Confusion matrix

```
[[2790  59]
 [178  589]]
```

<AxesSubplot: >



Nhận xét:

- ❖ Kết quả dự đoán đúng với thực tế Approved là 2790 dòng, với Cancelled là 59 dòng.
- ❖ Kết quả dự đoán sai với thực tế Approved là 178 dòng, với Cancelled là 589 dòng.
- Tính toán độ quan trọng

```
result = permutation_importance(SVM_model, X_test, y_test, n_repeats = 10, random_state=0)
```

✓ 9.2s

Python

- Xuất ra kết quả giá trị thuộc tính quan trọng

```
importance_scores = result.importances_mean
importance_scores
✓ 0.1s
```

array([0.1857854 , 0. , 0.15210177, -0.00080199, 0.00497788,
 -0.00063606, 0.00605642, 0.00412058, 0. , 0. , 0. , 0. , 0.])

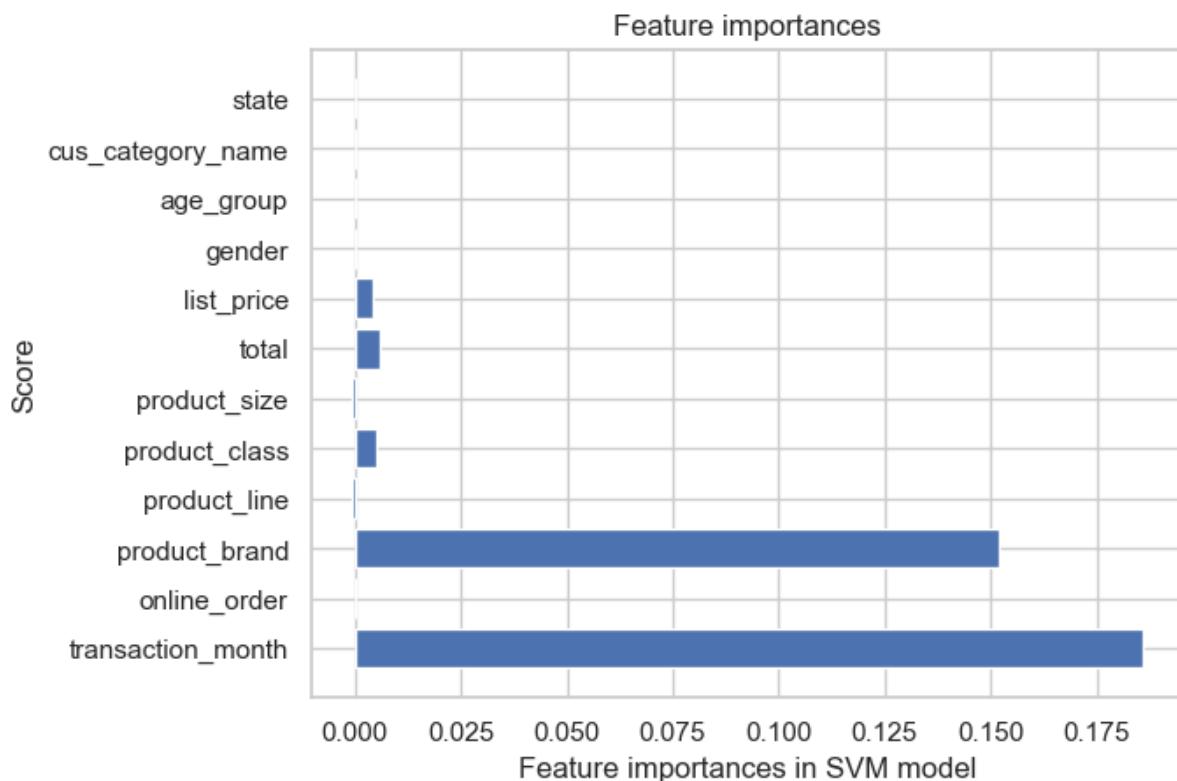
Python

- Vẽ biểu đồ biểu diễn các thuộc tính

```
plt.barh(features, importance_scores)
plt.title("Feature importances")
plt.xlabel("Feature importances in SVM model")
plt.ylabel("Score")

plt.show()
✓ 0.7s
```

Python



Các thuộc tính ảnh hưởng nhất là: transaction_month, online_order, total, list_price và product_class.

- Tạo dữ liệu để dự đoán

```
# Tạo dòng dữ liệu mới
new_data_svm = {
    'transaction_month':7, 'online_order':0, 'product_brand':3,
    'product_line':1, 'product_class':1, 'product_size':1, 'total':1458.17,
    'list_price':583.27, 'gender':1, 'age_group':3, 'cus_category_name':10, 'state':1
}
```

✓ 0.7s

Python

```
# Chuyển từ data về DataFrame
new_data_df_svm = pd.DataFrame([new_data_svm])
```

✓ 0.8s

Python

- Dự đoán nhãn

```
# Dự đoán nhãn
predicted_label_svm = SVM_model.predict(new_data_df_svm)

# In ra nhãn
print("transaction_status", predicted_label_svm)
```

✓ 0.1s

Python

transaction_status ['Approved']

5. Kết luận

Hai mô hình đều cho ra độ chính xác kết quả chính xác cao với 98% (Random Forest Classifier) và 93% (Support Vector Machine) để dự đoán tình trạng thanh toán (Approved) và hủy (Cancelled) đơn hàng.

Hai mô hình đều cho ra kết quả dự đoán nhãn giống nhau.