

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



ĐỒ ÁN MÔN HỌC
KHO DỮ LIỆU VÀ OLAP

ĐỀ TÀI
PHÂN TÍCH BỘ DỮ LIỆU DOANH SỐ BÁN XE ĐẠP

Giảng viên: ThS. Nguyễn Thị Kim Phụng

Lớp: IS217.O21.HTCL

Thành viên:

Trần Thị Kim Anh – 21520596

Lê Minh Chánh – 21521882

Thành phố Hồ Chí Minh, tháng 7 năm 2024

LỜI CẢM ƠN

Đầu tiên, nhóm chúng em xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới giảng viên ThS. Nguyễn Thị Kim Phụng – người đã giảng dạy và chia sẻ rất nhiều kiến thức cũng như các ví dụ thực tiễn trong các bài giảng. Cô đã hướng dẫn cho chúng em làm bài tập, sửa chữa và đóng góp nhiều ý kiến quý báu giúp chúng em hoàn thành tốt báo cáo môn học của mình.

Bộ môn Kho dữ liệu và OLAP là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Tuy nhiên, do vốn kiến thức chuyên môn còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ. Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài báo cáo khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, chúng em rất mong nhận được sự góp ý, chỉ bảo thêm của cô nhằm hoàn thiện những kiến thức của mình để nhóm chúng em có thể dùng làm hành trang thực hiện tiếp các đề tài khác trong tương lai cũng như là trong học tập và làm việc sau này.

Một lần nữa, nhóm xin gửi đến cô, bạn bè lời cảm ơn đặc biệt chân thành và tốt đẹp nhất!

Thành phố Hồ Chí Minh, tháng 7 năm 2024

Nhóm sinh viên thực hiện

Trần Thị Kim Anh

Lê Minh Chánh

NHẬN XÉT CỦA GIẢNG VIÊN

MỤC LỤC

LỜI CẢM ƠN	2
NHẬN XÉT CỦA GIẢNG VIÊN	3
DANH MỤC BẢNG & HÌNH ẢNH.....	8
Phần 1. GIỚI THIỆU TỔNG QUAN.....	13
1.1. Phát biểu về dữ liệu	13
1.1.1. Mô tả dataset.....	13
1.1.2. Thuộc tính kho dữ liệu	13
1.1.3. Mô tả chi tiết thuộc tính.....	14
1.1.4. Kho dữ liệu đã thực hiện tiền xử lý	17
1.2. Xây dựng kho dữ liệu.....	18
1.2.1. Sơ đồ dữ liệu xây dựng.....	18
1.2.2. Lược đồ kho dữ liệu	20
1.2.3. Mô tả chi tiết các bảng dữ liệu	20
Phần 2. QUÁ TRÌNH SSIS	24
2.1. Chuẩn bị công vụ và Data warehouse	24
2.2. Tạo database trong Microsoft SQL Server Management Studio	24
2.3. Quá trình SSIS.....	25
2.3.1. Khởi tạo Project.....	25
2.3.2. Khởi tạo Data Flow Task	26
2.3.3. Import Data.....	27
2.3.4. Tạo các bảng Dim	28
2.3.5. Tạo bảng Fact	32
2.3.6. Import dữ liệu vào database.....	33

Phần 3. PHÂN TÍCH BỘ DỮ LIỆU (QUÁ TRÌNH SSAS)	36
3.1. Tạo Project SSAS trong Vistual Studio	36
3.2. Connect đến Data Source	38
3.3. Tạo Data Source View	41
3.4. Tạo Cube và Dimensions	43
3.5. Thao tác trên các bảng Dim	49
3.5.1. Bảng Dim_Time	49
3.5.2. Bảng Dim_AgeGroup	51
3.5.3. Bảng Dim_Country và Dim_State	51
3.5.4. Bảng Dim_Category và Dim_SubCategory	52
3.5.5. Bảng Dim_Product	52
3.5.6. Bảng Dim_Gender	53
3.6. Deploy và Proces project	53
3.7. Thực hiện các câu truy vấn	55
3.7.1. Thống kê tổng số sản phẩm thuộc thể loại “Bikes”	57
3.7.2. Thống kê số lượng bán được của từng sản phẩm qua các năm ..	59
3.7.3. Thống kê doanh thu trên từng quốc gia của khách hàng trong tháng 8 năm 2015	61
3.7.4. Thống kê tổng số lượt bán qua từng năm của từng nhóm độ tuổi khách hàng	63
3.7.5. Thống kê top 5 bang thuộc nước Pháp có nhiều lượt mua từ khách hàng nhất	64
3.7.6. Thống kê danh sách các sản phẩm có trên 1000 lượt bán và tổng số lượng bán trên 50000 sản phẩm	66

3.7.7.	Thống kê danh sách và số lượng lượt bán của các sản phẩm trừ Water Bottle – 30oz. của các khách hàng là Nữ vào ngày 5/5/2015	68
3.7.8.	Thống kê top 3 danh mục phụ có lợi nhuận lớn nhất.	71
3.7.9.	Thống kê 3 tháng trong mỗi năm có nhiều lượt mua hàng nhất, tính tổng doanh thu của các tháng.	73
3.7.10.	Thống kê số lượng sản phẩm trên từng nhóm tuổi và giới tính..	75
Phần 4.	REPORT BẰNG CÔNG CỤ POWER BI (SSRS)	78
4.1.	Đỗ dữ liệu vào Project Power BI	78
4.2.	Thực hiện SSRS.....	80
4.2.1.	Thống kê tổng số sản phẩm thuộc thể loại “Bikes”	80
4.2.2.	Thống kê số lượng bán được của từng sản phẩm qua các năm ..	80
4.2.3.	Thống kê doanh thu trên từng quốc gia của khách hàng trong tháng 8 năm 2015	81
4.2.4.	Thống kê tổng số lượt bán qua từng năm của từng nhóm độ tuổi khách hàng	81
4.2.5.	Thống kê top 5 bang thuộc nước Pháp có nhiều lượt mua từ khách hàng nhất	82
4.2.6.	Thống kê danh sách các sản phẩm có trên 1000 lượt bán và tổng số lượng bán trên 50000 sản phẩm	82
4.2.7.	Thống kê danh sách và số lượng lượt bán của các sản phẩm trừ Water Bottle – 30oz. của các khách hàng là Nữ vào ngày 5/5/2015	83
4.2.8.	Thống kê top 3 danh mục phụ có lợi nhuận lớn nhất.	83
4.2.9.	Thống kê 3 tháng trong mỗi năm có nhiều lượt mua hàng nhất, tính tổng doanh thu của các tháng.	83
4.2.10.	Thống kê số lượng sản phẩm trên từng nhóm tuổi và giới tính..	84

Phần 5. DATA MINING	85
5.1. Giới thiệu.....	85
5.2. Các bước thực hiện	86
5.2.1. Tiền xử lý.....	86
5.2.2. Phân chia dữ liệu	86
5.2.3. Xây dựng thuật toán.....	87
5.2.3.1. ARIMA.....	87
5.2.3.2. Linear Regression.....	90
5.2.3.3. Holt-Winters	90
5.2.3.4. Random Forest	91
5.2.3.5. Long Short-Term Memory	93
5.2.4. Dự đoán 30 ngày kế tiếp.....	95
5.2.5. Vẽ đồ thị.....	96
5.2.5.1. ARIMA	96
5.2.5.2. Holt-Winters	97
5.2.5.3. Random Forest	98
5.2.5.4. LSTM	99
5.2.5.5. Linear Regression.....	100
5.3. Kết quả thực hiện	101
PHÂN CÔNG CÔNG VIỆC	102

DANH MỤC BẢNG & HÌNH ẢNH

<i>Bảng 1: Chi tiết các thuộc tính của dataset</i>	<i>13</i>
<i>Bảng 2: Mô tả chi tiết thuộc tính Age_Group.....</i>	<i>14</i>
<i>Bảng 3: Mô tả chi tiết thuộc tính Product_Category.....</i>	<i>15</i>
<i>Bảng 4: Mô tả chi tiết thuộc tính Sub_Category</i>	<i>15</i>
<i>Bảng 5: Mô tả chi tiết bảng Fact</i>	<i>20</i>
<i>Bảng 6: Mô tả chi tiết bảng Dim_Category.....</i>	<i>21</i>
<i>Bảng 7: Mô tả chi tiết bảng Dim_SubCategory.....</i>	<i>21</i>
<i>Bảng 8: Mô tả chi tiết bảng Dim_Product.....</i>	<i>22</i>
<i>Bảng 9: Mô tả chi tiết bảng Dim_Country</i>	<i>22</i>
<i>Bảng 10: Mô trả chi tiết bảng Dim_State</i>	<i>22</i>
<i>Bảng 11: Mô tả chi tiết bảng Dim_Gender.....</i>	<i>22</i>
<i>Bảng 12: Mô tả chi tiết bảng Dim_AgeGroup</i>	<i>23</i>
<i>Bảng 13: Mô tả chi tiết bảng Dim_Time.....</i>	<i>23</i>
<i>Bảng 14: Độ đo các model trong Data Mining</i>	<i>101</i>
<i>Hình 1.1: Dữ liệu sau khi được tiền xử lí.....</i>	<i>18</i>
<i>Hình 1.2: Lược đồ kho dữ liệu bông tuyết</i>	<i>20</i>
<i>Hình 2.1: Kết nối Database Engine trong DBMS.....</i>	<i>24</i>
<i>Hình 2.2: Tạo database mới trong SQL Server</i>	<i>25</i>
<i>Hình 2.3: Khởi tạo Integration Services Project trong Visual Studio</i>	<i>25</i>
<i>Hình 2.4: Đặt tên cho SSIS Project.....</i>	<i>26</i>
<i>Hình 2.5: Tạo Data Flow Task</i>	<i>26</i>
<i>Hình 2.6: Tạo Flat File Source để import dữ liệu</i>	<i>27</i>
<i>Hình 2.7: Import dữ liệu vào SSIS Project</i>	<i>27</i>
<i>Hình 2.8: Duyệt file dataset</i>	<i>28</i>
<i>Hình 2.9: Tạo Multicast và Sort.....</i>	<i>29</i>
<i>Hình 2.10: Cấu hình cho các Node Sort</i>	<i>29</i>

<i>Hình 2.11: Tạo OLE DB Destination.....</i>	30
<i>Hình 2.12: Câu lệnh cho OLE DB Destination</i>	30
<i>Hình 2.13: Thiết lập Server và Database để thực hiện SSIS</i>	31
<i>Hình 2.14: Tạo mới table cho database.....</i>	31
<i>Hình 2.15: Tạo thành công bảng Dim_Time</i>	32
<i>Hình 2.16: Kết quả sau khi tạo xong tất cả các bảng Dim</i>	32
<i>Hình 2.17: Tạo bảng Fact.....</i>	33
<i>Hình 2.18: Cấu hình Execute SQL Task</i>	34
<i>Hình 2.19: Điện SQL Statement.....</i>	34
<i>Hình 2.20: Kết quả sau khi thực hiện SSIS</i>	35
<i>Hình 3.1: Khởi tạo Analysis Services Multidimensional Project trong Visual Studio</i>	36
<i>Hình 3.2: Đặt tên cho SSAS Project</i>	37
<i>Hình 3.3: Kết quả sau khi khởi tạo SSAS Project</i>	37
<i>Hình 3.4: Khởi tạo một New Data Source</i>	38
<i>Hình 3.5: Chọn Create a data source based on an existing or new connection.....</i>	38
<i>Hình 3.6: Chọn Server và Database để thực hiện</i>	39
<i>Hình 3.7: Cung cấp tài khoản cho project</i>	39
<i>Hình 3.8: Kết quả khi tạo Data Source</i>	40
<i>Hình 3.9: Khởi tạo một New Data Source View</i>	41
<i>Hình 3.10: Chọn Data Source vừa tạo.....</i>	41
<i>Hình 3.11: Tích chọn tạo liên kết giữa các bảng</i>	42
<i>Hình 3.12: Chọn các bảng vào Source View</i>	42
<i>Hình 3.13: Kết quả sau khi tạo Data Source View</i>	43
<i>Hình 3.14: Tạo một New Cube.....</i>	43
<i>Hình 3.15: Tích chọn bảng đã được tạo</i>	44
<i>Hình 3.16: Chọn bảng chứa Measure.....</i>	45
<i>Hình 3.17: Chọn các Measure cho project</i>	46
<i>Hình 3.18: Chọn các Dimension cần thiết</i>	47

<i>Hình 3.19: Kiểm tra lại thông tin của Cube</i>	48
<i>Hình 3.20: Cube sau khi tạo thành công</i>	49
<i>Hình 3.21: Bên trong bảng Dim.....</i>	50
<i>Hình 3.22: Kéo các thuộc tính từ Data Source View sang Attributes</i>	50
<i>Hình 3.23: Bảng Dim_AgeGroup</i>	51
<i>Hình 3.24: Bảng Dim_Country và Dim_State</i>	51
<i>Hình 3.25: Bảng Dim_Category và Dim_SubCategory.....</i>	52
<i>Hình 3.26: Bảng Dim_Product</i>	52
<i>Hình 3.27: Bảng Dim_Gender</i>	53
<i>Hình 3.28: Bắt đầu thực thi project</i>	53
<i>Hình 3.29: Chạy cube</i>	54
<i>Hình 3.30: Kết quả sau khi deploy cube thành công</i>	55
<i>Hình 3.31: Cách thực hiện các câu query trong Cube</i>	55
<i>Hình 3.32: Cách để thực hiện các câu query bằng Pivot trong MS Excel</i>	56
<i>Hình 3.33: Cách để thực hiện các câu query bằng MDX trong SQL Server.....</i>	56
<i>Hình 3.34: Câu truy vấn 1 trong Visual Studio</i>	57
<i>Hình 3.35: Câu truy vấn 1 trong SQL Server</i>	58
<i>Hình 3.36: Câu truy vấn 1 trong MS Excel.....</i>	58
<i>Hình 3.37: Câu truy vấn 2 trong Visual Studio</i>	59
<i>Hình 3.38: Câu truy vấn 2 trong SQL Server</i>	60
<i>Hình 3.39: Câu truy vấn 2 trong MS Excel.....</i>	60
<i>Hình 3.40: Câu truy vấn 3 trong Visual Studio</i>	61
<i>Hình 3.41: Câu truy vấn 3 trong SQL Server</i>	62
<i>Hình 3.42: Câu truy vấn 3 trong MS Excel.....</i>	62
<i>Hình 3.43: Câu truy vấn 4 trong Visual Studio</i>	63
<i>Hình 3.44: Câu truy vấn 4 trong SQL Server</i>	64
<i>Hình 3.45: Câu truy vấn 4 trong MS Excel.....</i>	64
<i>Hình 3.46: Câu truy vấn 5 trong Visual Studio</i>	65
<i>Hình 3.47: Câu truy vấn 5 trong SQL Server</i>	66

<i>Hình 3.48: Câu truy vấn 5 trong MS Excel.....</i>	66
<i>Hình 3.49: Câu truy vấn 6 trong Visual Studio</i>	67
<i>Hình 3.50: Câu truy vấn 6 trong SQL Server</i>	68
<i>Hình 3.51: Câu truy vấn 6 trong MS Excel.....</i>	68
<i>Hình 3.52: Câu truy vấn 7 trong SQL Server</i>	69
<i>Hình 3.53: Câu truy vấn 7 trong SQL Server</i>	70
<i>Hình 3.54: Câu truy vấn 7 trong MS Excel.....</i>	71
<i>Hình 3.55: Câu truy vấn 8 trong Visual Studio</i>	71
<i>Hình 3.56: Câu truy vấn 8 trong SQL Server</i>	72
<i>Hình 3.57: Câu truy vấn 8 trong MS Excel.....</i>	73
<i>Hình 3.58: Câu truy vấn 9 trong Visua Studio</i>	74
<i>Hình 3.59: Câu truy vấn 9 trong SQL Server</i>	75
<i>Hình 3.60: Câu truy vấn 9 trong MS Excel.....</i>	75
<i>Hình 3.61: Câu truy vấn 9 trong Visual Studio</i>	76
<i>Hình 3.62: Câu truy vấn 9 trong SQL Server</i>	77
<i>Hình 3.63: Câu truy vấn 9 trong MS Excel.....</i>	77
<i>Hình 4.1: Liên kết Power BI đến Analysis Service</i>	78
<i>Hình 4.2: Điện thông tin Server và Database</i>	78
<i>Hình 4.3: Chọn Cube</i>	79
<i>Hình 4.4: Giao diện tạo Report trên Power BI.....</i>	79
<i>Hình 4.5: Câu truy vấn 1 trong Power BI.....</i>	80
<i>Hình 4.6: Câu truy vấn 2 trong Power BI.....</i>	80
<i>Hình 4.7: Câu truy vấn 3 trong Power BI.....</i>	81
<i>Hình 4.8: Câu truy vấn 4 trong Power BI.....</i>	81
<i>Hình 4.9: Câu truy vấn 5 trong Power BI.....</i>	82
<i>Hình 4.10: Câu truy vấn 6 trong Power BI.....</i>	82
<i>Hình 4.11: Câu truy vấn 7 trong Power BI.....</i>	83
<i>Hình 4.12: Câu truy vấn 7 trong Power BI.....</i>	83
<i>Hình 4.13: Câu truy vấn 10 trong Power BI.....</i>	84

<i>Hình 5.1: Tiết xử lý dữ liệu cho Data Mining</i>	86
<i>Hình 5.2: Chia tỉ lệ dữ liệu cho model ARIMA</i>	87
<i>Hình 5.3: Kiểm tra chuỗi dùng và xử lý cho ARIMA Model</i>	89
<i>Hình 5.4: Model autoARIMA</i>	90
<i>Hình 5.5: Dự báo tương lai bằng ARIMA Model</i>	90
<i>Hình 5.6: Chia tỉ lệ dữ li cho model Linear Regression</i>	90
<i>Hình 5.7: Chia tỉ lệ dữ liệu cho model Holt-Winters</i>	91
<i>Hình 5.8: Xây dựng model Holt-Winters</i>	91
<i>Hình 5.9: Chia tỉ lệ dữ liệu cho model Random Forest</i>	91
<i>Hình 5.10: Tạo dữ liệu đầu vào để train model</i>	92
<i>Hình 5.11: Xây dựng model Random Forest để dự báo tương lai</i>	92
<i>Hình 5.12: Chia tỉ lệ dữ liệu và tạo đầu vào cho model LSTM</i>	93
<i>Hình 5.13: Xây dựng model LSTM để dự báo</i>	94
<i>Hình 5.14: Dự báo bằng model LSTM</i>	94
<i>Hình 5.15: Dự báo 30 kế tiếp</i>	95
<i>Hình 5.16: Đò thi model ARIMA</i>	96
<i>Hình 5.17: Đò thi model Holt-Winters</i>	97
<i>Hình 5.18: Đò thi model Random Forest</i>	98
<i>Hình 5.19: Đò thi model LSTM</i>	99
<i>Hình 5.20: Đò thi model Linear Regression</i>	100

Phần 1. GIỚI THIỆU TỔNG QUAN

1.1. Phát biểu về dữ liệu

1.1.1. Mô tả dataset

Bộ dữ liệu Europe Bike Store Sales là bộ dữ liệu doanh số bán hàng xe đạp của Châu Âu được công bố trên trang Kaggle.com

Thông qua dữ liệu, người dùng có thể biết được ngày bán sản phẩm, độ tuổi và nhóm độ tuổi của khách hàng mua hàng cũng như là bang và quốc gia của khách hàng, thông tin, giá sản phẩm, số lượng mua hàng, lợi nhuận và doanh thu...

Dataset có 113.036 dòng dữ liệu với 18 thuộc tính.

Nguồn dữ liệu: [Europe Bike Store Sales \(kaggle.com\)](https://www.kaggle.com/datasets/nikhilnayak/europe-bike-store-sales)

1.1.2. Thuộc tính kho dữ liệu

Bảng 1: Chi tiết các thuộc tính của dataset

STT	Thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	Date	DATE	Ngày mua hàng
2	Day	INT	Giá trị ngày trong ngày mua hàng
3	Month	VARCHAR(50)	Giá trị tháng trong ngày mua hàng
4	Year	INT	Giá trị năm trong ngày mua hàng
5	Customer_Age	INT	Tuổi khách hàng
6	Age_Group	VARCHAR(50)	Phân nhóm tuổi khách hàng
7	Customer_Gender	VARCHAR(50)	Giới tính khách hàng
8	Country	VARCHAR(50)	Quốc gia của khách hàng

9	State	VARCHAR(50)	Tiểu bang của khách hàng
10	Product_Category	VARCHAR(50)	Danh mục sản phẩm
11	Sub_Category	VARCHAR(50)	Danh mục con
12	Product	VARCHAR(50)	Tên sản phẩm
13	Order_Quantity	INT	Số lượng mua hàng
14	Unit_Cost	INT	Giá của từng sản phẩm
15	Unit_Price	INT	
16	Profit	INT	Lợi nhuận bán sản phẩm
17	Cost	INT	Chi phí bán sản phẩm $Cost = Unit_Cost * Order_Quantity$
18	Revenue	INT	Doanh thu bán sản phẩm $Revenue = Profit + Cost$

1.1.3. Mô tả chi tiết thuộc tính

Dictionary thuộc tính Age_Group

Thuộc tính Age_Group thể hiện nhóm tuổi của khách hàng

Bảng 2: Mô tả chi tiết thuộc tính Age_Group

Age_Group		
STT	Giá trị	Ý nghĩa
1	Youth (<25)	Nhóm tuổi thiếu niên (nhỏ hơn 25)
2	Young Adults (25 - 34)	Nhóm tuổi thanh niên (từ 25 đến 34)

3	Adults (34 - 64)	Nhóm tuổi người lớn (từ 34 đến 64)
4	Seniors (64+)	Nhóm người lớn tuổi (trên 64)

Dictionary thuộc tính **Product_Category**

Thuộc tính **Product_Category** thể hiện danh mục sản phẩm chính

Bảng 3: Mô tả chi tiết thuộc tính *Product_Category*

Product_Category		
STT	Giá trị	Ý nghĩa
1	Bikes	Danh mục sản phẩm liên quan đến các loại xe đạp
2	Accessories	Danh mục sản phẩm liên quan đến các phụ kiện đi kèm
3	Clothing	Danh mục sản phẩm liên quan đến quần áo

Dictionary thuộc tính **Sub_Category**

Thuộc tính **Sub_Category** thể hiện danh mục phụ

Bảng 4: Mô tả chi tiết thuộc tính *Sub_Category*

Sub_Category		
STT	Giá trị	Ý nghĩa
1	Mountain Bikes	Xe đạp leo núi, thiết kế để chạy trên địa hình gò ghề Thuộc Product_Category: Bikes
2	Road Bikes	Xe đạp đường phố, thiết kế để chạy trên đường nhựa và đường phẳng. Thuộc Product_Category: Bikes

3	Bike Stands	Giá đỡ xe đạp, dùng để giữ xe đạp đứng yên khi không sử dụng. Thuộc Product_Category: Accessories
4	Bottles and Cages	Bình nước và giá đỡ bình, phụ kiện dùng để mang theo nước uống khi đạp xe. Thuộc Product_Category: Accessories
5	Helmets	Mũ bảo hiểm, dùng để bảo vệ đầu khi đạp xe. Thuộc Product_Category: Accessories
6	Jerseys	Áo thun thể thao dành cho việc đạp xe. Thuộc Product_Category: Clothing
7	Tires and Tubes	Lốp và ruột xe đạp, các thành phần thay thế khi lốp hoặc ruột xe bị hỏng. Thuộc Product_Category: Accessories
8	Touring Bikes	Xe đạp du lịch, thiết kế để chạy trên đường dài và chở theo hành lý. Thuộc Product_Category: Bikes
9	Caps	Mũ lưỡi trai, phụ kiện thời trang. Thuộc Product_Category: Clothing
10	Fenders	Vè chắn bùn, phụ kiện gắn vào xe đạp để chắn bùn và nước bắn lên. Thuộc Product_Category: Accessories

11	Gloves	Găng tay, giúp bảo vệ tay và cải thiện độ bám khi đạp xe. Thuộc Product_Category: Clothing
12	Socks	Tất, dùng để mang khi đạp xe. Thuộc Product_Category: Clothing
13	Hydration Packs	Ba lô nước, hệ thống cung cấp nước uống tiện lợi khi đạp xe. Thuộc Product_Category: Accessories
14	Vests	Áo gi-lê, thường dùng để giữ ấm hoặc cải thiện độ an toàn với chất liệu phản quang. Thuộc Product_Category: Clothing
15	Cleaners	Dụng cụ và chất tẩy rửa dùng để vệ sinh xe đạp. Thuộc Product_Category: Accessories
16	Bike Racks	Giá để xe đạp, dùng để chở hoặc lưu trữ xe đạp. Thuộc Product_Category: Accessories
17	Shorts	Quần ngắn, thường có đệm lót để tăng sự thoải mái khi đạp xe. Thuộc Product_Category: Clothing

1.1.4. Kho dữ liệu đã thực hiện tiền xử lý

- Nhóm đã thực hiện các bước tiền xử lý.
- Sau khi tiền xử lí, ta được 113.036 dòng và 17 cột dữ liệu cho việc phân tích đề tài.

The screenshot shows a Microsoft Excel spreadsheet with the title "EuropeBikeSales". The data is organized into 18 columns labeled A through Q. The columns represent various dimensions and facts such as Date, Day, Month, Year, Age Group, Customer Gender, Country, State, Product Category, Sub Category, Product, Order Quantity, Unit Cost, Unit Price, Profit, Cost, and Revenue. The data covers a period from January 1, 2011, to July 1, 2011, with entries for countries like Australia, Canada, United States, and the United Kingdom, and product types like Mountain Bikes and Road Bikes.

Hình 1.1: Dữ liệu sau khi được tiền xử lí

1.2. Xây dựng kho dữ liệu

1.2.1. Sơ đồ dữ liệu xây dựng

Các lược đồ dữ liệu đa chiều:

- Lược đồ đa chiều được thiết kế đặc biệt để mô hình hoá các hệ thống kho dữ liệu. Các lược đồ được thiết kế để giải quyết các nhu cầu của cơ sở dữ liệu rất lớn được thiết kế cho mục đích phân tích dữ liệu (OLAP).
- Có 2 loại lược đồ đa chiều:
 - a) **Lược đồ hình sao (Star Schema)**

Lược đồ hình sao là loại được đồ kho dữ liệu đơn giản nhất. Nó được gọi là được đồ hình sao vì cấu trúc giống như một ngôi sao. Trong lược đồ hình sao, tâm của ngôi sao có thể có một bảng sự kiện (Fact) và có các bảng chiều (Dimensions) được liên kết. Nó còn được gọi là Star Join Schema và được tối ưu hóa để truy vấn các tập dữ liệu lớn.

Ưu điểm:

- o Các truy vấn đơn giản hơn – login nối lược đồ sao thường đơn giản hơn logic nối được yêu cầu truy xuất dữ liệu từ lược đồ giao dịch được chuẩn hoá cao.

- Tăng hiệu suất truy vấn – các lược đồ sao có thể cung cấp các cải tiến hiệu suất cho các ứng dụng báo cáo chỉ đọc khi so sánh với các lược đồ được chuẩn hóa cao.
- Tổng hợp nhanh – các truy vấn đơn giản hơn đối với lược đồ sao có thể dẫn đến hiệu suất được cải thiện cho các hoạt động tổng hợp.

Nhược điểm:

- Tính toàn vẹn dữ liệu không được thực thi tốt vì nó không ở trạng thái chuẩn hóa cao.
- Lược đồ hình sao không linh hoạt về mặt nhu cầu phân tích như mô hình dữ liệu chuẩn hóa.

b) Lược đồ bông tuyết (Snowflake Schema)

Lược đồ Bông tuyết là một phần mở rộng của Lược đồ hình sao và nó thêm các kích thước bổ sung. Nó được gọi là bông tuyết vì sơ đồ của nó giống như một Bông tuyết.

Ưu điểm:

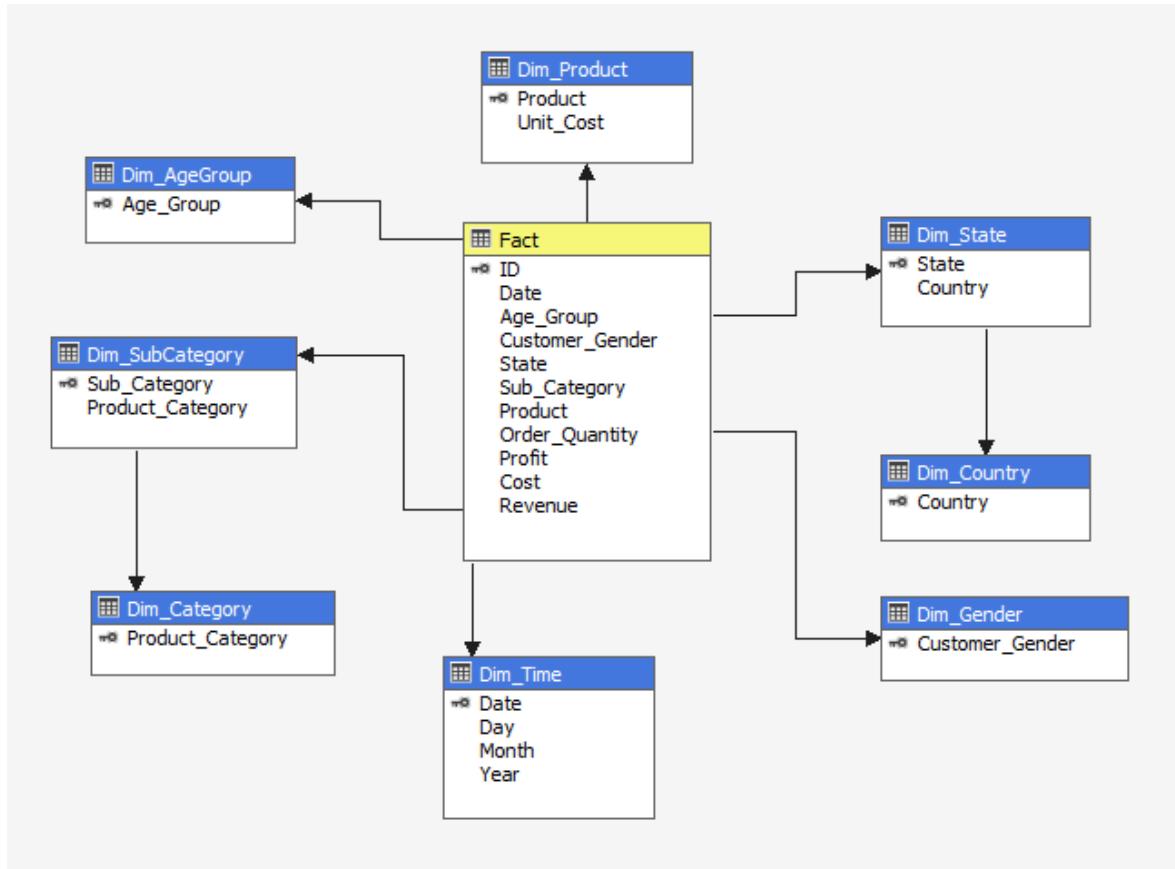
Lược đồ bông tuyết nằm trong cùng một họ với mô hình logic hình sao. Lược đồ hình sao được coi là một trường hợp đặc biệt của lược đồ bông tuyết. Và có những ưu điểm hơn so với lược đồ sao bao gồm:

- Một số công cụ mô hình hóa cơ sở dữ liệu đa chiều (OLAP) được tối ưu hóa cho các lược đồ bông tuyết.
- Đơn giản hóa các thuộc tính dẫn đến sự tiết kiệm, nhưng đánh đổi là sự phức tạp bổ sung trong các truy vấn nguồn.
- Một số chiều được phân cấp để thể hiện rõ ràng dạng chuẩn của bảng chiều.

Nhược điểm:

Mức chuẩn hóa thuộc tính bổ sung thêm độ phức tạp cho các phép truy vấn nguồn so với lược đồ hình sao.

1.2.2. Lược đồ kho dữ liệu



Hình 1.2: Lược đồ kho dữ liệu bông tuyết

1.2.3. Mô tả chi tiết các bảng dữ liệu

❖ Bảng Fact

Bảng 5: Mô tả chi tiết bảng Fact

Fact		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
ID	INT	Khoa chính
Date	VARCHAR(50)	Khoa ngoại
Age_Group	VARCHAR(50)	Khoa ngoại
Customer_Gender	VARCHAR(50)	Khoa ngoại

State	VARCHAR(50)	Khoá ngoại
Sub_Category	VARCHAR(50)	Khoá ngoại
Product	VARCHAR(50)	Khoá ngoại
Order_Quantity	INT	
Profit	INT	
Cost	INT	
Revenue	INT	

❖ **Bảng Dim_Category**

Bảng 6: Mô tả chi tiết bảng Dim_Category

Dim_Category		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
Category	VARCHAR(50)	Khoá chính

❖ **Bảng Dim_SubCategory**

Bảng 7: Mô tả chi tiết bảng Dim_SubCategory

Dim_SubCategory		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
Sub_Category	VARCHAR(50)	Khoá chính
Category	VARCHAR(50)	Khoá ngoại

❖ **Bảng Dim_Product**

Bảng 8: Mô tả chi tiết bảng Dim_Product

Dim_Product		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
Product	VARCHAR(50)	Khoá chính
Unit_Cost	INT	

❖ **Bảng Dim_Country**

Bảng 9: Mô tả chi tiết bảng Dim_Country

Dim_Country		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
Country	VARCHAR(50)	Khoá chính

❖ **Bảng Dim_State**

Bảng 10: Mô tả chi tiết bảng Dim_State

Dim_State		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
State	VARCHAR(50)	Khoá chính
Country	VARCHAR(50)	Khoá ngoại

❖ **Bảng Dim_Gender**

Bảng 11: Mô tả chi tiết bảng Dim_Gender

Dim_Gender		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
Customer_Gender	VARCHAR(50)	Khoá chính

❖ **Bảng Dim_AgeGroup**

Bảng 12: Mô tả chi tiết bảng Dim_AgeGroup

Dim_AgeGroup		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
Age_Group	VARCHAR(50)	Khoá chính

❖ **Bảng Dim_Time**

Bảng 13: Mô tả chi tiết bảng Dim_Time

Dim_Time		
Tên thuộc tính	Kiểu dữ liệu	Ràng buộc
Date	VARCHAR(50)	Khoá chính
Day	INT	
Month	VARCHAR(50)	
Year	INT	

Phần 2. QUÁ TRÌNH SSIS

2.1. Chuẩn bị công vụ và Data warehouse

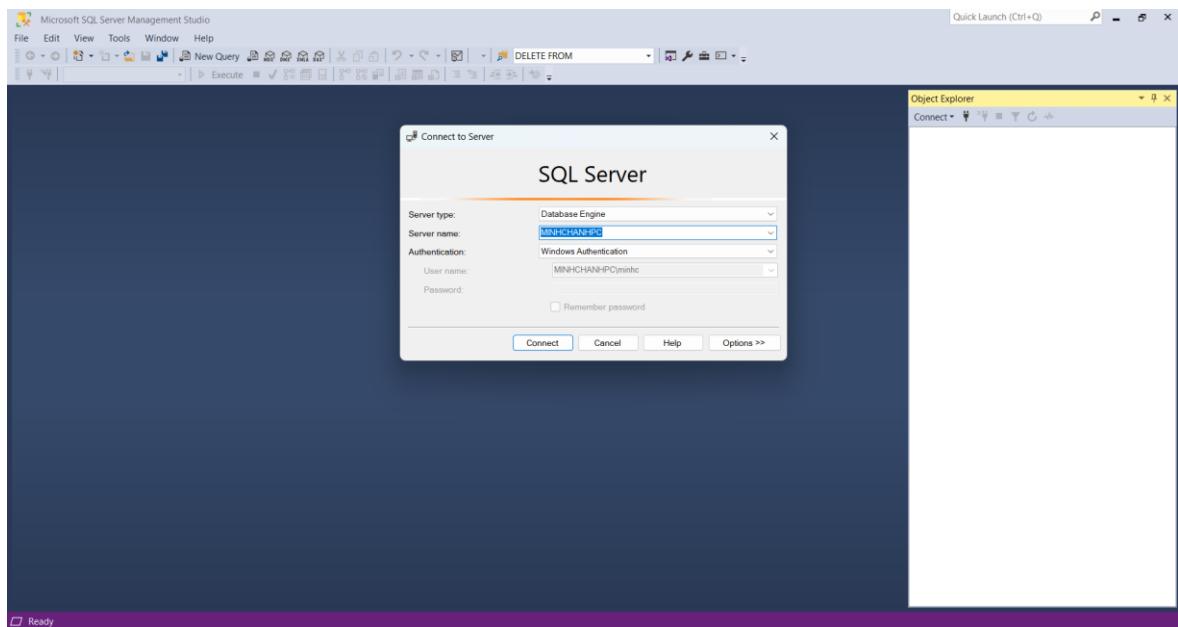
Tải Microsoft Visual Studio 2022.

Tải Microsoft SQL Server Management Studio.

Tải công cụ SQL Server Data Tools cho phiên bản Visual Studio 2022.

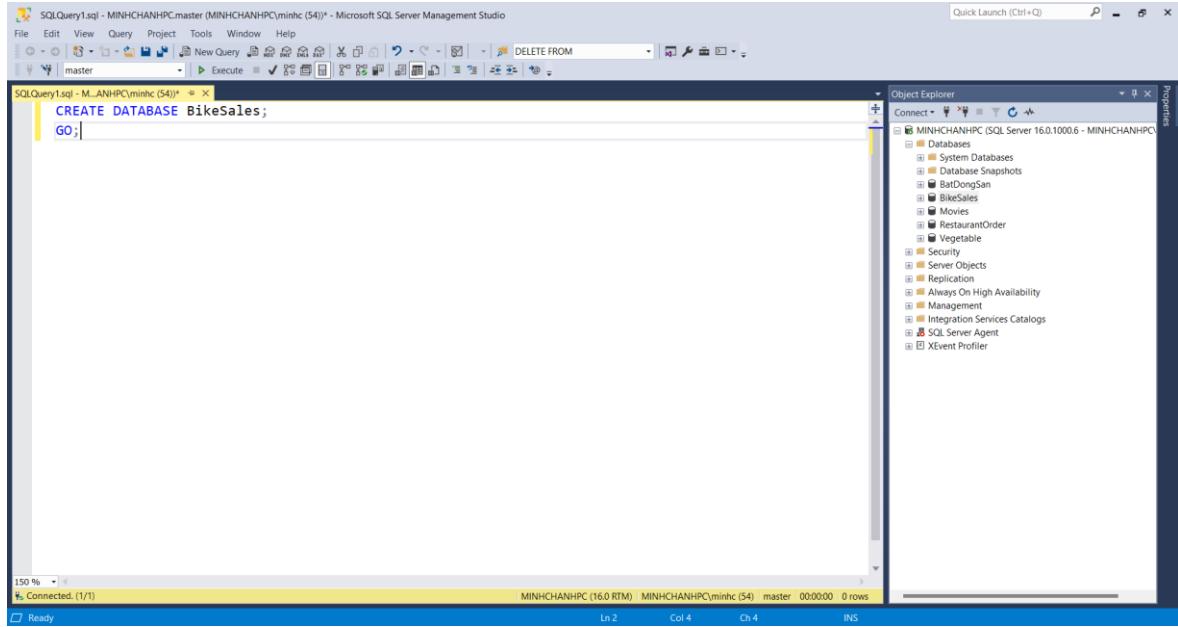
2.2. Tạo database trong Microsoft SQL Server Management Studio

Bước 1: Kết nối Database Engine



Hình 2.1: Kết nối Database Engine trong DBMS

Bước 2: Dùng lệnh tạo database trong cửa sổ Query

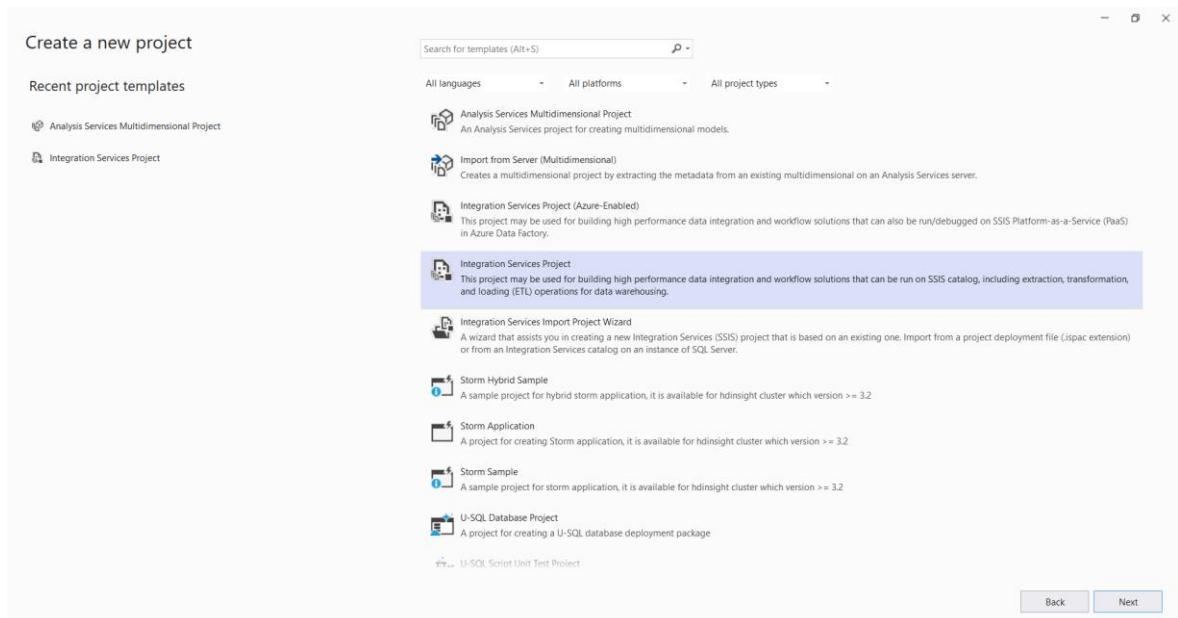


Hình 2.2: Tạo database mới trong SQL Server

2.3. Quá trình SSIS

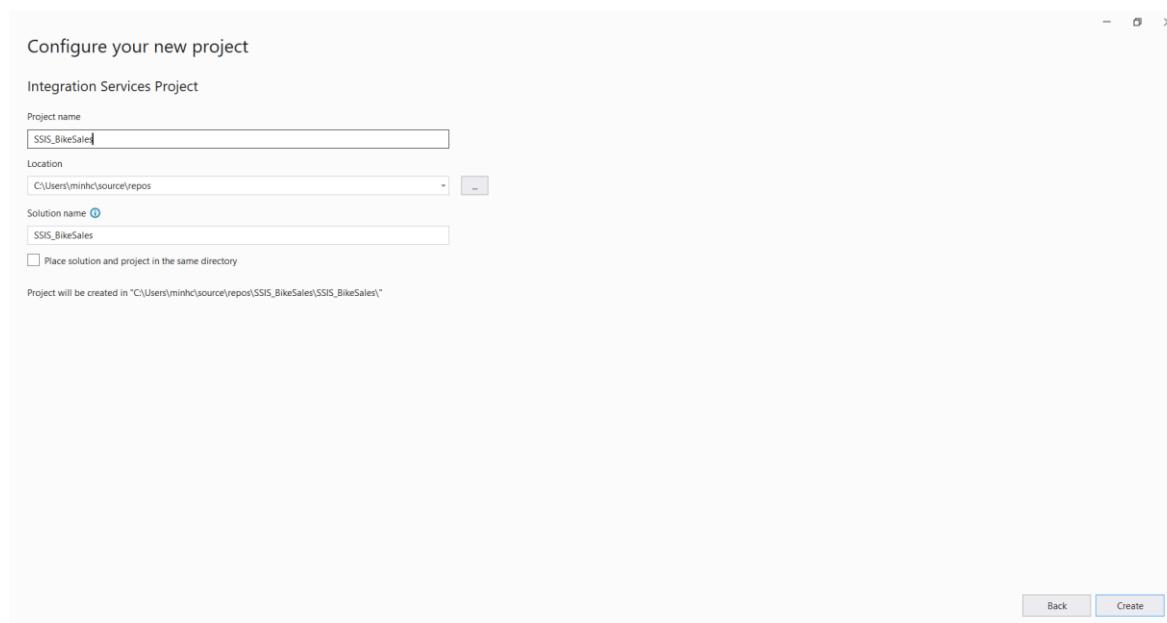
2.3.1. Khởi tạo Project

Bước 1: Tìm và nhấn chọn Integration Services Project sau đó nhấn Next



Hình 2.3: Khởi tạo Integration Services Project trong Visual Studio

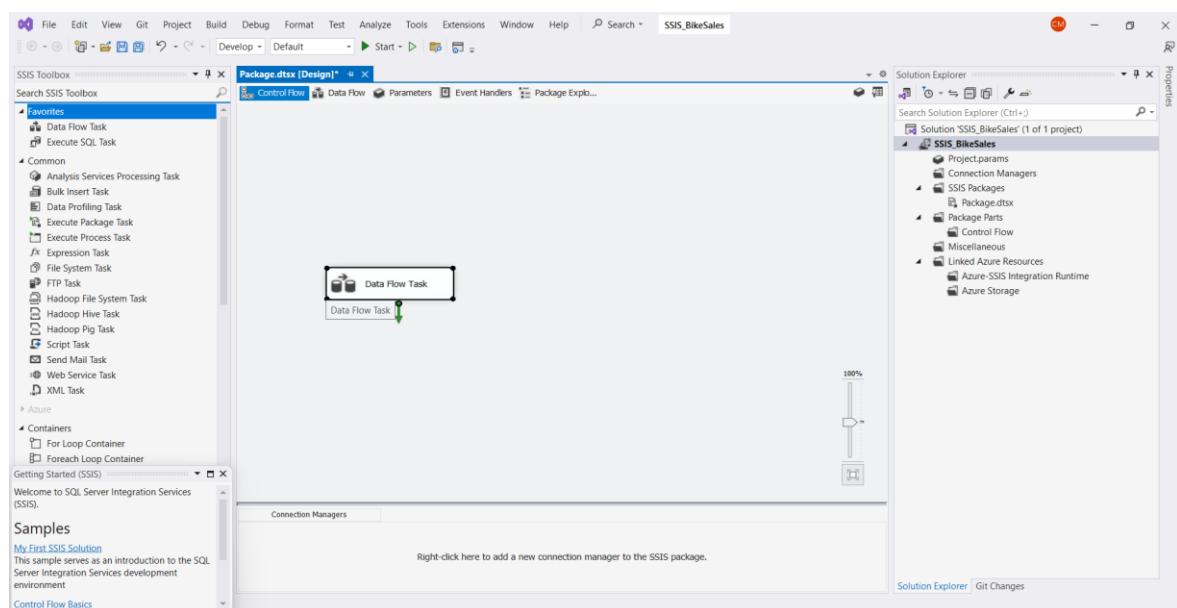
Bước 2: Đặt tên cho Project sau đó nhấn Create



Hình 2.4: Đặt tên cho SSIS Project

2.3.2. Khởi tạo Data Flow Task

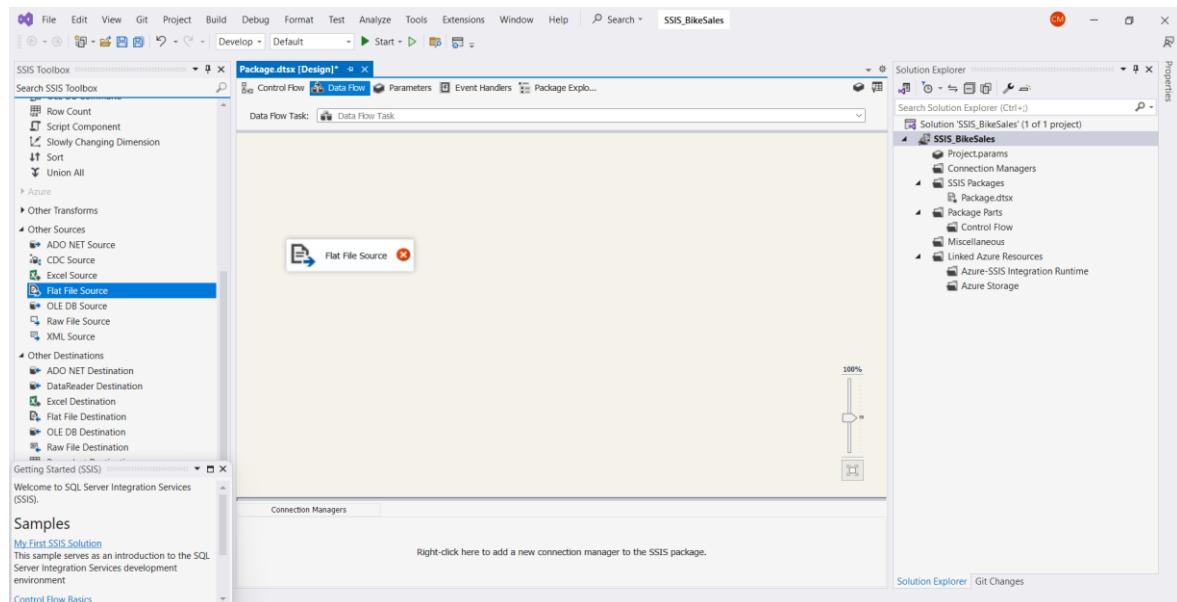
Tạo một **Data Flow Task** sau đó nhấn đúp vào để chuyển sang tab **Data Flow** để thực hiện cấu hình.



Hình 2.5: Tạo Data Flow Task

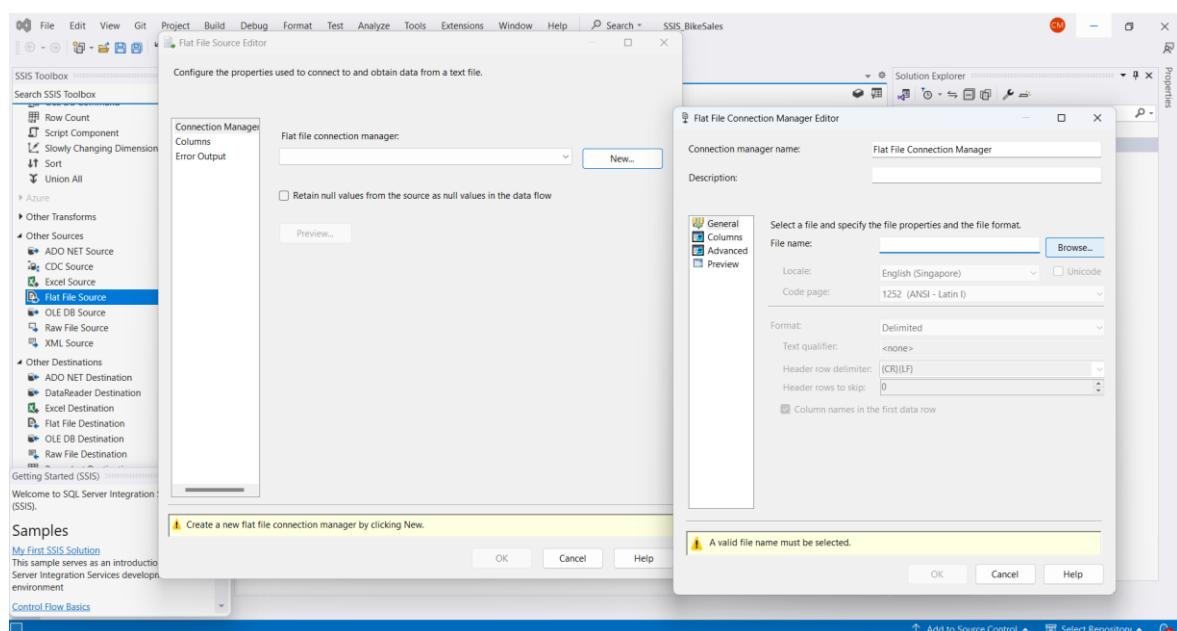
2.3.3. Import Data

Bước 1: Tại tab **Data Flow** thực hiện tìm và kéo ra một Node **Flat File Source** từ **SSIS Toolbox**



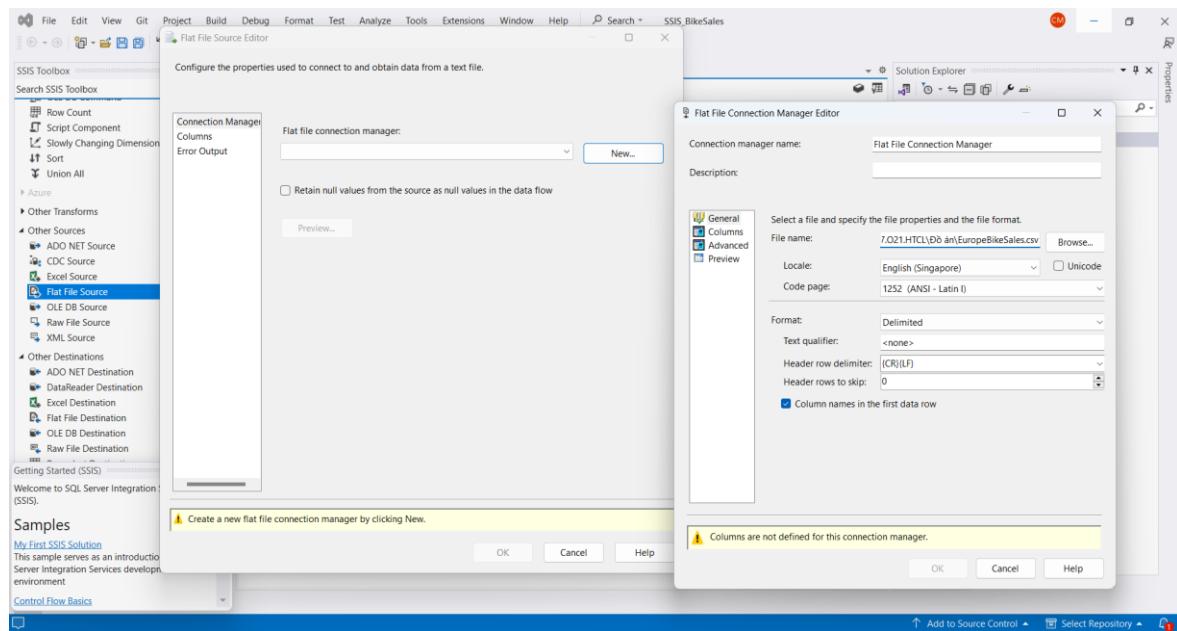
Hình 2.6: Tạo Flat File Source để import dữ liệu

Bước 2: Nhấn đúp vào Node **Flat File Source** để mở **Flat File Source Editor** tiến hành tạo mới kết nối. Nhấn **New**



Hình 2.7: Import dữ liệu vào SSIS Project

Bước 3: Nhấn Browse để duyệt đến file dữ liệu (EuropeBikeSales.csv)

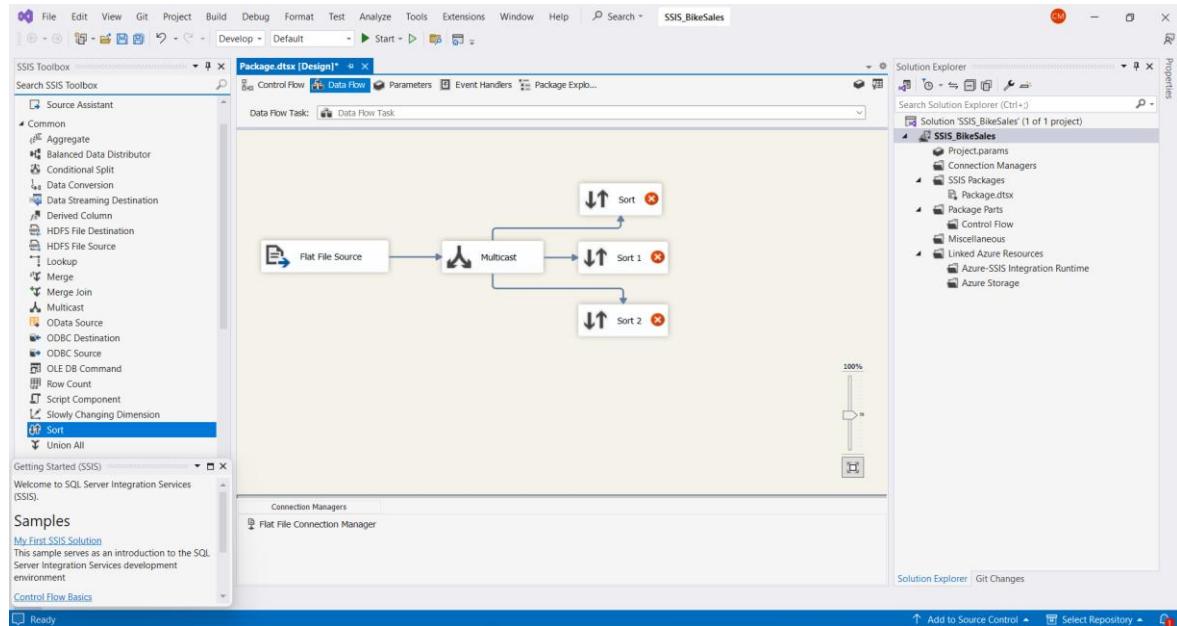


Hình 2.8: Duyệt file dataset

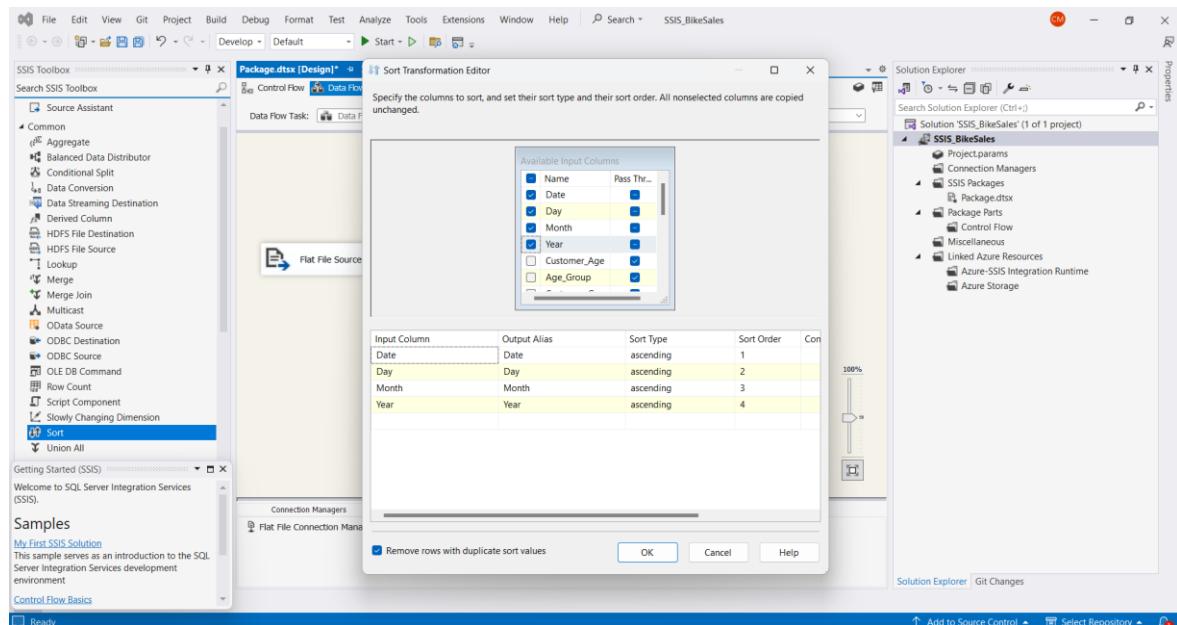
Tiến hành thay đổi các thông số (nếu cần) sao cho việc đọc file đầu vào là đúng yêu cầu nhất.

2.3.4. Tạo các bảng Dim

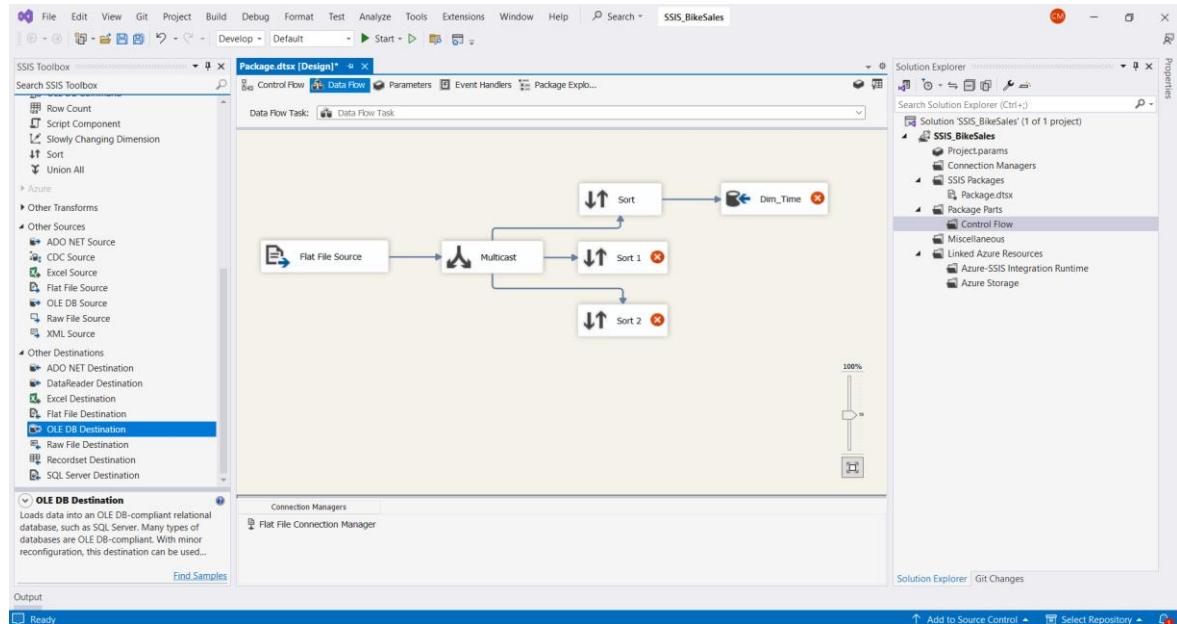
Bước 1: Tạo 1 Node **Muticast** và các Node **Sort** để sắp xếp dữ liệu cho các bảng Dim và thực hiện liên kết chúng với nhau

*Hình 2.9: Tạo Multicast và Sort*

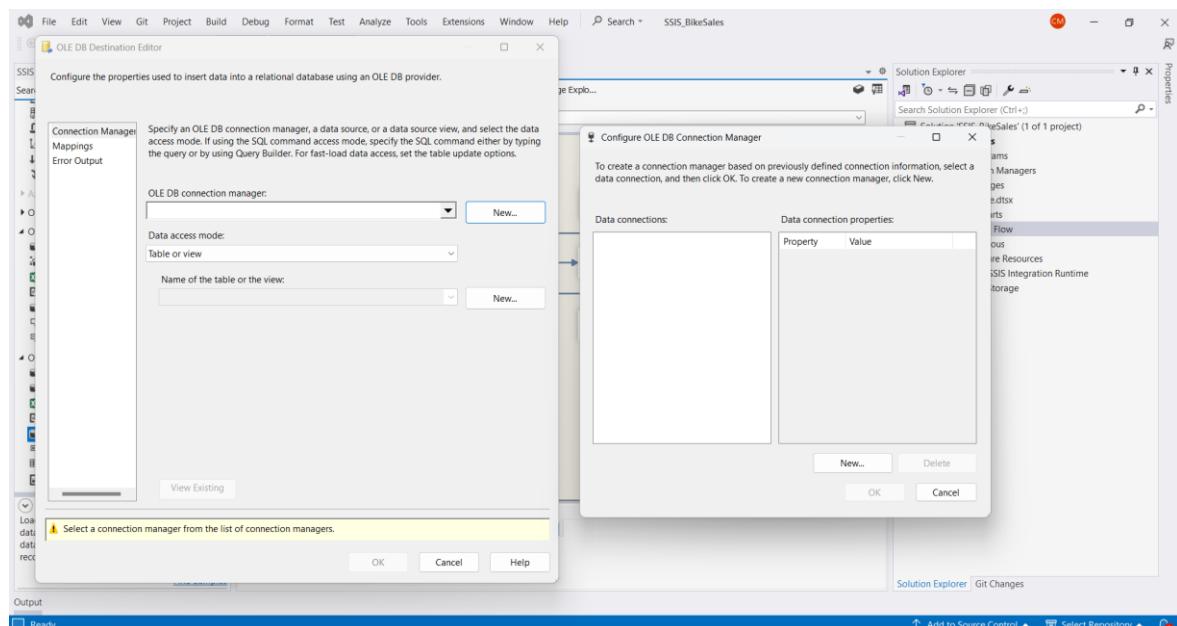
Bước 2: Cấu hình cho các Node **Sort** bằng cách nhấn đúp và chung. Tuỳ theo Sort cho bảng Dim nào mà sẽ tích vào các thuộc tính để đưa thuộc tính đó vào các bảng tương ứng và tích vào **Remove rows with duplicate sort values**.

*Hình 2.10: Cấu hình cho các Node Sort*

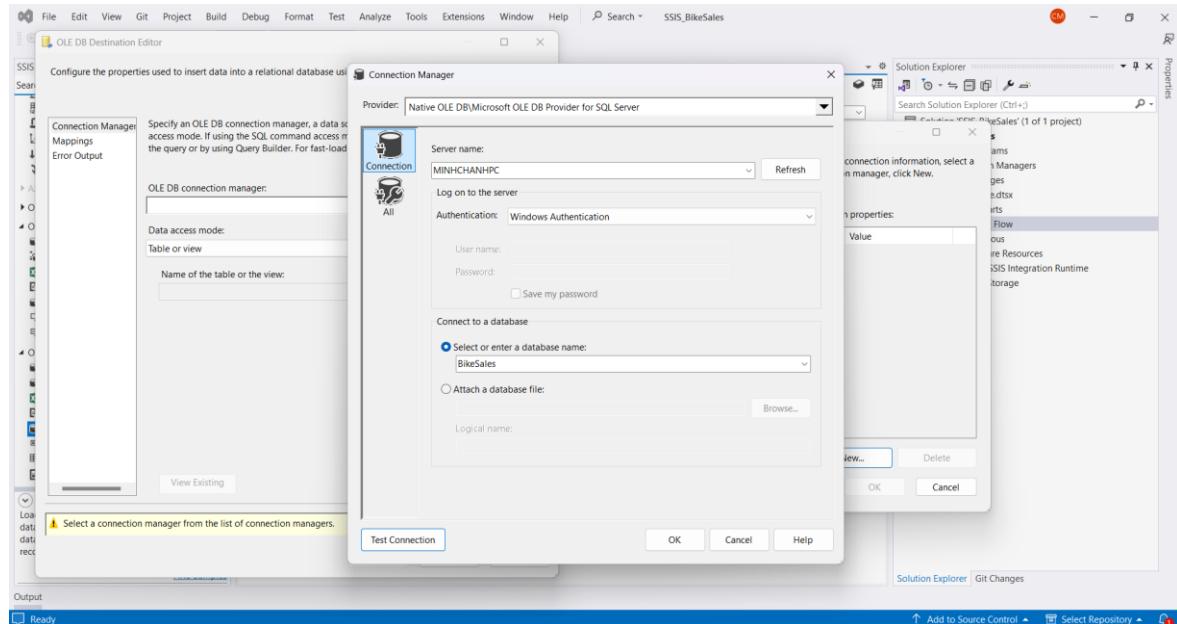
Bước 3: Tạo mới một Node OLE DB Destination từ SSIS Toolbox và thực hiện đổi tên Node sao cho trùng tên với tên bảng Dim cần tạo.

*Hình 2.11: Tạo OLE DB Destination*

Bước 4: Nhấn đúp vào Node để mở trình Editor và nhấn New để tạo mới một OLE DB Connect Manager → Cửa sổ Configure OLE DB Connection Manager mở ra sau đó nhấn New.

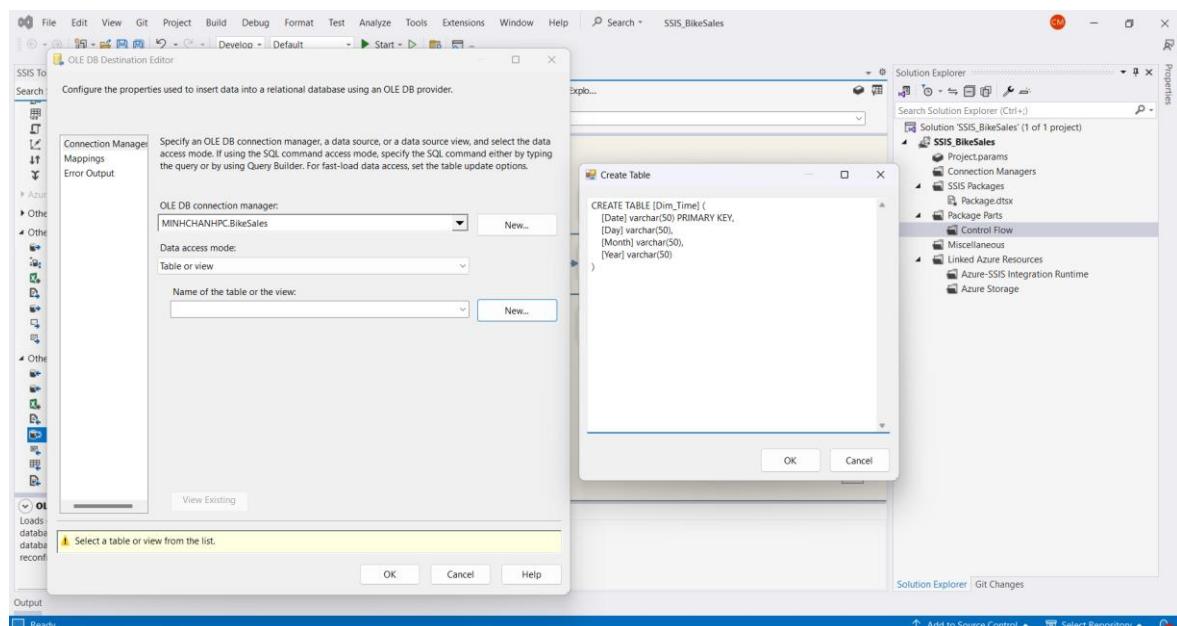
*Hình 2.12: Cấu hình cho OLE DB Destination*

Bước 5: Chọn đúng Provider, điền Server Name và chọn Database để kết nối và Test Connection để kiểm tra, sau đó nhấn OK để hoàn tất



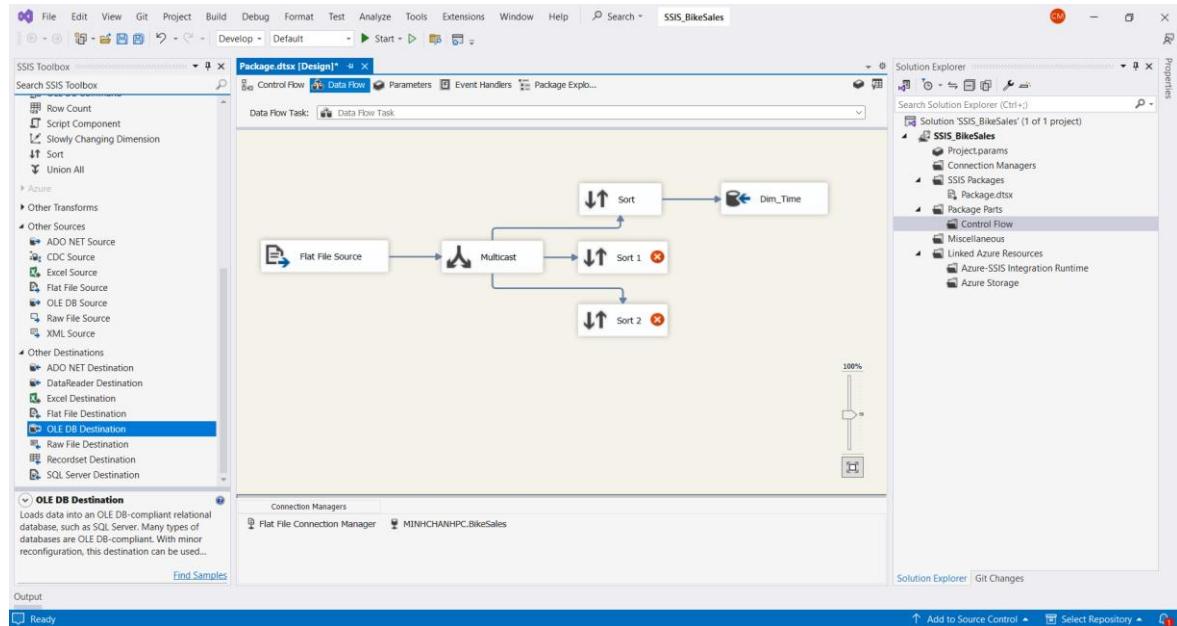
Hình 2.13: Thiết lập Server và Database để thực hiện SSIS

Bước 6: Sau khi có kết nối OLE DB Connection, tại giao diện cửa sổ OLE DB Destination Editor nhấn **New** để tạo mới Table. Sau đó điền Script để tạo table vào và nhấn **OK**.



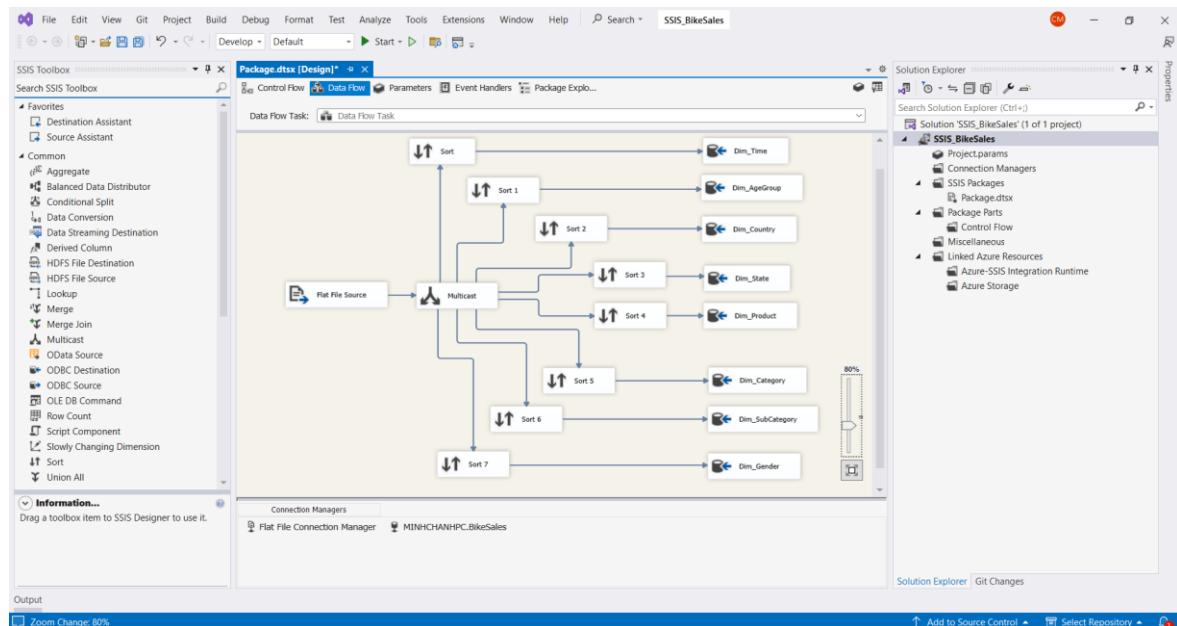
Hình 2.14: Tạo mới table cho database

Như vậy đã tạo thành luồng để tạo 1 bảng Dim.



Hình 2.15: Tạo thành công bảng Dim_Time

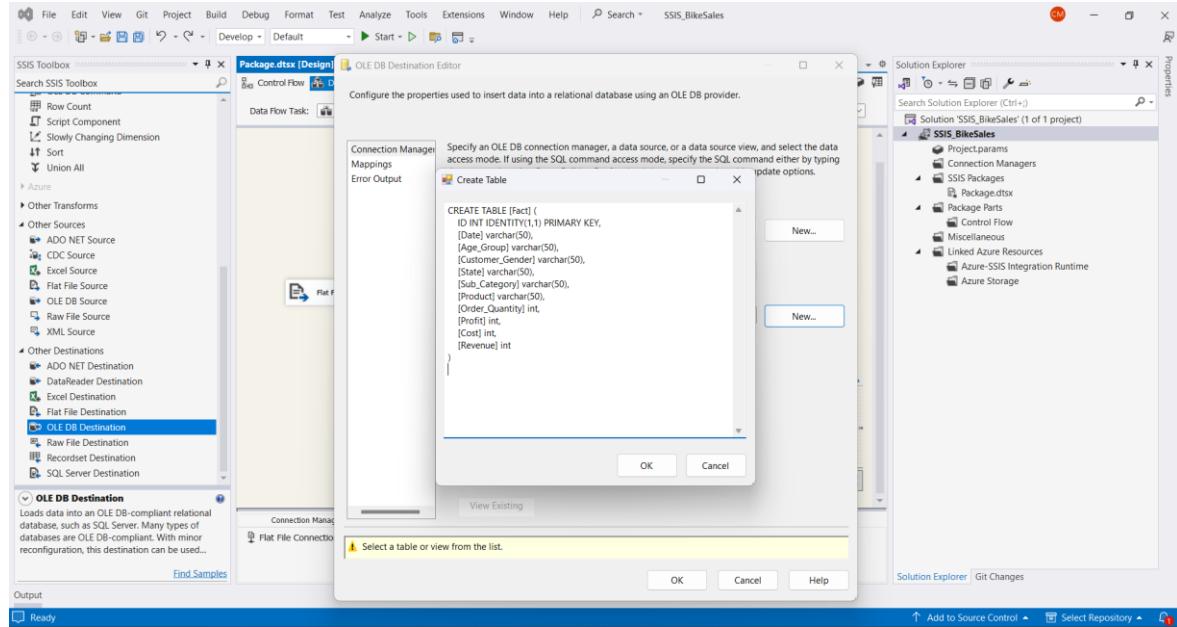
Tương tự như các bước ở trên, chúng ta sẽ tiến hành tạo tất cả các bảng Dim còn lại từ Multicast sẽ có được kết quả như sau:



Hình 2.16: Kết quả sau khi tạo xong tất cả các bảng Dim

2.3.5. Tạo bảng Fact

Tương tự như việc tạo bảng Dim, bảng Fact sẽ được lấy toàn bộ dữ liệu của file dataset đầu vào.



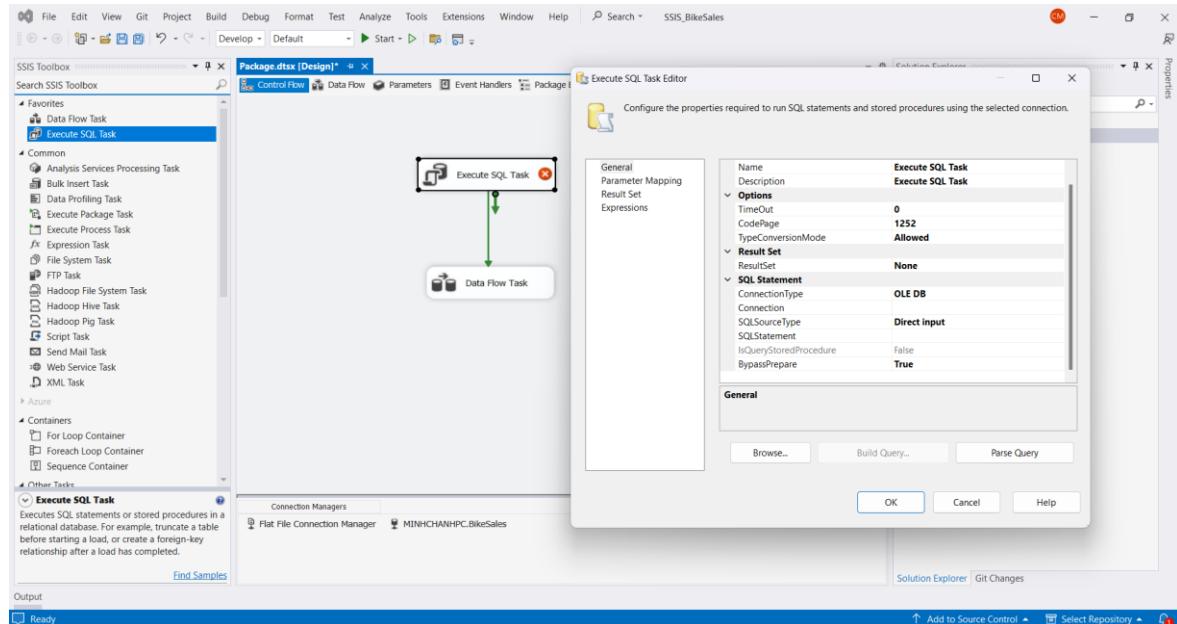
Hình 2.17: Tạo bảng Fact

Như vậy chúng ta đã tiến hành tạo thành công.

2.3.6. Import dữ liệu vào database

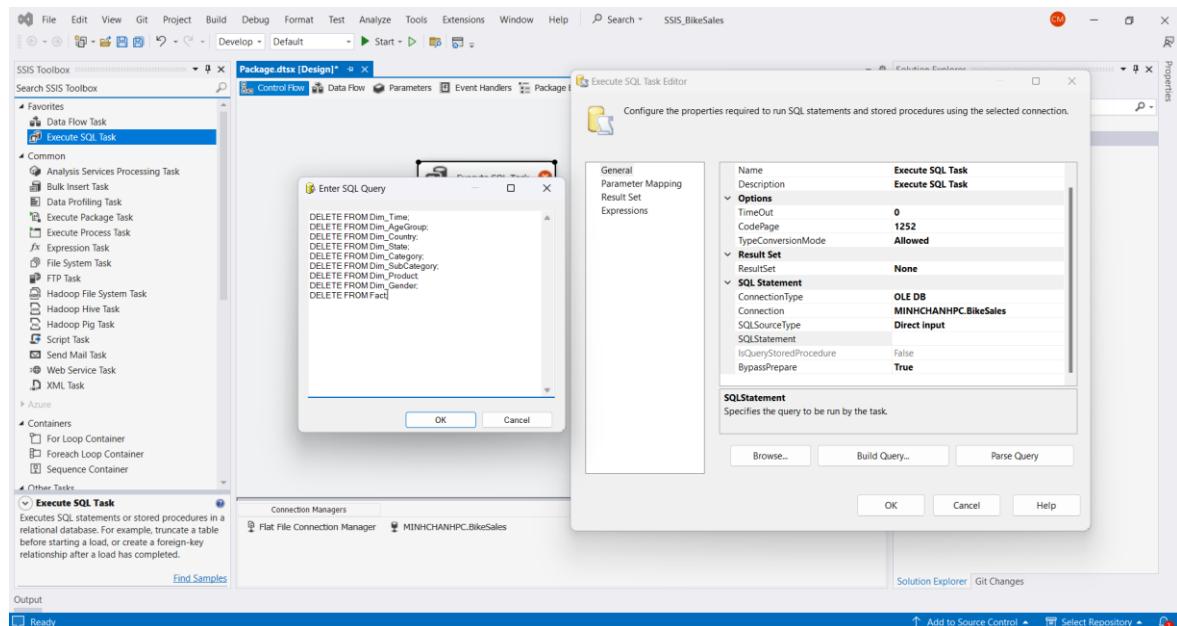
Để thực hiện quá trình import dữ liệu vào database, ta thực hiện các bước sau:

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



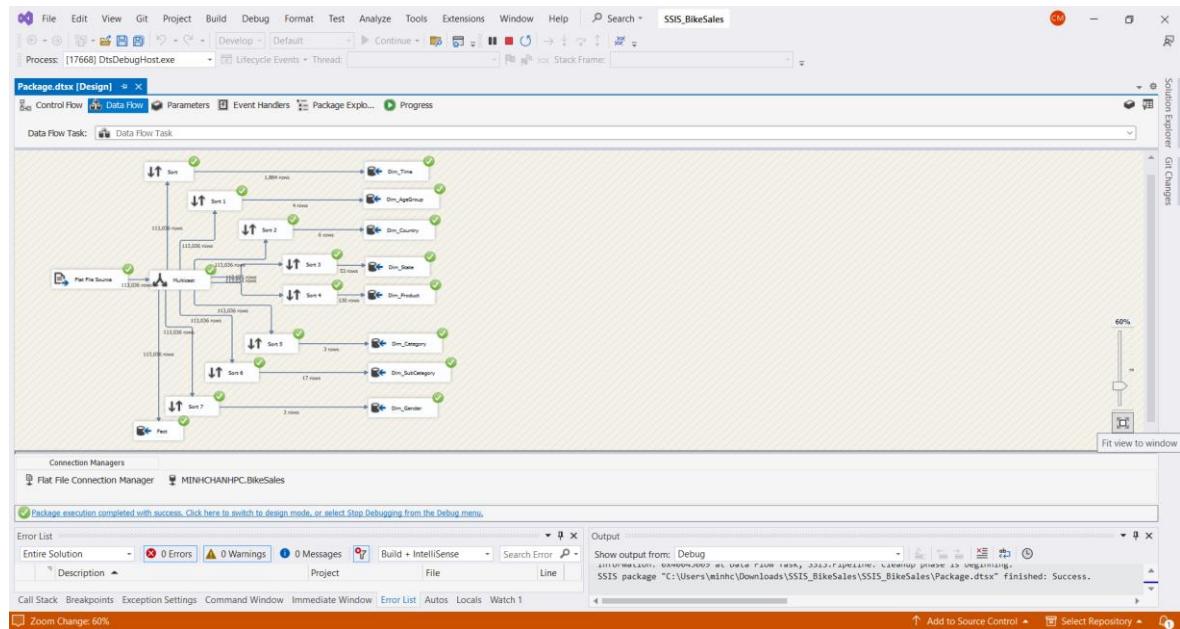
Hình 2.18: Cấu hình Execute SQL Task

Bước 2: Chọn các thông tin Connection và điền SQL Statement vào ô để xoá hết dữ liệu ở các bảng (nếu có) trước khi import dữ liệu



Hình 2.19: Điền SQL Statement

Kết quả luồng thực hiện của quá trình SSIS:



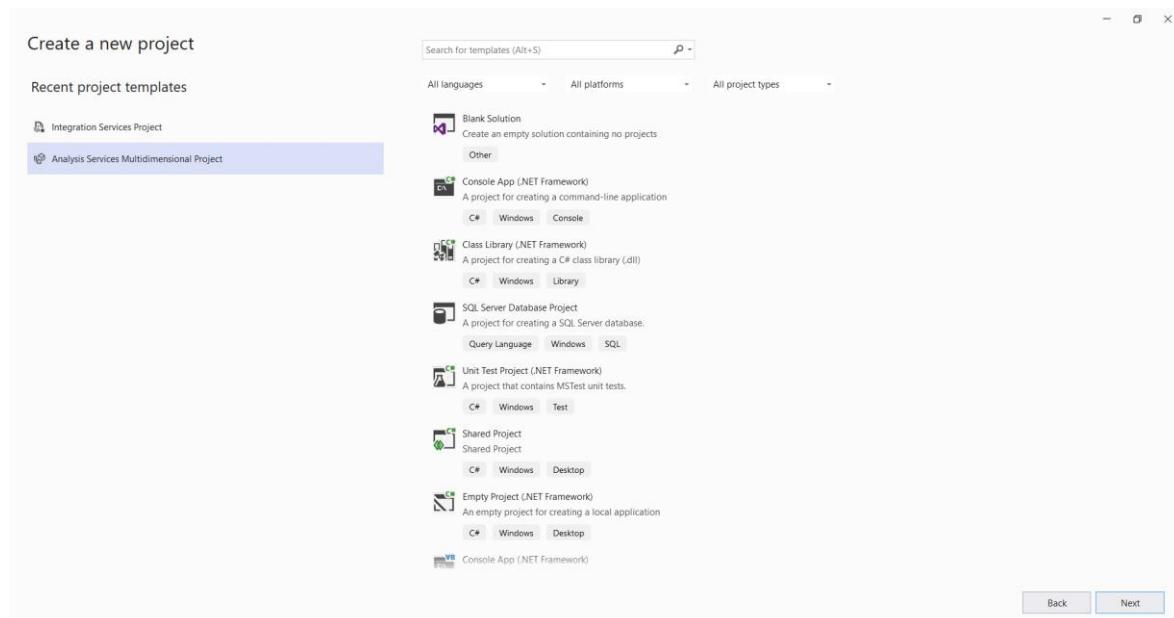
Hình 2.20: Kết quả sau khi thực hiện SSIS

Phần 3. PHÂN TÍCH BỘ DỮ LIỆU (QUÁ TRÌNH SSAS)

3.1. Tạo Project SSAS trong Vistual Studio

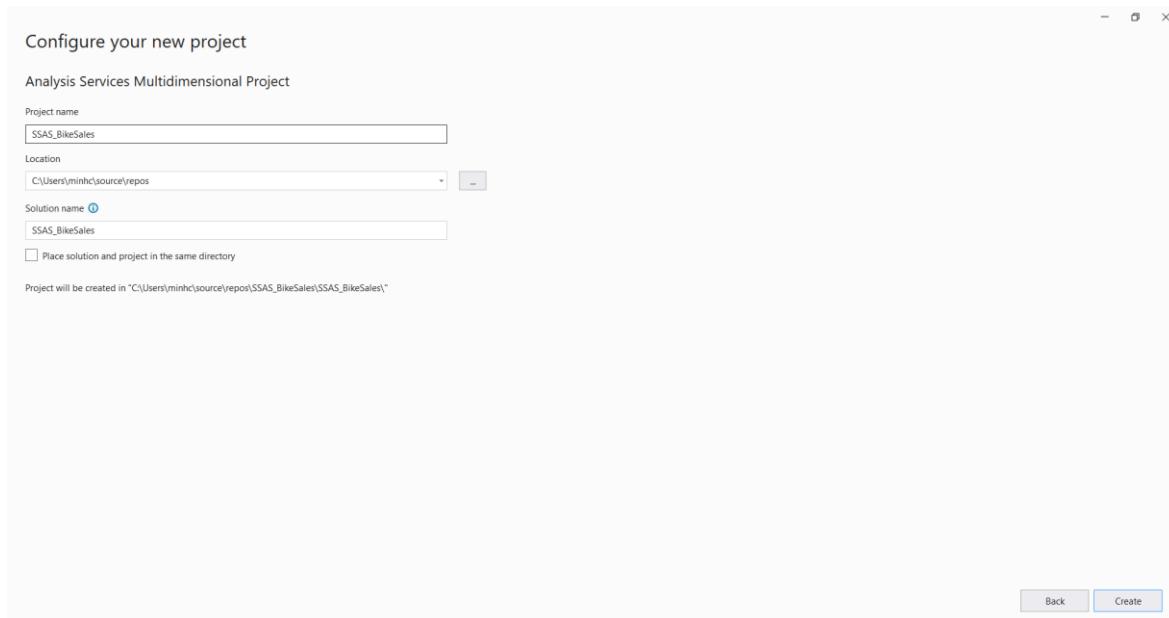
Bước 1: Khởi động Visual Studio 2022.

Bước 2: Tìm đến **Analysis Services Multidimensional Project** và chọn, sau đó nhấn next.



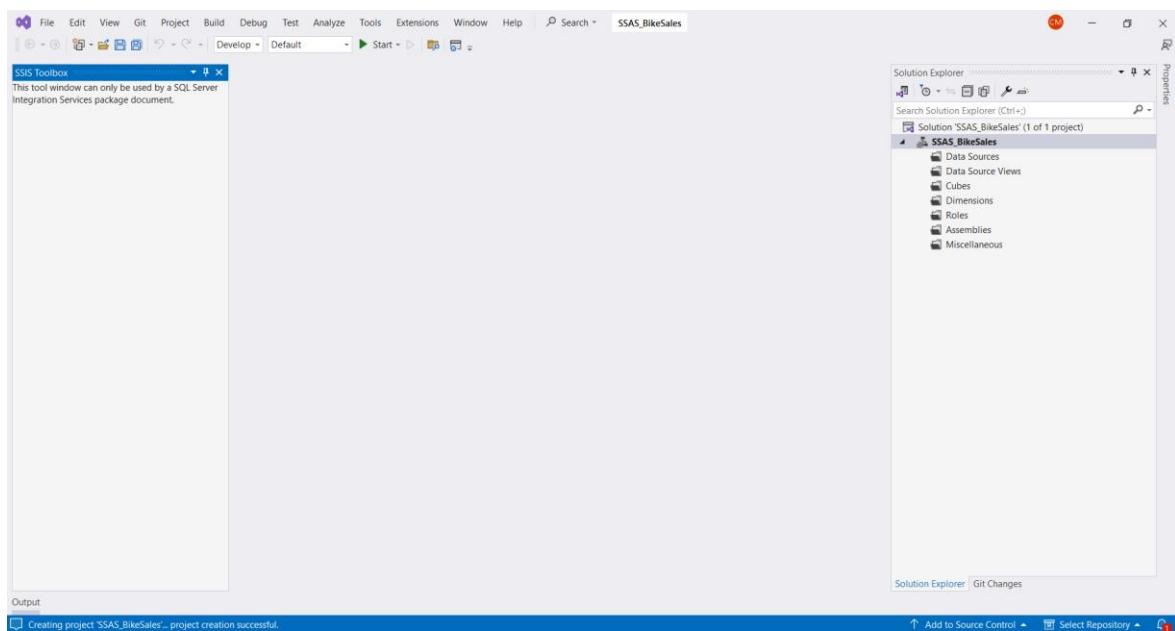
Hình 3.1: Khởi tạo Analysis Services Multidimensional Project trong Visual Studio

Bước 3: Đặt tên cho Project



Hình 3.2: Đặt tên cho SSAS Project

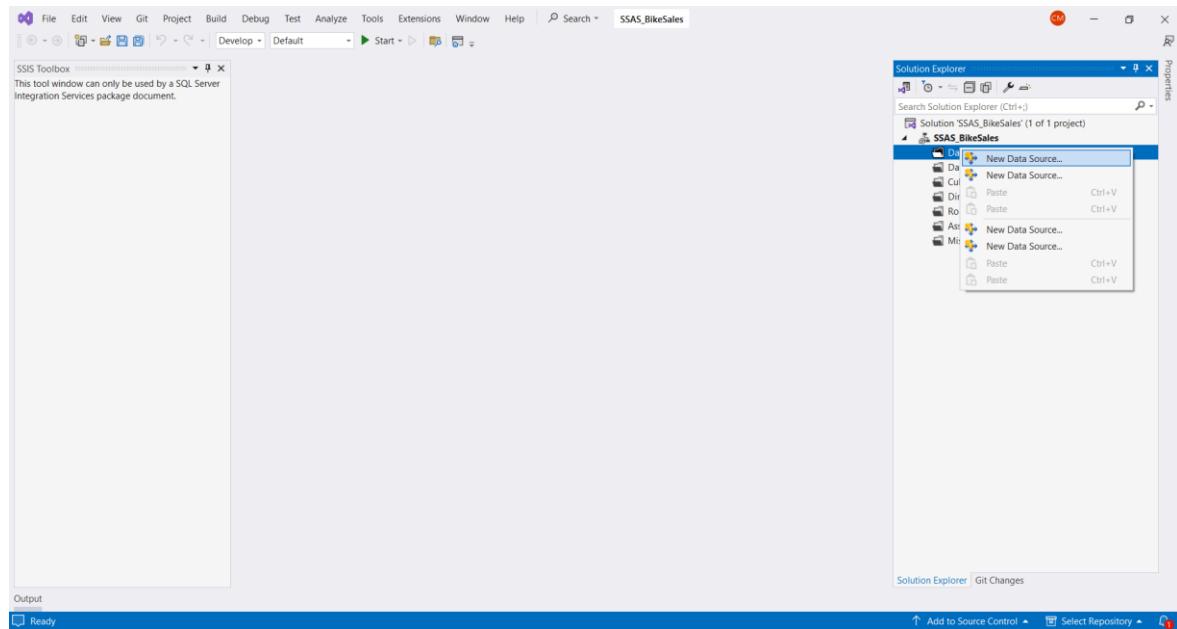
Kết quả sau khi khởi tạo Project SSAS



Hình 3.3: Kết quả sau khi khởi tạo SSAS Project

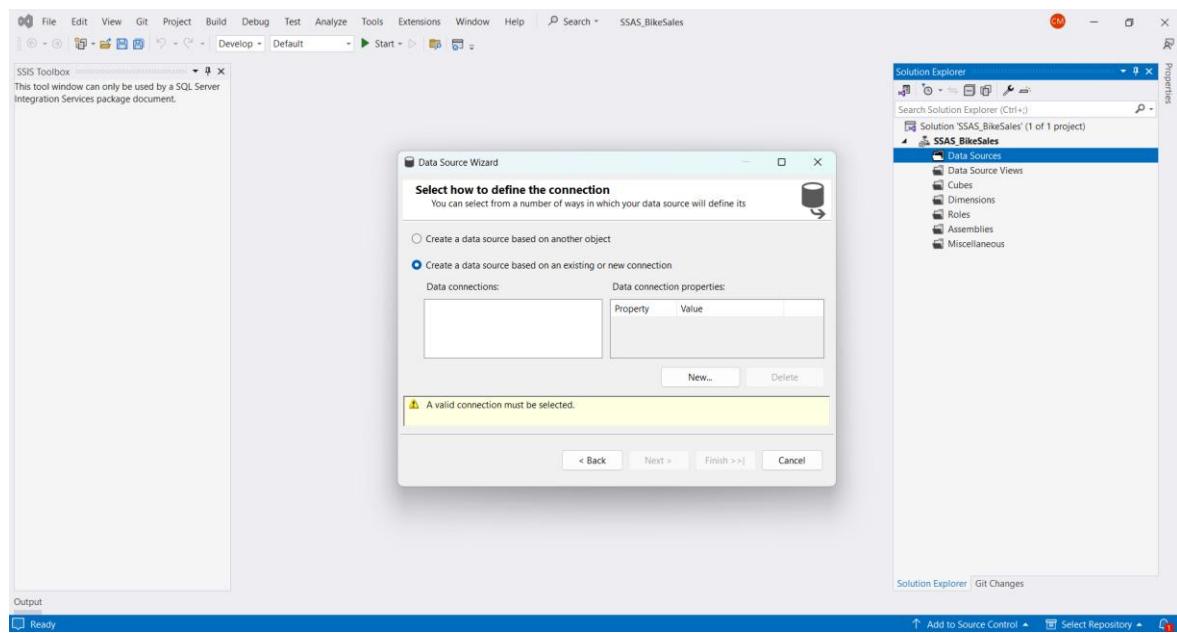
3.2. Connect đến Data Source

Bước 1: Chọn New Data Source



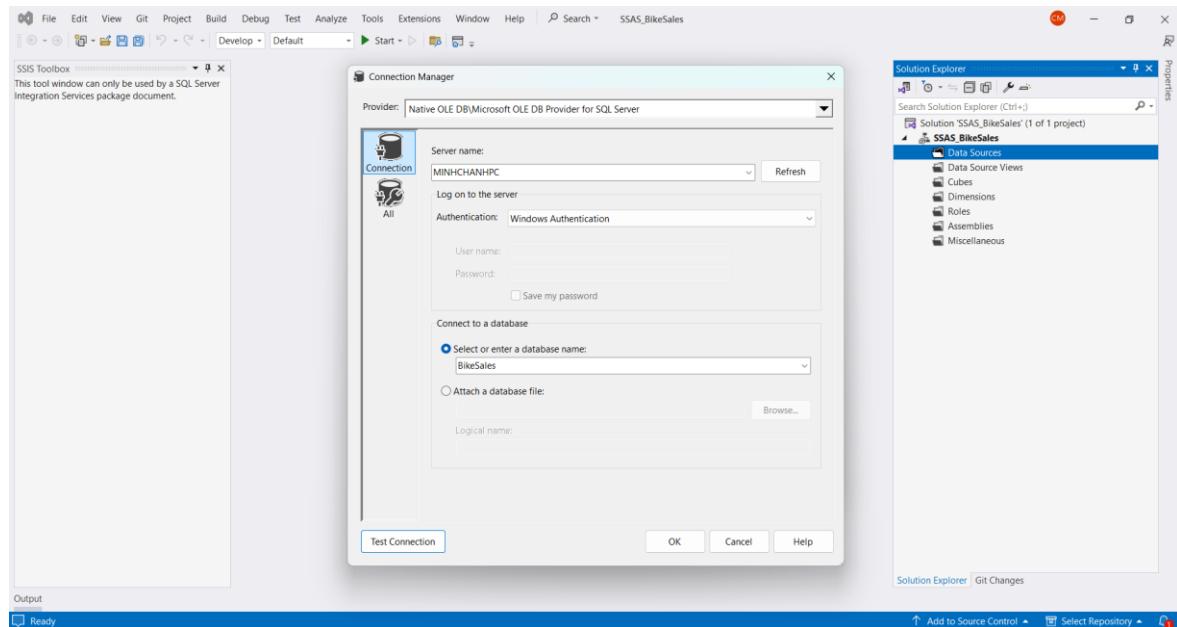
Hình 3.4: Khởi tạo một New Data Source

Bước 2: Chọn Create a data source based on an existing or new connection



Hình 3.5: Chọn Create a data source based on an existing or new connection

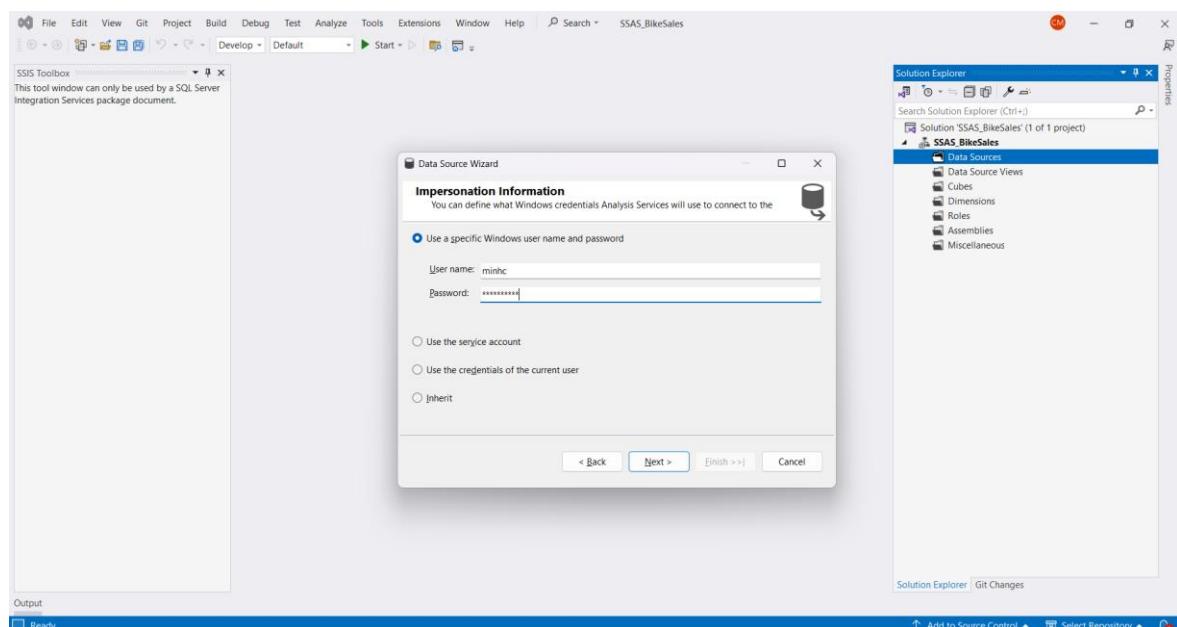
Bước 3: Click New, sau đó điền thông tin của **Connection Manager** → Nhấn **OK**



Hình 3.6: Chọn Server và Database để thực hiện

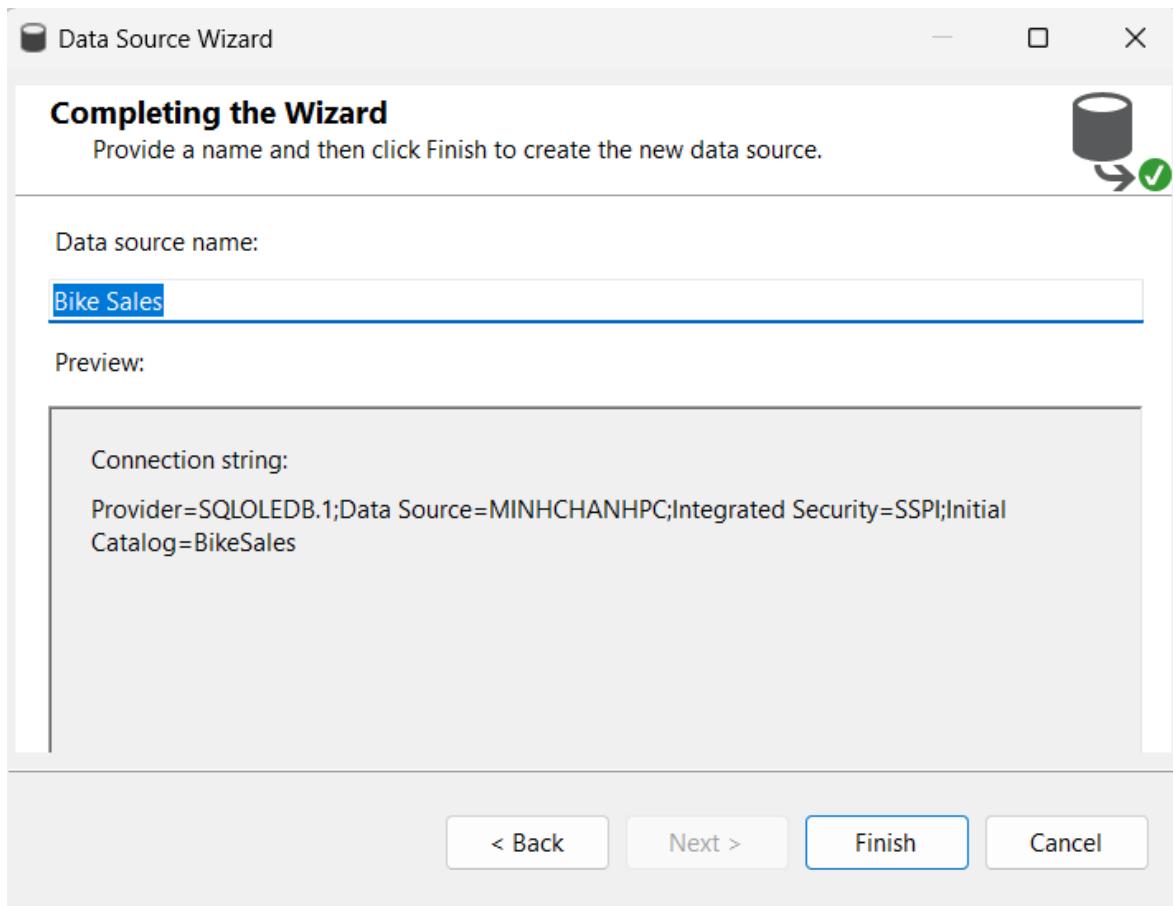
Tạo Connection thành công → Nhấn **Next**

Bước 4: Chọn **Use a specific Windows user name and password**, sau đó điền **Username** và **Password** của tài khoản Microsoft trên máy → Nhấn **Next**



Hình 3.7: Cung cấp tài khoản cho project

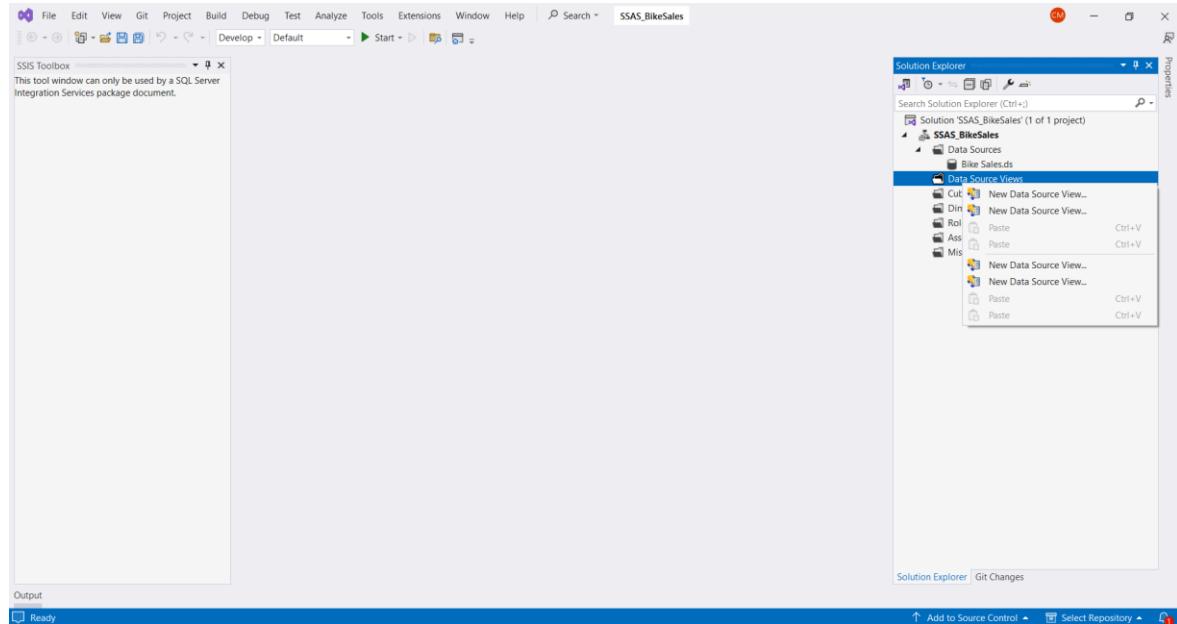
Bước 5: Điều chỉnh tên của Data source → Nhấn **Finish**. Vậy là đã tạo thành công Data Source.



Hình 3.8: Kết quả khi tạo Data Source

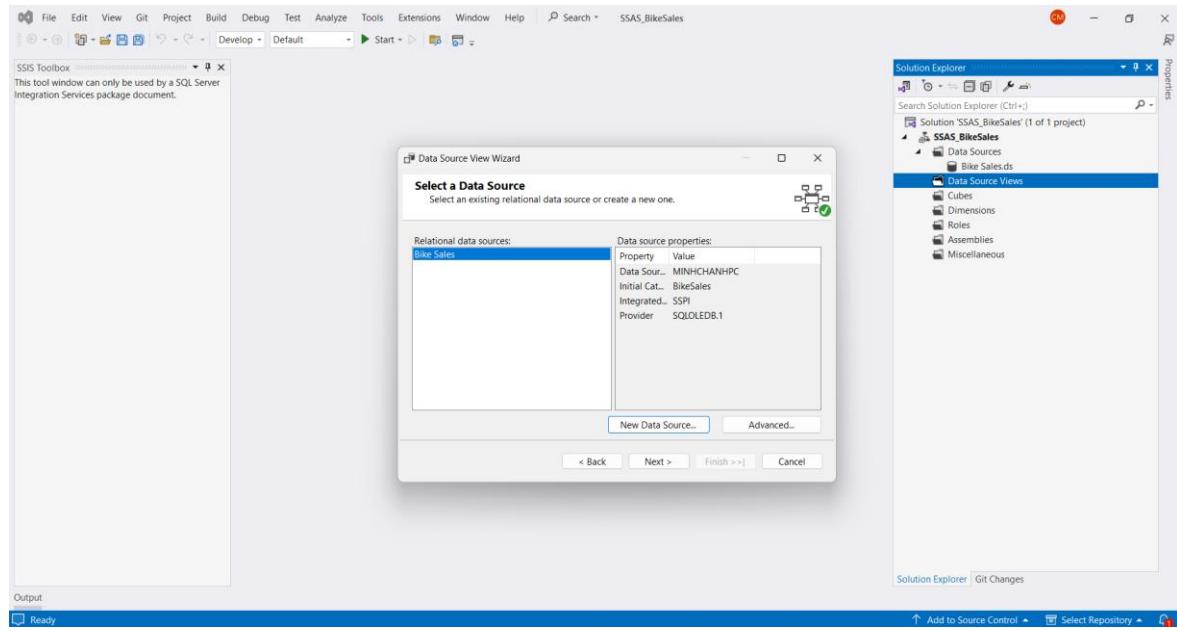
3.3. Tạo Data Source View

Bước 1: Tương tự như Data Source, ta nhấn chọn New Data Source View



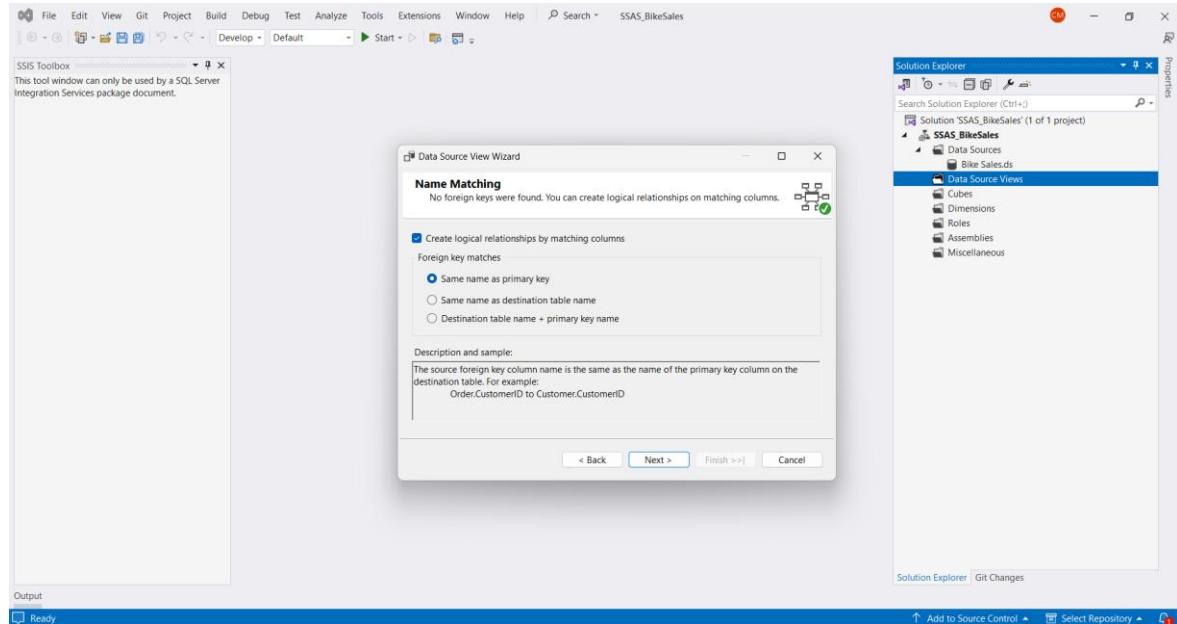
Hình 3.9: Khởi tạo một New Data Source View

Bước 2: Chọn Data Source vừa được tạo ở phần trên, sau đó nhấn Next



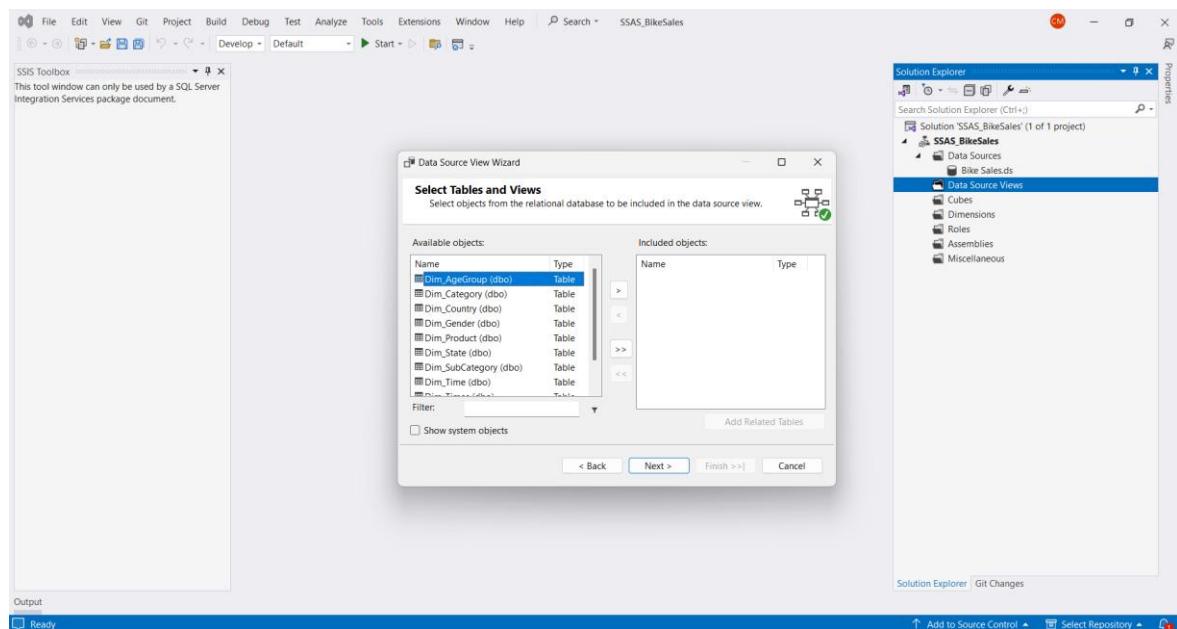
Hình 3.10: Chọn Data Source vừa tạo

Bước 3: Chọn **Create logical relationships by matching columns** để tạo liên kết cho các bảng tự động với **Same name as primary key**.



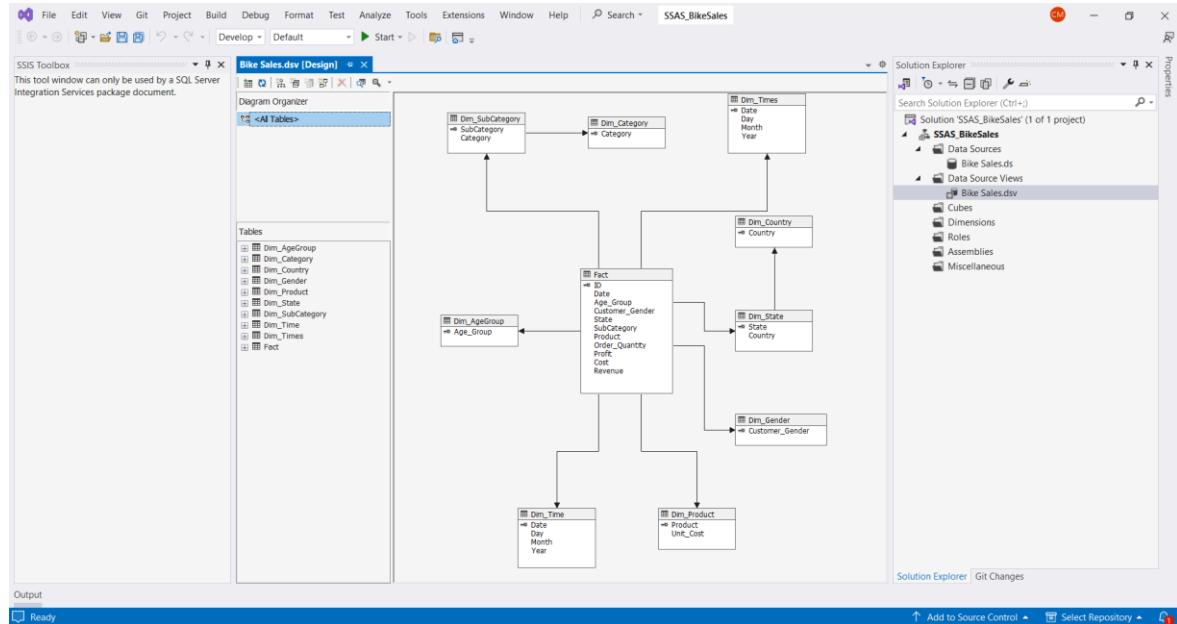
Hình 3.11: Tích chọn tạo liên kết giữa các bảng

Bước 4: Lần lượt Chọn bảng Fact sau đó nhấn nút > để đẩy bảng qua. Sau đó nhấn **Add Related Tables** để đẩy các bảng liên quan đến Fact qua, nếu là lượt đồ hình bông tuyết thì nhấn thêm **Add Related Tables** một lần nữa → Nhấn Next.



Hình 3.12: Chọn các bảng vào Source View

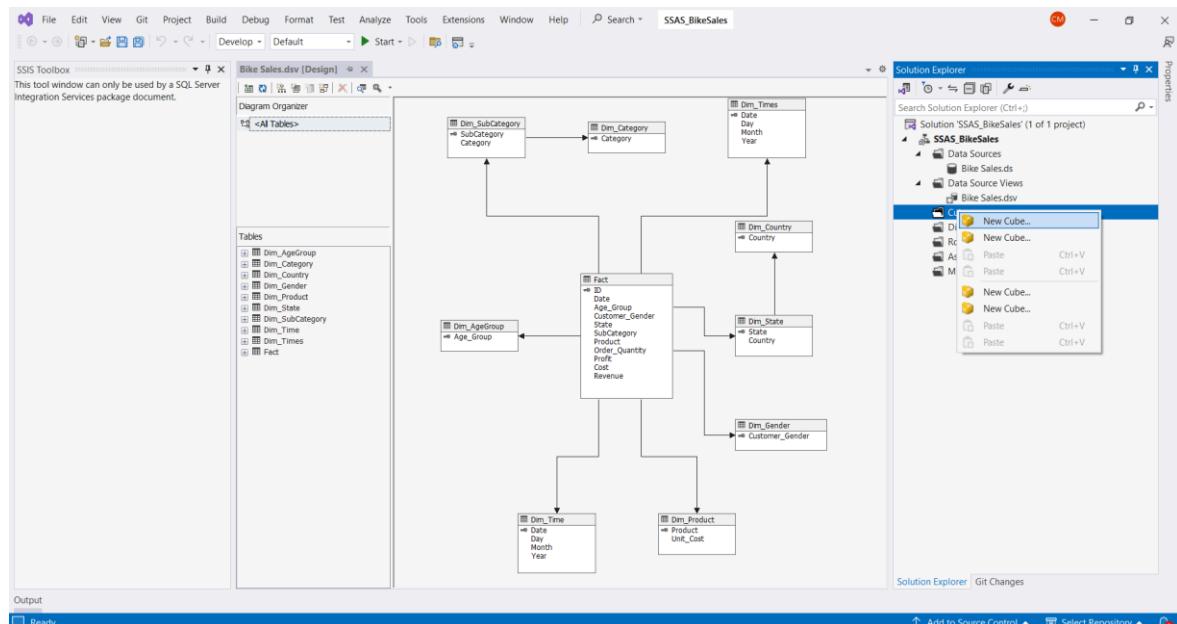
Bước 5: Kiểm tra lại số lượng bảng, sau đó nhấn Finish để hoàn thành. Như vậy đã tạo thành công Data Source View



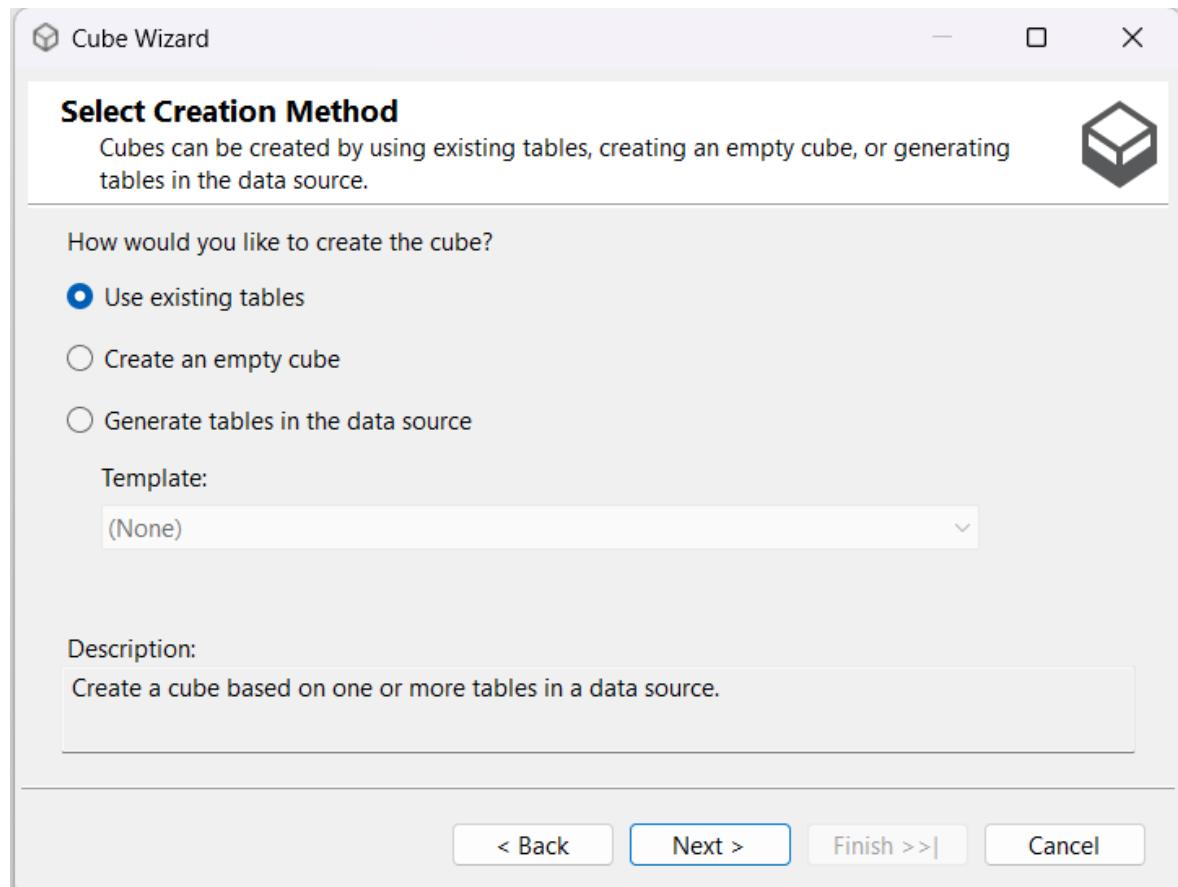
Hình 3.13: Kết quả sau khi tạo Data Source View

3.4. Tạo Cube và Dimensions

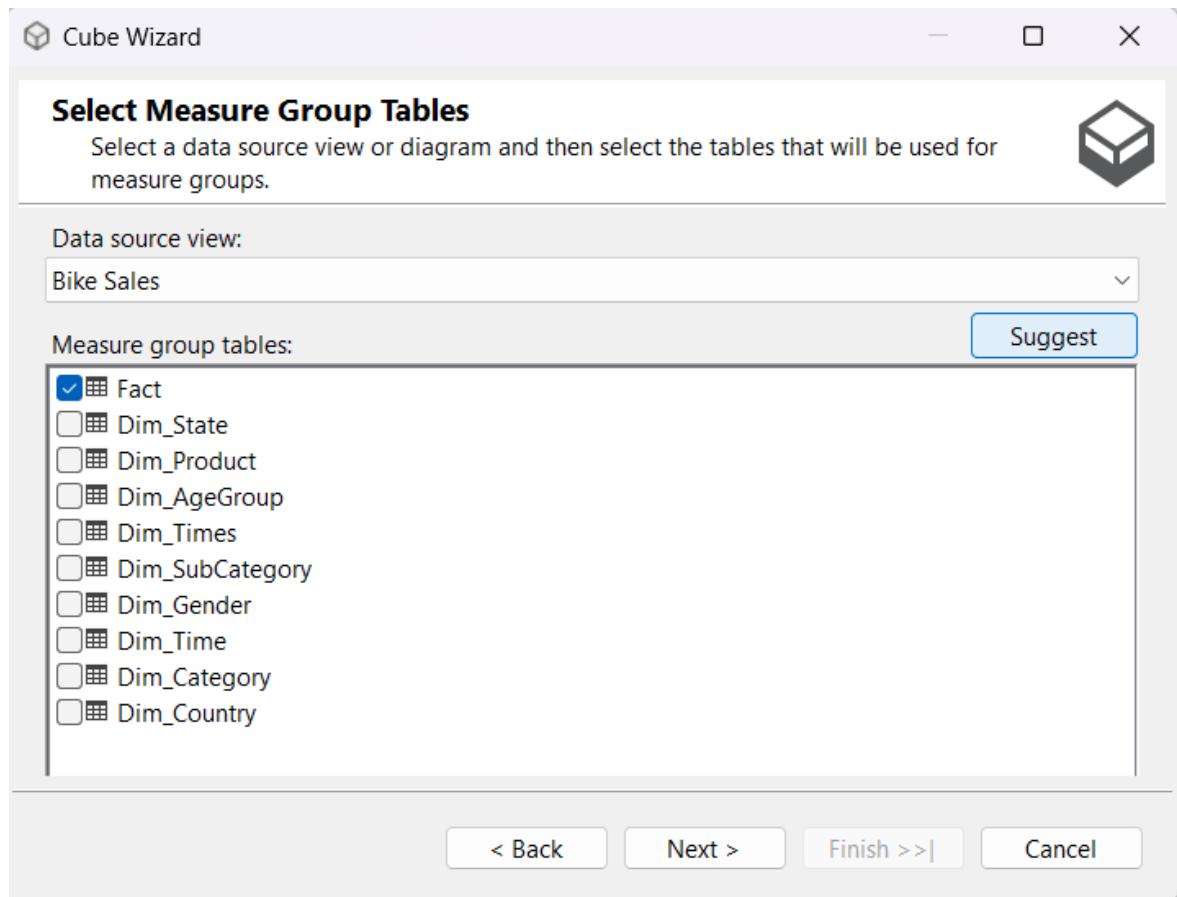
Bước 1: Chọn Cubes sau đó chọn New Cubes



Hình 3.14: Tạo một New Cube

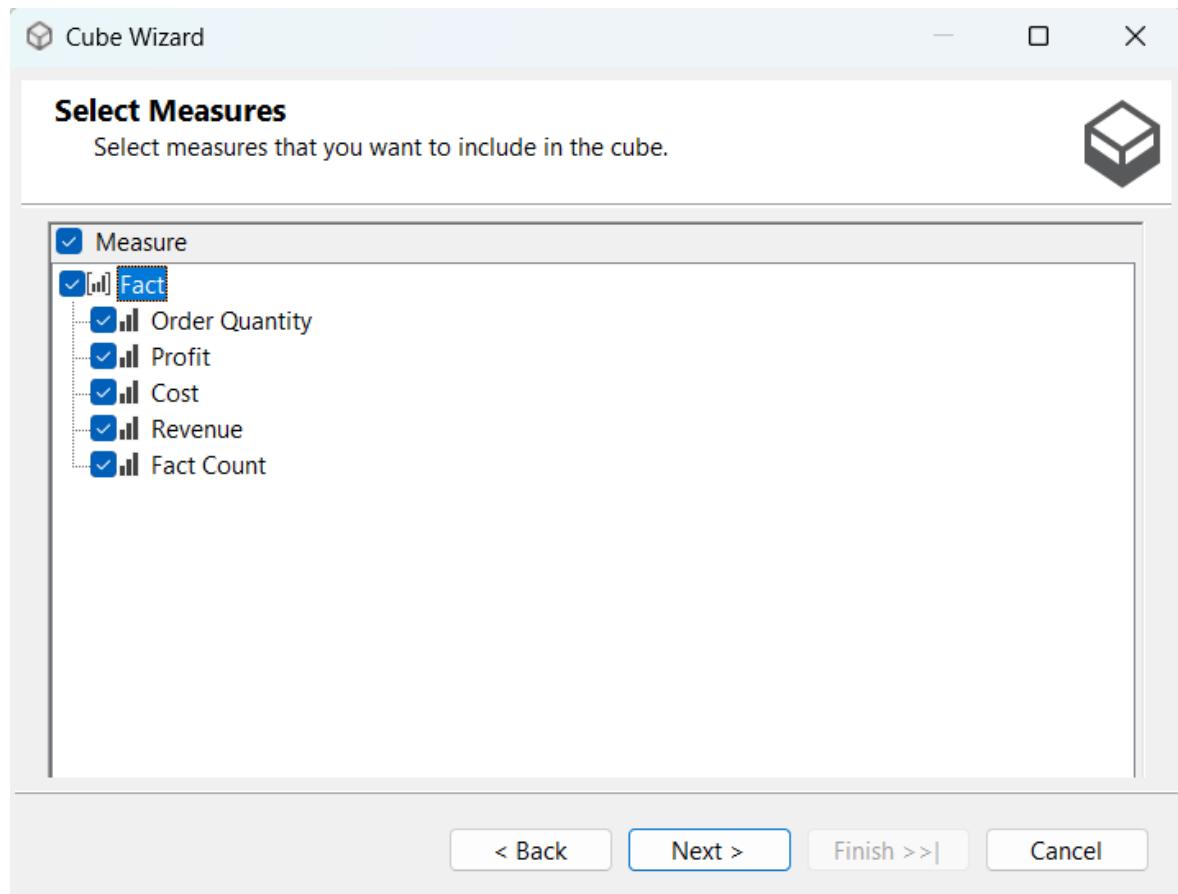
Bước 2: Chọn **Use an existing table**. Nhấn **Next**.*Hình 3.15: Tích chọn bảng đã được tạo*

Bước 3: Chọn **Measure Group Table** là bảng **Fact**, hoặc có thể nhấp **Suggest** để tự động chọn → Nhấn **Next**.

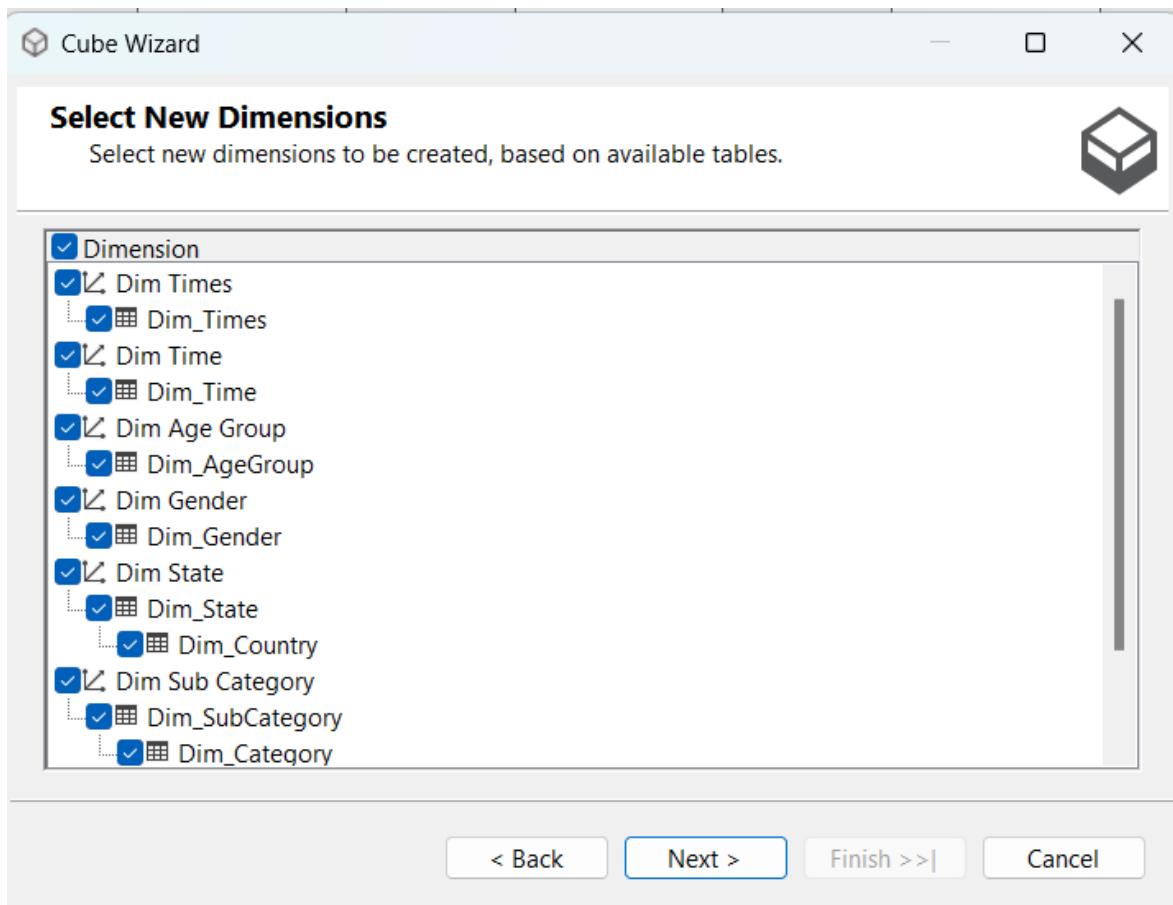


Hình 3.16: Chọn bảng chứa Measure

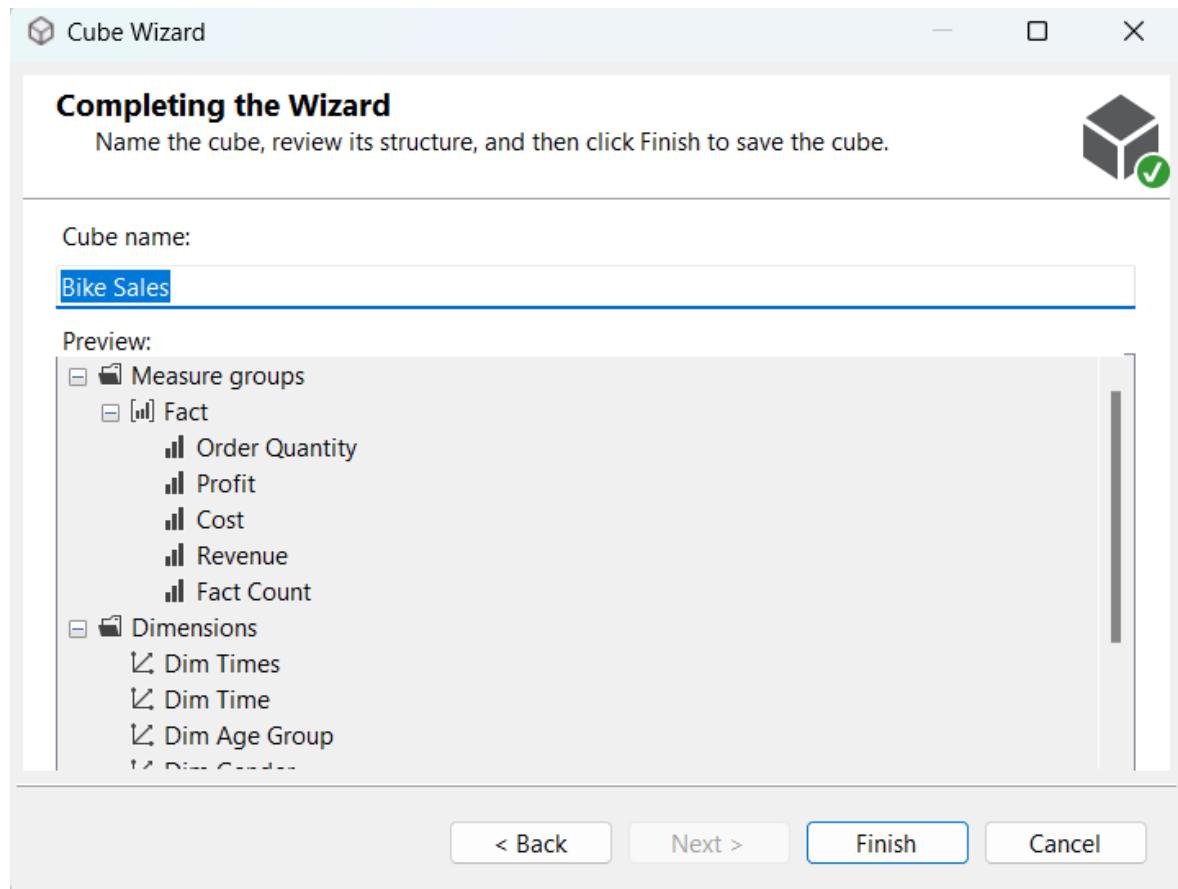
Bước 4: Chọn các thuộc tính trong bảng **Fact** để tạo các **Measure** → Nhấn Next.



Hình 3.17: Chọn các Measure cho project

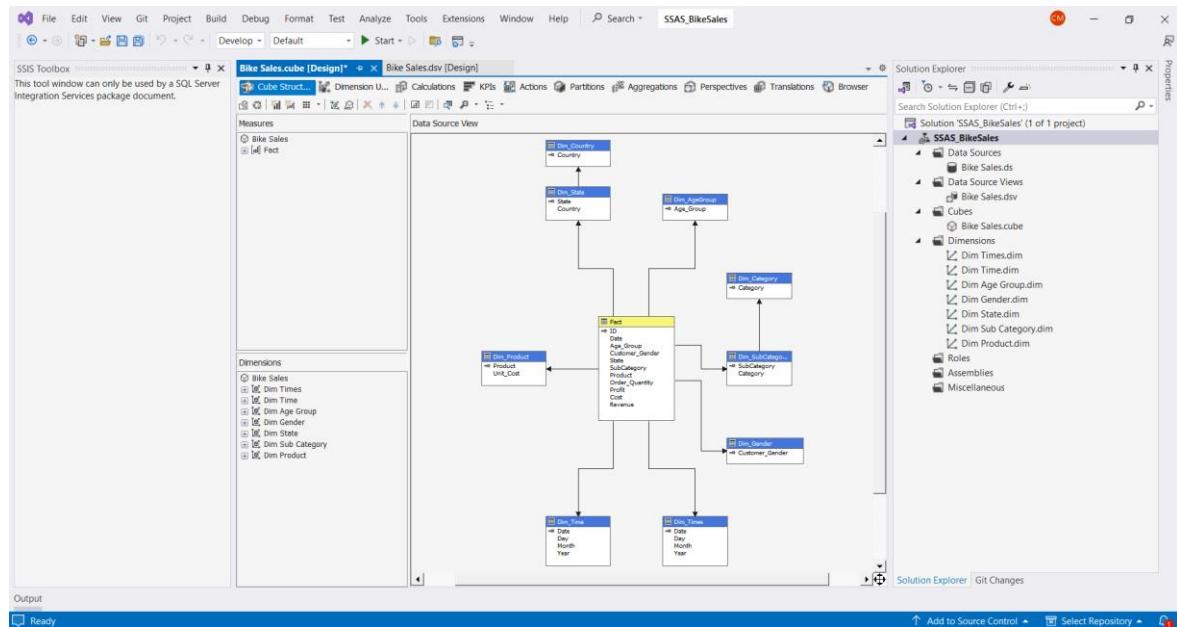
Bước 5: Chọn những Dimension cần thiết → Nhấn Next.

Hình 3.18: Chọn các Dimension cần thiết

Bước 5: Kiểm tra các thông tin và nhấn **Finish**.

Hình 3.19: Kiểm tra lại thông tin của Cube

Kết quả sau khi tạo Cube thành công.



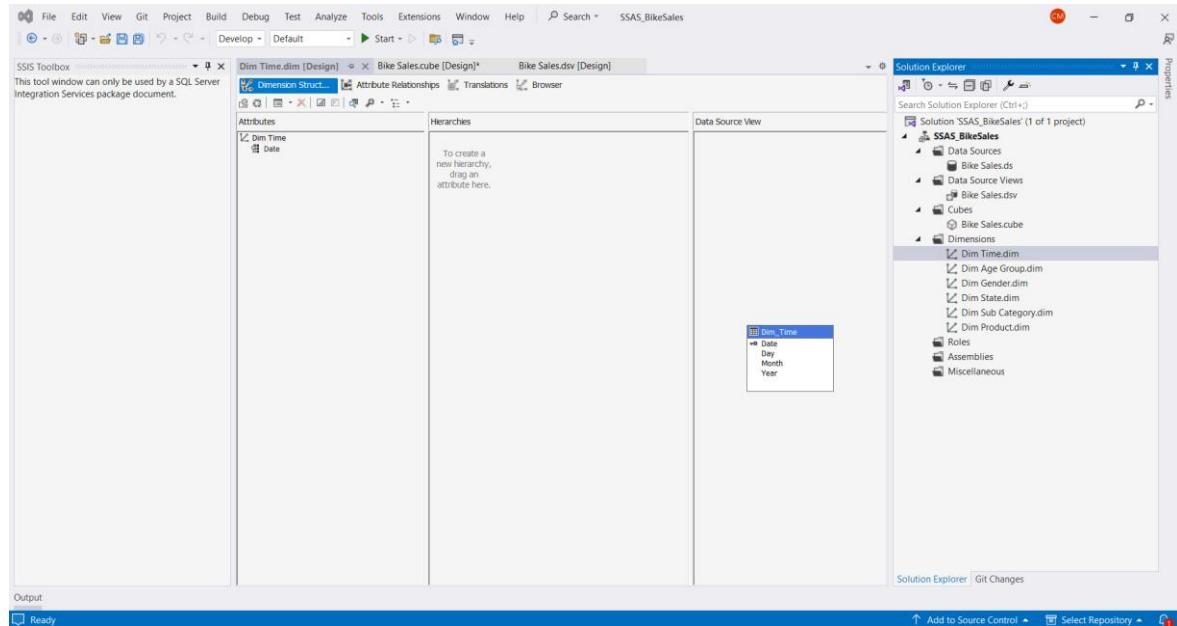
Hình 3.20: Cube sau khi tạo thành công

3.5. Thao tác trên các bảng Dim

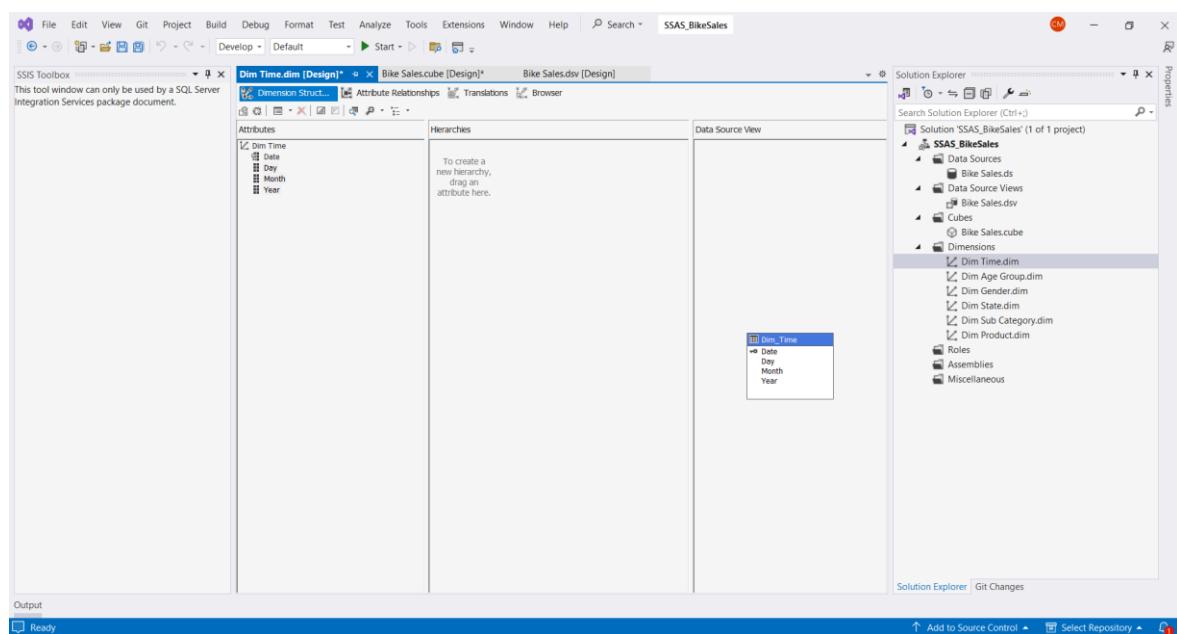
3.5.1. Bảng Dim_Time

Bước 1: Chọn bảng **Dim_Time** từ khung **Solution Explorer**.

Bước 2: Chọn các thuộc tính cần thiết từ **Data Source View** chưa được xuất hiện vào **Attributes**.

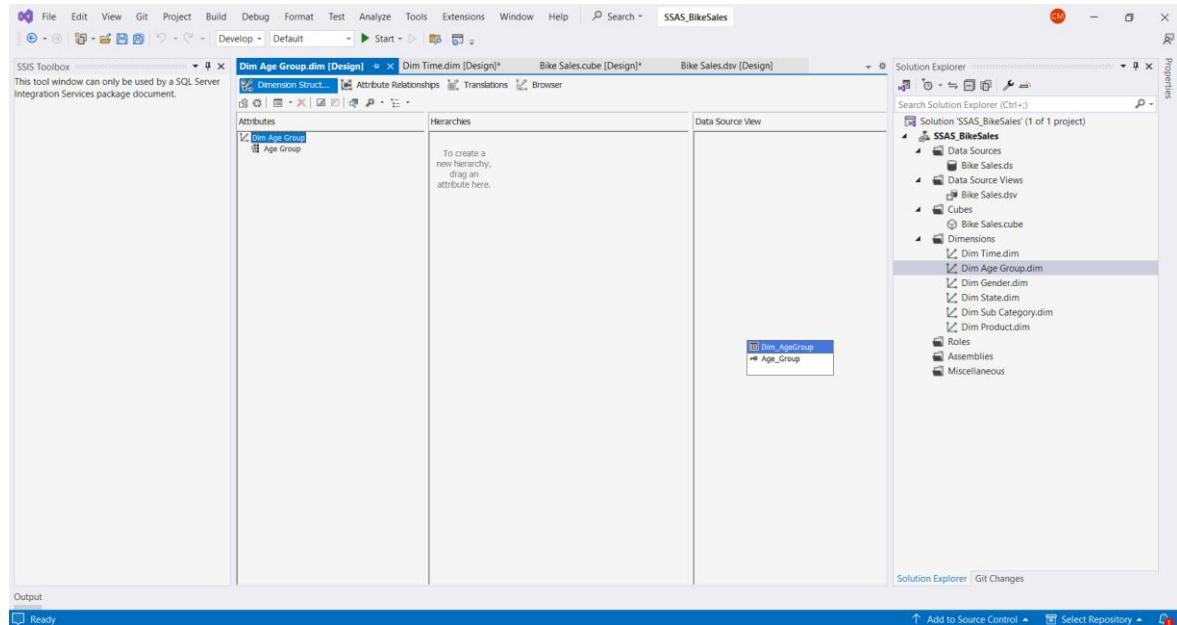
*Hình 3.21: Bên trong bảng Dim*

Bước 3: Kéo các thuộc tính đã chọn từ **Data Source View** sang cột **Attributes**.

*Hình 3.22: Kéo các thuộc tính từ Data Source View sang Attributes*

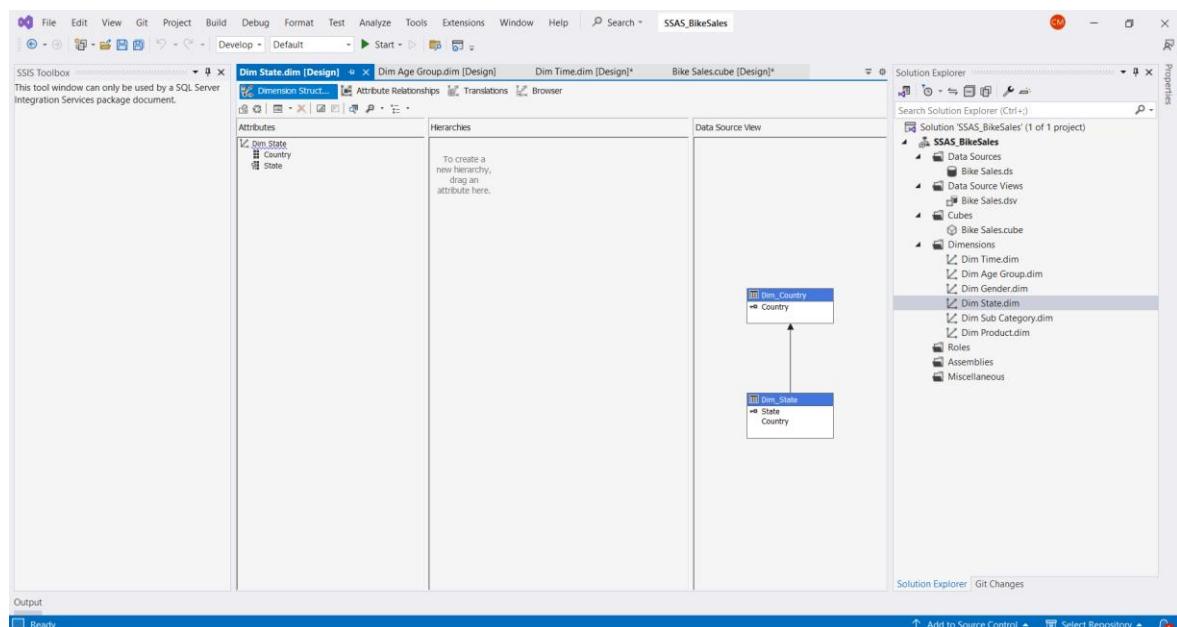
Tương tự, ta thực hiện với tất cả các bảng còn lại.

3.5.2. Bảng Dim_AgeGroup



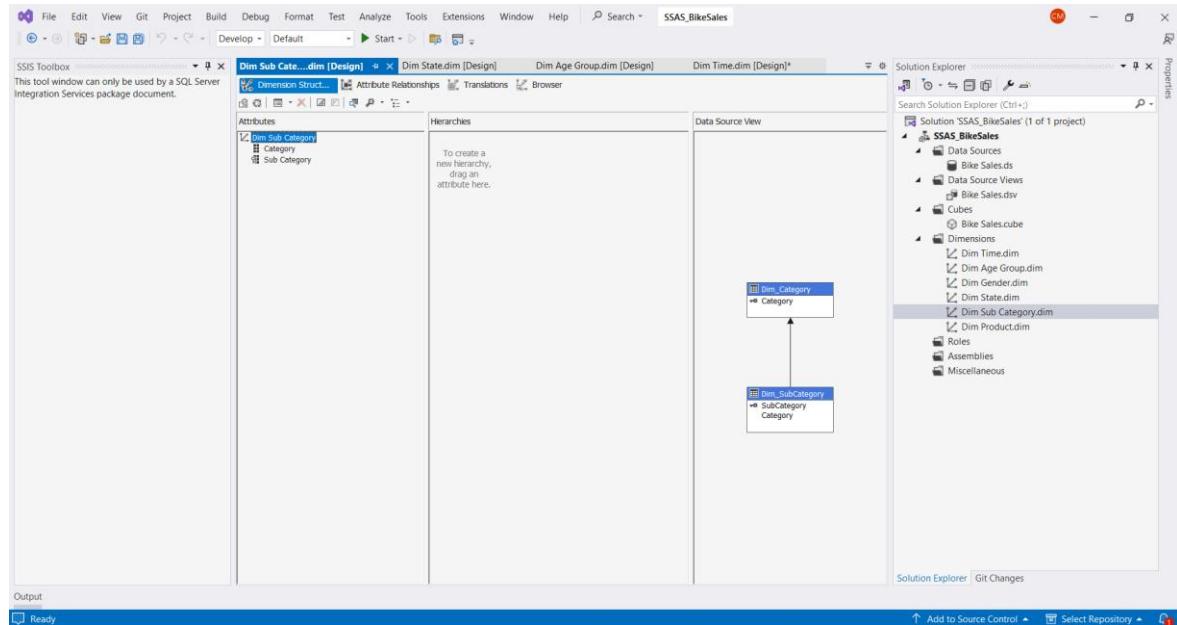
Hình 3.23: Bảng Dim_AgeGroup

3.5.3. Bảng Dim_Country và Dim_State



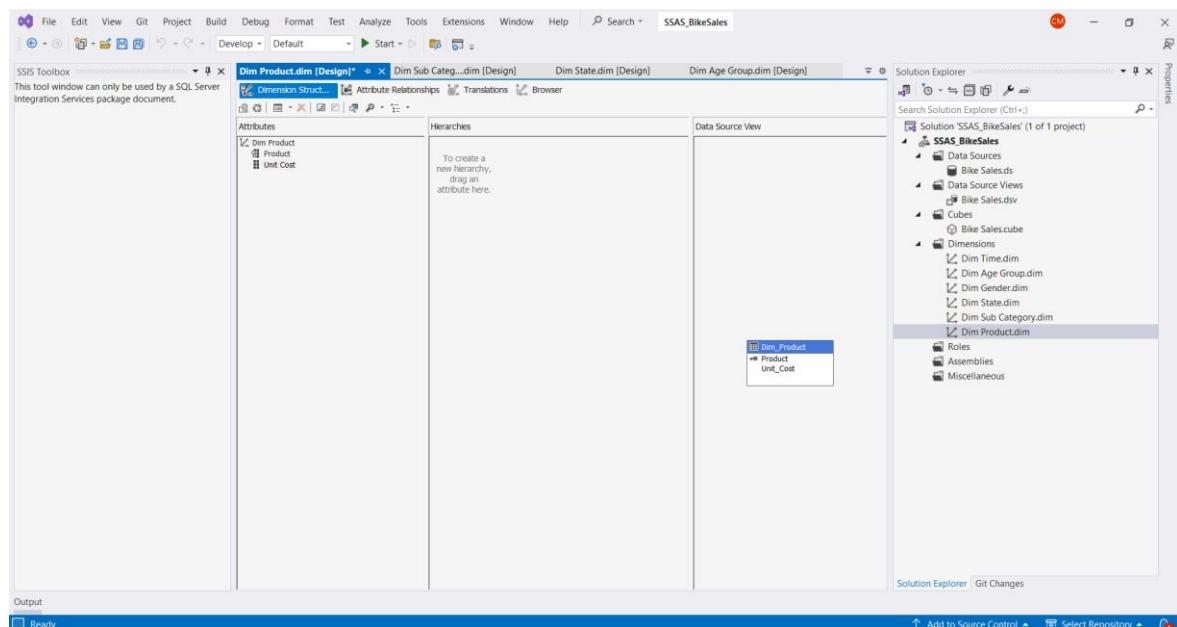
Hình 3.24: Bảng Dim_Country và Dim_State

3.5.4. Bảng Dim_Category và Dim_SubCategory



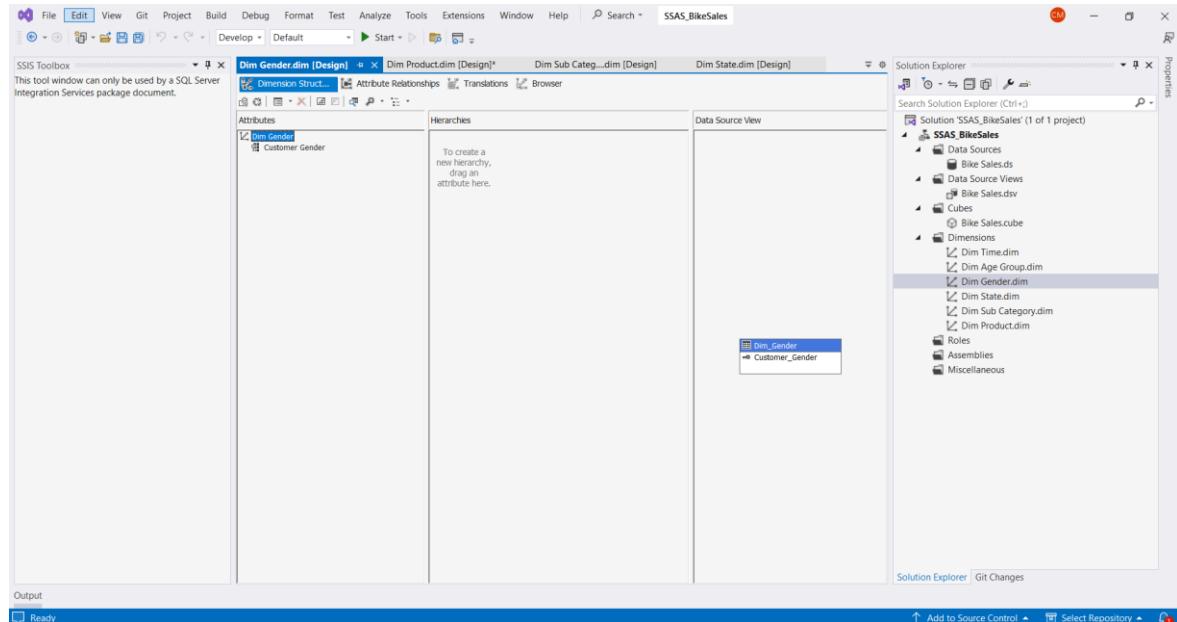
Hình 3.25: Bảng Dim_Category và Dim_SubCategory

3.5.5. Bảng Dim_Product



Hình 3.26: Bảng Dim_Product

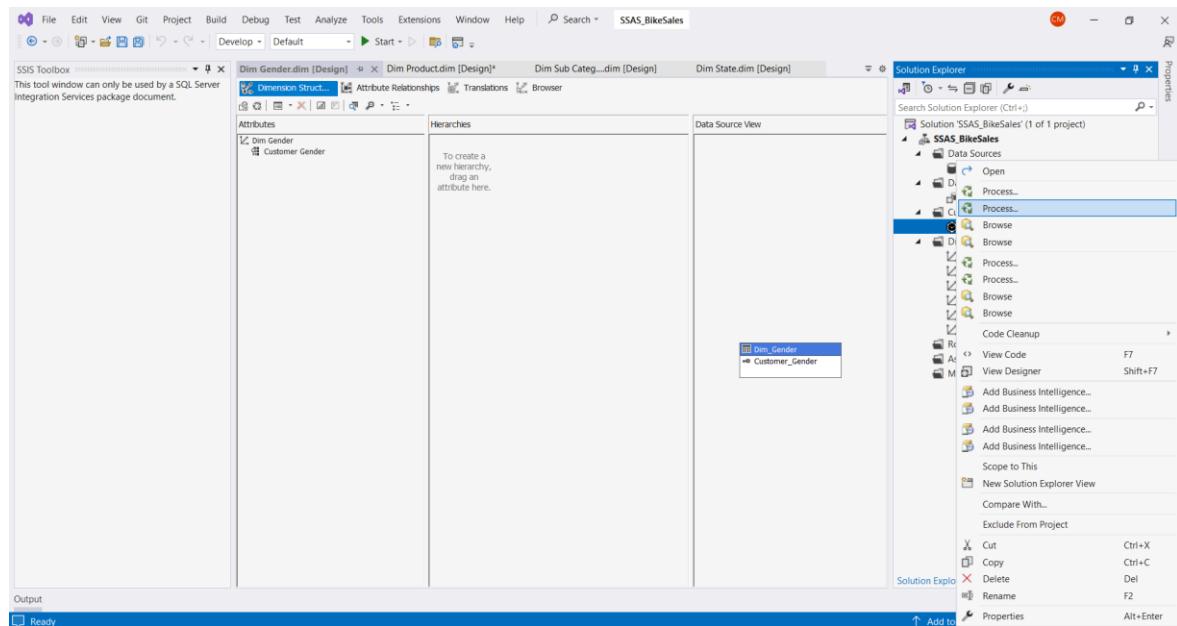
3.5.6. Bảng Dim_Gender



Hình 3.27: Bảng Dim_Gender

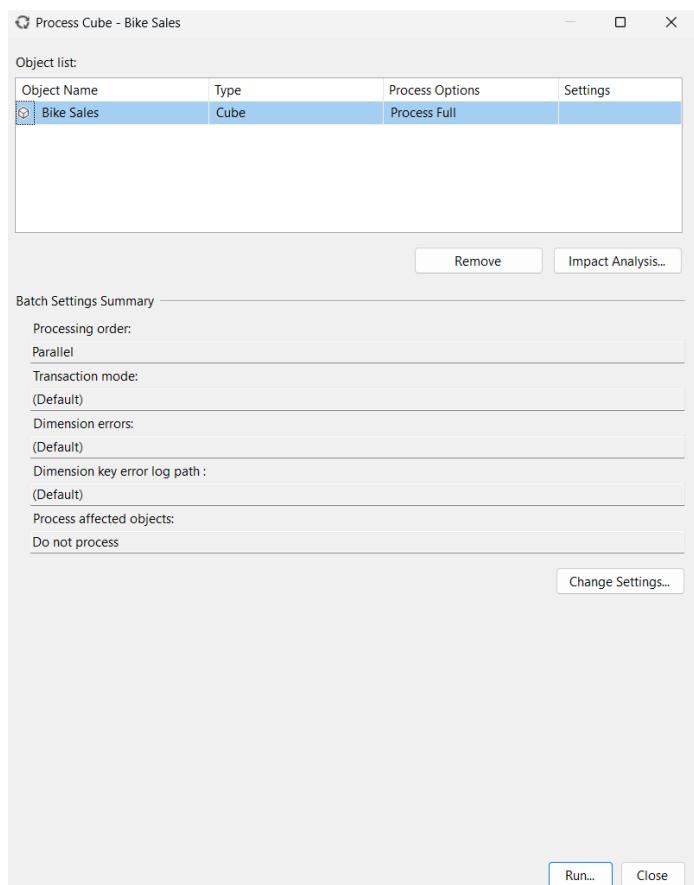
3.6. Deploy và Proces project

Bước 1: Tại Cubes, chuột phải vào tên cubes và chọn Process.



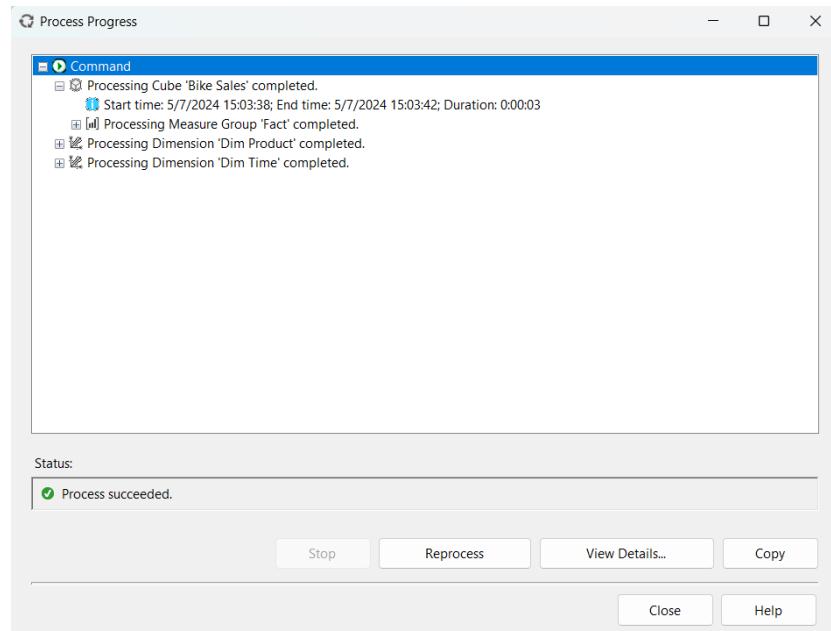
Hình 3.28: Bắt đầu thực thi project

Bước 2: Xuất hiện 1 popup thông báo, sau đó chọn Yes. Sau đó hộp thoại Process Cube xuất hiện → Nhấn Run để bắt đầu chạy.



Hình 3.29: Chạy cube

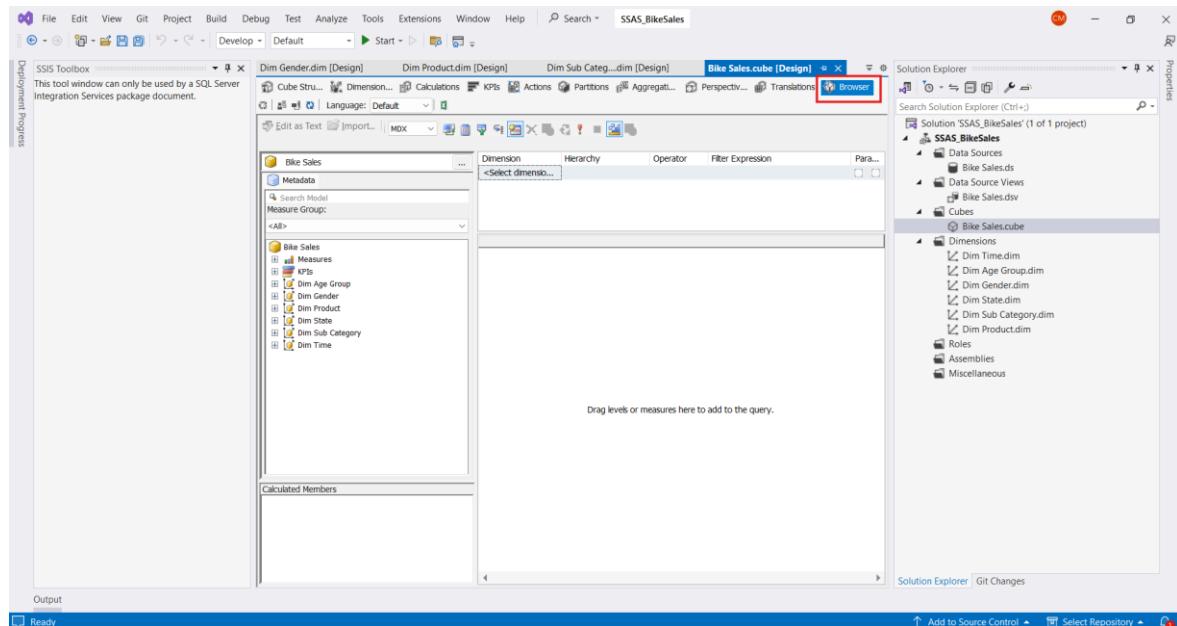
Bước 3: Sau khi chạy thành công → Nhấn Close để hoàn tất.



Hình 3.30: Kết quả sau khi deploy cube thành công

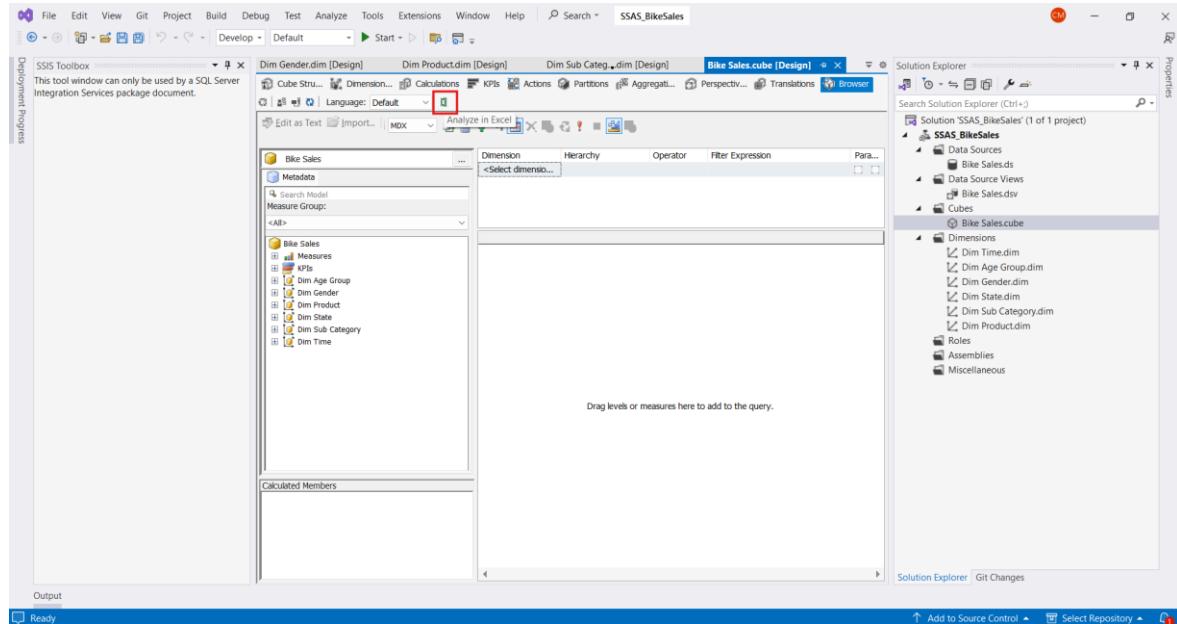
3.7. Thực hiện các câu truy vấn

Để thực hiện các câu truy vấn bằng thủ công, chúng ta vào bên trong Cube, nhấn qua tab Brower và thực hiện.

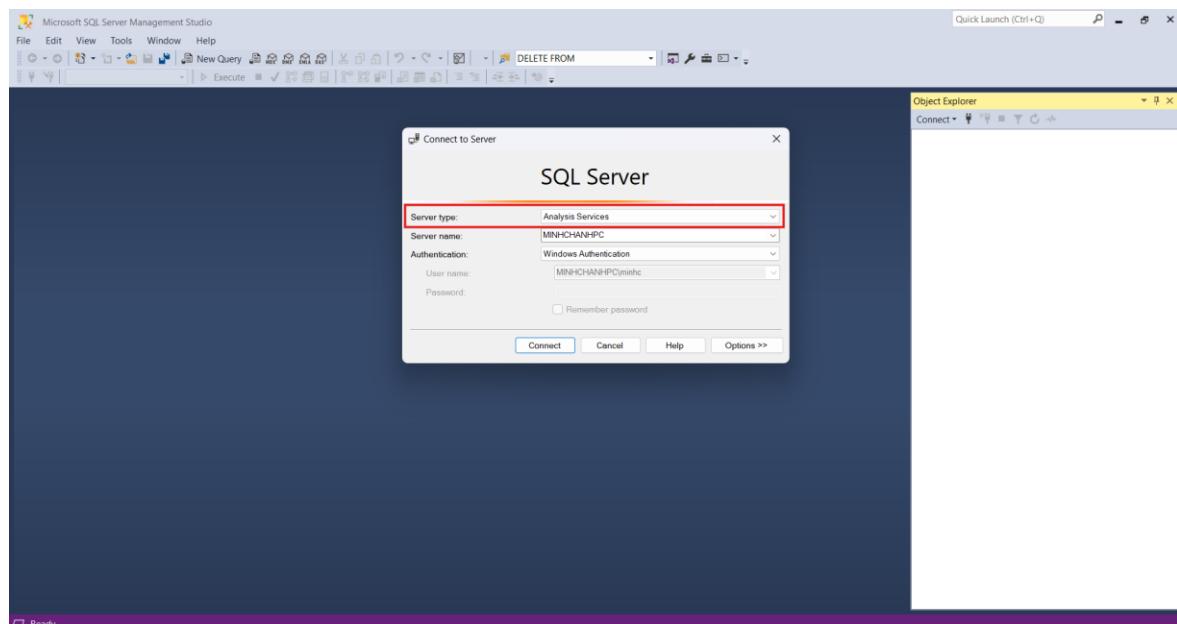


Hình 3.31: Cách thực hiện các câu query trong Cube

Để thực hiện Pivot trong MS Excel, chúng ta sẽ nhấp tìm và nhấp vào nút Analyze in Excel trong Brower của Cube.



Hình 3.32: Cách để thực hiện các câu query bằng Pivot trong MS Excel
 Để thực thi các câu MDX, chúng ta vào SQL Server, sau đó kết nối với Analysis Service và tạo mới 1 cửa sổ soạn thảo MDX và thực hiện.



Hình 3.33: Cách để thực hiện các câu query bằng MDX trong SQL Server

3.7.1. Thống kê tổng số sản phẩm thuộc thể loại “Bikes”

❖ SSAS:

The screenshot shows the Microsoft Analysis Services (SSAS) Cube Designer interface. The top menu bar includes options like Cube Structure, Dimension Usage, Calculations, KPIs, Actions, Partitions, Aggregations, Perspectives, Translations, and Browser. The main workspace displays a cube named 'Bike Sales.cube [Design]'. On the left, there's a tree view of the cube's structure under 'Bike Sales' and 'Metadata'. The central area shows a table with one row: 'Product Category' (Bikes) and 'Order Quantity' (36411). A filter expression 'Bikes' is applied to the Product Category dimension. The bottom pane shows a section for 'Calculated Members'.

Hình 3.34: Câu truy vấn 1 trong Visual Studio

❖ MDX: Thực hiện câu truy vấn và kết quả

Query:

```
SELECT NON EMPTY {[Measures].[Order Quantity]} ON COLUMNS,
{[Dim Sub Category].[Product Category].&[Bikes]} ON ROWS
FROM [Bike Sales]
```

Kết quả:

The screenshot shows the SSMS interface with the following details:

- MDXQuery1.mdx - ...HCHANHPC\minhc***: The current query file.
- Cube: Bike Sales**: The selected cube.
- Measure Group: <All>**: The selected measure group.
- Results Window Content:**

```
SELECT NON EMPTY {[Measures].[Fact Count]} ON COLUMNS,
        {[Dim Sub Category].[Product Category].&[Bikes]} ON Rows
    From [Bike Sales]
```

Messages Results

	Fact Count
Bikes	25982
- Status Bar:** Query executed successfully. MINHCHANHPC | MINHCHANHPC\minhc SSAS_BikeSales 00:00:01

Hình 3.35: Câu truy vấn 1 trong SQL Server

❖ Pivot Table tên MS Excel:

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable Fields pane on the right indicates that 'Order Quantity' is selected under the 'Values' section. The main table displays the following data:

Thống kê tổng số sản phẩm thuộc thể loại "Bikes"	
Product Category	Order Quantity
Bikes	36411

Hình 3.36: Câu truy vấn 1 trong MS Excel

3.7.2. Thống kê số lượng bán được của từng sản phẩm qua các năm

❖ SSAS:

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface with the 'Bike Sales.cube [Design]' window open. The window has a toolbar at the top with various icons for cube structure, dimension usage, calculations, KPIs, actions, partitions, aggregations, perspectives, and translations. Below the toolbar is a menu bar with options like File, Edit, View, Tools, etc. The main area is divided into several panes: a left pane for 'Metadata' showing dimensions like Dim Age Group, Dim Gender, Dim Product, Dim State, Dim Sub Category, and Dim Time; a central pane for 'Calculated Members' displaying a table with columns 'Product', 'Year', and 'Order Quantity'; and a right pane for 'Dimensions' with sections for 'Hierarchy', 'Operator', and 'Filter Expression'. The 'Order Quantity' table data is as follows:

Product	Year	Order Quantity
All-Purpose Bike Stand	2013	663
All-Purpose Bike Stand	2014	582
All-Purpose Bike Stand	2015	617
All-Purpose Bike Stand	2016	538
AWC Logo Cap	2013	14336
AWC Logo Cap	2014	19825
AWC Logo Cap	2015	13913
AWC Logo Cap	2016	19242
Bike Wash - Dissolver	2013	5969
Bike Wash - Dissolver	2014	8016
Bike Wash - Dissolver	2015	5795
Bike Wash - Dissolver	2016	7799
Classic Vest; L	2013	970
Classic Vest; L	2014	1378
Classic Vest; L	2015	934
Classic Vest; L	2016	1369
Classic Vest; M	2013	1278
Classic Vest; M	2014	1217
Classic Vest; M	2015	1219
Classic Vest; M	2016	1162
Classic Vest; S	2013	1160
Classic Vest; S	2014	1366
Classic Vest; S	2015	1100
Classic Vest; S	2016	1317

Hình 3.37: Câu truy vấn 2 trong Visual Studio

❖ MDX:

Query:

```
SELECT NON EMPTY {[Dim Time].[Year].Children} ON COLUMNS,
       NON EMPTY {[Dim Product].[Product].Children} ON ROWS
  FROM [Bike Sales]
 WHERE [Measures].[Fact Count]
```

Kết quả:

The screenshot shows the SSAS_BikeSales cube in the MDXQuery1.mdx editor. The query is:

```
SELECT NON EMPTY {[Dim Time].[Year].Children} ON COLUMNS,
NON EMPTY {[Dim Product].[Product].Children} ON ROWS
FROM [Bike Sales]
WHERE [Measures].[Fact Count]
```

The results grid displays data for various products across years from 2011 to 2016. The columns are labeled 2011, 2012, 2013, 2014, 2015, and 2016. The rows list products like All-Purpose Bike Stand, AWC Logo Cap, Bike Wash - Dissolver, etc.

	2011	2012	2013	2014	2015	2016
All-Purpose Bike Stand	(null)	117	109	117	109	109
AWC Logo Cap	(null)	884	1295	884	1295	517
Bike Wash - Dissolver	(null)	385	517	385	517	517
Classic Vest; L	(null)	61	98	61	98	98
Classic Vest; M	(null)	69	91	69	91	91
Classic Vest; S	(null)	74	88	74	88	88
Fender Set - Mountain	(null)	882	1134	882	1134	1134
Half-Finger Gloves; L	(null)	185	258	185	258	258
Half-Finger Gloves; M	(null)	164	237	164	237	237
Half-Finger Gloves; S	(null)	231	266	231	266	266
Hitch Rack - 4-Bike	(null)	139	159	139	159	159
HL Mountain Tire	(null)	403	497	403	497	497
HL Road Tire	(null)	217	286	217	286	286
Hydration Pack - 70 oz	(null)	304	362	304	362	362
LL Mountain Tire	(null)	181	240	181	240	240
LL Road Tire	(null)	372	471	372	471	471

Hình 3.38: Câu truy vấn 2 trong SQL Server

❖ Pivot Table tên MS Excel

The screenshot shows a PivotTable in Excel with the following structure:

		Thống kê số lượng bán được của từng sản phẩm qua các năm						
		Order Quantity		Column Labels				
Row Labels		2011	2012	2013	2014	2015	2016	Grand Total
All-Purpose Bike Stand		663	582	617	538	538	2400	
AWC Logo Cap		14336	19825	13913	19242	19242	67316	
Bike Wash - Dissolver		5969	8016	5795	7799	7799	27579	
Classic Vest; L		970	1378	934	1369	1369	4651	
Classic Vest; M		1278	1217	1219	1162	1162	4876	
Classic Vest; S		1160	1366	1100	1317	1317	4943	
Fender Set - Mountain		14154	17350	13789	16825	16825	62118	
Half-Finger Gloves; L		2864	3812	2762	3715	3715	13153	
Half-Finger Gloves; M		2366	3698	2301	3629	3629	11994	
Half-Finger Gloves; S		3549	3987	3420	3838	3838	14794	
Hitch Rack - 4-Bike		1051	1388	1005	1309	1309	4753	
HL Mountain Tire		6439	7571	6234	7318	7318	27562	
HL Road Tire		3420	4472	3342	4376	4376	15610	
Hydration Pack - 70 oz.		4752	5314	4619	5172	5172	19857	

The PivotTable Fields pane on the right shows the following field settings:

- Report Filter: Year (selected)
- Columns: Fact (selected)
- Rows: Product (selected)
- Values: Order Quantity (selected)

Hình 3.39: Câu truy vấn 2 trong MS Excel

3.7.3. Thống kê doanh thu trên từng quốc gia của khách hàng trong tháng 8 năm 2015

❖ SSAS:

Country	Year	Month	Revenue
Australia	2015	August	454392
Canada	2015	August	256797
France	2015	August	210855
Germany	2015	August	191591
United Kingdom	2015	August	254379
United States	2015	August	772567

Hình 3.40: Câu truy vấn 3 trong Visual Studio

❖ MDX: Thực hiện câu truy vấn và kết quả
Query:

```

SELECT
    NON EMPTY { [Measures].[Revenue] } ON COLUMNS,
    NON EMPTY {
        ([Dim State].[Country].[Country].ALLMEMBERS *
        [Dim Time].[Year].[Year].ALLMEMBERS *
        [Dim Time].[Month].[Month].ALLMEMBERS )
    } ON ROWS
FROM
    (
        SELECT (
            { [Dim Time].[DateOrder].[Month].&[2015]&[August] }
            ON COLUMNS
            FROM [Bike Sales]
        )
WHERE
    ( [Dim Time].[DateOrder].[Month].&[2015]&[August] );

```

Kết quả:

The screenshot shows the SQL Server Management Studio interface. On the left, the Object Explorer displays a cube named 'Bike Sales' with various dimensions and measures. In the center, the Results pane shows the output of an MDX query:

```

SELECT NON EMPTY {[Measures].[Revenue]} ON 0,
NON EMPTY CrossJoin ([Dim State].[Country].Children,
[Dim Time].[Month].Children, [Dim Time].[Year].Children ) ON 1
FROM (SELECT ({[Dim Time].[Year]&[2015]}) ON 0
FROM ( SELECT ({[Dim Time].[Month]&[August]}) ON 0
FROM [Bike Sales])

```

The Results pane displays a table with the following data:

Country	Month	Year	Revenue
Australia	August	2015	454392
Canada	August	2015	256797
France	August	2015	210855
Germany	August	2015	191591
United Kingdom	August	2015	254379
United States	August	2015	772567

At the bottom, a message indicates: "Query executed successfully."

Hình 3.41: Câu truy vấn 3 trong SQL Server

❖ Pivot Table tên MS Excel:

The screenshot shows a Microsoft Excel spreadsheet titled "Thống kê doanh thu trên từng quốc gia của khách hàng trong tháng 8 năm 2015". The PivotTable Fields pane on the right shows the following fields:

- Report Filter:** Month (August), Year (2015)
- Columns:** None
- Rows:** Country
- Values:** Revenue

The PivotTable displays the following data:

Country	Revenue
Australia	454392
Canada	256797
France	210855
Germany	191591
United Kingdom	254379
United States	772567
Grand Total	2140581

Hình 3.42: Câu truy vấn 3 trong MS Excel

3.7.4. Thống kê tổng số lượt bán qua từng năm của từng nhóm độ tuổi khách hàng

❖ SSAS:

Age Group	Year	Fact Count
Adults (35-64)	2011	1207
Adults (35-64)	2012	1207
Adults (35-64)	2013	11236
Adults (35-64)	2014	15469
Adults (35-64)	2015	11236
Adults (35-64)	2016	15469
Seniors (64+)	2011	7
Seniors (64+)	2012	7
Seniors (64+)	2013	131
Seniors (64+)	2014	227
Seniors (64+)	2015	131
Seniors (64+)	2016	227
Young Adults (25-34)	2011	974
Young Adults (25-34)	2012	974
Young Adults (25-34)	2013	8785
Young Adults (25-34)	2014	9566
Young Adults (25-34)	2015	8785
Young Adults (25-34)	2016	9568
Youth (<25)	2011	489
Youth (<25)	2012	489
Youth (<25)	2013	4291
Youth (<25)	2014	4134
Youth (<25)	2015	4291
Youth (<25)	2016	4134

Hình 3.43: Câu truy vấn 4 trong Visual Studio

❖ MDX: Thực hiện câu truy vấn và kết quả
Query:

```

SELECT
    NON EMPTY {[Dim Age Group].[Age Group].Children} ON COLUMNS,
    NON EMPTY {[Dim Time].[Year].Children} ON ROWS
FROM [Bike Sales]
WHERE
    [Measures].[Fact Count]
  
```

Kết quả:

MDXQuery1.mdx ...HCHANHPC\minhc* | NON EMPTY {[Dim Age Group].[Age Group].Children} ON COLUMNS,
NON EMPTY {[Dim Time].[Year].Children} ON ROWS
FROM [Bike Sales]
WHERE
[Measures].[Fact Count]

	Adults (35-64)	Seniors (64+)	Young Adults (25-34)	Youth (<25)
2011	1207	7	974	489
2012	1207	7	974	489
2013	11236	131	8785	4291
2014	15469	227	9568	4134
2015	11236	131	8785	4291
2016	15469	227	9568	4134

Query executed successfully. MINHCHANHPC | MINHCHANHPC\minhc | SSAS_BikeSales | 00:00:01

Hình 3.44: Câu truy vấn 4 trong SQL Server

❖ Pivot Table tên MS Excel:

	Fact Count	Column Labels	Row Labels	Adults (35-64)	Seniors (64+)	Young Adults (25-34)	Youth (<25)	Grand Total
5	2011			1207	7	974	489	2677
6	2012			1207	7	974	489	2677
7	2013			11236	131	8785	4291	24443
8	2014			15469	227	9568	4134	29398
9	2015			11236	131	8785	4291	24443
10	2016			15469	227	9568	4134	29398
11		Grand Total		55824	730	38654	17828	113036

Hình 3.45: Câu truy vấn 4 trong MS Excel

3.7.5. Thống kê top 5 bang thuộc nước Pháp có nhiều lượt mua từ khách hàng nhất

❖ SSAS:

Script Organizer: Top5StateInFrance

TopCount(
DrilldownLevel({[Dim State].[Country].&[France]}) *

```
[Dim State].[State].Children),
5,
[Measures].[Fact Count]
```

The screenshot shows the Microsoft Analysis Services (AS) Management Studio interface. The title bar says "Bike Sales.cube [Design]". The main area has tabs for "Edit as Text" and "MDX". The MDX query is displayed above a results grid.

Dimension Filter:

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim State	State	Custom	[Top5StateInFrance]	

Results Grid:

Country	State	Fact Count
France	Essonne	994
France	Hauts de Seine	1084
France	Nord	1670
France	Seine (Paris)	2328
France	Seine Saint Denis	1684

The left sidebar shows the cube structure with dimensions like Bike Sales, Measures, Fact, and Dim State, and various measures like Cost, Profit, Revenue, etc.

Hình 3.46: Câu truy vấn 5 trong Visual Studio

❖ **MDX: Thực hiện câu truy vấn và kết quả**
Query:

```
SELECT [Measures].[Fact Count] ON 0,
       ORDER(
           TopCount(
               [Dim State].[Country].&[France] *
               [Dim State].[State].Children,
               5,
               [Measures].[Fact Count]
           ), [Measures].[Fact Count], DESC
       ) ON 1
FROM [Bike Sales]
```

Kết quả:

The screenshot shows the SSMS interface with an MDX query window and a results grid.

```

SELECT [Measures].[Fact Count] ON 0,
       ORDER(
           TopCount(
               [Dim State].[Country].&[France] *
               [Dim State].[State].Children,
               5,
               [Measures].[Fact Count]
           ), [Measures].[Fact Count], DESC
       ) ON 1
      FROM [Bike Sales]
  
```

Results:

		Fact Count
France	Seine (Paris)	2328
France	Seine Saint Denis	1684
France	Nord	1670
France	Hauts de Seine	1084
France	Essonne	994

Query executed successfully. | MINHCHANHPC | MINHCHANHPC\minhc | SSAS_BikeSales | 00:00:01

Hình 3.47: Câu truy vấn 5 trong SQL Server

❖ Pivot Table tên MS Excel:

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable Fields pane is open on the right side.

PivotTable Fields pane:

- Choose fields to add to report: Country
- Filters: Country
- Columns: State
- Rows: State
- Values: Fact Count

PivotTable Data:

Country	Fact Count
Esso	994
Hauts de Seine	1084
Nord	1670
Seine (Paris)	2328
Seine Saint Denis	1684
Grand Total	7760

Hình 3.48: Câu truy vấn 5 trong MS Excel

3.7.6. Thống kê danh sách các sản phẩm có trên 1000 lượt bán và tổng số lượng bán trên 50000 sản phẩm

❖ SSAS:

Script Organizer: TopProduct

```

INTERSECT(
    FILTER([Dim Product].[Product].Children,
          [Measures].[Fact Count] > 1000),
    ...
)
  
```

```

    FILTER([Dim Product].[Product].Children,
          [Measures].[Order Quantity]>50000)
)

```

Product	Fact Count	Order Quantity
AWC Logo Cap	4358	67316
Fender Set - Mountain	4032	62118
Mountain Tire Tube	6816	102792
Patch Kit/8 Patches	10416	157583
Road Tire Tube	4176	62296
Sport-100 Helmet; Black	4152	62105
Sport-100 Helmet; Blue	3788	55895
Sport-100 Helmet; Red	4220	63663
Touring Tire Tube	3702	56802
Water Bottle - 30 oz.	10794	164086

Hình 3.49: Câu truy vấn 6 trong Visual Studio

❖ **MDX: Thực hiện câu truy vấn và kết quả**
Query:

```

SELECT
  {[Measures].[Fact Count], [Measures].[Order Quantity]} ON 0,
  {[Dim Product].[Product].Children} ON 1
FROM
  (
    SELECT
      (
        INTERSECT(
          FILTER(
            [Dim Product].[Product].Children,
            [Measures].[Fact Count] > 1000
          ),
          FILTER(
            [Dim Product].[Product].Children,
            [Measures].[Order Quantity] > 50000
          ))) ON 0
  FROM [Bike Sales])

```

Kết quả:

The screenshot shows the SSMS interface with an MDX query window and a results grid.

```

MDXQuery1.mdx -...HCHANHPC\minhc*
Cube: Bike Sales
SELECT { [Measures].[Fact Count], [Measures].[Order Quantity] } ON 0,
{([Dim Product].[Product].Children) } ON 1
FROM ( SELECT
( INTERSECT(
    FILTER([Dim Product].[Product].Children,
        [Measures].[Fact Count] > 1000),
    FILTER([Dim Product].[Product].Children,
        [Measures].[Order Quantity]>50000)
)) ON 0
FROM [Bike Sales])
    
```

The results grid displays the following data:

	Fact Count	Order Quantity
AWC Logo Cap	4358	67316
Fender Set - Mountain	4032	62118
Mountain Tire Tube	6816	102792
Patch Kit/8 Patches	10416	157583
Road Tire Tube	4176	62296
Sport-100 Helmet; Black	4152	62105
Sport-100 Helmet; Blue	3788	55895
Sport-100 Helmet; Red	4220	63663
Touring Tire Tube	3702	56802
Water Bottle - 30 oz.	10794	164086

Message bar: Query executed successfully.

Hình 3.50: Câu truy vấn 6 trong SQL Server

❖ Pivot Table tên MS Excel:

The screenshot shows an Excel spreadsheet with a PivotTable report titled "Thống kê danh sách các sản phẩm có trên 1000 lượt bán và tổng số lượng bán trên 50000 sản phẩm".

	Fact Count	Order Quantity
AWC Logo Cap	4358	67316
Fender Set - Mountain	4032	62118
Mountain Tire Tube	6816	102792
Patch Kit/8 Patches	10416	157583
Road Tire Tube	4176	62296
Sport-100 Helmet; Black	4152	62105
Sport-100 Helmet; Blue	3788	55895
Sport-100 Helmet; Red	4220	63663
Touring Tire Tube	3702	56802
Water Bottle - 30 oz.	10794	164086
Grand Total	56454	854656

PivotTable Fields pane:

- Rows: Product
- Values: Fact Count, Order Quantity
- Fields: Fact, Cost, Customer Age, Order Quant..., Profit, Revenue, Unit Cost, Unit Price, Dim Age Group, Dim Gender, Customer Ge..., Dim Product, Product, Dim State, Country

Hình 3.51: Câu truy vấn 6 trong MS Excel

3.7.7. Thống kê danh sách và số lượng lượt bán của các sản phẩm trừ Water Bottle – 30oz. của các khách hàng là Nữ vào ngày 5/5/2015

❖ SSAS:

The screenshot shows the Microsoft Analysis Services (SSAS) Management Studio interface. The title bar says "Bike Sales.cube [Design] - Dim Time.dim [Design]". The left pane is a tree view of the cube structure, showing dimensions such as Dim Gender, Dim Product, Dim State, and Dim Time. The Dim Time dimension is currently selected. The right pane has two main sections: a "Filter Expression" grid and a preview table. The filter expression grid shows three rows: Dim Gender (Customer Gender, Equal, {F}), Dim Time (DateOrder, Equal, {5}), and Dim Product (Product, Not Equal, {Water Bottle - 30 oz.}). Below this is a table titled "<Select dimension>" with columns State, Product, Customer Gender, Day, and Fact Count. The table contains four rows of data:

State	Product	Customer Gender	Day	Fact Count
California	Mountain-200 Silver; 42	F	5	1
New South Wales	Mountain-100 Black; 42	F	5	1
Victoria	Road-150 Red; 48	F	5	1
Washington	Mountain-200 Black; 46	F	5	1

Hình 3.52: Câu truy vấn 7 trong SQL Server

❖ MDX: Thực hiện câu truy vấn và kết quả

Query:

```

SELECT
    NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
    NON EMPTY {
        (
            [Dim State].[State].[State].ALLMEMBERS *
            [Dim Product].[Product].[Product].ALLMEMBERS *
            [Dim Gender].[Customer Gender].[Customer
Gender].ALLMEMBERS *
            [Dim Time].[Day].[Day].ALLMEMBERS
        )
    } ON ROWS
FROM
    (
        SELECT (
            -{ [Dim Product].[Product].&[Water Bottle - 30 oz.] }
        ) ON COLUMNS
        FROM
            (
                SELECT (
                    { [Dim Time].[DateOrder].[Day].&[2015]&[May]&[5] }
                ) ON COLUMNS
                FROM

```

```

(
    SELECT (
        { [Dim Gender].[Customer Gender].&[F] }
    ) ON COLUMNS
    FROM [Bike Sales]
)
)
)
WHERE ( [Dim Time].[DateOrder].[Day].&[2015]&[May]&[5] ) ;

```

Kết quả:

The screenshot shows a Microsoft SQL Server Management Studio (SSMS) interface. On the left, there is a cube browser pane titled "MDX_BikeSales.m...HCHANHPC\minhc". It lists various dimensions and measures, such as Bike Sales, Dim Age Group, Dim Product, Dim State, Dim Sub Category, and Dim Time. The main pane displays an MDX query:

```

SELECT (
    { [Dim Gender].[Customer Gender].&[F] }
) ON COLUMNS
FROM [Bike Sales]
)
)
)
WHERE ( [Dim Time].[DateOrder].[Day].&[2015]&[May]&[5] ) ;

```

Below the query, the results pane shows a table with four rows of data:

			Fact Count
California	Mountain-200 Silver, 42	F	5
New South Wales	Mountain-100 Black, 42	F	5
Victoria	Road-150 Red, 48	F	5
Washington	Mountain-200 Black, 46	F	5

At the bottom of the window, a status bar indicates: "Query executed successfully." and "MINHCHANHPC | MINHCHANHPC\minhc | SSAS_BikeSales | 00:00:01".

Hình 3.53: Câu truy vấn 7 trong SQL Server

❖ Pivot Table tên MS Excel:

	Fact Count
Row Labels	
California	
Mountain-200 Silver; 42	1
New South Wales	
Mountain-100 Black; 42	1
Victoria	
Road-150 Red; 48	1
Washington	
Mountain-200 Black; 46	1
Grand Total	4

Hình 3.54: Câu truy vấn 7 trong MS Excel

3.7.8. Thống kê top 3 danh mục phụ có lợi nhuận lớn nhất.

❖ SSAS:

Script Organizer: Top3SubCate

```
TopCount(
    [Dim Sub Category].[Sub Category].CHILDREN,
    3,
    [Measures].[Profit]
)
```

Sub Category	Profit
Helmets	3380203
Mountain Bikes	8160463
Road Bikes	10078875

Hình 3.55: Câu truy vấn 8 trong Visual Studio

❖ **MDX: Thực hiện câu truy vấn và kết quả**
Query:

```

SELECT
    [Measures].[Profit] ON 0,
    NON EMPTY
        TopCount(
            [Dim Sub Category].[Sub Category].CHILDREN,
            3,
            [Measures].[Profit]
        ) ON 1
FROM [Bike Sales];

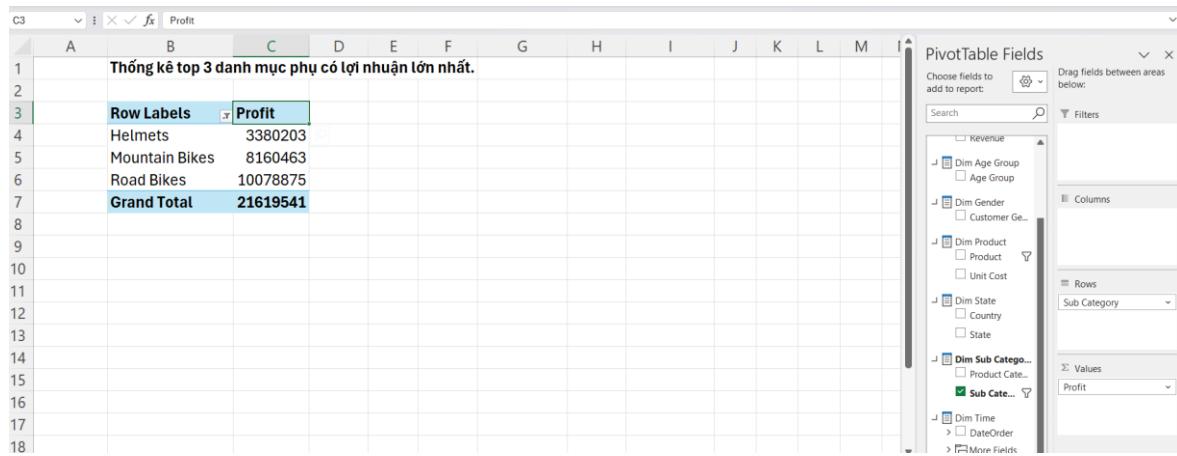
```

Kết quả:

Product Category	Profit
Road Bikes	10078875
Mountain Bikes	8160463
Helmets	3380203

Hình 3.56: Câu truy vấn 8 trong SQL Server

❖ Pivot Table tên MS Excel:



Hình 3.57: Câu truy vấn 8 trong MS Excel

3.7.9. Thống kê 3 tháng trong mỗi năm có nhiều lượt mua hàng nhất, tính tổng doanh thu của các tháng.

Script Organizer: **TopMonthYear**

```
Generate(
    [Dim Time].[DateOrder].[Year].MEMBERS,
    TopCount(
        Descendants([Dim Time].[DateOrder].CURRENTMEMBER,
        [Dim Time].[DateOrder].[Month]),
        3,
        [Measures].[Revenue]
    )
)
```

❖ SSAS:

Year	Month	Revenue
2011	August	880037
2011	December	1132666
2011	July	882219
2012	August	927504
2012	December	1110101
2012	July	863284
2013	December	2995953
2013	November	2232077
2013	October	2041128
2014	April	2204586
2014	June	2786120
2014	May	2587491
2015	December	3848211
2015	November	2836222
2015	October	2529084
2016	April	2756864
2016	June	3586300
2016	May	3264343

Hình 3.58: Câu truy vấn 9 trong Visua Studio

❖ MDX: Thực hiện câu truy vấn và kết quả

Query:

```

SELECT
    NON EMPTY {[Measures].[Revenue]} ON COLUMNS,
    NON EMPTY
        Generate(
            [Dim Time].[DateOrder].[Year].MEMBERS,
            TopCount(
                Descendants([Dim Time].[DateOrder].CURRENTMEMBER,
                [Dim Time].[DateOrder].[Month]),
                3,
                [Measures].[Revenue]
            )
        ) ON ROWS
    FROM [Bike Sales]

```

Kết quả:

The screenshot shows the SSMS interface with an MDX query window and a results grid.

```

SELECT
    NON EMPTY { [Measures].[Revenue] } ON COLUMNS,
    NON EMPTY
        Generate(
            [Dim Time].[DateOrder].[Year].MEMBERS,
            TopCount(
                Descendants([Dim Time].[DateOrder].CURRENTMEMBER, [Dim Time].[DateOrder].[Month]),
                3,
                [Measures].[Revenue]
            )
        ) ON ROWS
    FROM [Bike Sales]
  
```

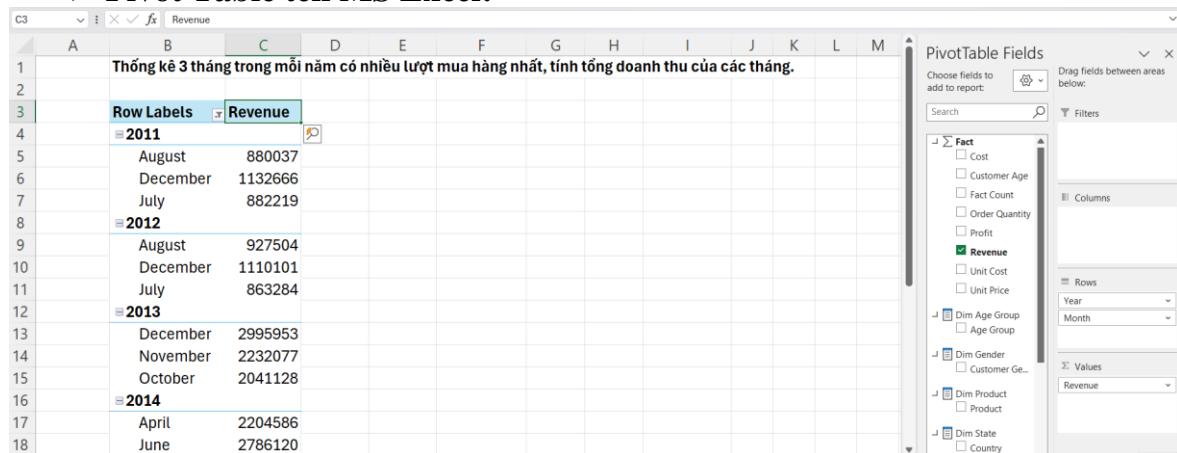
Results Grid:

	Revenue
December	1132666
July	882219
August	880037
December	1110101
August	927504
July	863284
December	2995953
November	2232077
October	2041128
June	2786120
May	2587491
April	2204586
December	3848211
November	2836222
October	2529084
June	3586300
...	...

Query executed successfully. | MINHCHANHPC | MINHCHANHPC\minhc | SSAS_BikeSales | 00:00:01

Hình 3.59: Câu truy vấn 9 trong SQL Server

❖ Pivot Table tên MS Excel:



Hình 3.60: Câu truy vấn 9 trong MS Excel

3.7.10. Thống kê số lượng sản phẩm trên từng nhóm tuổi và giới tính.

❖ SSAS:

The screenshot shows the Microsoft Analysis Services (SSAS) Management Studio interface. The title bar reads "Dim Time.dim [Design] Bike Sales.cube [Design]". The main window has several tabs at the top: Cube Structure, Dimension Usage, Calculations, KPIs, Actions, Partitions, Aggregations, Perspectives, Translations, and Browser. Below the tabs is a toolbar with icons for Edit as Text, Import..., MDX, and various navigation and search functions. On the left, there's a navigation pane titled "Bike Sales" containing "Measures", "Fact", "Customer Age", "Order Quantity", "Profit", "Revenue", "Unit Cost", "Unit Price", "KPIs", "Dim Age Group", "Age Group", "Dim Gender", "Customer Gender", "Dim Product", "Top3ProductEachYear", "TopProduct", "Product", "Dim State", "Country", "State", and "Dim Sub Category". A "Calculated Members" section is also present. The right side of the interface displays a results grid with three columns: "Age Group", "Customer Gender", and "Order Quantity". The data in the grid is as follows:

Age Group	Customer Gender	Order Quantity
Adults (35-64)	F	328479
Adults (35-64)	M	343938
Seniors (64+)	F	4506
Seniors (64+)	M	5371
Young Adults (25-34)	F	214441
Young Adults (25-34)	M	232434
Youth (<25)	F	98307
Youth (<25)	M	117840

Hình 3.61: Câu truy vấn 9 trong Visual Studio

❖ MDX: Thực hiện câu truy vấn và kết quả

Query:

```

SELECT
    NON EMPTY {
        [Dim Gender].[Customer Gender].[Customer Gender].MEMBERS
    } ON COLUMNS,
    NON EMPTY {
        [Dim Age Group].[Age Group].[Age Group].MEMBERS
    } ON ROWS
FROM [Bike Sales]
WHERE ([Measures].[Order Quantity])
  
```

Kết quả:

The screenshot shows the SQL Server Management Studio interface. On the left, the Object Explorer displays a cube named 'Bike Sales' with various dimensions like Dim Age Group, Dim Gender, etc., and measures like Order Quantity. In the center, the Results pane shows the output of an MDX query:

```

SELECT
    NON EMPTY {
        [Dim Gender].[Customer Gender].[Customer Gender].MEMBERS
    } ON COLUMNS,
    NON EMPTY {
        [Dim Age Group].[Age Group].[Age Group].MEMBERS
    } ON ROWS
FROM [Bike Sales]
WHERE ([Measures].[Order Quantity])
  
```

The results are presented in a table:

	F	M
Adults (35-64)	328479	343938
Seniors (64+)	4506	5371
Young Adults (25-34)	214441	232434
Youth (<25)	98307	117840

At the bottom, a message indicates "Query executed successfully." and shows the session details: MINHCHANHPC / MINHCHANHPC\minhc / SSAS_BikeSales 00:00:01.

Hình 3.62: Câu truy vấn 9 trong SQL Server

❖ Pivot Table tên MS Excel:

The screenshot shows a Microsoft Excel spreadsheet with a PivotTable. The table has the following structure:

Thống kê số lượng sản phẩm trên từng nhóm tuổi và giới tính.			
	Order Quantity	Column Labels	
Row Labels		F	M
Adults (35-64)		328479	343938
Seniors (64+)		4506	5371
Young Adults (25-34)		214441	232434
Youth (<25)		98307	117840
Grand Total	645733	699583	1345316

To the right of the table, the PivotTable Fields pane is open, showing the following field settings:

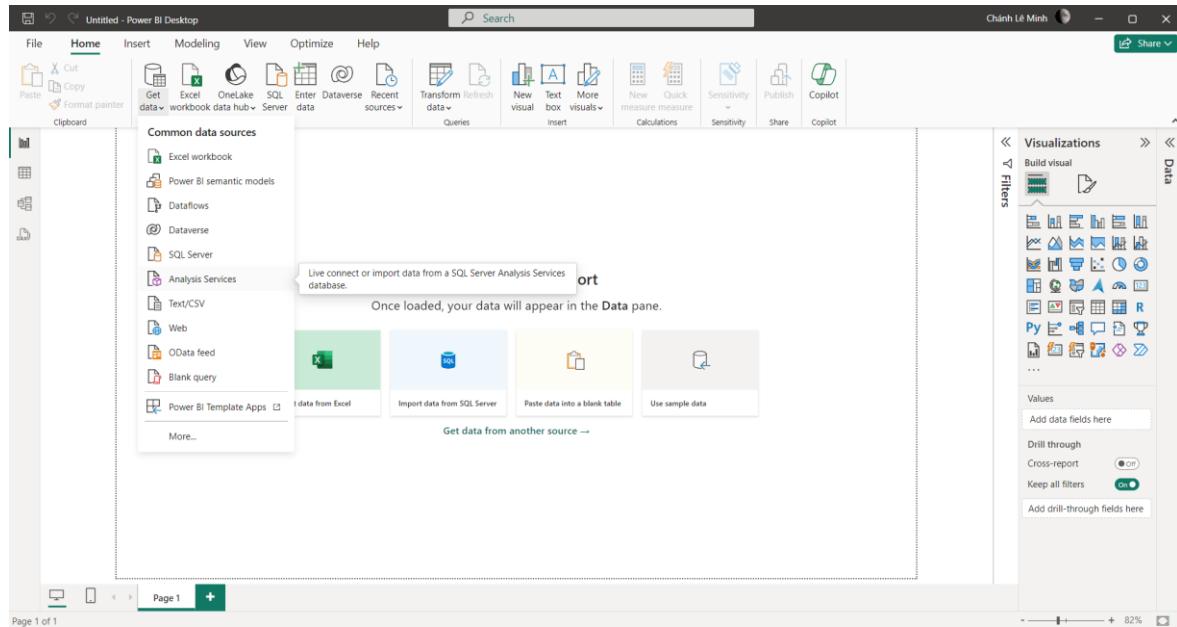
- Values:** Order Quantity
- Columns:** Customer Gender (selected)
- Rows:** Age Group (selected)
- Report Filter:** Order Quantity

Hình 3.63: Câu truy vấn 9 trong MS Excel

Phần 4. REPORT BẰNG CÔNG CỤ POWER BI (SSRS)

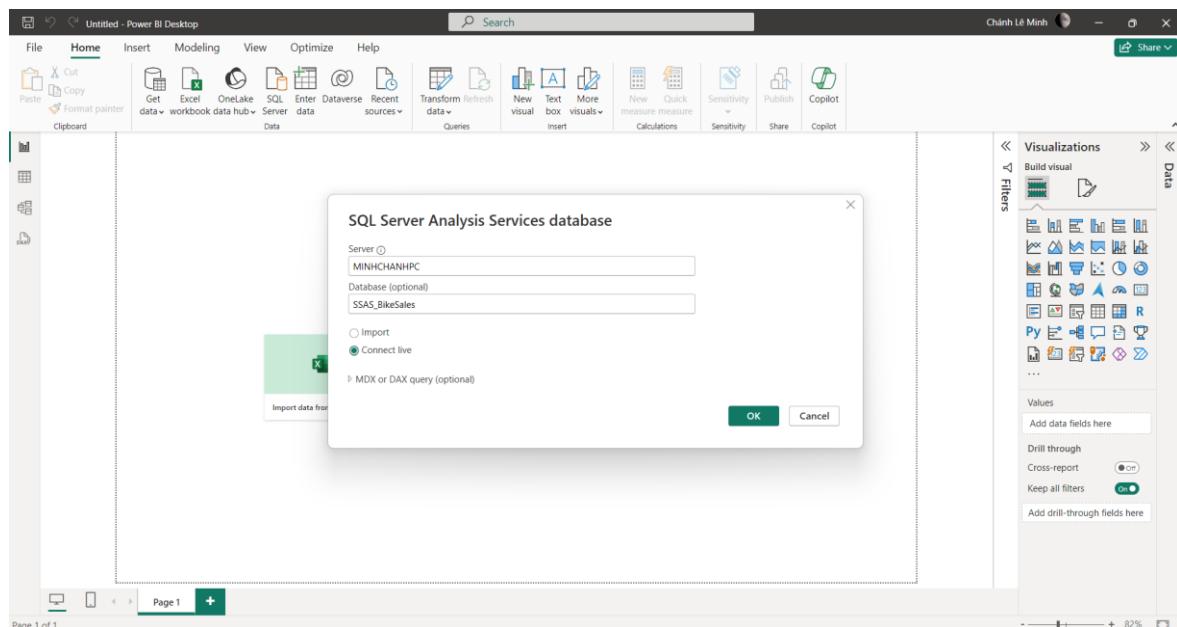
4.1. Đỗ dữ liệu vào Project Power BI

Bước 1: Chọn Get data → Chọn Analysis Server



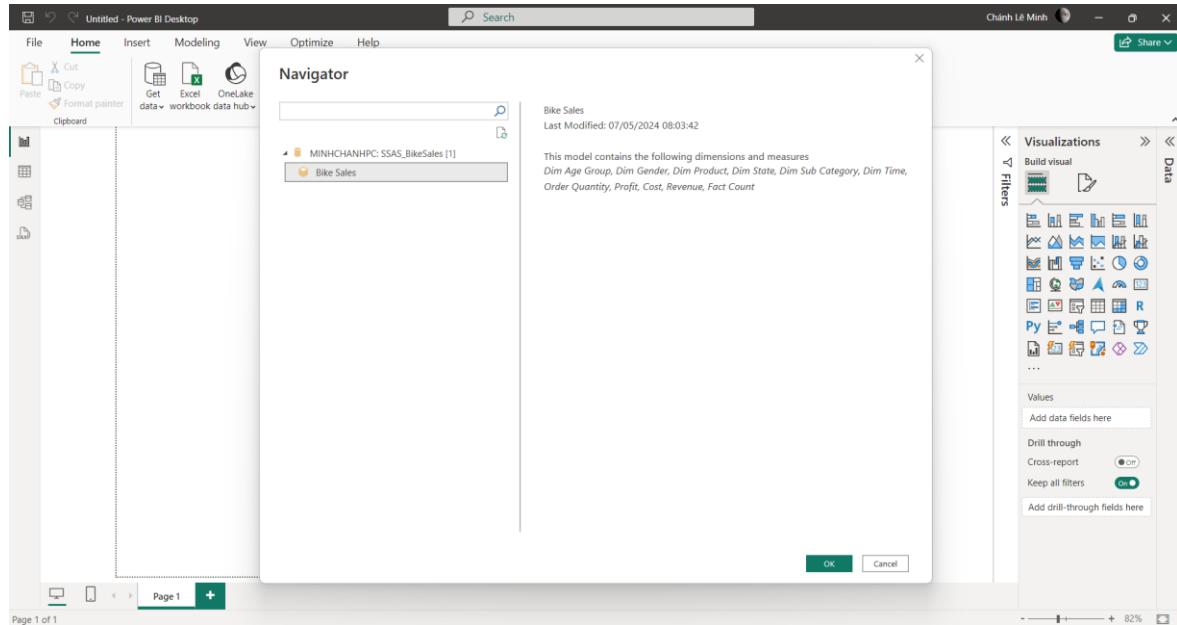
Hình 4.1: Liên kết Power BI đến Analysis Service

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



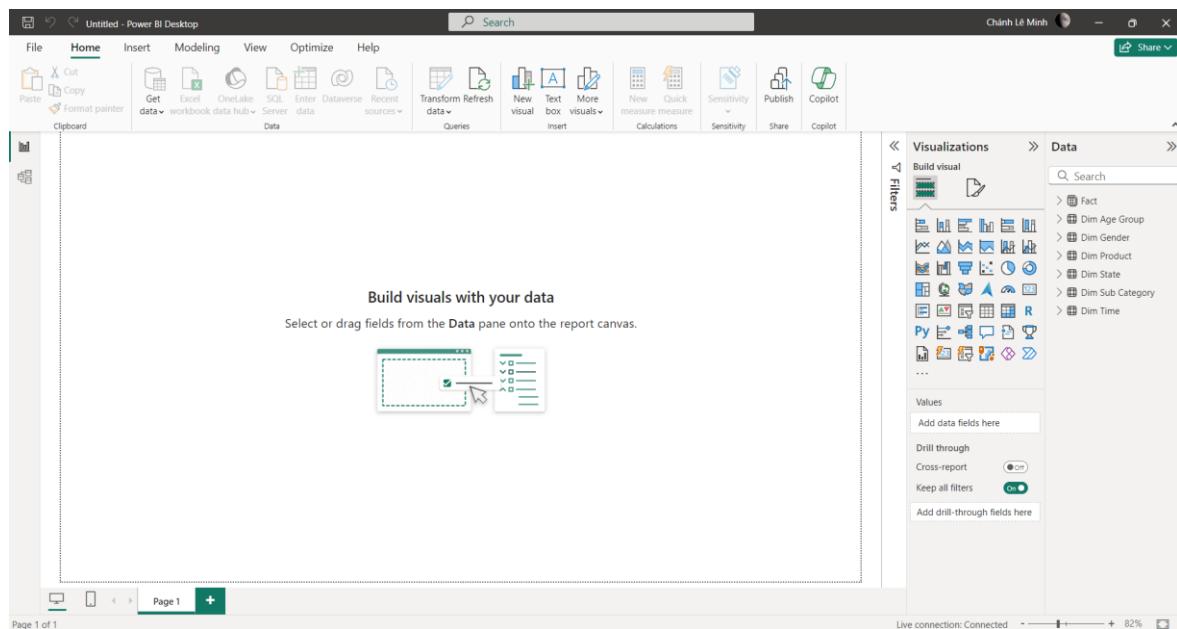
Hình 4.2: Điền thông tin Server và Database

Bước 3: Chọn Cube chứa thông tin để lấy dữ liệu → Nhận OK



Hình 4.3: Chọn Cube

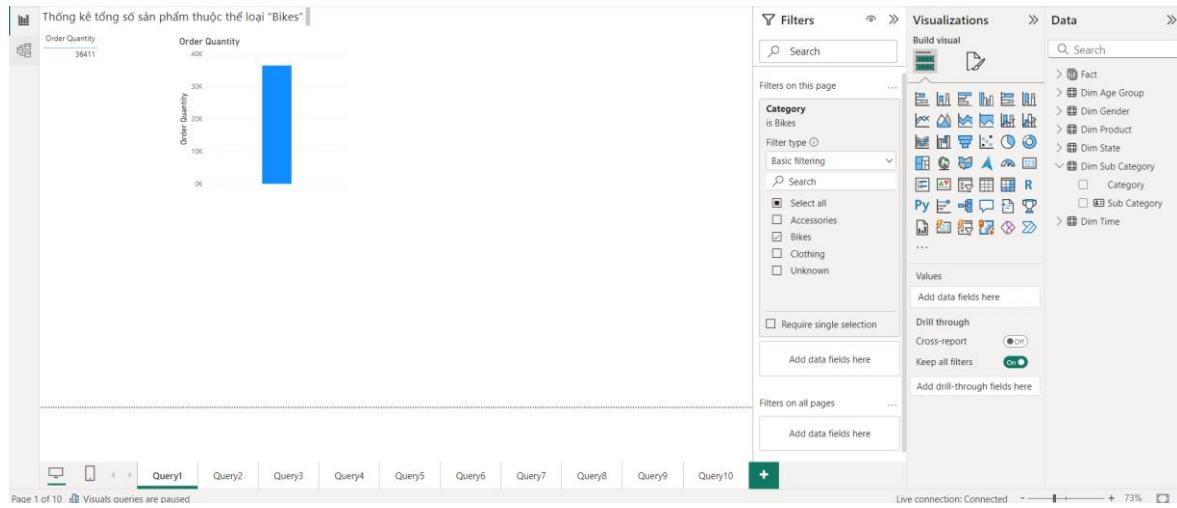
Như vậy chúng ta đã có thể thực hiện tạo các report trên Power BI tại đây



Hình 4.4: Giao diện tạo Report trên Power BI

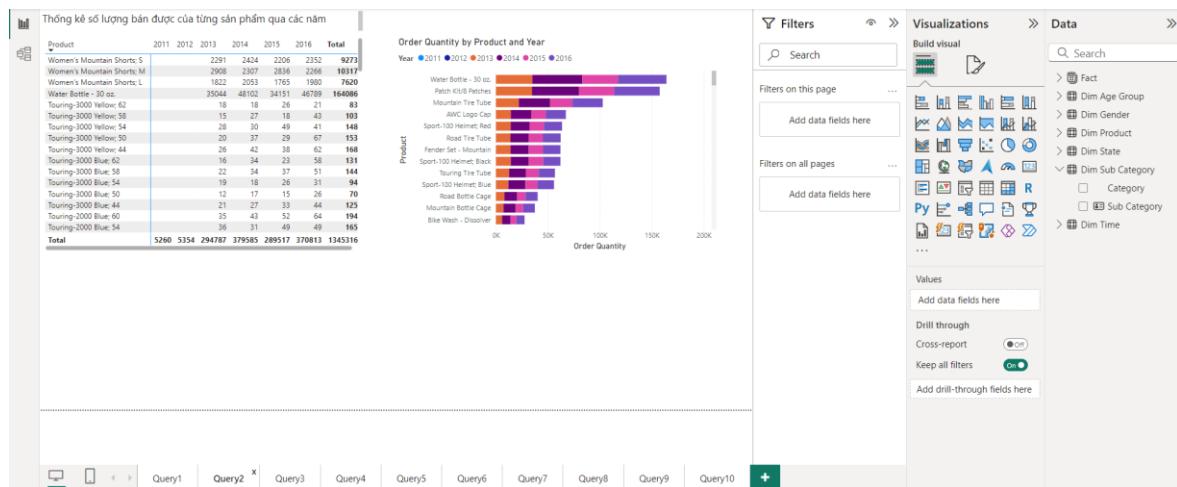
4.2. Thực hiện SSRS

4.2.1. Thống kê tổng số sản phẩm thuộc thể loại “Bikes”



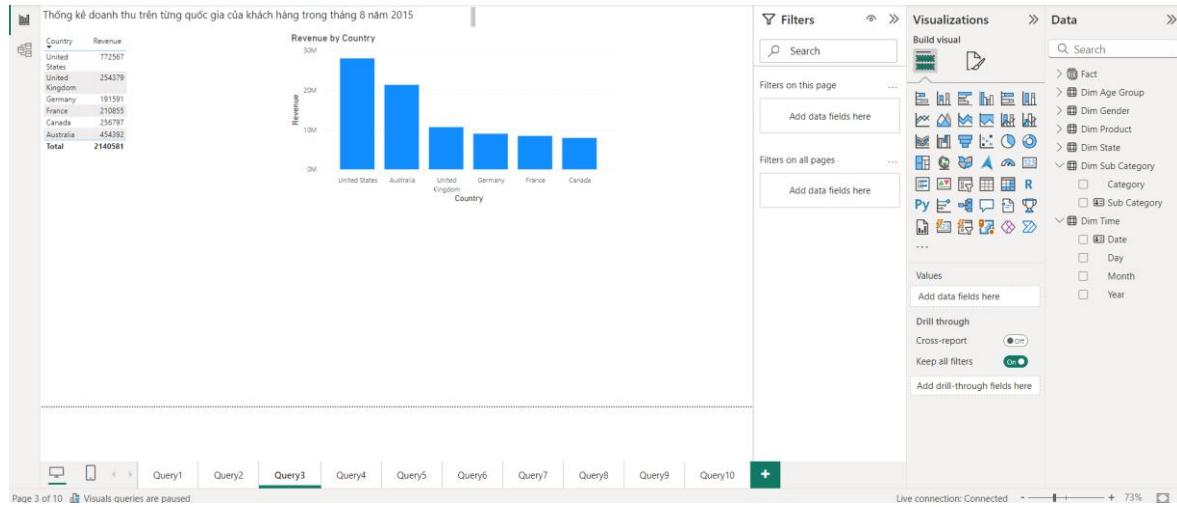
Hình 4.5: Câu truy vấn 1 trong Power BI

4.2.2. Thống kê số lượng bán được của từng sản phẩm qua các năm



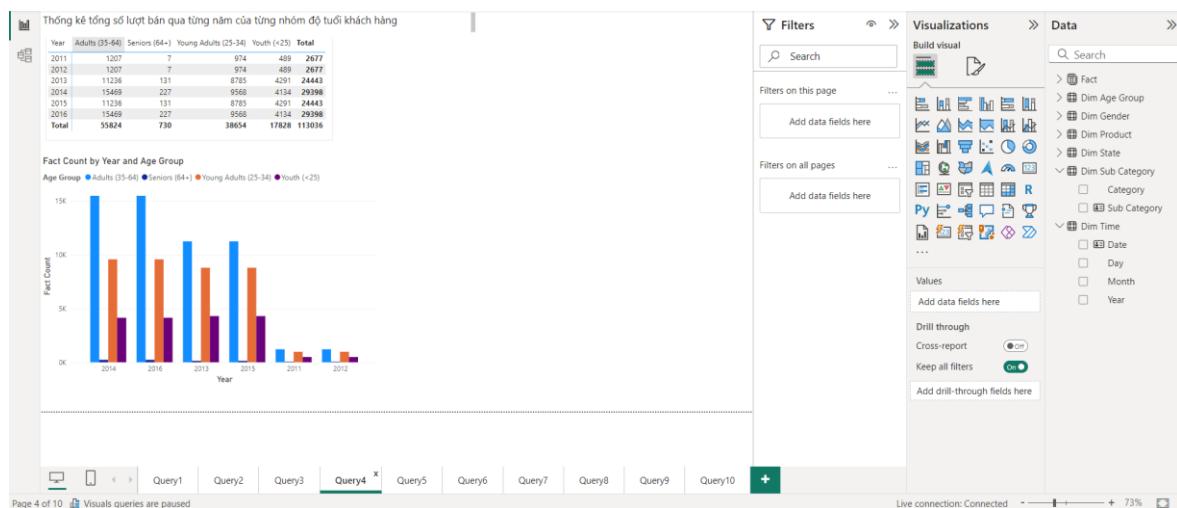
Hình 4.6: Câu truy vấn 2 trong Power BI

4.2.3. Thống kê doanh thu trên từng quốc gia của khách hàng trong tháng 8 năm 2015



Hình 4.7: Câu truy vấn 3 trong Power BI

4.2.4. Thống kê tổng số lượt bán qua từng năm của từng nhóm độ tuổi khách hàng



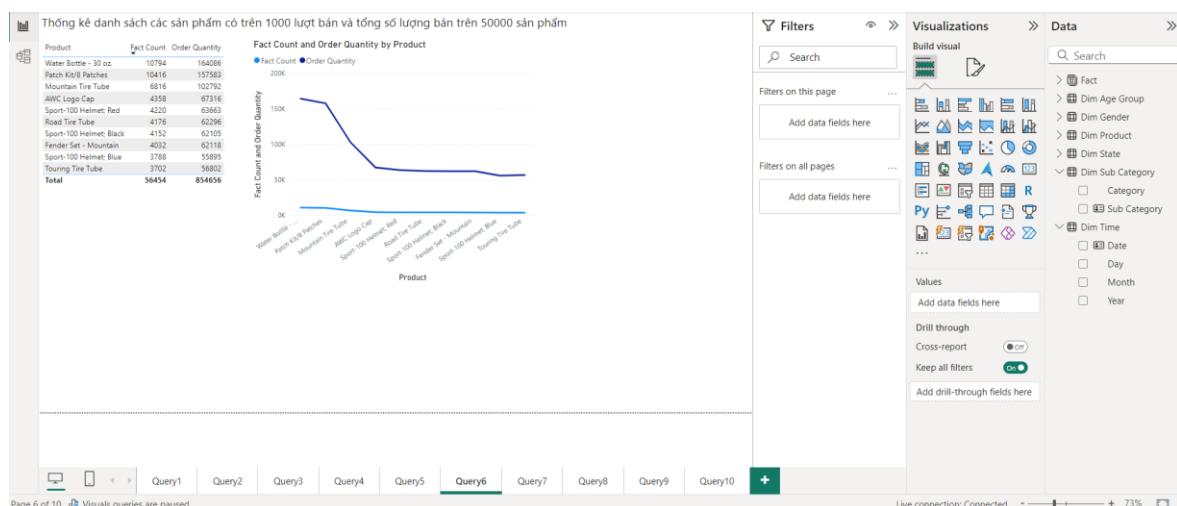
Hình 4.8: Câu truy vấn 4 trong Power BI

4.2.5. Thống kê top 5 bang thuộc nước Pháp có nhiều lượt mua từ khách hàng nhất



Hình 4.9: Câu truy vấn 5 trong Power BI

4.2.6. Thống kê danh sách các sản phẩm có trên 1000 lượt bán và tổng số lượng bán trên 50000 sản phẩm



Hình 4.10: Câu truy vấn 6 trong Power BI

4.2.7. Thống kê danh sách và số lượng lượt bán của các sản phẩm trừ Water Bottle – 30oz. của các khách hàng là Nữ vào ngày 5/5/2015

State	Product	Customer Gender	Fact Count	
California	Mountain-200 Silver	42	F	1
New South Wales	Mountain-100 Black	42	F	1
Victoria	Road-150 Red	48	F	1
Washington	Mountain-200 Black	46	F	1
Total				4

Hình 4.11: Câu truy vấn 7 trong Power BI

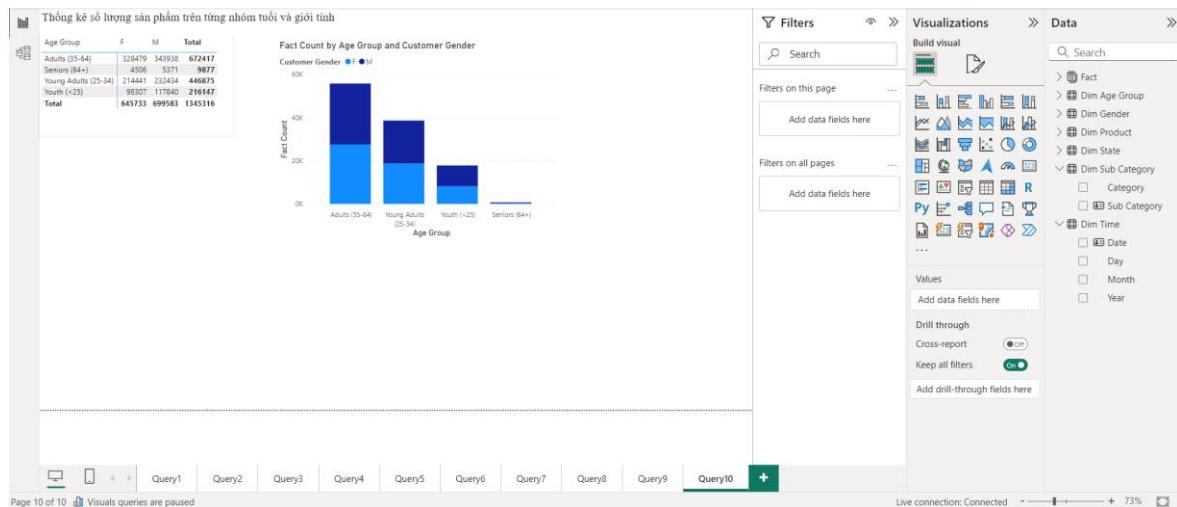
4.2.8. Thống kê top 3 danh mục phụ có lợi nhuận lớn nhất.

Sub Category	Profit
Helmets	3380203
Mountain Bikes	8160463
Road Bikes	10078875
Total	21619541

Hình 4.12: Câu truy vấn 7 trong Power BI

4.2.9. Thống kê 3 tháng trong mỗi năm có nhiều lượt mua hàng nhất, tính tổng doanh thu của các tháng.

4.2.10. Thống kê số lượng sản phẩm trên từng nhóm tuổi và giới tính.



Hình 4.13: Câu truy vấn 10 trong Power BI

Phần 5. DATA MINING

5.1. Giới thiệu

Trong bối cảnh hiện nay, dự đoán doanh thu là một phần quan trọng của việc quản lý kinh doanh, đặc biệt là trong ngành bán lẻ. Bộ dữ liệu Europe Bike Store Sales, một bộ dữ liệu doanh số bán hàng xe đạp tại Châu Âu được công bố trên trang Kaggle.com, cung cấp một cơ hội tuyệt vời để áp dụng các thuật toán dự đoán nhằm nâng cao hiệu quả kinh doanh. Chúng tôi đã triển khai năm thuật toán dự đoán khác nhau để dự đoán doanh thu trong 30 ngày sắp tới dựa trên bộ dữ liệu này. Các thuật toán được sử dụng bao gồm Linear Regression, Holt-Winter, Random Forest (RF), Long Short-Term Memory (LSTM), và ARIMA.

- **Linear Regression** là một trong những phương pháp đơn giản nhất nhưng rất hiệu quả trong việc dự đoán xu hướng. Thuật toán này giúp xác định mối quan hệ tuyến tính giữa doanh thu và các yếu tố khác trong dữ liệu, từ đó dự đoán doanh thu tương lai.
- **Holt-Winter** là một phương pháp dự báo theo mùa thích ứng với các thành phần xu hướng, mùa và ngẫu nhiên của dữ liệu. Đây là một thuật toán hữu ích khi dữ liệu doanh thu có tính mùa vụ rõ rệt, giúp dự đoán doanh thu chính xác hơn trong các khoảng thời gian có tính chu kỳ.
- **Random Forest (RF)** là một phương pháp học máy mạnh mẽ dựa trên tập hợp các cây quyết định. Bằng cách xây dựng nhiều cây quyết định và tổng hợp kết quả, RF có thể nắm bắt các mối quan hệ phức tạp trong dữ liệu và cải thiện độ chính xác của dự đoán.
- **Long Short-Term Memory (LSTM)** là một loại mạng nơ-ron hồi quy sâu (RNN) có khả năng ghi nhớ các giá trị trong các chuỗi thời gian dài. LSTM rất hữu ích trong việc dự đoán chuỗi thời gian vì nó có thể học và dự đoán các mẫu dài hạn và ngắn hạn trong dữ liệu doanh thu.
- **ARIMA (AutoRegressive Integrated Moving Average)** là một trong những phương pháp dự báo chuỗi thời gian phổ biến nhất. ARIMA kết hợp các thành

phân tự hồi quy, tích hợp và trung bình di chuyển để tạo ra các dự báo chính xác dựa trên dữ liệu lịch sử.

Mỗi thuật toán này có những ưu điểm riêng và được chọn để mang lại cái nhìn toàn diện về khả năng dự đoán doanh thu từ bộ dữ liệu Europe Bike Store Sales. Kết quả từ các phương pháp này sẽ được so sánh và phân tích để xác định phương pháp nào mang lại độ chính xác cao nhất và phù hợp nhất cho việc dự đoán doanh thu trong tương lai.

5.2. Các bước thực hiện

5.2.1. Tiền xử lý

Tiền hành nhập dữ liệu file csv, đổi định dạng cột ‘Date’ thành timestamp sau đó sắp xếp DataFrame theo cột ‘Date’.

```
data = pd.read_csv('EuropeBikeSales.csv')
data = data[['Date', 'Revenue']]

# Group by Date và tính tổng Revenue
grouped_df = data.groupby('Date').sum().reset_index()

# Chuyển đổi cột 'Date' thành kiểu datetime với định dạng 'dd/mm/yyyy'
grouped_df['Date'] = pd.to_datetime(grouped_df['Date'], format='%d/%m/%Y')

# Sắp xếp DataFrame theo cột 'Date'
df_sorted = grouped_df.sort_values(by='Date')

grouped_df.set_index('Date', inplace=True)

data = grouped_df.sort_index()
```

Hình 5.1: Tiền xử lí dữ liệu cho Data Mining

5.2.2. Phân chia dữ liệu

Phân chia tập dataset thành 2 tập train - test theo tỉ lệ 7:3.

```
train_size = round(len(data) * 0.7)
test_size = round(len(data) - train_size)

len(data), train_size, test_size
train_data = data[:train_size]
test_data = data[train_size:]

train_data.shape, test_data.shape
```

Hình 5.2: Chia tỉ lệ dữ liệu cho model ARIMA

5.2.3. Xây dựng thuật toán

5.2.3.1. ARIMA

Kiểm tra chuỗi dùng và xử lý

```
from statsmodels.tsa.stattools import adfuller, kpss

# Kiểm định ADF
adf_result = adfuller(df_sorted['Revenue'])
print('ADF Statistic: %f' % adf_result[0])
print('p-value: %f' % adf_result[1])
print('Critical Values:')
for key, value in adf_result[4].items():
    print('\t%s: %.3f' % (key, value))

# Diễn giải kết quả ADF
if adf_result[1] > 0.05:
    print("Chuỗi không dừng (không bác bỏ H0)")
else:
    print("Chuỗi dừng (bác bỏ H0)")

# Kiểm định KPSS
kpss_result = kpss(df_sorted['Revenue'], regression='c')
print('KPSS Statistic: %f' % kpss_result[0])
print('p-value: %f' % kpss_result[1])
print('Critical Values:')
for key, value in kpss_result[3].items():
    print('\t%s: %.3f' % (key, value))

# Diễn giải kết quả KPSS
if kpss_result[1] < 0.05:
    print("Chuỗi không dừng (bác bỏ H0)")
else:
    print("Chuỗi dừng (không bác bỏ H0)")
```

```
# Lấy sai phân bậc 1 của chuỗi thời gian
revenue_diff = df_sorted['Revenue'].diff().dropna()

# Kiểm tra lại tính dừng với ADF và KPSS
adf_result_diff = adfuller(revenue_diff)
kpss_result_diff = kpss(revenue_diff, regression='c')
```

```
# Kiểm định ADF trên chuỗi đã lấy sai phân
print('ADF Statistic: %f' % adf_result_diff[0])
print('p-value: %f' % adf_result_diff[1])
print('Critical Values:')
for key, value in adf_result_diff[4].items():
    print('\t%s: %.3f' % (key, value))

# Diễn giải kết quả ADF cho chuỗi đã lấy sai phân
if adf_result_diff[1] > 0.05:
    print("Chuỗi không dừng (không bác bỏ H0)")
else:
    print("Chuỗi dừng (bác bỏ H0)")

# Kiểm định KPSS trên chuỗi đã lấy sai phân
print('KPSS Statistic: %f' % kpss_result_diff[0])
print('p-value: %f' % kpss_result_diff[1])
print('Critical Values:')
for key, value in kpss_result_diff[3].items():
    print('\t%s: %.3f' % (key, value))

# Diễn giải kết quả KPSS cho chuỗi đã lấy sai phân
if kpss_result_diff[1] < 0.05:
    print("Chuỗi không dừng (bác bỏ H0)")
else:
    print("Chuỗi dừng (không bác bỏ H0)")
```

Hình 5.3: Kiểm tra chuỗi dừng và xử lý cho ARIMA Model

Sử dụng model autoARIMA để dự đoán tương lai.

```
# Training process
x_train = np.array(train_data.index).reshape(-1, 1)
y_train = np.array(train_data['Revenue'])

# Find the best ARIMA model using auto_arima
model = auto_arima(y_train, trace=True, error_action='ignore', suppress_warnings=True)

# Fit the model
model.fit(y_train)
```

Hình 5.4: Model autoARIMA

```
# make predictions on the testing set
x_test = np.array(test_data.index).reshape(-1, 1)
y_test = np.array(test_data['Revenue'])
y_pred = model.predict(n_periods=len(y_test))

✓ 0.0s
```

Hình 5.5: Dự báo tương lai bằng ARIMA Model

5.2.3.2. Linear Regression

Chia tập dữ liệu thành 2 tập train-test theo tỉ lệ 7:3.

```
train_data = data[:round(len(data)*70/100)]
test_data = data[round(len(data)*70/100):]

x = data['Date'].apply(lambda x: x.timestamp()).values.reshape(-1, 1)
y = data['Revenue'].values

train_size = round(len(x) * 0.7)
x_train, x_test = x[:train_size], x[train_size:]
y_train, y_test = y[:train_size], y[train_size:]

model = LinearRegression()
model.fit(x_train, y_train)
```

Hình 5.6: Chia tỉ lệ dữ liệu cho model Linear Regression

5.2.3.3. Holt-Winters

Chia tập dữ liệu thành 2 tập train-test theo tỉ lệ 7:3.

```

train_size = round(len(data) * 0.7)
test_size = round(len(data) - train_size)

len(data), train_size, test_size
train_data = data[:train_size]
test_data = data[train_size:]

train_data.shape, test_data.shape

```

Hình 5.7: Chia tỉ lệ dữ liệu cho model Holt-Winters

Sử dụng model Holt-Winters để dự đoán tương lai.

```

model = ExponentialSmoothing(np.asarray(train_data['Revenue'])).fit()
test_data['Revenue Predicted'] = model.forecast(len(test_data))

```

Hình 5.8: Xây dựng model Holt-Winters

5.2.3.4. Random Forest

Chia tập dữ liệu thành 2 tập train-test theo tỉ lệ 7:3.

```

# Scaler data
price = data['Revenue'].values
price = price.reshape(-1, 1)
scaler = MinMaxScaler(feature_range=(0, 1))
price_scaled = scaler.fit_transform(price)

# Chia dữ liệu thành train test validate
train_size = int(len(price_scaled) * 0.7)
test_size = int(len(price_scaled) * 0.3)

train_data = price_scaled[:train_size]
test_data = price_scaled[train_size:]

```

Hình 5.9: Chia tỉ lệ dữ liệu cho model Random Forest

```
# convert an array of values into a dataset matrix
def create_dataset(dataset, time_step=1):
    dataX, dataY = [], []
    for i in range(len(dataset)-time_step-1):
        a = dataset[i:(i+time_step), 0] ####i=0, X=0,1,2,3----99   Y=100
        dataX.append(a)
        dataY.append(dataset[i + time_step, 0])
    return np.array(dataX), np.array(dataY)
```

Python

```
# Reshape into X=t,t+1,t+2..t+99 and Y=t+100
time_step = 30
X_train, y_train = create_dataset(train_data, time_step)
X_test, y_test = create_dataset(test_data, time_step)
```

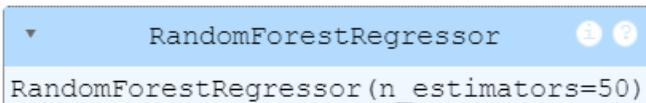
Python

```
# Reshape input to be [samples, time steps, features] which is required for LSTM
X_train = X_train.reshape(X_train.shape[0], X_train.shape[1])
X_test = X_test.reshape(X_test.shape[0], X_train.shape[1])
X_train, y_train
```

Hình 5.10: Tạo dữ liệu đầu vào để train model

Sử dụng model Random Forest để dự đoán tương lai.

```
model = RandomForestRegressor(n_estimators=50)
model.fit(X_train,y_train)
```



```
# Dự báo
test_predict = model.predict(X_test)
test_predict = scaler.inverse_transform(test_predict.reshape(-1, 1))
y_test = scaler.inverse_transform([y_test])
test_predict
```

Hình 5.11: Xây dựng model Random Forest để dự báo tương lai

5.2.3.5. Long Short-Term Memory

Chia tập dữ liệu thành 2 tập train-test theo tỉ lệ 7:3.

```
# 3. Scaler data
scaler=MinMaxScaler(feature_range=(0,1))
df=scaler.fit_transform(np.array(df).reshape(-1,1))

✓ 0.0s
```

```
# 4. Chia train test
train_size = round(0.7 * len(df))
test_size = round(len(df) - train_size)

train_data = df[:train_size]
test_data = df[train_size:]

✓ 0.0s
```

```
# 5. Hàm Create Dataset
# convert an array of values into a dataset matrix
def create_dataset(dataset, time_step=1):
    dataX, dataY = [], []
    for i in range(len(dataset)-time_step-1):
        a = dataset[i:(i+time_step), 0] ####i=0, X=0,1,2,3----99   Y=100
        dataX.append(a)
        dataY.append(dataset[i + time_step, 0])
    return np.array(dataX), np.array(dataY)
```

Hình 5.12: Chia tỉ lệ dữ liệu và tạo đầu vào cho model LSTM

Sử dụng model LSTM để dự đoán tương lai.

```
#6. Reshape into X=t,t+1,t+2..t+99 and Y=t+100
time_step = 100
x_train, y_train = create_dataset(train_data, time_step)
x_test, ytest = create_dataset(test_data, time_step)
✓ 0.0s

# 7. Reshape input to be [samples, time steps, features] which is required for LSTM
x_train = x_train.reshape(x_train.shape[0], x_train.shape[1], 1)
x_test = x_test.reshape(x_test.shape[0], x_test.shape[1], 1)
✓ 0.0s

# 8. Define LSTM Model
model=Sequential()
model.add(LSTM(50,return_sequences=True,input_shape=(100,1)))
model.add(LSTM(50,return_sequences=True))
model.add(LSTM(50))
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam')
✓ 0.0s

d:\LearnPython\Lib\site-packages\keras\src\layers\rnn\rnn.py:204: UserWarning: Do not pass an `input_
super().__init__(**kwargs)

# 9. Fit mô hình với dữ liệu train
model.fit(x_train,y_train,epochs=100,batch_size=64,verbose=1)
```

Hình 5.13: Xây dựng model LSTM để dự báo

```
# 10. Dự báo dữ liệu test, val
y_pred=model.predict(x_test)
✓ 1.0s
```

15/15 ————— **1s** 41ms/step

```
# 11. Chuẩn hóa dữ liệu y_pred, y_pred_val
ytest=scaler.inverse_transform(ytest.reshape(-1,1))
y_pred=scaler.inverse_transform(y_pred)
```

Hình 5.14: Dự báo bằng model LSTM

5.2.4. Dự đoán 30 ngày kế tiếp

```
# 13. Dự báo 30 ngày tiếp theo
def predict_future(model, data, time_step, future_days):
    x_input = data[-time_step: ].reshape(1, -1)
    temp_input = x_input[0].tolist()
    lst_output = []
    i = 0
    while i < future_days:
        if len(temp_input) > time_step:
            x_input = np.array(temp_input[-time_step: ])
            x_input = x_input.reshape((1, time_step, 1))
            yhat = model.predict(x_input, verbose=0)
            temp_input.extend(yhat[0].tolist())
            temp_input = temp_input[1:]
            lst_output.extend(yhat.tolist())
            i += 1
        else:
            x_input = np.array(temp_input).reshape((1, time_step, 1))
            yhat = model.predict(x_input, verbose=0)
            temp_input.extend(yhat[0].tolist())
            lst_output.extend(yhat.tolist())
            i += 1
    return lst_output

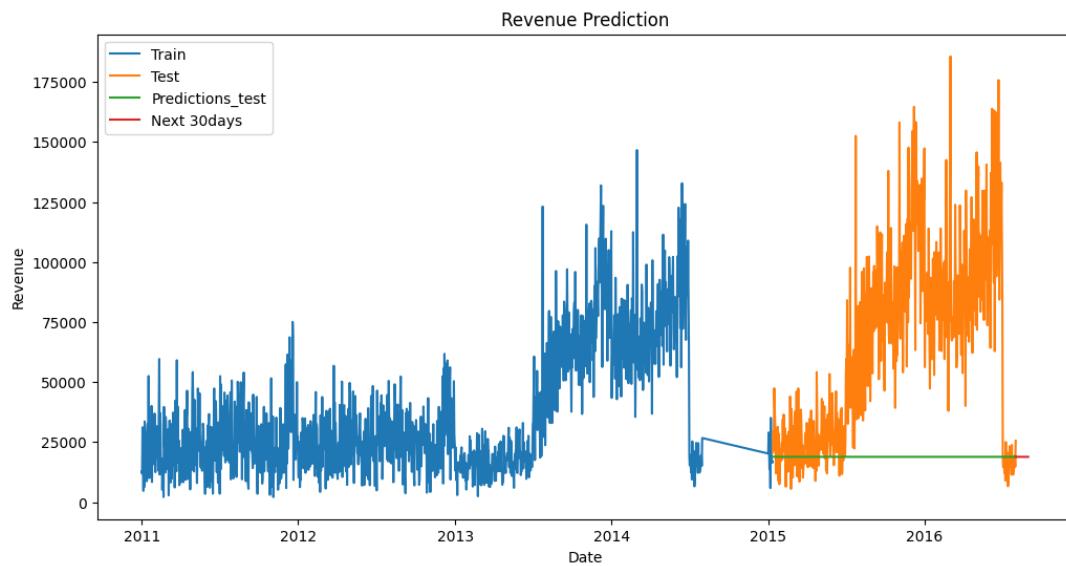
n_steps = 100

pred_30_days = predict_future(model, test_data, n_steps, 30)
pred_30_days = scaler.inverse_transform(np.array(pred_30_days).reshape(-1, 1))
```

Hình 5.15: Dự báo 30 kế tiếp

5.2.5. Vẽ đồ thị

5.2.5.1. ARIMA



Hình 5.16: Đồ thị model ARIMA

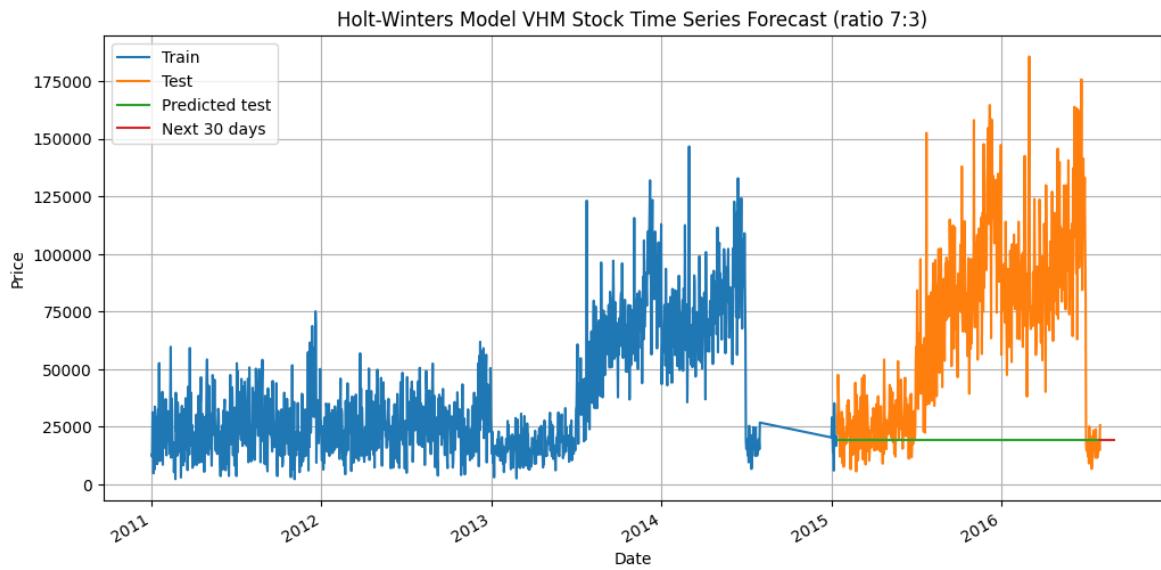
Với thuật toán ARIMA, thông số độ sai lệch của thuật toán là:

RMSE: 61674.68

MAPE: 62.23

MAE: 48702.18

5.2.5.2. Holt-Winters



Hình 5.17: Đồ thị model Holt-Winters

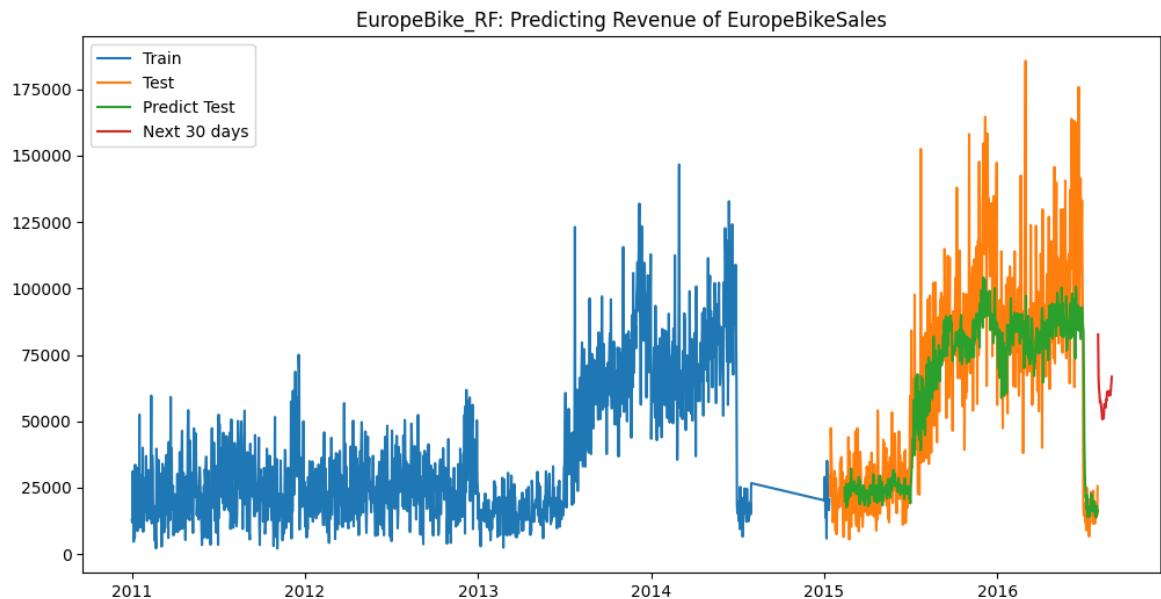
Với thuật toán Holt-Winters, thông số độ sai lệch của thuật toán là:

RMSE: 61465.45

MAPE: 62.10

MAE: 48528.88

5.2.5.3. Random Forest



Hình 5.18: Đồ thị model Random Forest

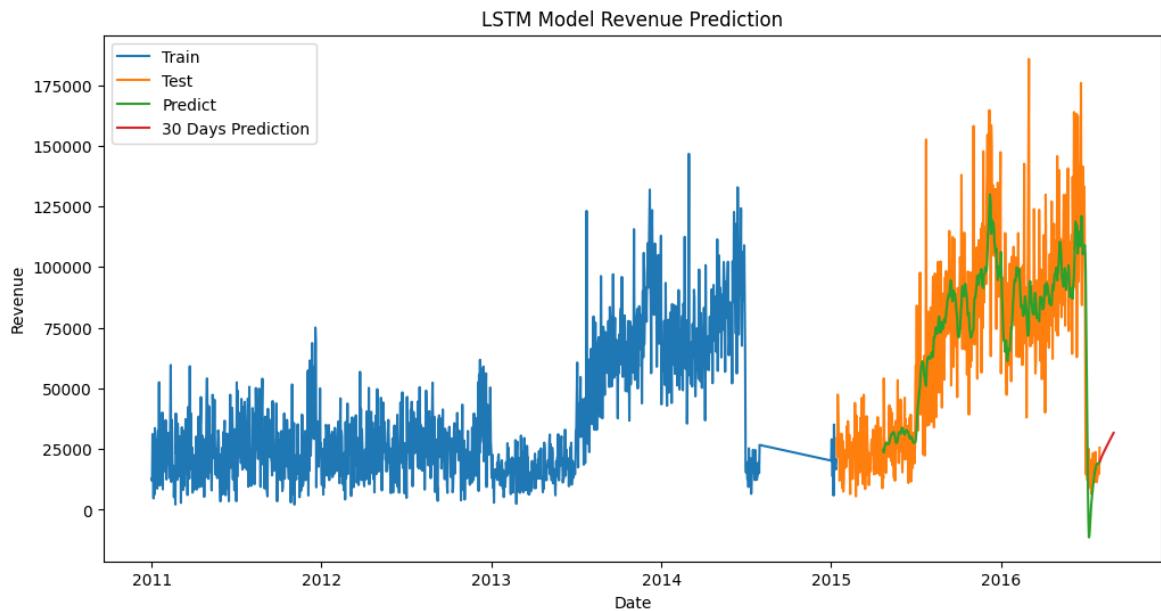
Với thuật toán Random Forest, thông số độ sai lệch của thuật toán là:

RMSE: 22727.59

MAPE: 28.61

MAE: 16452.97

5.2.5.4. LSTM



Hình 5.19: Đồ thị model LSTM

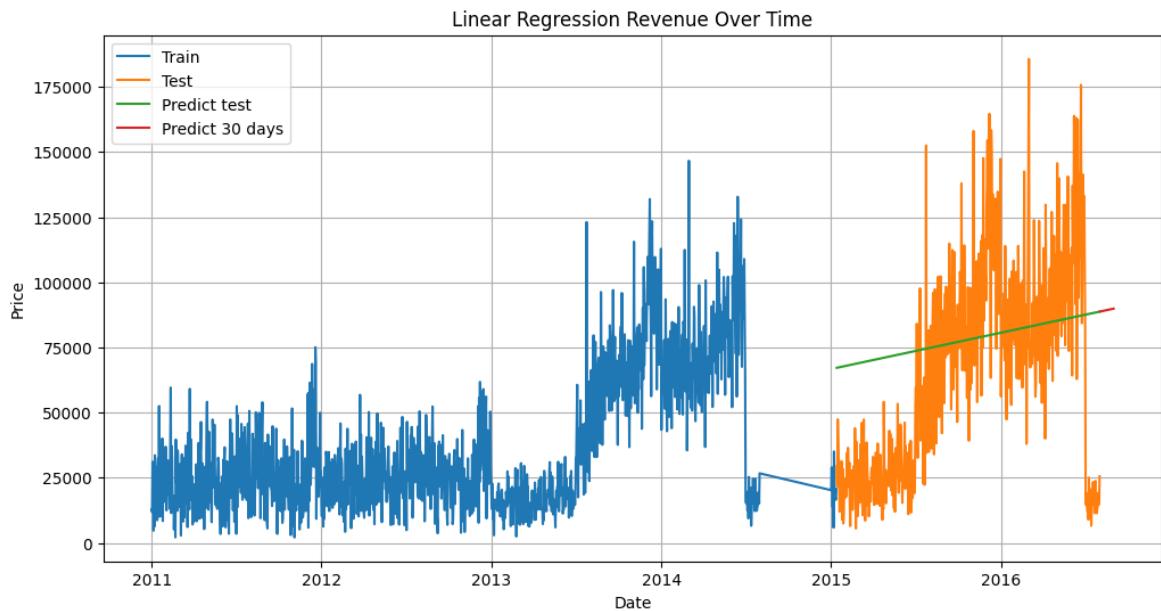
Với thuật toán LSTM, thông số độ sai lệch của thuật toán là:

RMSE: 22970.12

MAPE: 30.17

MAE: 17201.11

5.2.5.5. Linear Regression



Hình 5.20: Đồ thị model Linear Regression

Với thuật toán Linear Regression, thông số độ sai lệch của thuật toán là:

RMSE: 38056.90

MAPE: 114.41

MAE: 31424.34

5.3. Kết quả thực hiện

Bảng 14: Độ đo các model trong Data Mining

	RMSE	MAPE	MAE
ARIMA	61674.68	62.234	48702.18
HoltWinters	61465.45	62.10	48528.88
Random Forest	22722.75	28.61	16452.97
LSTM	22970.12	30.17	17201.11
Linear Regression	38056.90	114.41	31424.34

Với bảng trên, ta thấy thuật toán Random Forest có độ dự đoán chính xác nhất so với các thuật toán còn lại mà nhóm sử dụng.

PHÂN CÔNG CÔNG VIỆC

STT	MSSV	Họ và tên	Công việc	Mức độ hoàn thành
1	21520596	Trần Thị Kim Anh	Thực hiện các quá trình SSIS, SSAS, Pivot, Power BI, 4 câu MDX, Data Mining. Viết báo cáo và chỉnh sửa.	100%
2	21521882	Lê Minh Chánh	Thực hiện các quá trình SSIS, SSAS, Pivot, 6 câu MDX. Quay video demo và viết báo cáo.	100%