

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



BÁO CÁO ĐỒ ÁN
KHO DỮ LIỆU VÀ OLAP

ĐỀ TÀI:
XÂY DỰNG KHO DỮ LIỆU ĐẶT PHÒNG
KHÁCH SẠN

GIẢNG VIÊN HƯỚNG DẪN

ThS. Nguyễn Thị Kim Phụng

SINH VIÊN THỰC HIỆN

Trần Phương Anh – 21520595

Trần Thị Luyến – 21521107

Thành phố Hồ Chí Minh, tháng 07 năm 2024

LỜI CẢM ƠN

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến toàn thể giảng viên trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM cũng như là nhà trường vì đã giúp nhóm chúng em có những kiến thức cơ bản làm nền tảng để thực hiện đồ án môn học này.

Đặc biệt hơn nữa, chúng em xin gửi lời cảm ơn chân thành đến cô Nguyễn Thị Kim Phụng – Giảng viên môn Kho dữ liệu và OLAP. Thời gian vừa qua, cô đã trực tiếp giảng dạy, truyền đạt kinh nghiệm, hướng dẫn chúng em một cách tận tình giúp nhóm chúng em hoàn thành tốt đồ án môn học của mình. Chúc cô sẽ luôn dồi dào sức khỏe, tràn đầy nhiệt huyết để có thể tiếp tục giảng dạy, dìu dắt những thế hệ sinh viên tiếp theo.

Và để đồ án này được hoàn thành thì không thể nào không nhắc đến những người đã làm ra nó. Cảm ơn thành viên trong nhóm đã chăm chỉ và chịu khó hoàn thành nhiệm vụ đúng tiến độ để đồ án của chúng ta được hoàn thành đúng hạn.

Mặc dù đã vận dụng tối đa những gì đã học được nhưng chúng em vẫn khó có thể tránh khỏi những sai sót. Chính vì vậy, nhóm chúng em rất mong nhận được sự góp ý từ phía cô để có thể hoàn thiện một cách tốt nhất có thể. Qua đó cũng tích lũy và học hỏi kinh nghiệm để làm hành trang cho tương lai.

Lời cuối cùng, chúng em một lần nữa xin được chân thành cảm ơn.

Thành phố Hồ Chí Minh, tháng 06 năm 2024

Nhóm sinh viên thực hiện

NHẬN XÉT CỦA GIẢNG VIÊN

[illegible]

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU	1
1.1. Tổng quan đề tài.....	1
1.2. Phát biểu về dữ liệu.....	1
1.2.1. Mô tả dữ liệu.....	1
1.2.2. Thuộc tính kho dữ liệu	2
1.3. Xây dựng kho dữ liệu.....	4
1.3.1. Lược đồ bông tuyết	4
1.3.2. Dim_Guest	4
1.3.3. Dim_Country	5
1.3.4. Dim_Reservation	5
1.3.5. Dim_Payment	5
1.3.6. Dim_Room	6
1.3.7. Dim_Date	6
1.3.8. Fact	6
CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)	8
2.1. Tạo Project và thực hiện kết nối.....	8
2.1.1. Tạo Project.....	8
2.1.2. Tạo cơ sở dữ liệu	9
2.2. Chuẩn bị dữ liệu.....	10
2.3. Quá trình lọc dữ liệu.....	13
2.4. Quá trình tạo các bảng Dimension.....	16
2.5. Quá trình tạo bảng Fact	20
2.6. Quy trình và lược đồ dữ liệu.....	23
2.6.1. Quy trình ETL.....	23
2.6.2. Lược đồ bông tuyết	24
CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS)	24
3.1. Quá trình SSAS	24

3.1.1. Quá trình thực hiện trên Visual Data Studio	24
3.1.2. Điều chỉnh và phân cấp các Dimension	34
3.2. Truy vấn MDX	38
3.2.1. Doanh thu từng loại phòng qua các năm	38
3.2.2. Doanh thu từng nguồn đặt phòng của từng quý qua các năm	38
3.2.3. Danh sách khách hàng đặt phòng loại ‘Standard’	39
3.2.4. Tổng doanh thu theo loại phòng và tháng năm 2023	40
3.2.5. Tổng số lượng khách theo quốc gia và loại phòng năm 2023	40
3.2.6. Số lượng đặt phòng theo từng tháng và từng loại phòng	41
3.2.7. Thông tin khách từ Thụy Sĩ (Switzerland) , chỉ hiển thị các cột có giá trị tổng số tiền chi tiêu lớn hơn 1000	42
3.2.8. Số lượng khách nữ từ Thụy Sĩ (Switzerland) với tổng số tiền chi tiêu lớn hơn 1000.....	43
3.2.9. Số lượng đặt phòng qua các nguồn đặt phòng qua các năm	44
3.2.10. Thông tin khách hàng nam ở phòng loại ‘Deluxe’, chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 740.....	45
3.2.11. Tìm các khách hàng đặt phòng có số lượng đêm ở lớn hơn 4 trong quý 4/2023.....	46
3.2.12. Mỗi quý trong năm 2023, cho biết tháng nào có tổng doanh thu cao nhất.....	47
3.2.13. Doanh thu của loại phòng ‘Deluxe’ qua từng năm.....	48
3.2.14. Phần trăm số lượng đặt phòng có sử dụng bữa sáng	48
3.2.15. Phần trăm doanh thu từ mỗi nguồn đặt phòng năm 2023	50
3.2.16. Top 3 doanh thu theo quốc gia từ các đơn đặt phòng có sử dụng dịch vụ đưa đón sân bay	52

3.2.17. Số lượng đặt phòng của từng loại phòng theo mỗi nguồn đặt phòng.....	53
3.2.18. Tổng số lượng đặt phòng theo từng loại phòng và tầng	54
3.2.19. Trung bình số đêm ở theo từng tháng trong năm 2023.	55
3.2.20. Tổng doanh thu theo loại phòng.....	56
3.2.21. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng.....	57
3.2.22. Liệt kê top 2 loại phòng có doanh thu nhiều nhất trong cả hai năm 2022 và năm 2023	58
3.3. Thực hiện trên Excel.....	59
3.3.1. Tổng doanh thu theo loại phòng qua các năm (group by)	59
3.3.2. Tổng số đơn đặt phòng theo nguồn đặt phòng (group by)	59
3.3.3. Số lượng khách từ Thụy Sĩ (Switzerland) , chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 1000 (table)	60
3.3.4. Tạo ma trận hiển thị số lượng khách hàng theo loại phòng và nguồn đặt phòng (matrix).....	62
3.3.5. Tạo biểu đồ hiển thị tổng doanh thu theo tháng năm 2023 (chart)	63
3.3.6. Tạo biểu đồ hiển thị top 2 doanh thu theo tháng năm 2023 (chart)	63
3.3.7. Số lượng đặt phòng theo từng tháng và từng loại phòng trong năm 2023	65
3.3.8. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng.....	67
3.4. Thực hiện trên Power BI.....	69
3.4.1. Tổng doanh thu theo loại phòng qua các năm	70
3.4.2. Tổng số lượng khách hàng theo nguồn đặt phòng.....	72
3.4.3. Số lượng khách từ Thụy Sĩ (Switzerland), chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 1000.....	74

3.4.4. Tạo ma trận hiển thị số lượng đặt phòng theo loại phòng và nguồn đặt phòng	75
3.4.5. Tạo biểu đồ hiển thị tổng doanh thu theo tháng năm 2023	76
3.4.6. Tạo biểu đồ hiển thị top 2 doanh thu theo tháng năm 2023	79
3.4.7. Số lượng đặt phòng theo từng tháng và từng loại phòng trong năm 2023	80
3.4.8. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng.....	81
CHƯƠNG 4. QUÁ TRÌNH DATA MINING	84
4.1. Tổng quan đề tài.....	84
4.2. Lý thuyết mô hình phân cụm.....	85
4.2.1. K-Means	85
4.2.2. DBSCAN (Density-Based Clustering)	86
4.2.3. Agglomerative Clustering	87
4.2.4. Thống kê mô tả.....	88
4.3. Tiền xử lý dữ liệu.....	90
4.3.1. Xử lý giá trị thiếu	90
4.3.2. Tính toán các chỉ số RFM	91
4.3.3. Khám phá và xử lý dữ liệu RFM.....	92
4.3.4. Chuẩn hóa dữ liệu	94
4.4. Xây dựng mô hình phân cụm.....	95
4.4.1. Xác định số lượng cụm (k) tối ưu bằng kỹ thuật Elbow .	95
4.4.2. Xây dựng các mô hình	96
4.4.3. Đánh giá mô hình và trực quan hóa kết quả.....	96
4.4.4. Phân tích & phân cụm khách hàng.....	97

DANH MỤC CHỮ VIẾT TẮT

Từ viết tắt	Từ đầy đủ	Ý nghĩa
DB	Database	Cơ sở dữ liệu
SSMS	SQL Server Management Studio	
SSIS	SQL Server Integration Services	
SSAS	SQL Server Analysis Services	
ETL	Extract, Transform, Load	Trích xuất, Chuyển đổi, Nạp dữ liệu
MDX	Multidimensional Expressions	Ngôn ngữ truy vấn trong SSAS

BẢNG PHÂN CÔNG NHIỆM VỤ

Công việc	Trần Phương Anh	Trần Thị Luyện
Xây dựng bố cục bài báo cáo	✓	
Chương 1. Giới thiệu	✓	✓
Chương 2. SSIS	✓	
Chương 3. SSAS	✓	
Chương 4. Data Mining		✓
Tổng hợp và chỉnh sửa báo cáo	✓	✓
Hoàn thành	100%	100%

NỘI DUNG

CHƯƠNG 1. GIỚI THIỆU

1.1. Tổng quan đề tài

Du lịch và lữ hành chiếm hơn 10% GDP trên toàn thế giới và đang có xu hướng chiếm được phần lớn hơn trong chiếc bánh toàn cầu. Đồng thời, đây là ngành tạo ra khối lượng dữ liệu khổng lồ và việc tận dụng lợi thế của nó có thể giúp các doanh nghiệp nổi bật giữa đám đông.

Đề tài "Xây dựng Kho dữ liệu OLAP về đặt phòng khách sạn" nhấn mạnh vào việc áp dụng công nghệ thông tin để cải thiện quản lý thông tin đặt phòng trong ngành du lịch và khách sạn. Trong bối cảnh ngày nay, việc quản lý thông tin khách hàng, dự đoán nhu cầu và tối ưu hóa dịch vụ là rất quan trọng để cung cấp trải nghiệm tốt nhất cho khách hàng và tăng cường hiệu suất kinh doanh cho các doanh nghiệp trong ngành.

Hệ thống đặt phòng khách sạn truyền thống thường gặp phải nhiều thách thức, bao gồm việc xử lý lượng dữ liệu lớn từ các giao dịch đặt phòng, theo dõi xu hướng và sở thích của khách hàng, cũng như dự đoán nhu cầu đặt phòng trong tương lai. Để giải quyết những thách thức này, việc xây dựng một Kho dữ liệu OLAP (Online Analytical Processing) là cần thiết.

Chúng em sẽ sử dụng Microsoft SQL Server Integration Services (SSIS) để triển khai quy trình ETL (Extract, Transform, Load) để lấy dữ liệu từ các nguồn khác nhau và chuyển đổi chúng thành định dạng phù hợp cho việc phân tích. Sau đó, sử dụng Microsoft SQL Server Analysis Services (SSAS) để xây dựng các cube và dimension, tạo ra một môi trường phân tích dữ liệu linh hoạt và mạnh mẽ. Bên cạnh đó, chúng em sử dụng công cụ Power BI để phát triển các báo cáo và trực quan hóa dữ liệu giúp người dùng dễ dàng tìm hiểu và phân tích thông tin.

Cuối cùng, khám phá ứng dụng của kỹ thuật Data Mining trong việc phân tích dữ liệu đặt phòng khách sạn, từ việc phát hiện mẫu tự nhiên trong dữ liệu đến việc dự đoán xu hướng và hành vi của khách hàng.

1.2. Phát biểu về dữ liệu

1.2.1. Mô tả dữ liệu

Bộ dữ liệu cung cấp dữ liệu đặt phòng trong hai mùa liên tiếp (2021 - 2023) của một khách sạn hạng sang

Bộ dữ liệu gồm 32 thuộc tính và 9974 dòng

Link dữ liệu: <https://www.kaggle.com/dimitrisangelide/hotel-reservations-data>

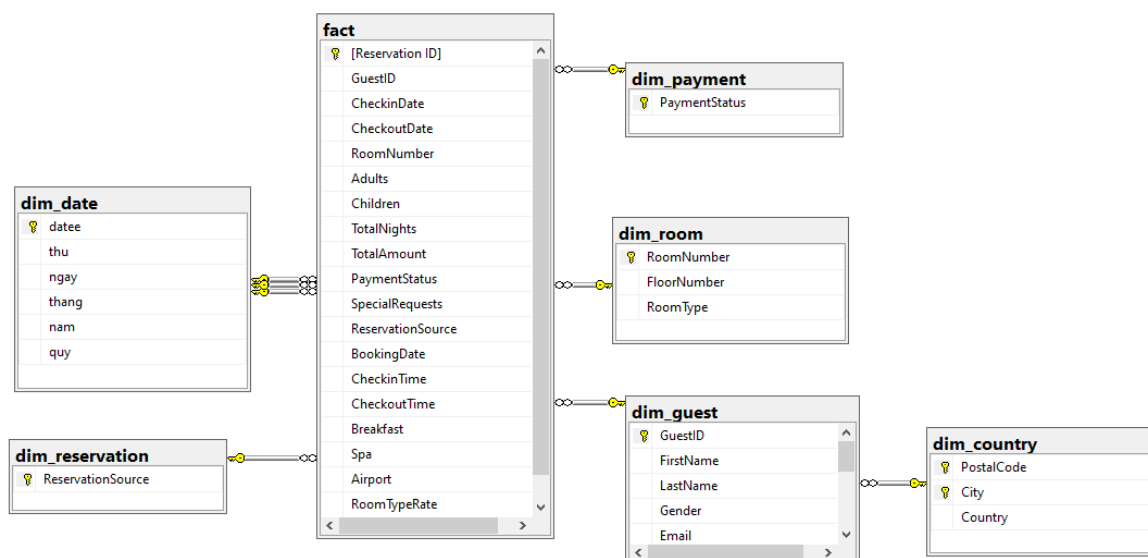
1.2.2. Thuộc tính kho dữ liệu

STT	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	Reservation ID	int	Mã định danh duy nhất cho đặt phòng
2	Guest ID	int	Mã định danh duy nhất của người đã đặt phòng
3	First Name	string	Tên của khách
4	Last Name	string	Họ của khách
5	Gender	string	Giới tính
6	Email	string	Địa chỉ email
7	Phone	string	Số điện thoại
8	Nationality	string	Quốc tịch
9	Birthdate	date	Ngày sinh
10	Address	string	Địa chỉ nhà
11	City	string	Thành phố mà khách sống
12	Postal Code	int	Mã bưu chính
13	Country	string	Quốc gia cư trú
14	Check-in Date	date	Ngày dự kiến nhận phòng
15	Check-out Date	date	Ngày dự kiến trả phòng
16	Room Number	int	Số phòng đã đặt
17	Floor Number	int	Tầng của phòng đã đặt
18	Room Type	string	Loại phòng đã đặt
19	Adults	int	Số người lớn ở trong phòng
20	Children	int	Số trẻ em ở trong phòng
21	Total Nights	int	Tổng số đêm đã đặt
22	Total Amount	int	Tổng số tiền đã thanh toán cho đặt phòng
23	Payment Status	string	Trạng thái thanh toán vào thời điểm đặt phòng
24	Special Requests	string	Yêu cầu đặc biệt của khách
25	Reservation Source	string	Nguồn thông tin mà khách sử dụng để đặt phòng
26	Booking Date	date	Ngày đặt phòng diễn ra

27	Check-in Time	time	Thời gian dự kiến nhận phòng
28	Check-out Time	time	Thời gian dự kiến trả phòng
29	Breakfast Included	boolean	Nếu bữa sáng đã bao gồm trong đặt phòng
30	Spa Package Included	string	Nếu gói dịch vụ spa đã bao gồm trong đặt phòng.
31	Airport Pickup Included	string	Nếu khách yêu cầu đưa đón sân bay
32	Room Type Rate	int	Giá của mỗi loại phòng tại thời điểm đặt phòng

1.3. Xây dựng kho dữ liệu

1.3.1. Lược đồ bông tuyết



1.3.2. Dim_Guest

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	GuestID	int	Primary key	Mã số của khách
2	FirstName	string		Tên của khách
3	LastName	string		Họ của khách
4	Gender	string		Giới tính
5	Email	string		Địa chỉ email
6	Phone	string		Số điện thoại
7	Nationality	string		Quốc tịch
8	Birthdate	date		Ngày sinh
9	Address	string		Địa chỉ nhà
10	City	string		Thành phố mà khách sống
11	PostalCode	int		Mã bưu chính

1.3.3. Dim_Country

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	City	string	Primary key	Thành phố mà khách sống
2	PostalCode	int	Primary key	Mã bưu chính
3	Country	string		Quốc gia cư trú

1.3.4. Dim_Reservation

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	ReservationSource	string	Primary key	Nguồn thông tin mà khách sử dụng để đặt phòng

1.3.5. Dim_Payment

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	PaymentStatus	string	Primary key	Trạng thái thanh toán vào thời điểm đặt phòng

1.3.6. Dim_Room

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	RoomNumber	int	Primary key	Số phòng đã đặt
2	FloorNumber	int		Tầng của phòng đã đặt
3	RoomType	string		Loại phòng đã đặt

1.3.7. Dim_Date

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	Date_ngay	date	Primary key	Ngày dạng dd/MM/yyyy
2	Thu	string		Thứ trong tuần của ngày
3	Ngày	int		Ngày
4	Thang	int		Tháng
5	Nam	int		Năm
6	Quy	int		Quý

1.3.8. Fact

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	ReservationID	int	Primary key	Mã định danh duy nhất cho đặt phòng
2	GuestID	int	Foreign Key	Mã số của khách
3	CheckinDate	date	Foreign Key	Ngày dự kiến nhận phòng
4	CheckoutDate	date	Foreign Key	Ngày dự kiến trả phòng
5	RoomNumber	int	Foreign Key	Số phòng đã đặt
6	Adults	int		Số người lớn ở trong phòng
7	Children	int		Số trẻ em ở trong phòng
8	TotalNights	int		Tổng số đêm đã đặt
9	TotalAmount	int		Tổng số tiền đã thanh toán cho đặt phòng
10	PaymentStatus	string	Foreign Key	Trạng thái thanh toán vào thời điểm đặt phòng
11	SpecialRequests	string		Yêu cầu đặc biệt của khách
12	ReservationSource	string	Foreign Key	Nguồn thông tin mà khách sử dụng để đặt phòng
13	BookingDate	date	Foreign Key	Ngày đặt phòng diễn ra
14	CheckinTime	time		Thời gian dự kiến nhận phòng
15	CheckoutTime	time		Thời gian dự kiến trả phòng
16	Breakfast	boolean		Nếu bữa sáng đã bao gồm trong đặt phòng
17	Spa	string		Nếu gói dịch vụ spa đã bao gồm trong đặt phòng.
18	Airport	string		Nếu khách yêu cầu đưa đón sân bay

19	RoomTypeRate	int		Giá của mỗi loại phòng tại thời điểm đặt phòng
----	--------------	-----	--	--

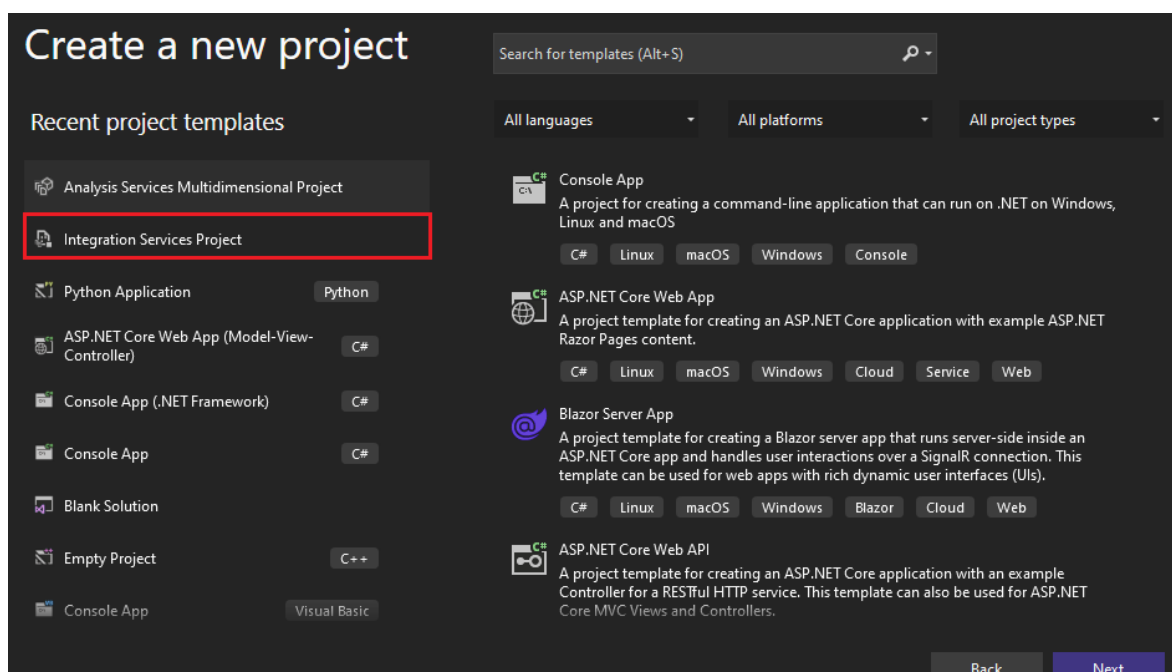
CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)

2.1. Tạo Project và thực hiện kết nối

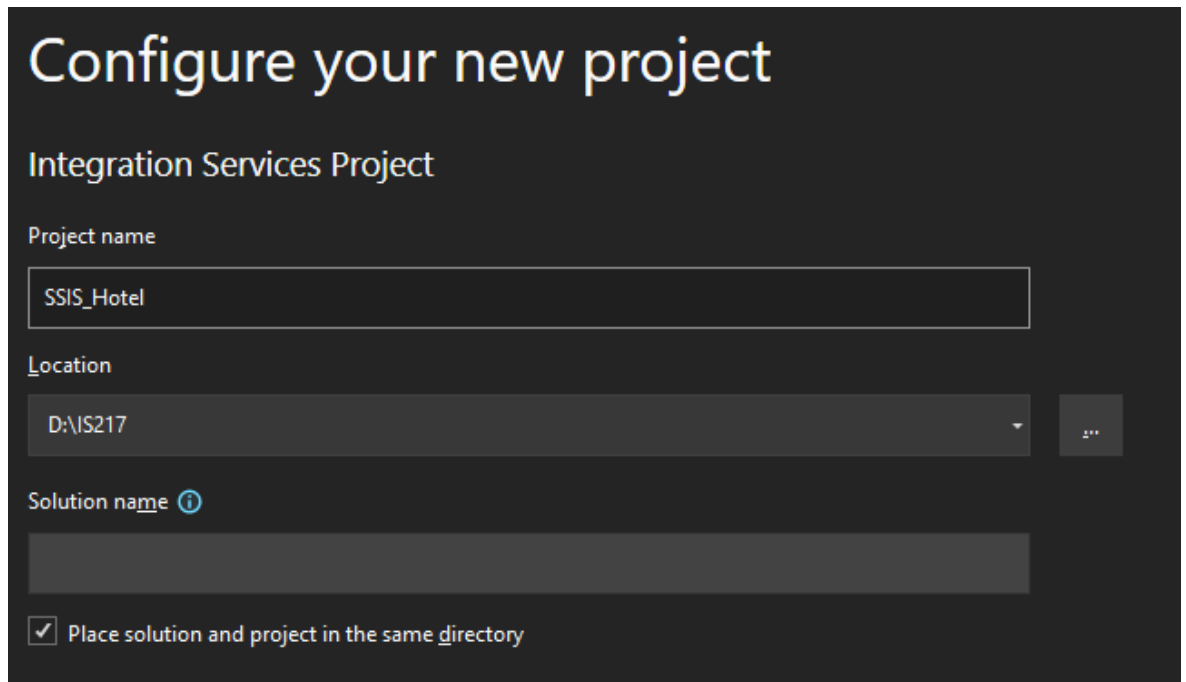
2.1.1. Tạo Project

Bước 1: Mở Visual Studio (2022) và tạo mới dự án

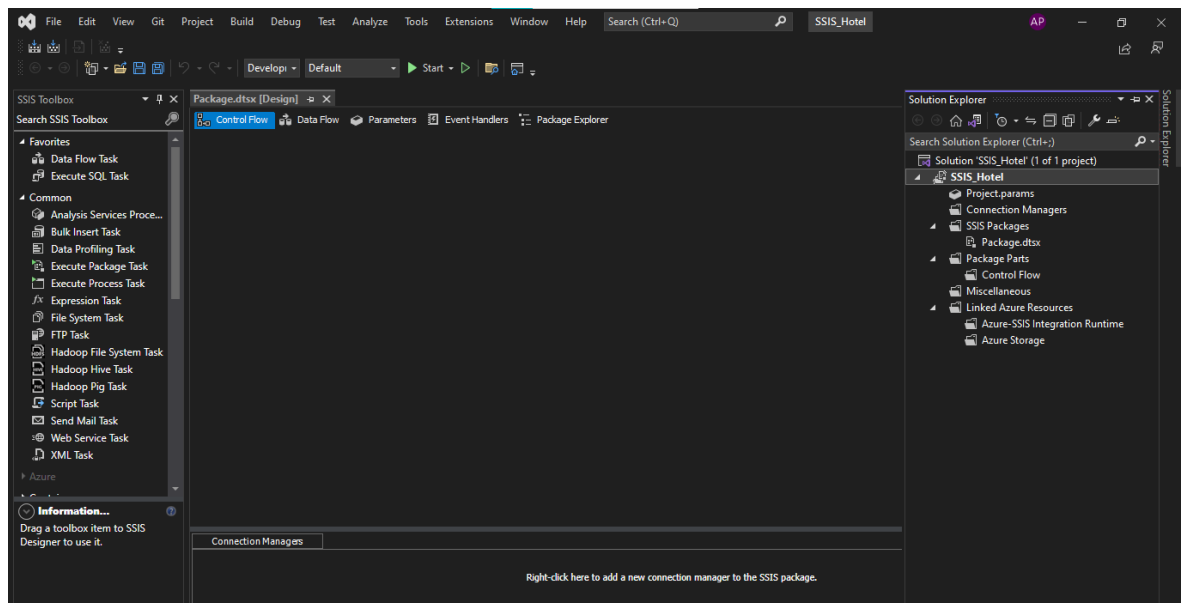
Bước 2: Để tạo dự án SSIS, ta sử dụng công cụ “Integration Services Project”



Bước 3: Điền thông tin về dự án và “Create” để tạo mới



Bước 4: Sau khi hoàn thành, ta có giao diện của SSIS



2.1.2. Tạo cơ sở dữ liệu

Ta sử dụng công cụ SQL Server Management Studio (SSMS) để tạo Database cho dự án

Tên Database: SSIS

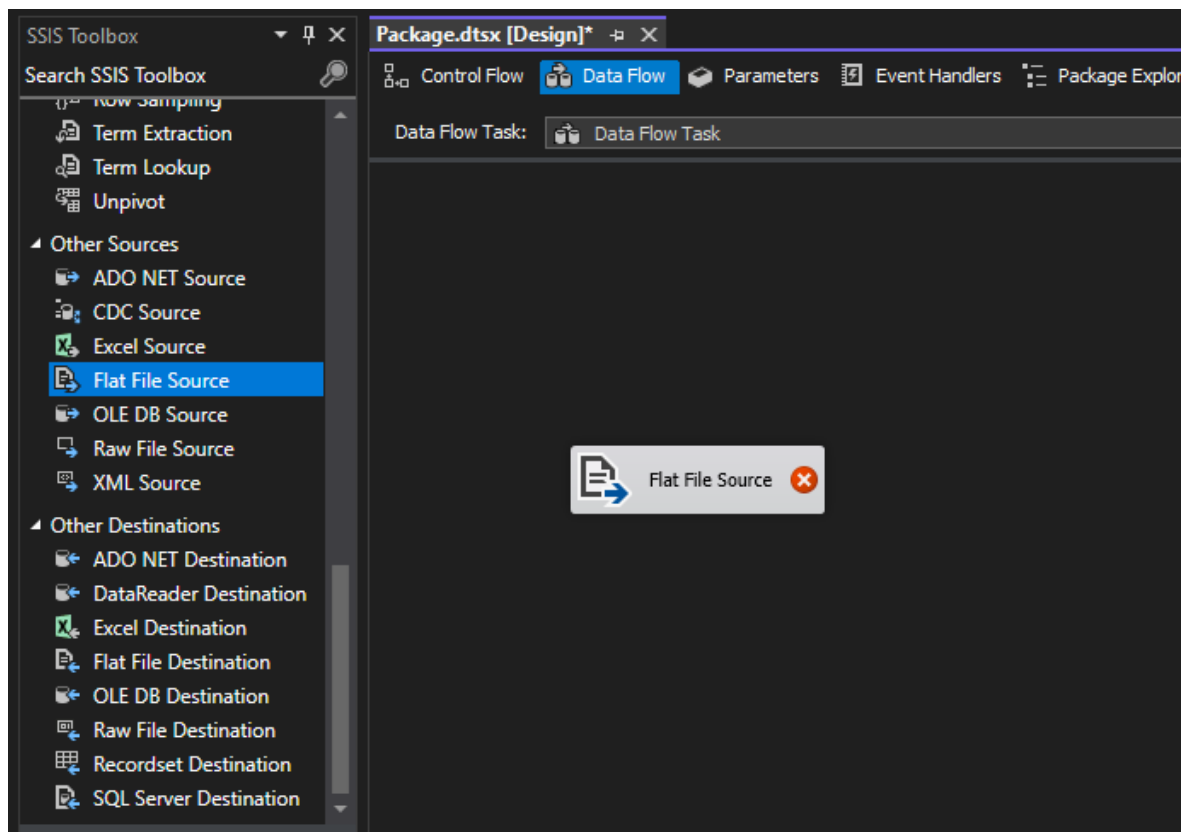
SSIS là nơi lưu trữ dữ liệu của các bảng Dim và Fact

- Fact
- Dim_Guest

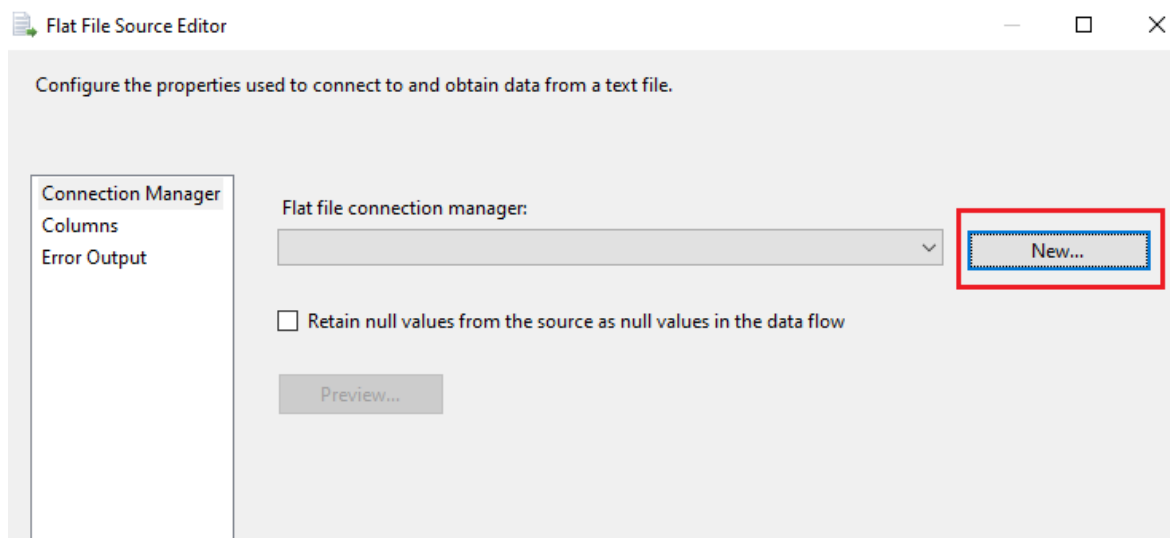
- Dim_Date
- Dim_Room
- Dim_Country
- Dim_Payment
- Dim_Reservation

2.2. Chuẩn bị dữ liệu

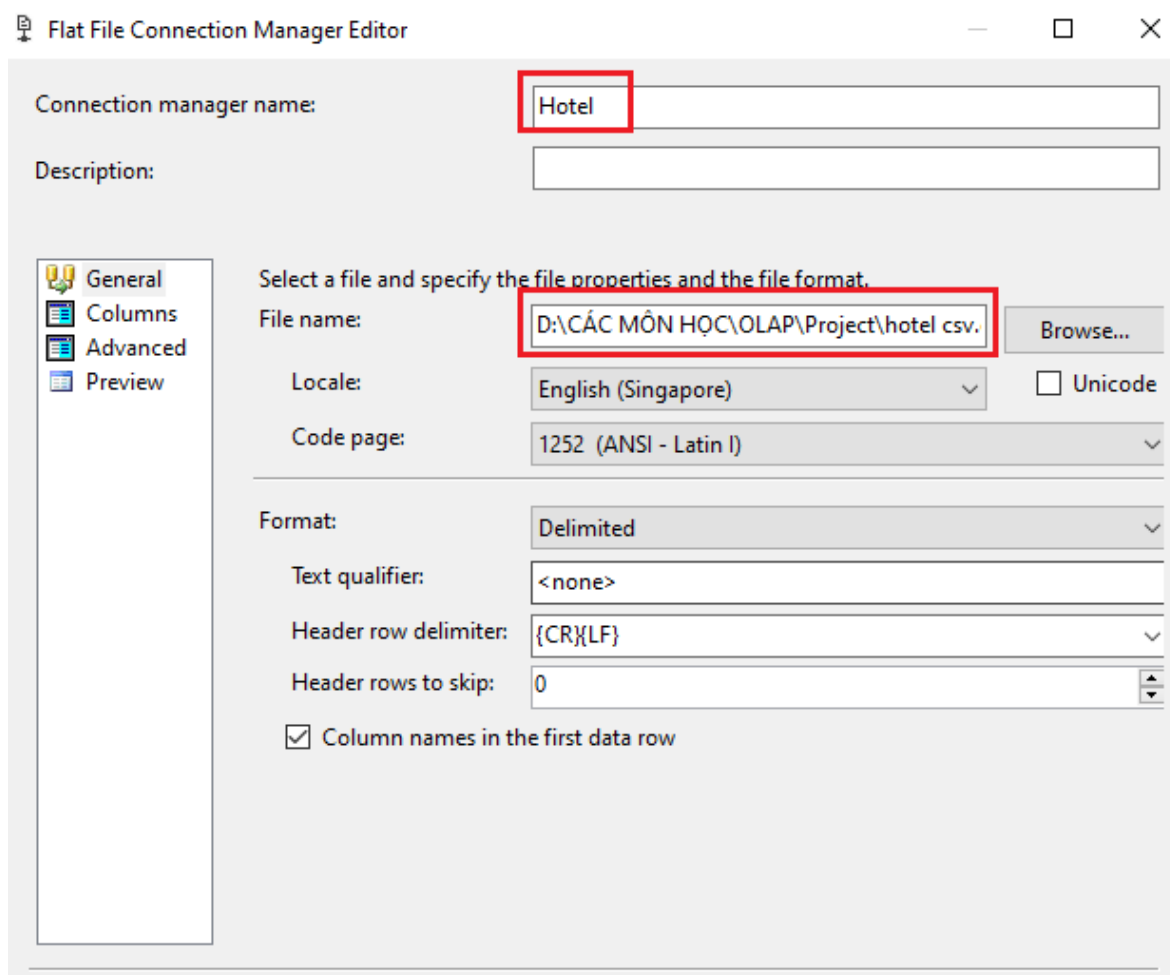
Bước 1: Trong SSIS Toolbox ở phía trái màn hình, tìm mục Other Sources
→ chọn Flat File Source và kéo vào Data Flow



Bước 2: Nhấp chuột phải vào Flat File Source vừa kéo, chọn Edit, hiển thị Editor để đổ dữ liệu vào và chọn New để thêm kết nối với DB



Bước 3: Điền tên Connection và chọn file dữ liệu (CSV)



Bước 4: Ở mục Advanced, chỉnh sửa kiểu dữ liệu cho khớp với dữ liệu trong file CSV (hệ thống mặc định kiểu dữ liệu là string, độ dài 50 ký tự)

Flat File Connection Manager Editor

Connection manager name:

Description:

General Columns **Advanced** Preview

Configure the properties of each column.

ReservationID
GuestID
FirstName
LastName
Gender
Email
Phone
Nationality
Birthdate
Address
City
PostalCode
Country
CheckinDate
CheckoutDate
RoomNumber
FloorNumber
RoomType

New ▼ Delete Suggest Types...

Misc

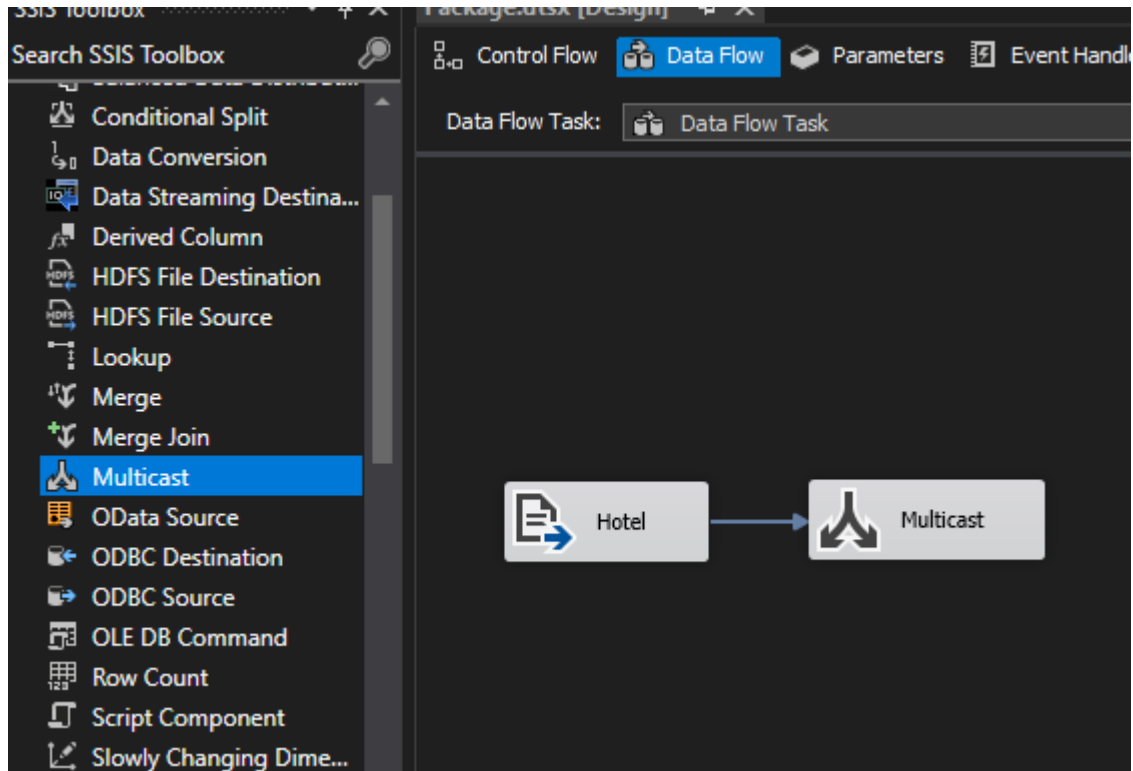
Name	ReservationID
ColumnDelimiter	Comma {,}
ColumnType	Delimited
InputColumnWidth	0
DataPrecision	0
DataScale	0
DataType	string [DT_STR]
OutputColumnWidth	50
TextQualified	True

Name

OK Cancel Help

Bước 5: Trong mục “Common”, chọn thành phần Multicast

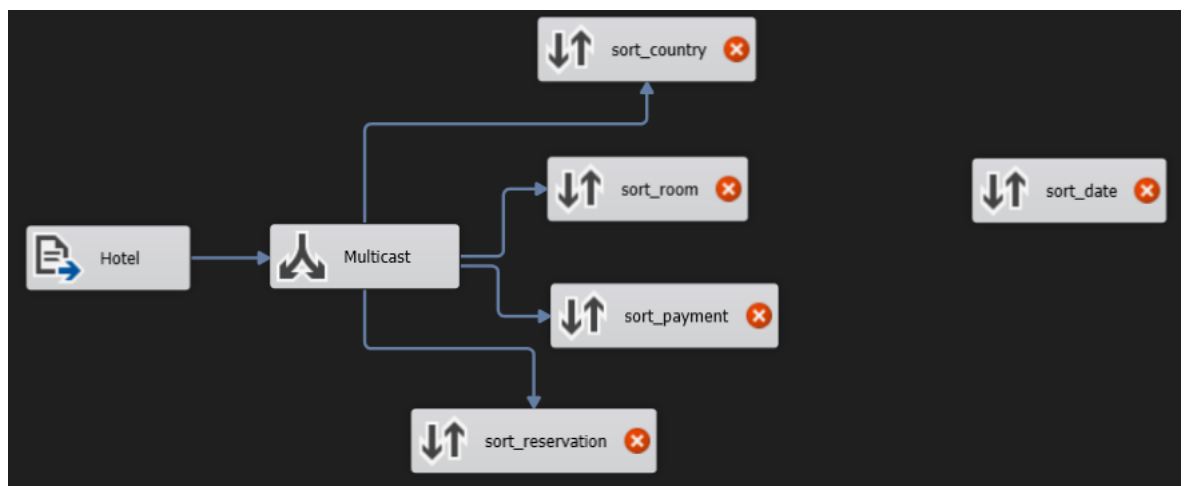
Multicast giúp sao chép dữ liệu từ một nguồn đến nhiều đích khác nhau, giúp tạo ra nhiều nhánh luồng dữ liệu từ một luồng dữ liệu duy nhất (flat file source)



Nhóm em xử lý thuộc tính Ngày (CheckinDate, CheckoutDate, BookingDate) là gộp tất cả hàng/giá trị của 3 thuộc tính lại, tách riêng thành 1 sheet và lưu thành 1 file CSV riêng nên sẽ đổ dữ liệu bằng một Flat File Source khác, cách làm cũng tương tự như trên.

2.3. Quá trình lọc dữ liệu

Bước 1: Trong mục “Common” kéo thành phần Sort để thực hiện sắp xếp và lọc dữ liệu, ở đây ta sẽ tạo 5 Sort cho 5 Dimension (Country, Room, Payment, Reservation và Date).



Bước 2: Nhấp chuột phải vào thành phần Sort muốn xử lý (ở đây chọn sort_payment) và chọn “Edit”

Sort Transformation Editor

Specify the columns to sort, and set their sort type and their sort order. All nonselected columns are copied unchanged.

Available Input Columns

<input type="checkbox"/>	Name	Pass Throu...
<input type="checkbox"/>	ReservationID	<input checked="" type="checkbox"/>
<input type="checkbox"/>	GuestID	<input checked="" type="checkbox"/>
<input type="checkbox"/>	FirstName	<input checked="" type="checkbox"/>
<input type="checkbox"/>	LastName	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Gender	<input checked="" type="checkbox"/>

Input Column	Output Alias	Sort Type	Sort Order	Con

At least one column must be selected for sorting.

☐ Remove rows with duplicate sort values

OK Cancel Help

Bước 3: Chọn thuộc tính (PaymentStatus) muốn sắp xếp (tăng/giảm dần) và lọc dữ liệu trùng (Remove rows with duplicates)

Available Input Columns

<input checked="" type="checkbox"/>	Name	Pass Throu...
<input type="checkbox"/>	TotalAmount	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	PaymentSta...	<input checked="" type="checkbox"/>
<input type="checkbox"/>	SpecialRequ...	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Reservation...	<input checked="" type="checkbox"/>
<input type="checkbox"/>	BookingDate	<input checked="" type="checkbox"/>

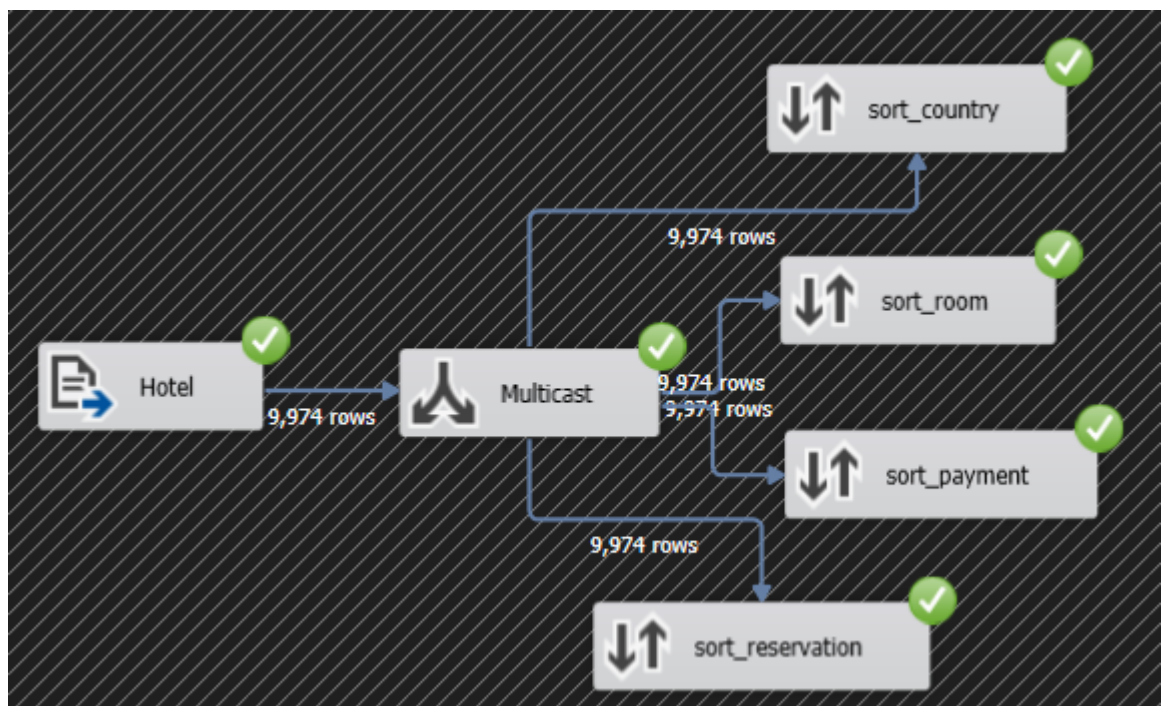
Input Column	Output Alias	Sort Type	Sort Order	Con
PaymentStatus	PaymentStatus	ascending	1	

☒ Remove rows with duplicate sort values

OK Cancel Help

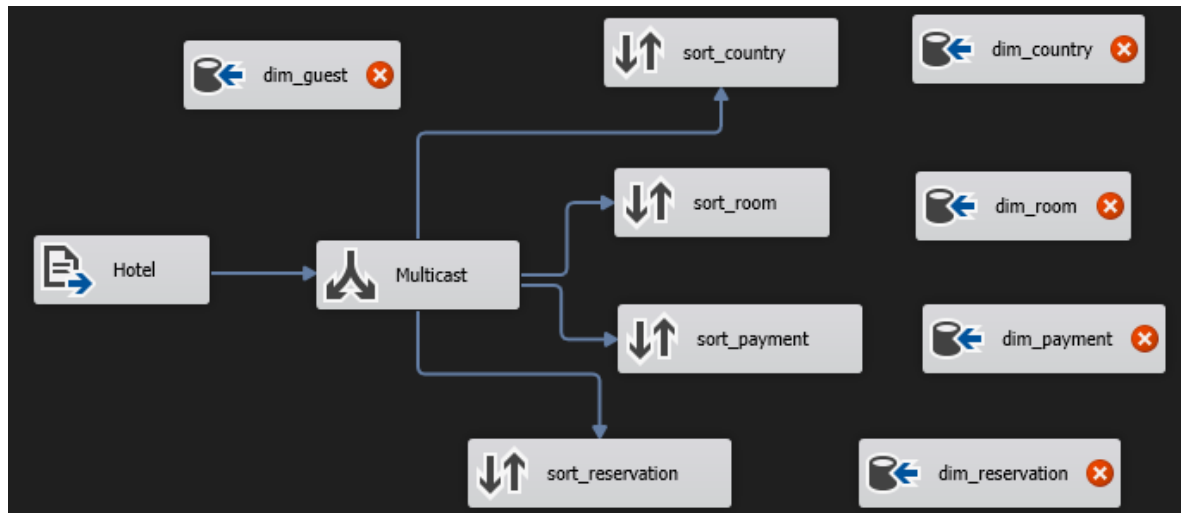
Bước 4: Thực hiện tương tự với các Sort khác

Bước 5: Nhấn “Start” để tiến hành đổ dữ liệu vào

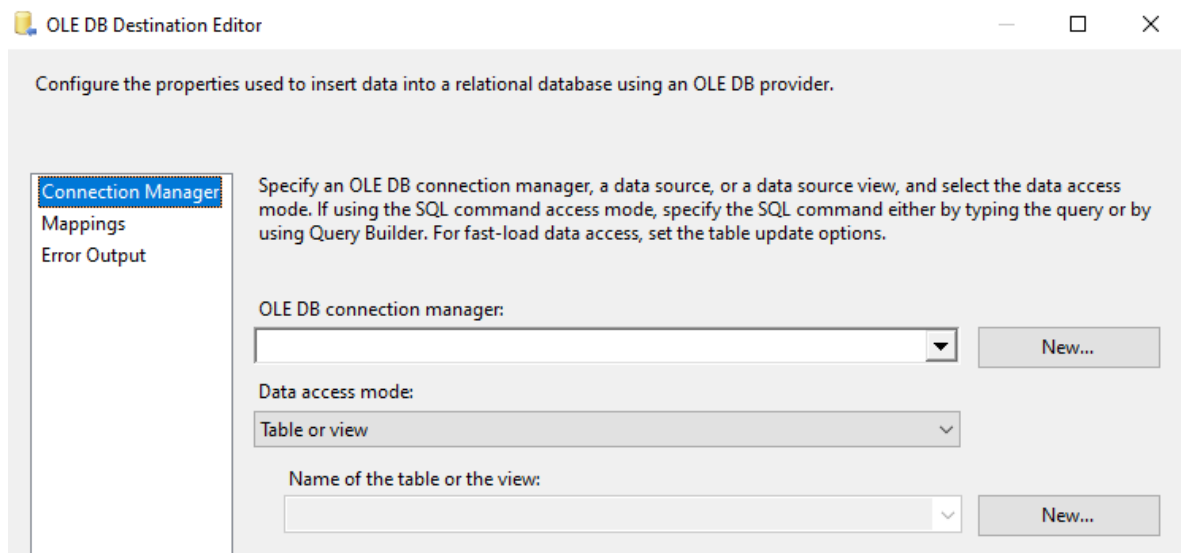


2.4. Quá trình tạo các bảng Dimension

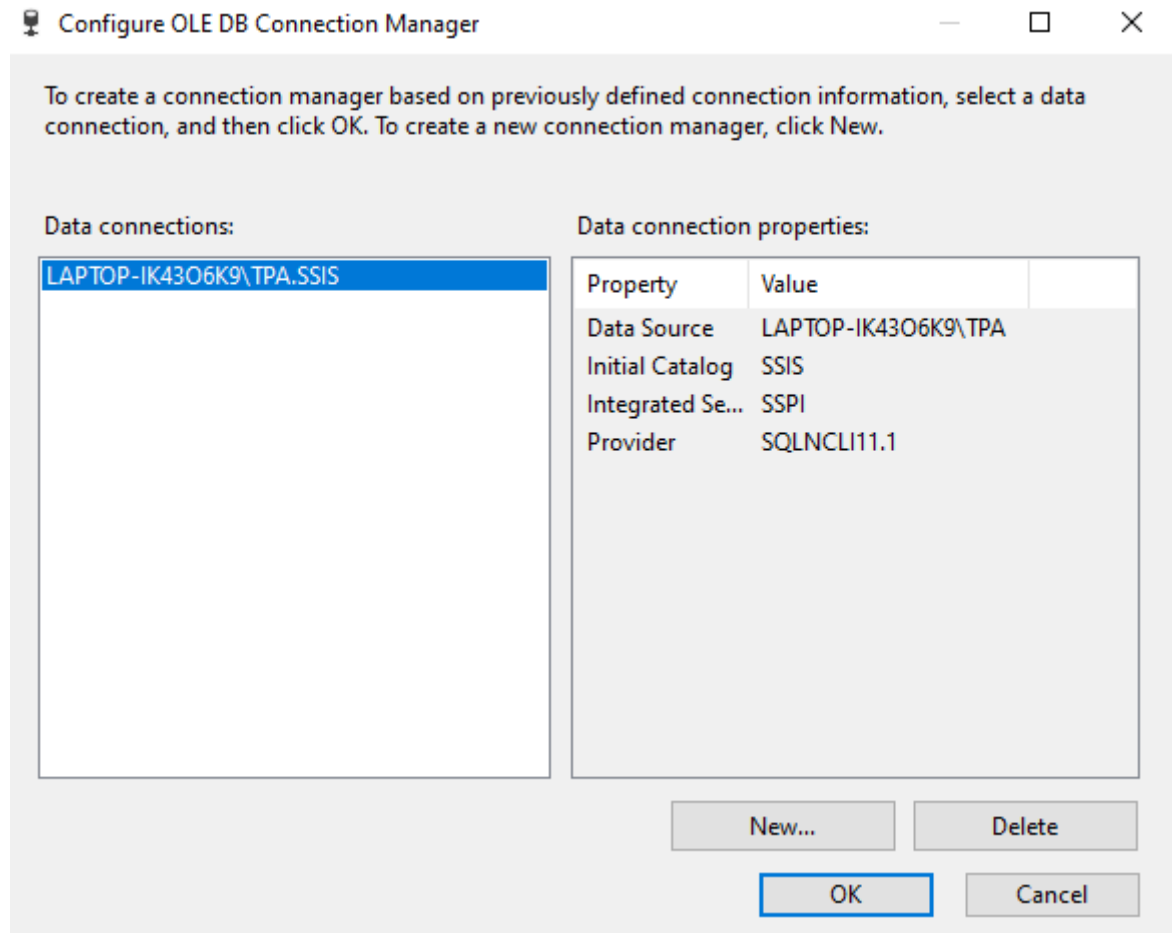
Bước 1: Trong mục “Other Destinations”, chọn thành phần “OLE DB Destination”. Đây là thành phần được sử dụng để nạp dữ liệu vào các bảng đích (bảng Dim) trong cơ sở dữ liệu



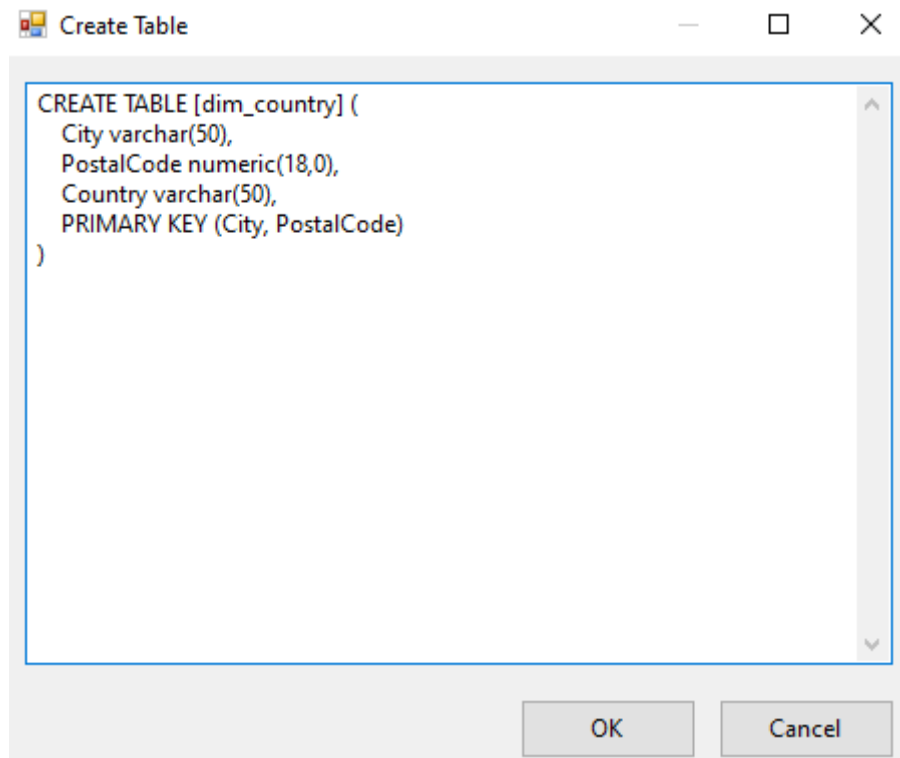
Bước 2: Kéo mũi tên xanh từ Sort qua Dim để đổ dữ liệu từ Sort. Nhấp chuột phải vào bảng Dim muốn chỉnh sửa (ở đây chọn dim_country) và chọn “Edit”



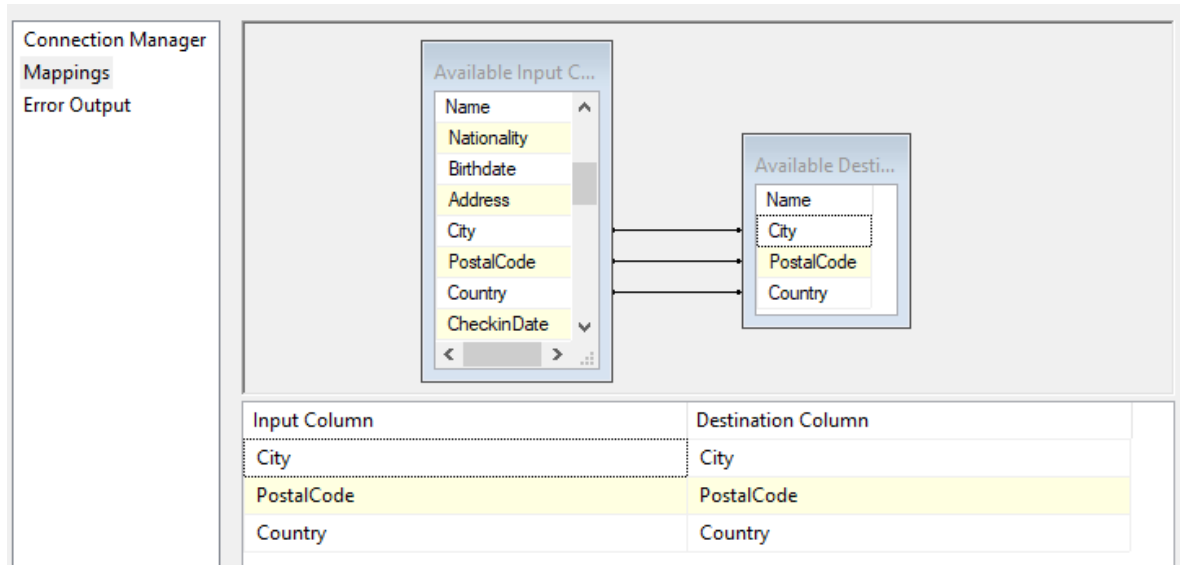
Bước 3: Chọn OLE DB connection manager (nhấn “New”) và chọn Connection vừa tạo



Bước 4: Tạo bảng dim_country

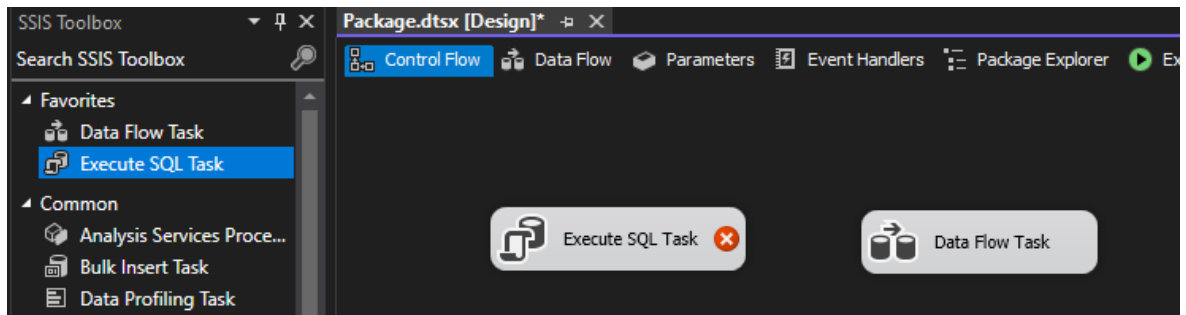


Bước 5: Kiểm tra mapping và chọn “OK” để hoàn tất

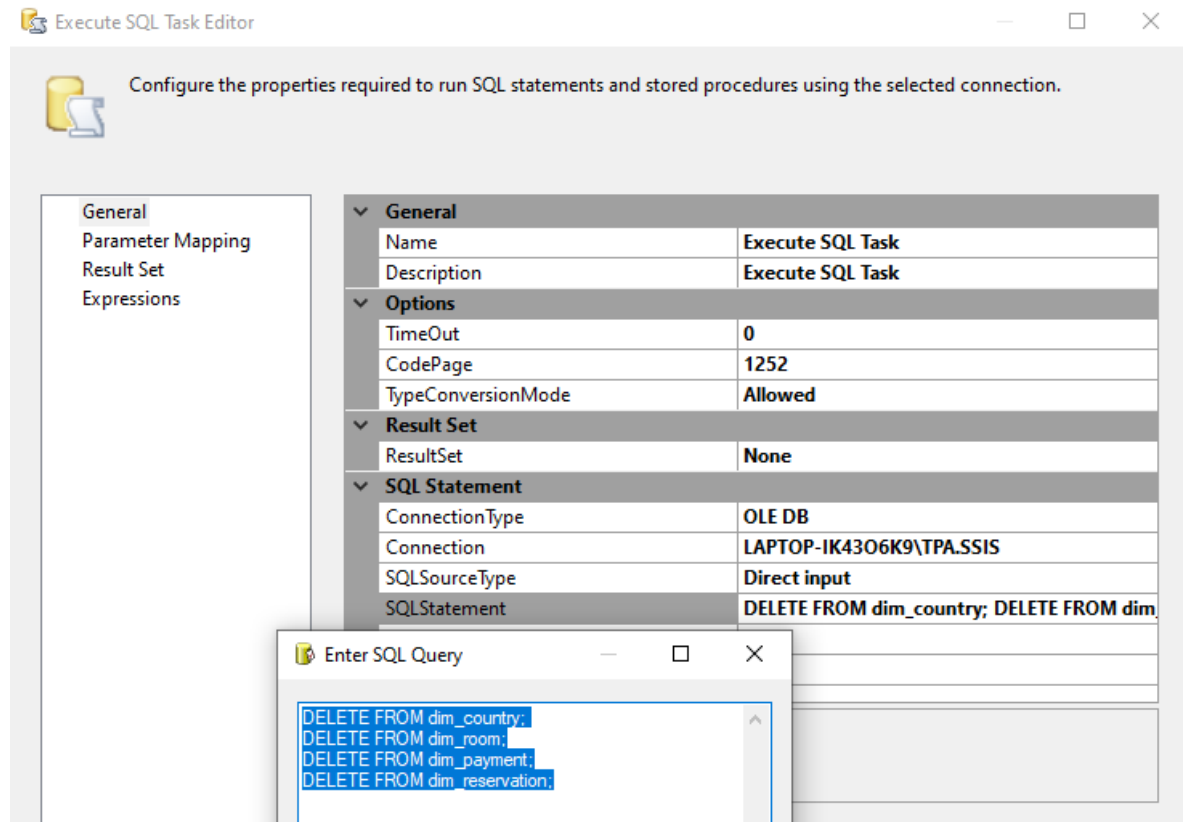


Bước 6: Thực hiện tương tự với các bảng Dim khác

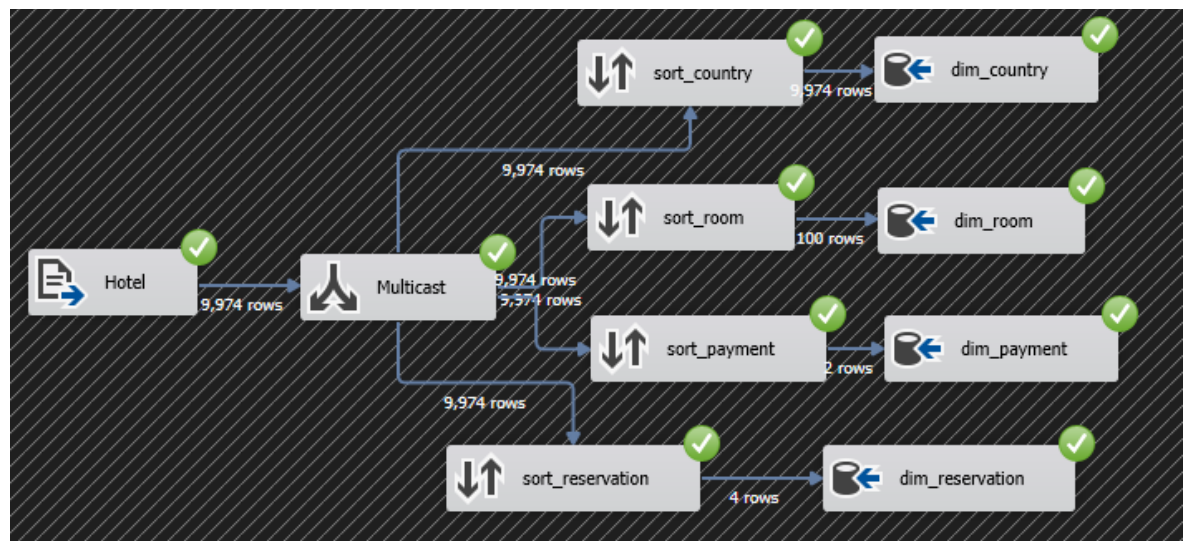
Bước 7: Kéo thả thành phần Execute SQL Task vào không gian Control Flow



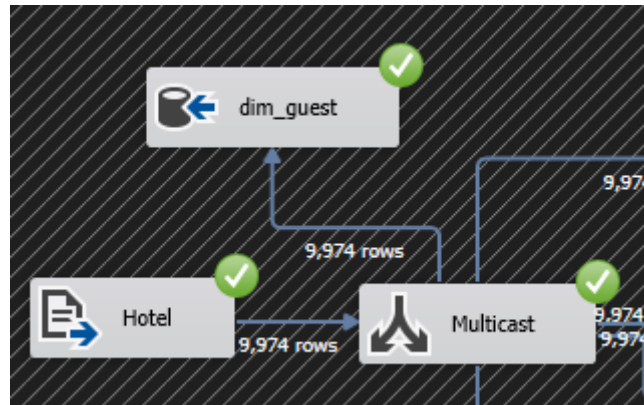
Bước 8: Nhấp phải chọn “Edit”, điền thông tin cần thiết và chọn “Build Query” cho câu SQL vừa tạo (DELETE FROM)



Bước 9: Nhấn “Start” để đổ dữ liệu vào bảng Dim



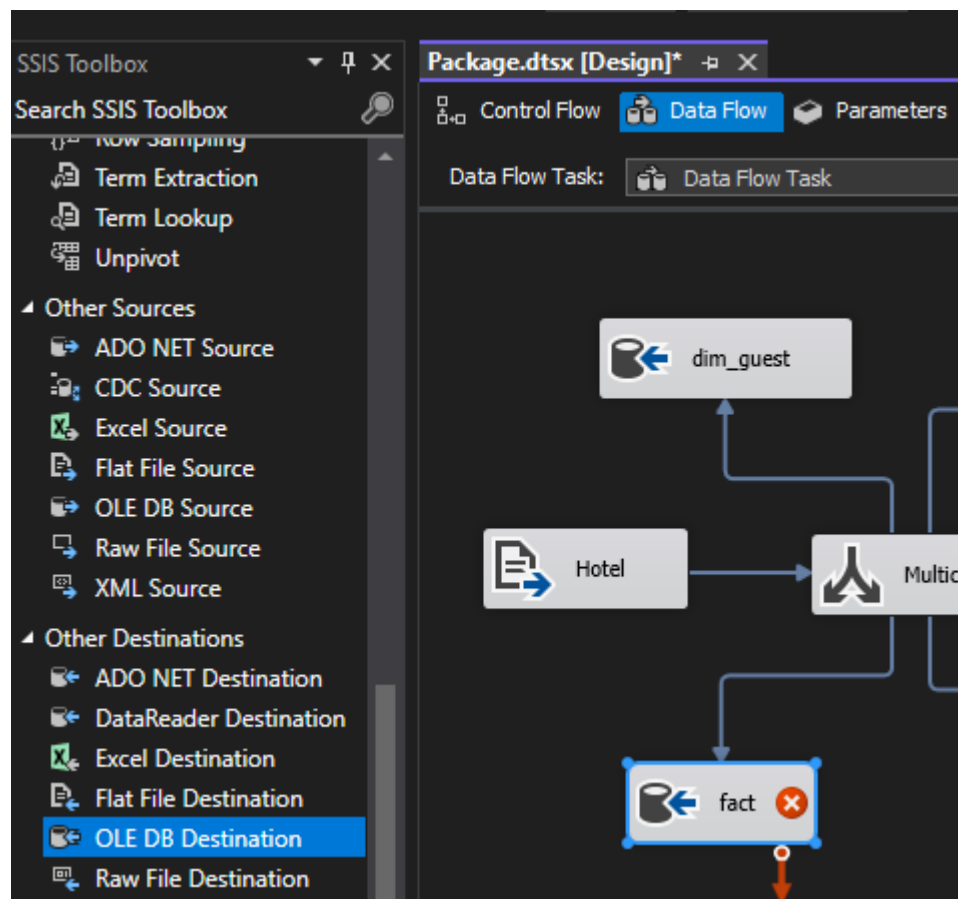
Do bảng dim_guest có khóa ngoại từ các bảng Dim khác nên phải tạo thành công các bảng Dim đó mới tạo được dim_guest



Thực hiện tương tự với bảng dim_date

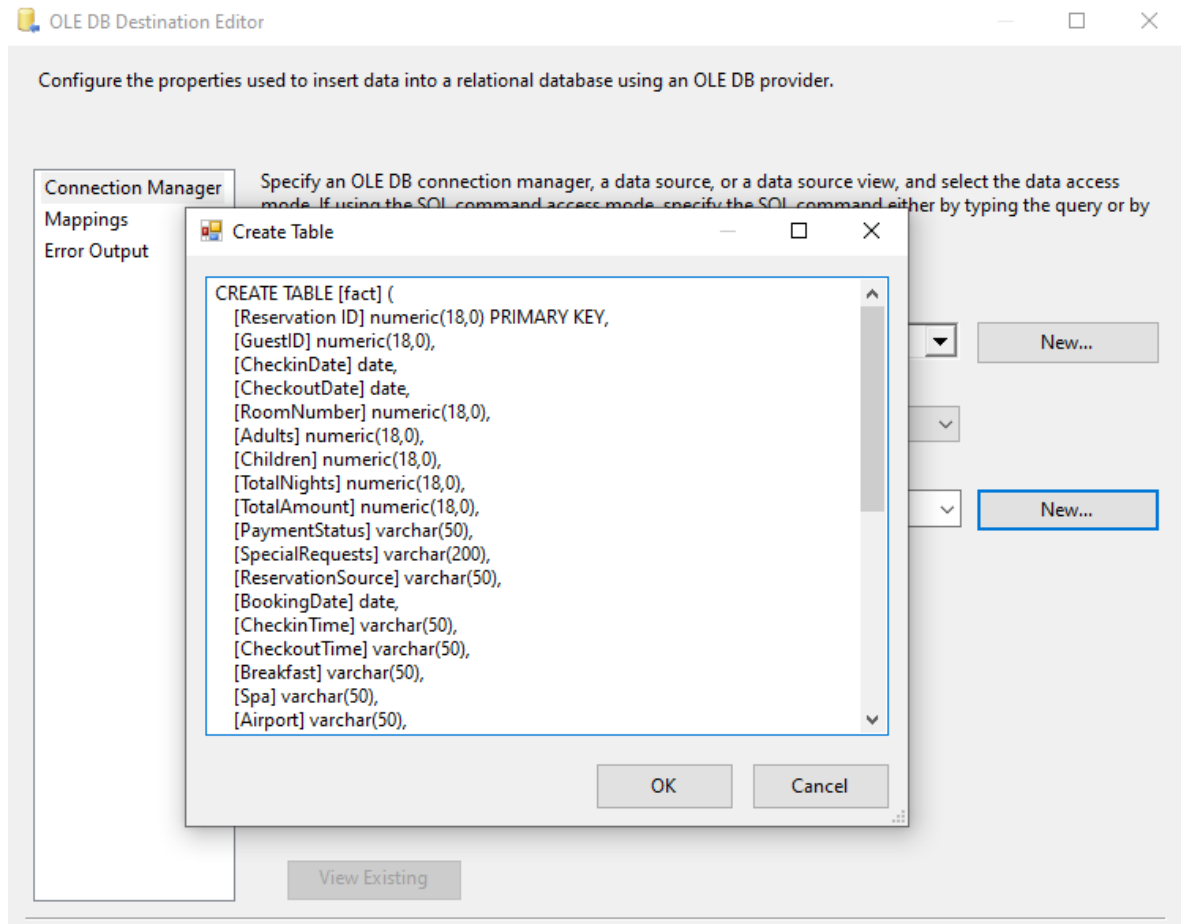
2.5. Quá trình tạo bảng Fact

Bước 1: Kéo thành phần OLE DB Destination vào Data Flow

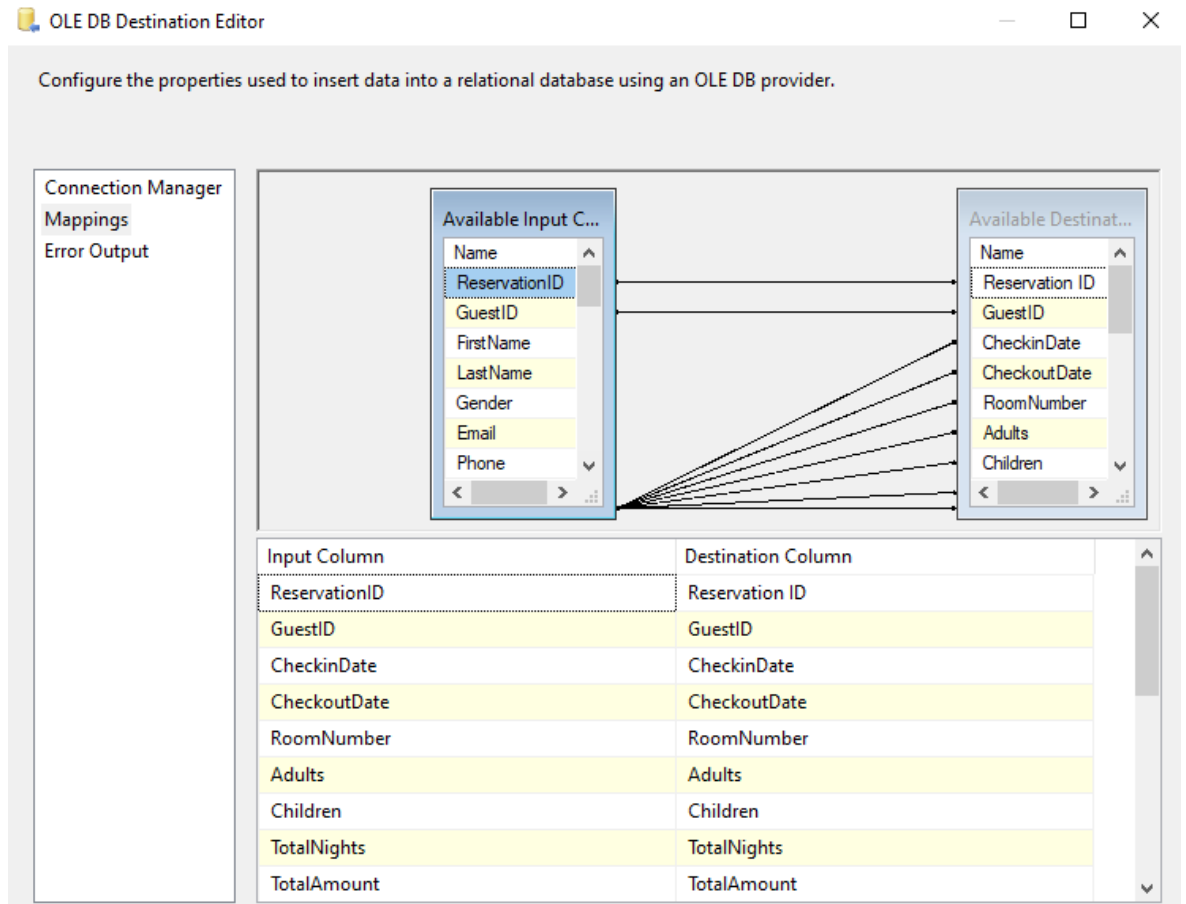


Bước 2: Kết nối bằng Connection Manager đến Database

Bước 3: Tạo bảng Fact bằng câu lệnh CREATE TABLE



Bước 4: Kiểm tra Mappings



Bước 5: Chỉnh sửa câu lệnh SQL trong Execute SQL Task và “Build Query”

Configure the properties required to run SQL statements and stored procedures using the selected connection.

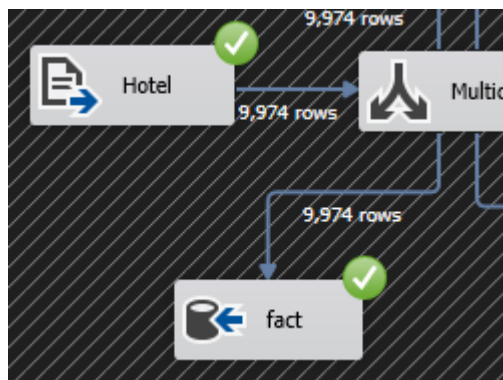
General
Parameter Mapping
Result Set
Expressions

General	
Name	Execute SQL Task
Description	Execute SQL Task
Options	
TimeOut	0
CodePage	1252
TypeConversionMode	Allowed
Result Set	
ResultSet	None
SQL Statement	
ConnectionType	OLE DB
Connection	LAPTOP-1K4306K9\TPA.SSIS
SQLSourceType	Direct input
SQLStatement	DELETE FROM fact
IsQueryStoredProcedure	False
BypassPrepare	True

SQLStatement
Specifies the query to be run by the task.

Browse... Build Query... Parse Query

Bước 6: Nhấn “Start” để tiến hành đổ dữ liệu vào bảng Fact

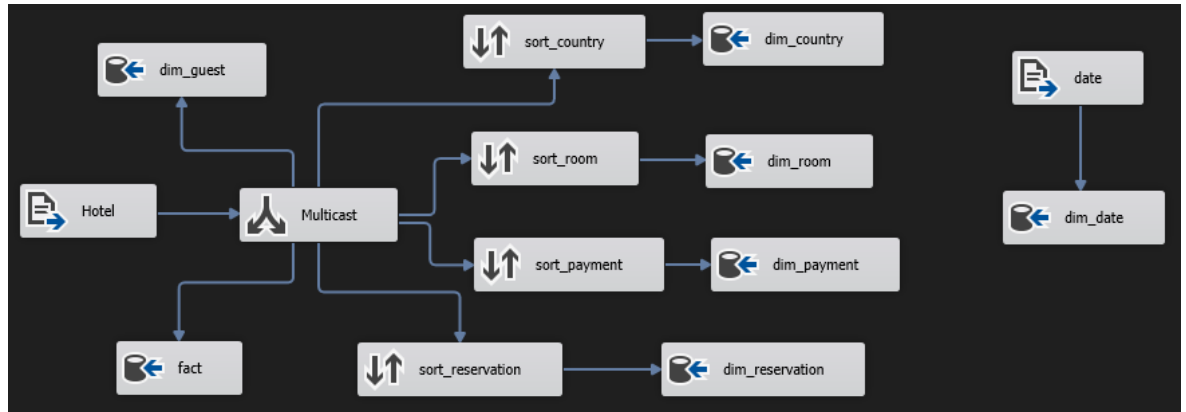


2.6. Quy trình và lược đồ dữ liệu

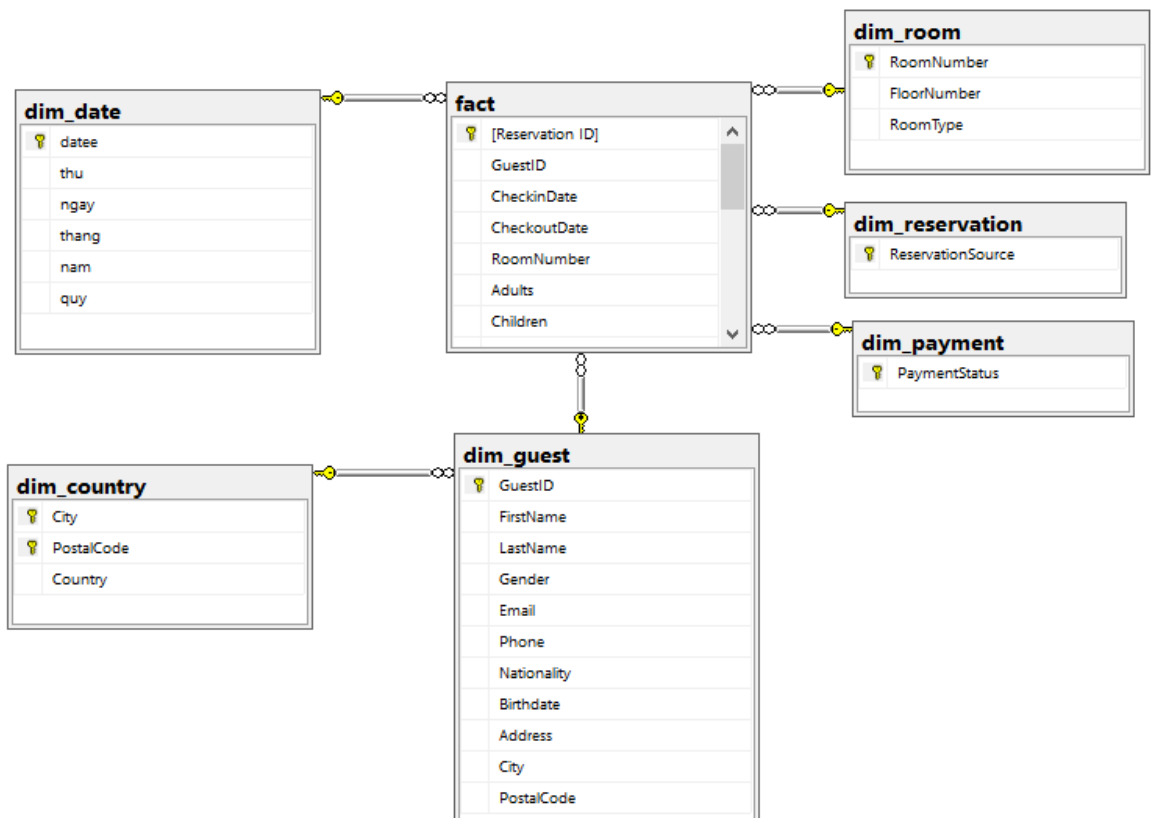
2.6.1. Quy trình ETL

- **Extract** (Trích xuất): Sử dụng thành phần Flat File Source để đọc dữ liệu từ flat file
- **Transform** (Chuyển đổi):
 - Dữ liệu được làm sạch, chuyển đổi và chuẩn hóa để phù hợp với mô hình dữ liệu
 - Các bước chuyển đổi bao gồm: lọc và sắp xếp dữ liệu, loại bỏ giá trị trùng lặp

- **Load (Nạp):** Sử dụng thành phần OLE DB Destination để nạp dữ liệu vào các bảng trong cơ sở dữ liệu đích gồm bảng fact (sự kiện) hoặc bảng dimension (chiều)



2.6.2. Lược đồ bông tuyết



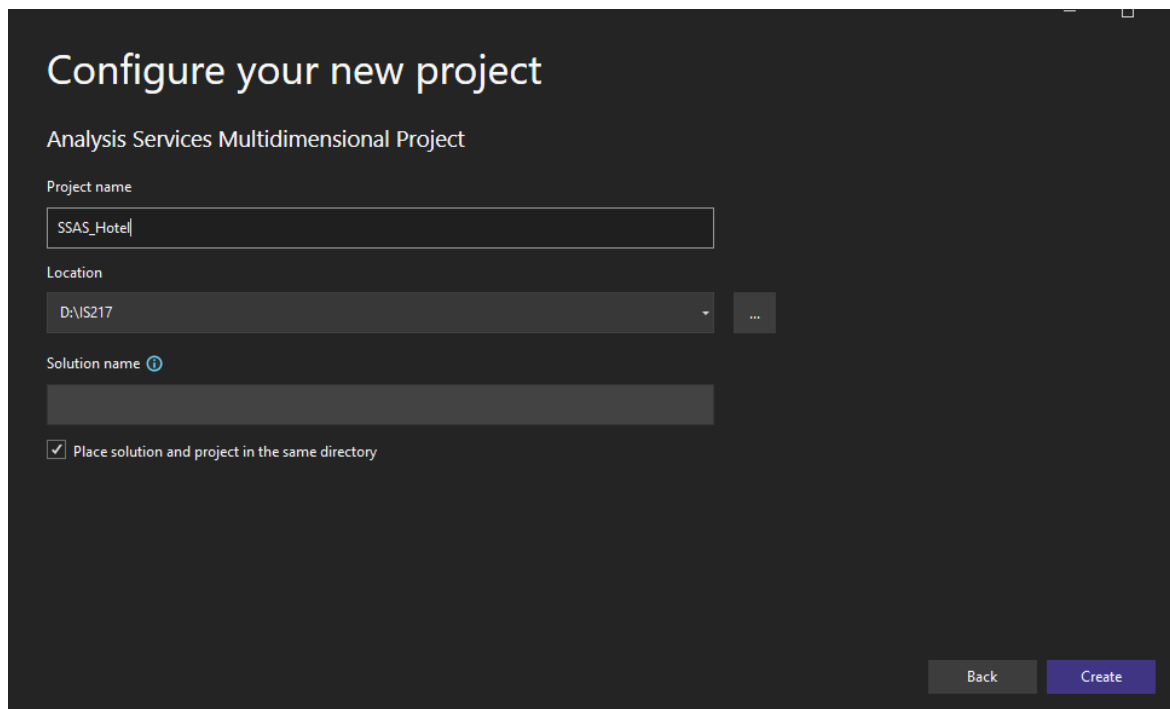
CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS)

3.1. Quá trình SSAS

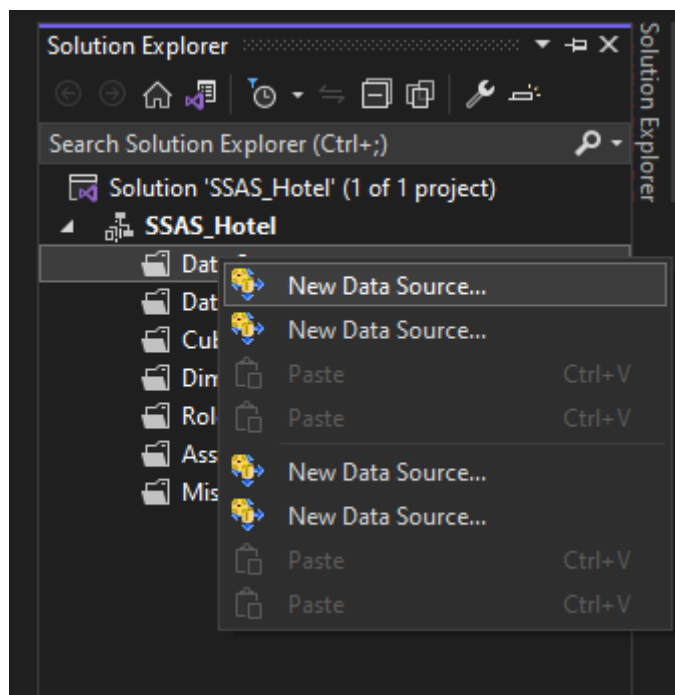
3.1.1. Quá trình thực hiện trên Visual Data Studio

Bước 1: Tạo project SSAS trong Visual Studio 2022

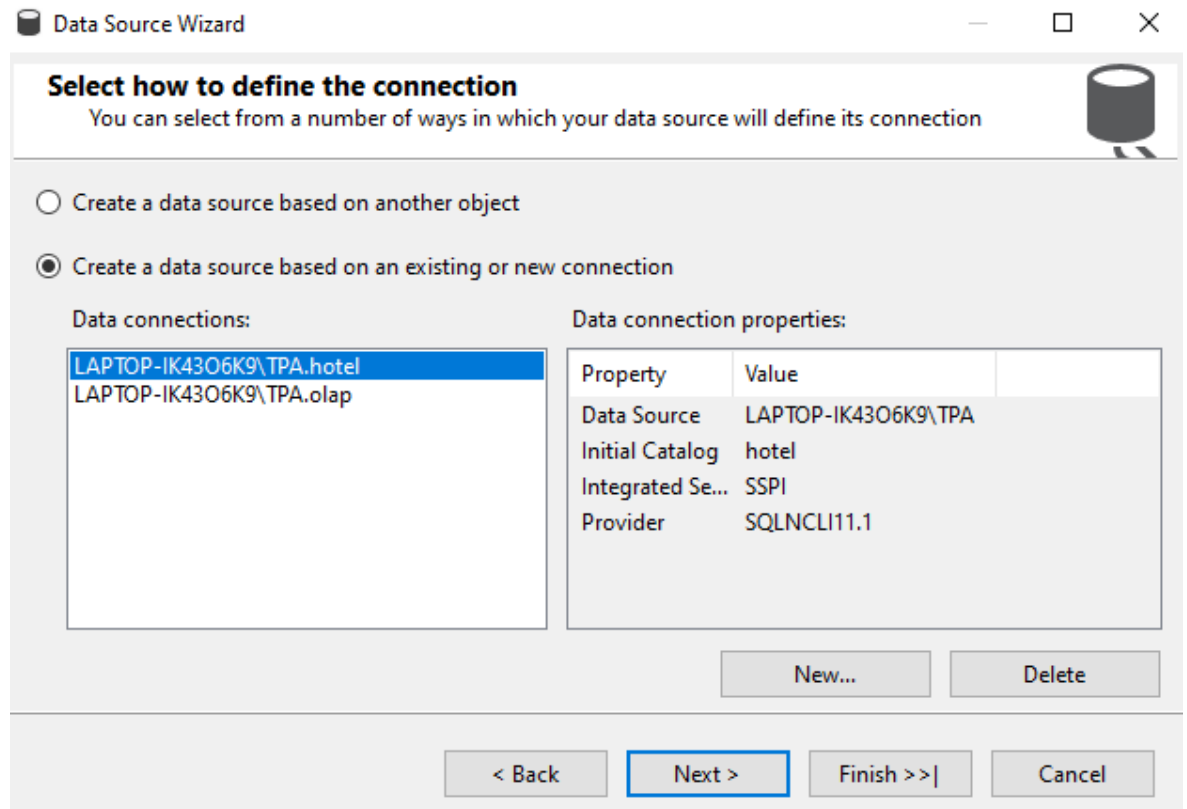
Bước 2: Chọn Analysis Services Multidimensional Project và đặt tên



Bước 3: Chọn Data Source để chọn Database đã tạo ở quá trình SSIS



Bước 4: Chọn “Create a data source based on an existing or new connection” và chọn “New” để tạo Data Connection mới



Bước 5: Điền tên Server, tên Database muốn kết nối và chọn “OK”

Connection Manager

Provider: Native OLE DB\SQL Server Native Client 11.0

Server name: LAPTOP-IK43O6K9\TPA Refresh

Log on to the server

Authentication: Windows Authentication

User name: Password: Save my password

Connect to a database

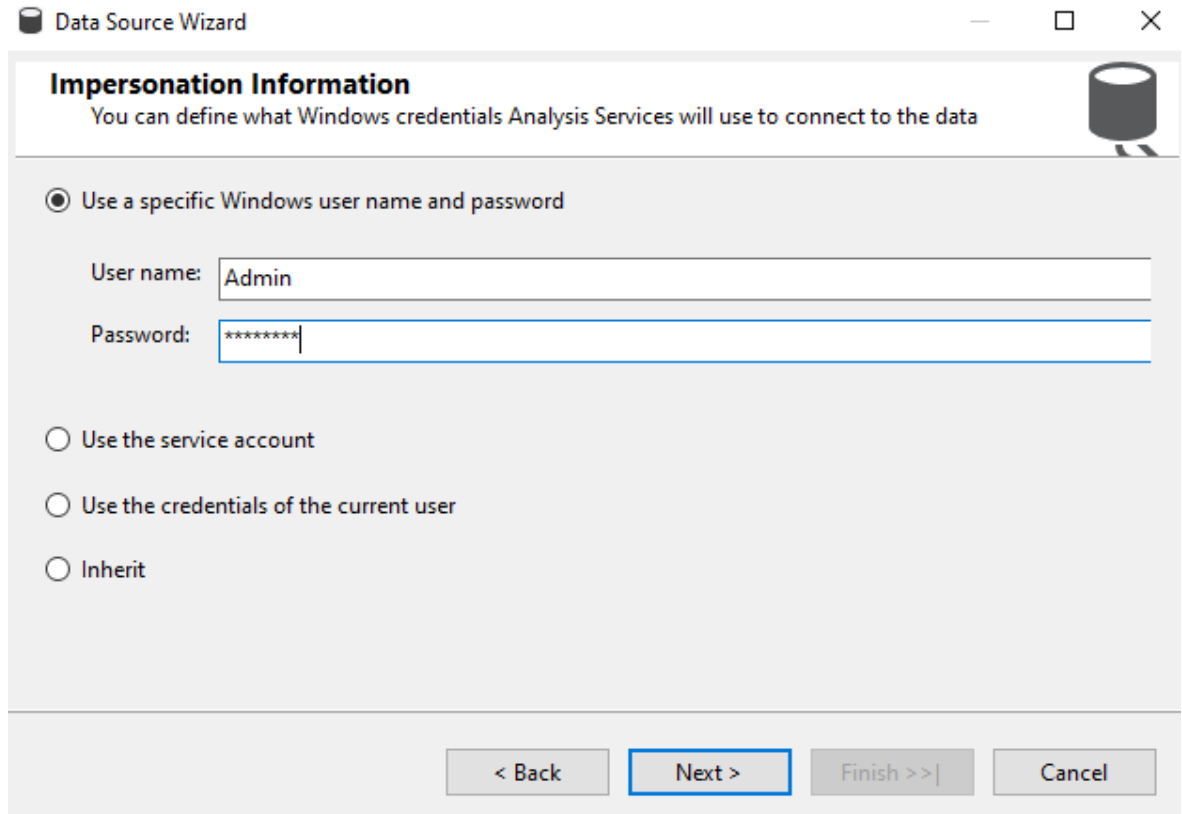
☒ Select or enter a database name: SSIS

☐ Attach a database file: Browse...

Logical name:

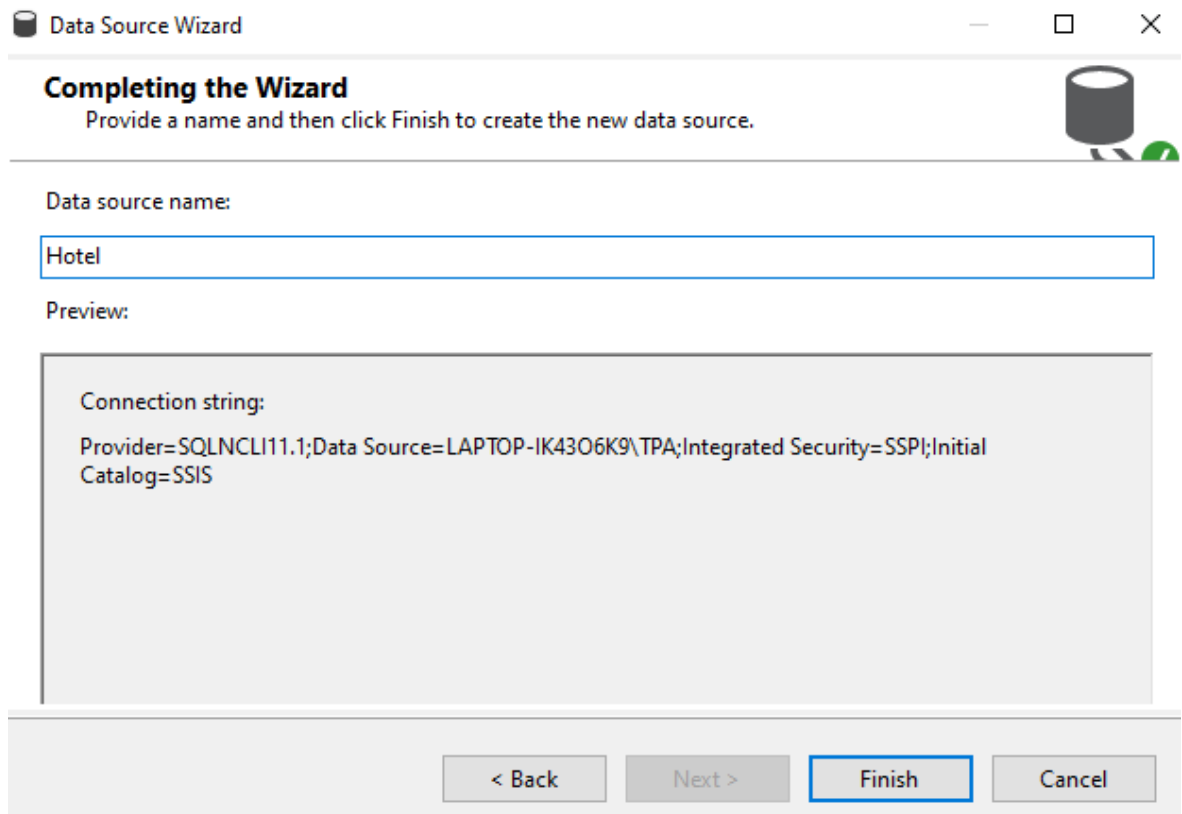
Test Connection OK Cancel Help

Bước 6: Điền username và password của Windows để kết nối



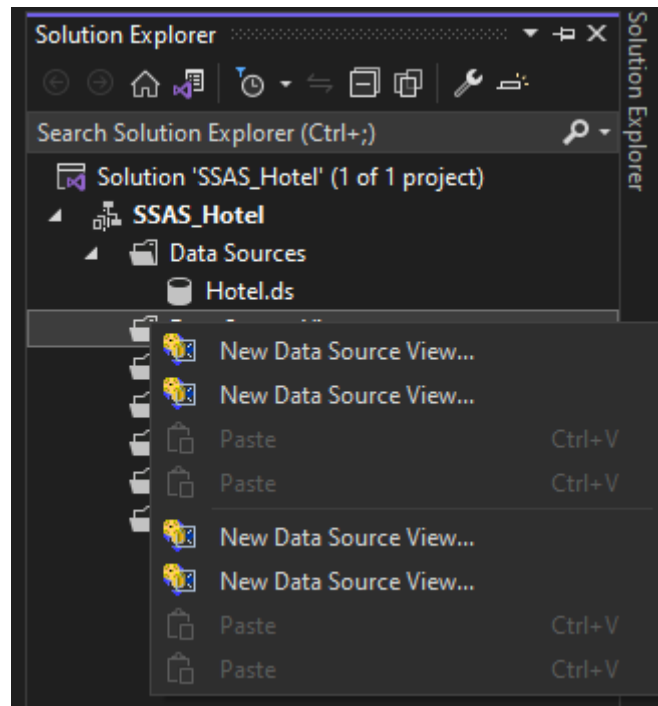
The screenshot shows the 'Data Source Wizard' window with the title bar 'Data Source Wizard'. The main heading is 'Impersonation Information' with a subtitle 'You can define what Windows credentials Analysis Services will use to connect to the data'. There are four radio button options: 'Use a specific Windows user name and password' (selected), 'Use the service account', 'Use the credentials of the current user', and 'Inherit'. The 'User name' field contains 'Admin' and the 'Password' field contains '*****'. At the bottom, there are four buttons: '< Back', 'Next >' (highlighted with a blue border), 'Finish >>|', and 'Cancel'.

Bước 7: Đặt tên cho Data source và chọn “Finish”

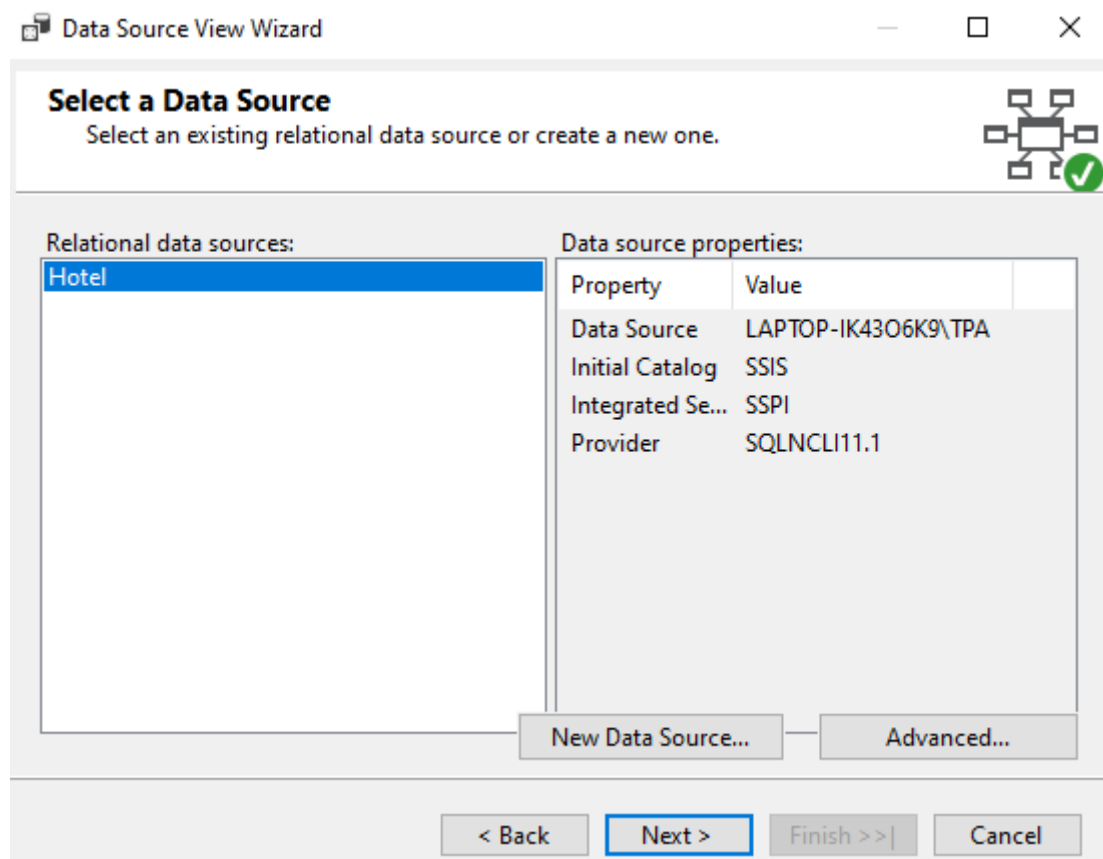


The screenshot shows the 'Data Source Wizard' window with the title bar 'Data Source Wizard'. The main heading is 'Completing the Wizard' with a subtitle 'Provide a name and then click Finish to create the new data source.' The 'Data source name' field contains 'Hotel'. Below it, the 'Preview' section shows the 'Connection string' as 'Provider=SQLNCLI11.1;Data Source=LAPTOP-1K43O6K9\TPA;Integrated Security=SSPI;Initial Catalog=SSIS'. At the bottom, there are four buttons: '< Back', 'Next >', 'Finish' (highlighted with a blue border), and 'Cancel'.

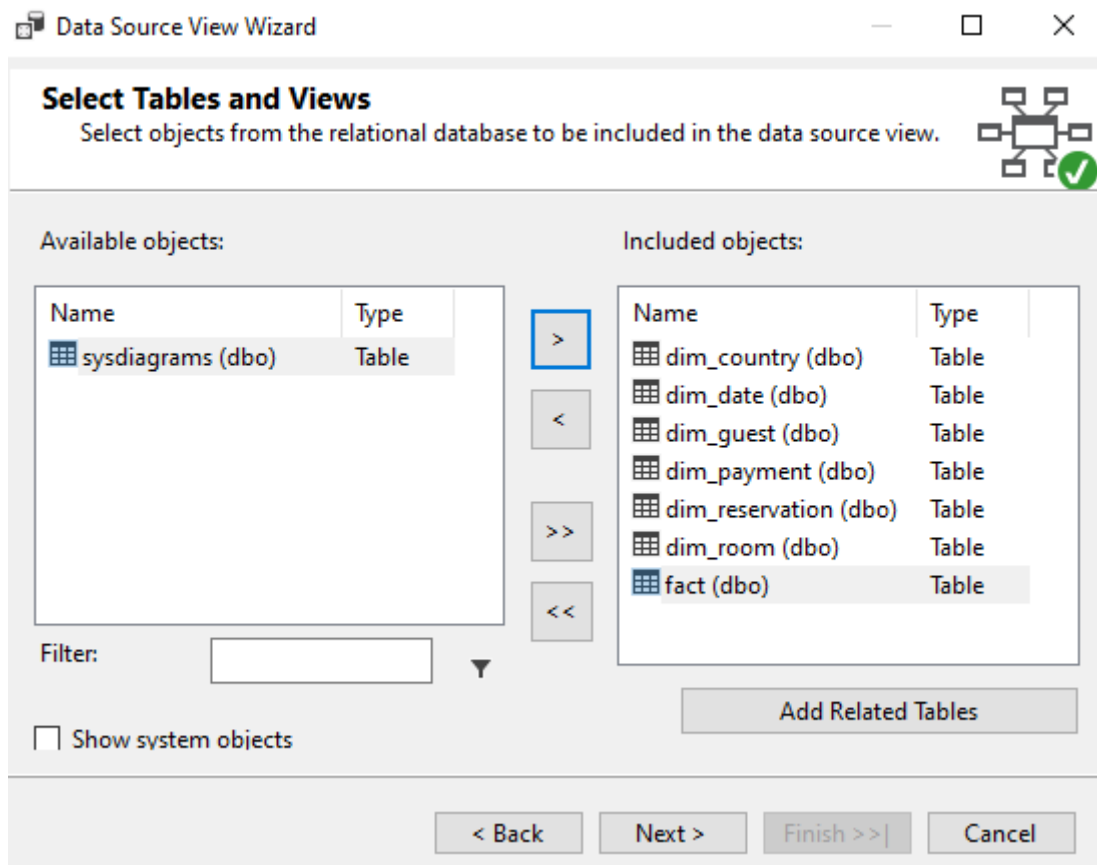
Bước 8: Trong mục Data Source Views, chọn “New Data source view”



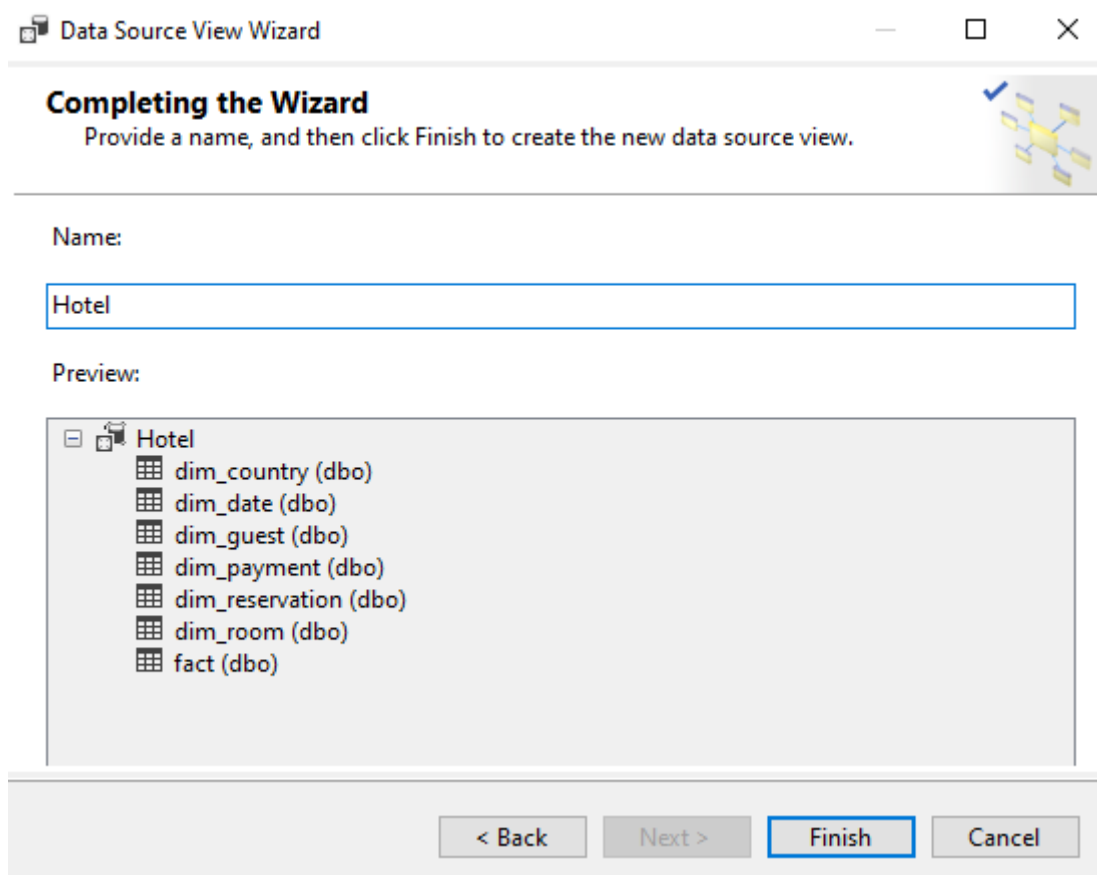
Bước 9: Chọn Data source vừa tạo và bấm “Next”



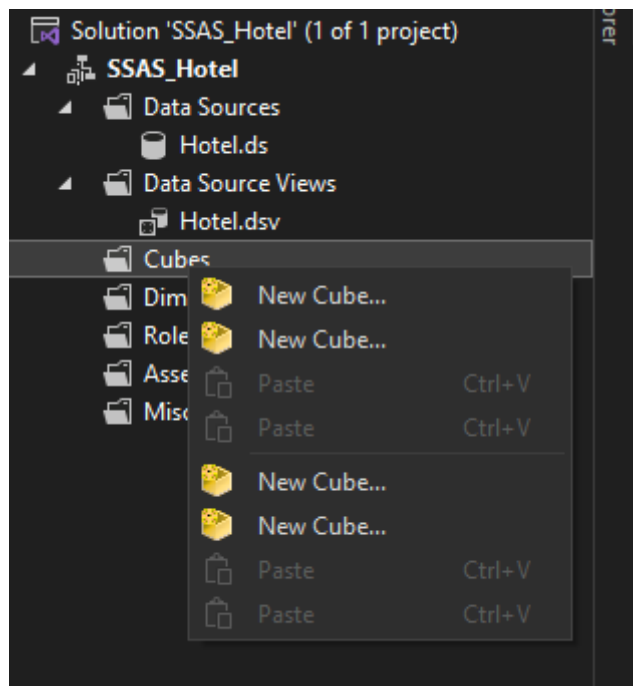
Bước 10: Chọn các bảng Dim và Fact cần sử dụng và chọn “Next”



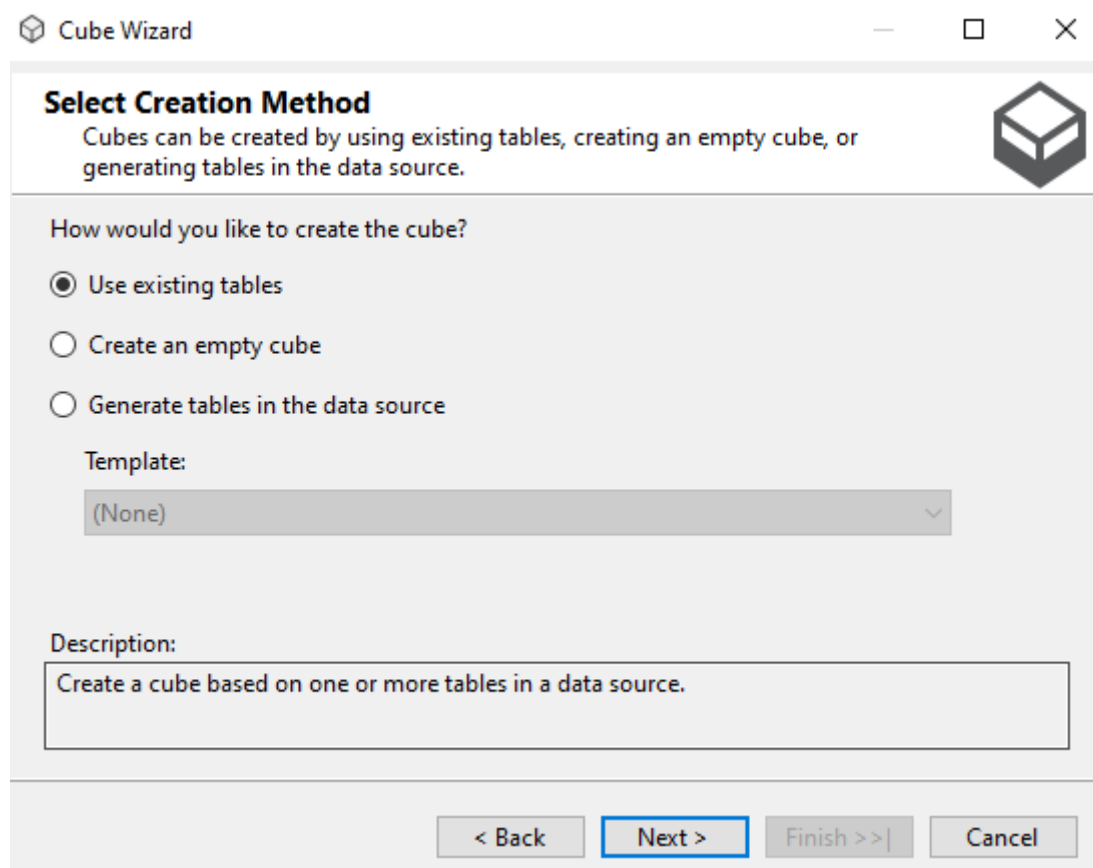
Bước 11: Đặt tên tùy thích và “Finish”



Bước 12: Trong mục Cubes, chọn “New cube”



Bước 13: Chọn “Use existing tables”



Bước 14: Chọn bảng có chứa Measures dùng để tính toán, ở đây là bảng Fact

Cube Wizard

Select Measure Group Tables

Select a data source view or diagram and then select the tables that will be used for measure groups.

Data source view:
Hotel

Measure group tables:

- ☐ dim_country
- ☐ dim_date
- ☐ dim_guest
- ☐ dim_payment
- ☐ dim_reservation
- ☐ dim_room
- ☒ fact

Suggest

< Back Next > Finish >>| Cancel

Bước 15: Chọn các thuộc tính dùng để tính toán, mặc định sẽ chọn hết

Cube Wizard

Select Measures

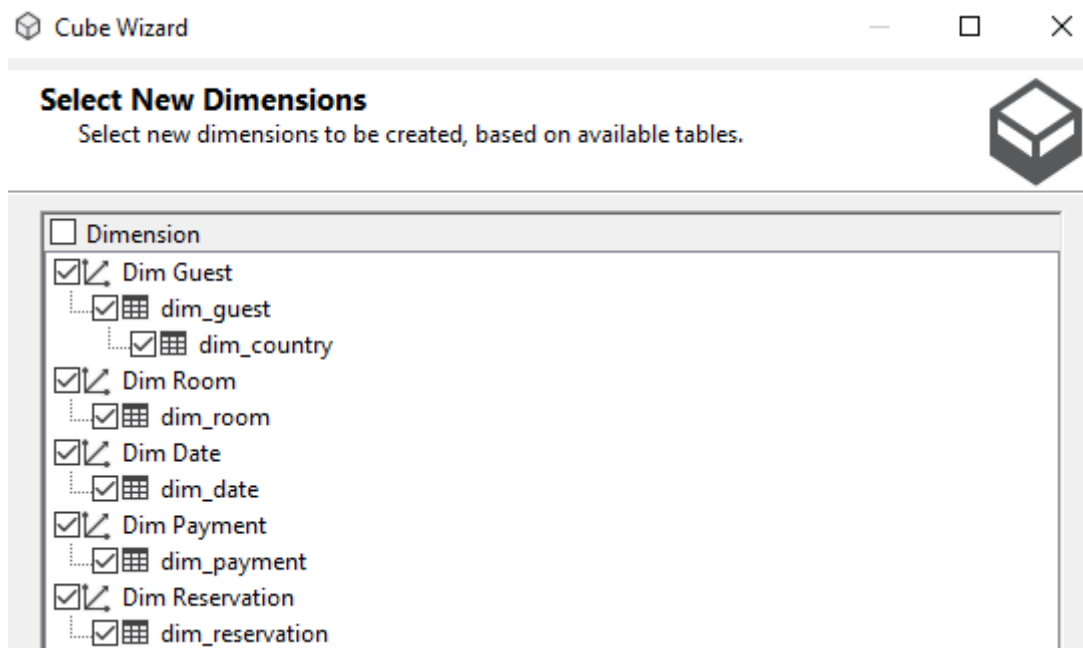
Select measures that you want to include in the cube.

☒ Measure

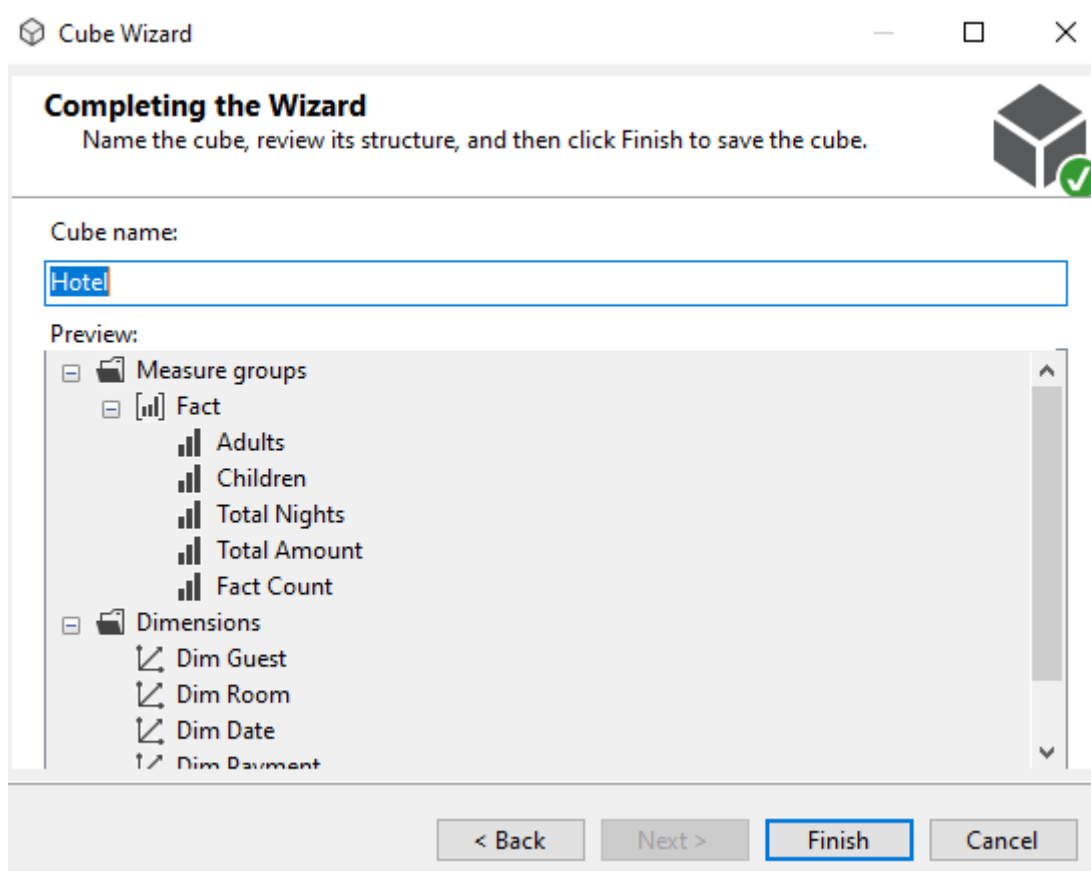
- ☒ Fact
 - ☒ Adults
 - ☒ Children
 - ☒ Total Nights
 - ☒ Total Amount
 - ☒ Fact Count

< Back Next > Finish >>| Cancel

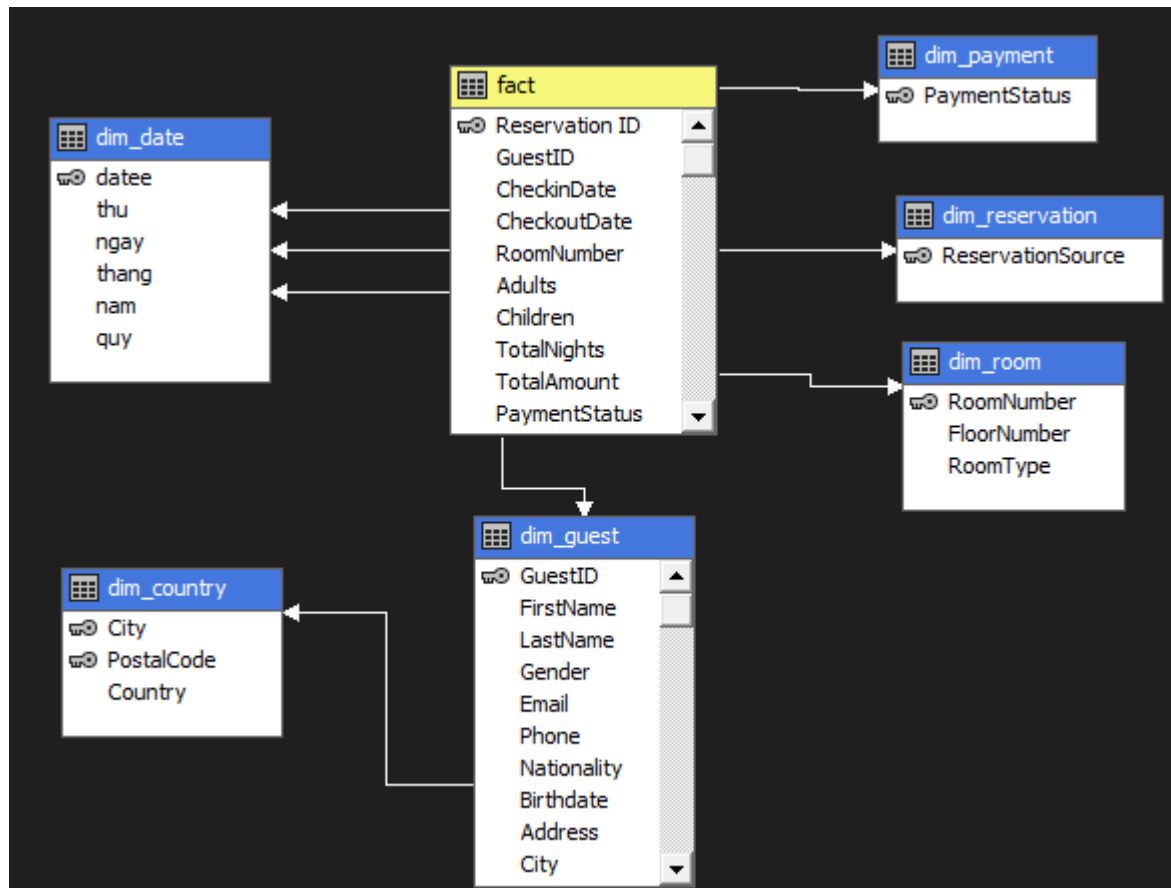
Bước 16: Chọn các bảng Dim



Bước 17: Đặt tên cho Cube và xem lại cấu trúc

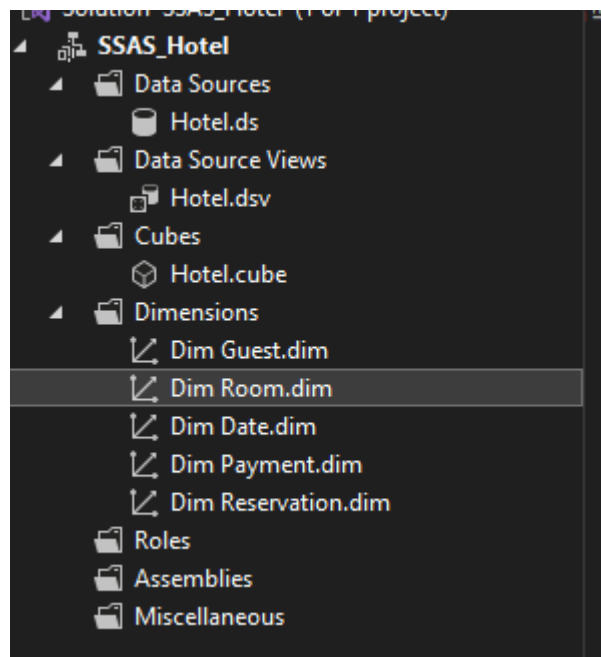


Bước 18: Sau khi hoàn thành, màn hình sẽ hiển thị lược đồ Data Warehouse của Database



3.1.2. Điều chỉnh và phân cấp các Dimension

Bước 1: Tại Dim Room, nhấp chuột hai lần để mở Dim Room

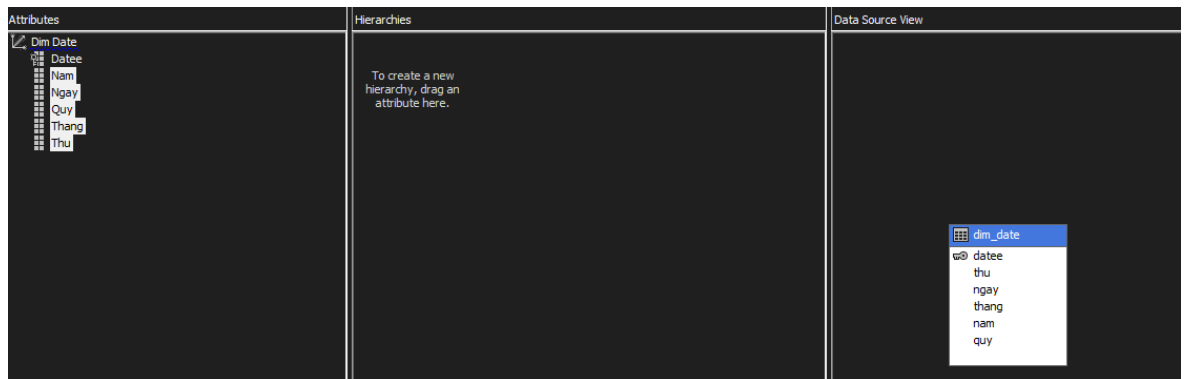


Bước 2: Kéo thả các thuộc tính của Dim Room vào Attributes

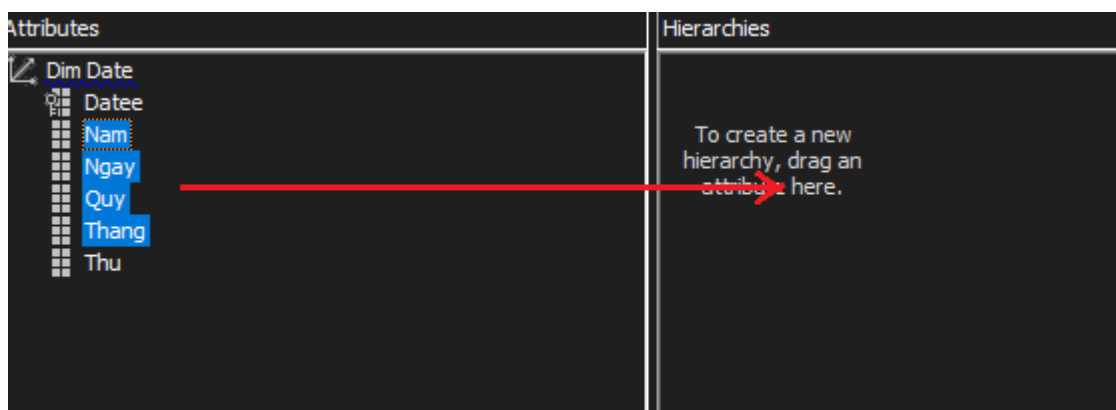


Bước 3: Thực hiện tương tự với bảng dim_guest, dim_country, dim_payment, dim_reservation

Bước 4: Với bảng dim_date, kéo thả các thuộc tính vào Attributes (tương tự bước 2)

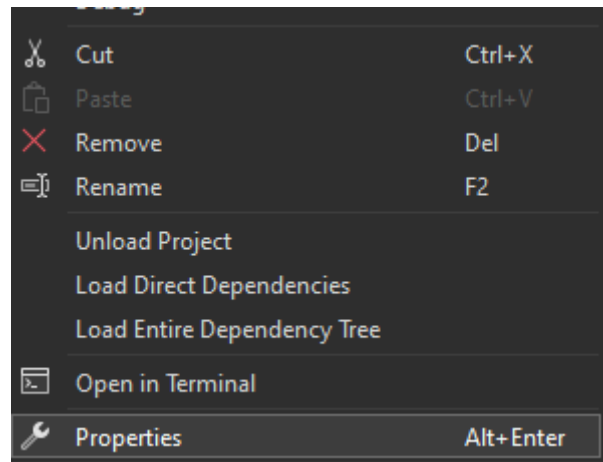


Bước 5: Kéo các thuộc tính vừa để vào trong Attributes vào Hierachies để tạo cây phân cấp

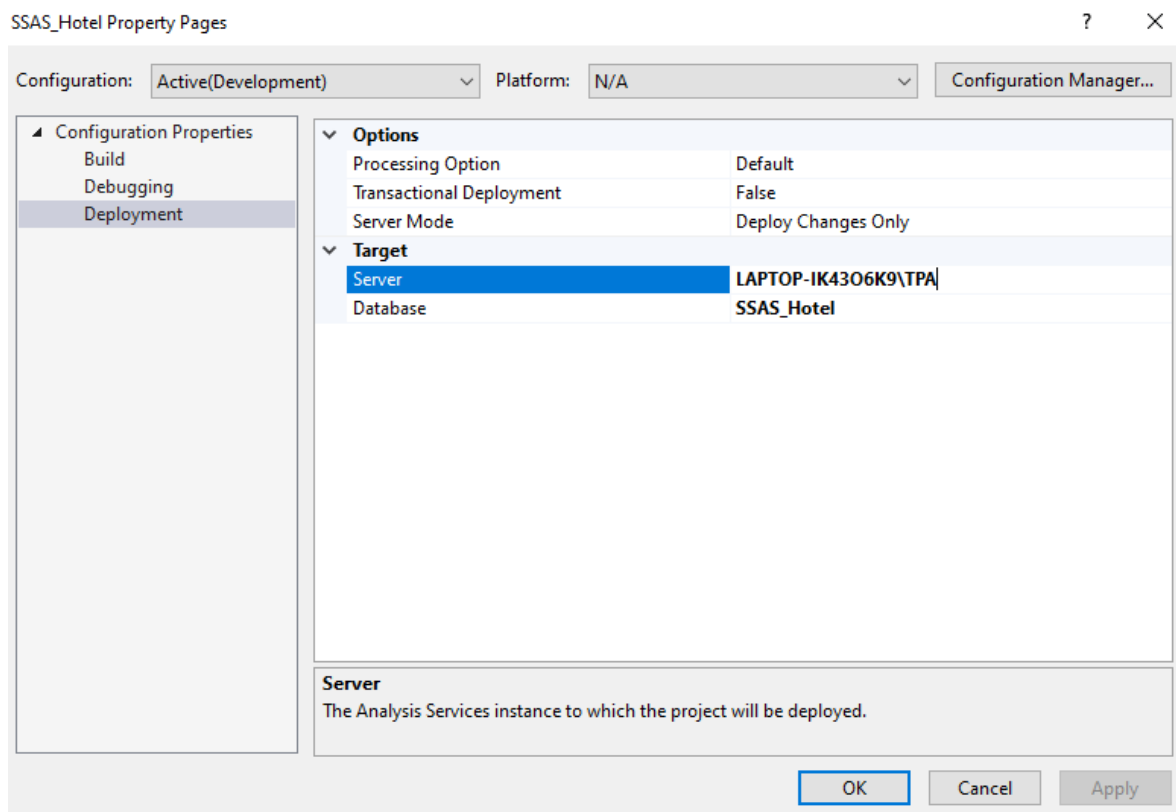


Khi kéo thả các thuộc tính vào trong Hierarchies để tạo cây phân cấp, hệ thống sẽ có thể xác định rõ ràng từng mức phân cấp mà không bị nhầm lẫn bởi các giá trị trùng lặp. Việc tạo mới "Ngày, Tháng, Quy, Nam" nhằm đảm bảo tính duy nhất của các giá trị trong cây phân cấp, tránh sự trùng lặp và đảm bảo hệ thống có thể phân tích dữ liệu một cách chính xác và hiệu quả

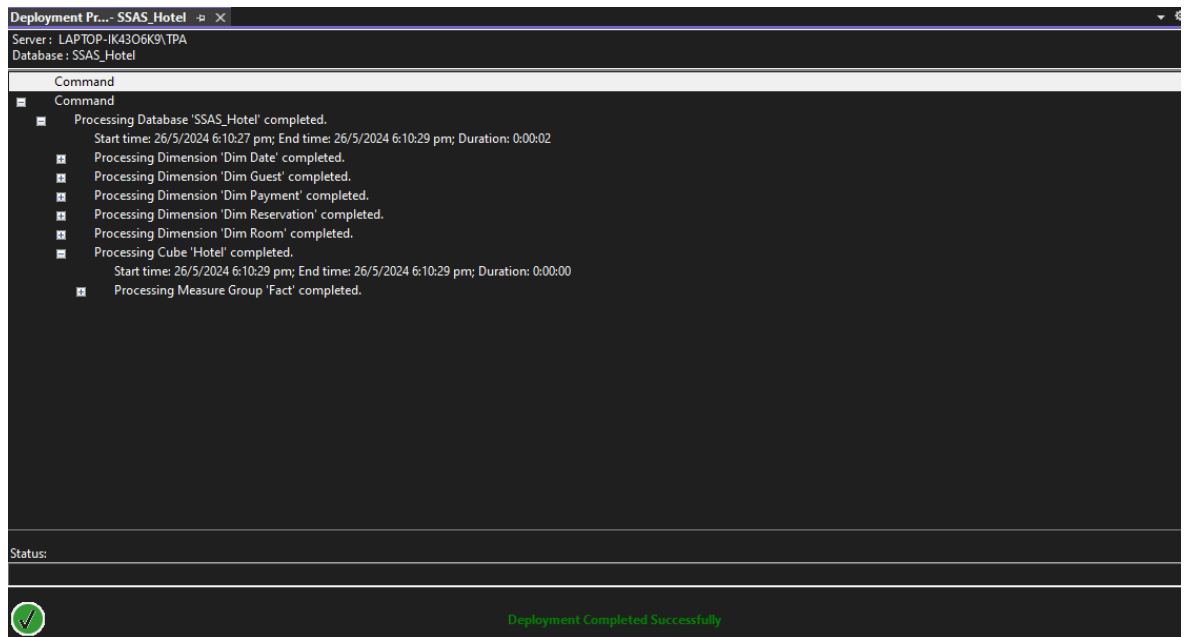
Bước 6: Nhấp chuột phải vào thư mục gốc “SSAS_Hotel” và chọn Properties



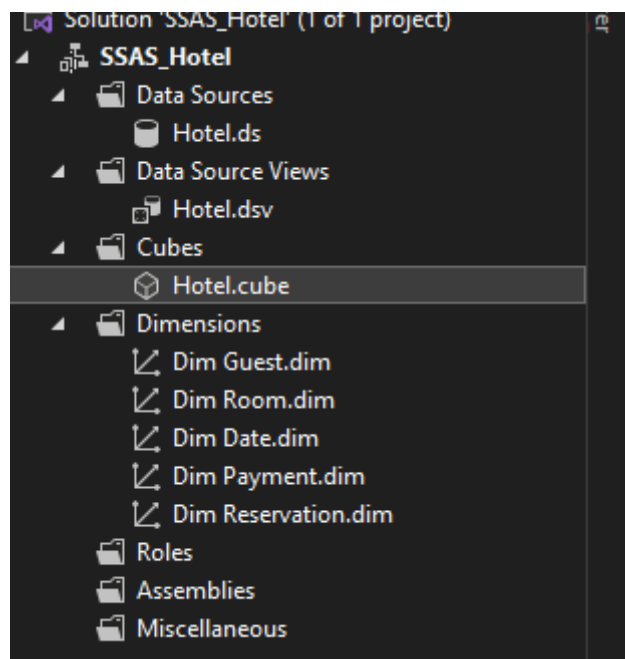
Bước 7: Trong mục Deployment, sửa tên Server và chọn “OK”



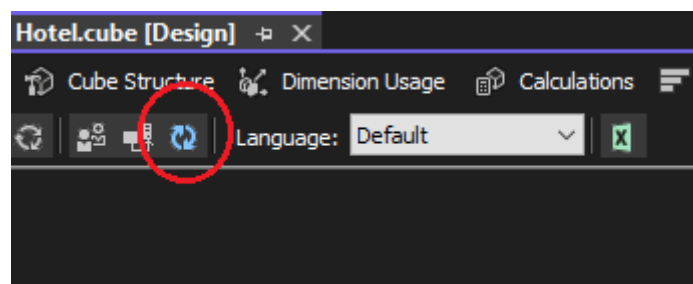
Bước 8: Nhấn chuột phải và chọn Deploy trong “SSAS_Hotel” để triển khai dự án và kiểm tra các cube, dimension



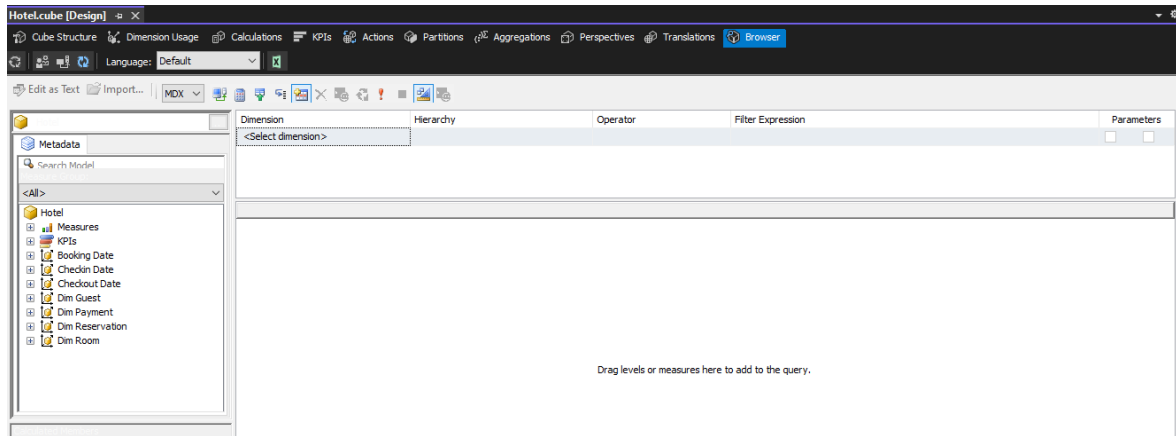
Bước 9: Trong Hotel.cube, chuột phải chọn Browse



Bước 10: Mặc định màn hình sẽ không hiển thị gì hết, cần load lại trang để hiện giao diện dùng để truy vấn



Sau khi load, hiển thị giao diện dùng để thực hiện các công việc như kéo thả, truy vấn MDX



3.2. Truy vấn MDX

3.2.1. Doanh thu từng loại phòng qua các năm

- Sử dụng Roll up
- Câu lệnh truy vấn MDX:

```
SELECT NON EMPTY [Measures].[Total Amount] ON COLUMNS,  
NON EMPTY CROSSJOIN ([Dim Room].[Room Type].children, [Checkin  
Date].[Nam].children) ON ROWS  
FROM [Hotel];
```

- Kết quả:

```
SELECT NON EMPTY [Measures].[Total Amount] ON COLUMNS,  
NON EMPTY CROSSJOIN ([Dim Room].[Room Type].children, [Checkin Date].[Nam].children) ON ROWS  
FROM [Hotel];
```

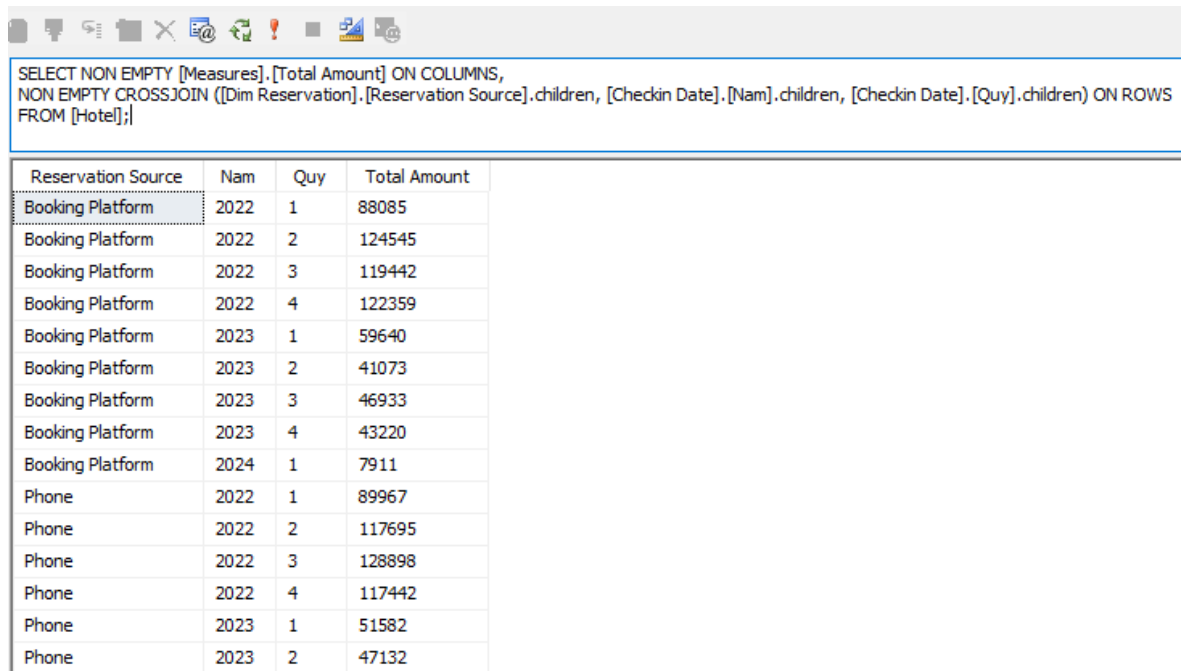
Room Type	Nam	Total Amount
Deluxe	2022	588244
Deluxe	2023	245827
Deluxe	2024	14360
Standard	2022	307262
Standard	2023	142798
Standard	2024	6319
Suite	2022	943462
Suite	2023	413713
Suite	2024	22053

3.2.2. Doanh thu từng nguồn đặt phòng của từng quý qua các năm

- Sử dụng Drill down
- Câu lệnh truy vấn MDX:

```
SELECT NON EMPTY [Measures].[Total Amount] ON COLUMNS,  
NON EMPTY CROSSJOIN ([Dim Reservation].[Reservation  
Source].children, [Checkin Date].[Nam].children, [Checkin  
Date].[Quy].children) ON ROWS  
FROM [Hotel];
```

- Kết quả:



The screenshot shows a BI tool window with a query editor at the top containing the same MDX query as above. Below the editor is a table with the following data:

Reservation Source	Nam	Quy	Total Amount
Booking Platform	2022	1	88085
Booking Platform	2022	2	124545
Booking Platform	2022	3	119442
Booking Platform	2022	4	122359
Booking Platform	2023	1	59640
Booking Platform	2023	2	41073
Booking Platform	2023	3	46933
Booking Platform	2023	4	43220
Booking Platform	2024	1	7911
Phone	2022	1	89967
Phone	2022	2	117695
Phone	2022	3	128898
Phone	2022	4	117442
Phone	2023	1	51582
Phone	2023	2	47132

3.2.3. Danh sách khách hàng đặt phòng loại ‘Standard’

- Sử dụng Slice
- Câu lệnh truy vấn MDX:

```
SELECT NON EMPTY { [Measures].[Adults], [Measures].[Children],  
[Measures].[Total Nights] } ON COLUMNS,  
NON EMPTY { ([Fact].[Reservation ID].[Reservation ID].ALLMEMBERS  
* [Dim Guest].[Guest ID].[Guest ID].ALLMEMBERS ) } ON ROWS  
FROM [Hotel]  
WHERE [Dim Room].[Room Type].&[Standard];
```

- Kết quả:


```
SELECT NON EMPTY { [Measures].[Adults], [Measures].[Children], [Measures].[Total Nights] } ON COLUMNS,
NON EMPTY { ([Fact].[Reservation ID].[Reservation ID].ALLMEMBERS * [Dim Guest].[Guest ID].[Guest ID].ALLMEMBERS ) } ON ROWS
FROM [Hotel]
WHERE [Dim Room].[Room Type].&[Standard];
```

Reservation ID	Guest ID	Adults	Children	Total Nights
1003	103	1	0	2
1010	110	1	1	1
1020	120	2	2	4
1038	138	2	1	2
1041	141	3	0	1
1044	144	3	0	2
1045	145	2	2	1
1050	150	1	0	3
1051	151	2	1	3
1054	154	4	1	1
1056	156	3	2	5
1058	158	4	0	5
1061	161	1	2	5
1063	163	2	1	3

3.2.4. Tổng doanh thu theo loại phòng và tháng năm 2023

- Sử dụng Pivot
- Câu lệnh truy vấn MDX và kết quả:

```
SELECT
    NON EMPTY ([Dim Room].[Room Type].children) ON ROWS ,
    [Checkin Date].[Thang].children ON COLUMNS
FROM
    [Hotel]
WHERE
    ([Measures].[Total Amount], [Checkin Date].[Nam].[Nam].&[2023]);
```

	1	10	11	12	2	3	4	5	6	7	8	9
Deluxe	39027	18972	22066	22005	14617	19419	16306	20542	20276	15017	20593	16987
Standard	21996	13128	9667	13456	10456	11048	11730	10433	10341	9896	12251	8396
Suite	65767	38560	33758	34049	20054	29086	28678	31617	35319	36986	29823	30016

3.2.5. Tổng số lượng khách theo quốc gia và loại phòng năm 2023

- Tạo Calculate member **[SumAdultsChildren]**: Tính tổng số lượng khách gồm người lớn (Adults) và trẻ em (Children)

Name:

Parent Properties

Parent hierarchy:

Parent member:

Expression

✓ No issues found

➤ Câu lệnh truy vấn MDX:

```
WITH
  MEMBER [Measures].[SumAdultsChildren] AS
    [Measures].[Adults] + [Measures].[Children]
SELECT
  NON EMPTY ([Dim Room].[Room Type].children) ON COLUMNS ,
  NON EMPTY ([Dim Guest].[Country].children) ON ROWS
FROM [Hotel]
WHERE
  ([Measures].[SumAdultsChildren], [Checkin
Date].[Nam].[Nam].&[2023]);
```

➤ Kết quả:

	Deluxe	Standard	Suite
Argentina	7	20	24
Australia	23	11	22
Austria	65	30	37
Belgium	78	97	59
Brazil	47	31	38
Bulgaria	122	127	111
Canada	22	28	53
China	12	21	13
Croatia	74	81	87
Cyprus	32	55	68
Czech Republic	60	50	42
Denmark	24	23	37
Estonia	30	61	46
Finland	47	29	56
France	45	45	34
Germany	53	69	66

3.2.6. Số lượng đặt phòng theo từng tháng và từng loại phòng

➤ Sử dụng Pivot

- Câu lệnh truy vấn MDX:

```
SELECT
    NON EMPTY [Measures].[Fact Count] ON COLUMNS,
    NON EMPTY CROSSJOIN(
        [Checkin Date].[Thang].[Thang].Members,
        [Dim Room].[Room Type].[Room Type].Members
    ) ON ROWS
FROM [Hotel]
WHERE ([Checkin Date].[Nam].&[2023]);
```

- Kết quả:

```
SELECT
    NON EMPTY [Measures].[Fact Count] ON COLUMNS,
    NON EMPTY CROSSJOIN(
        [Checkin Date].[Thang].[Thang].Members,
        [Dim Room].[Room Type].[Room Type].Members
    ) ON ROWS
FROM
    [Hotel]
WHERE ([Checkin Date].[Nam].&[2023]);
```

Thang	Room Type	Fact Count
1	Deluxe	113
1	Standard	107
1	Suite	109
10	Deluxe	48
10	Standard	56
10	Suite	58
11	Deluxe	60
11	Standard	50
11	Suite	45
12	Deluxe	56
12	Standard	51
12	Suite	52
2	Deluxe	33
2	Standard	42
2	Suite	34
3	Deluxe	54

3.2.7. Thông tin khách từ Thụy Sĩ (Switzerland) , chỉ hiển thị các cột có giá trị tổng số tiền chi tiêu lớn hơn 1000

- Sử dụng Slice
➤ Câu lệnh truy vấn MDX:

```
SELECT
    { [Measures].[Total Amount] } ON COLUMNS,
    FILTER(
```

```
NONEMPTY(  
    [Dim Guest].[Guest ID].[Guest ID].MEMBERS * [Dim  
Guest].[First Name].[First Name].MEMBERS * [Dim Guest].[Last  
Name].[Last Name].MEMBERS,  
    [Measures].[Total Amount]  
),  
    [Measures].[Total Amount] > 1000  
) ON ROWS  
FROM [Hotel]  
WHERE ([Dim Guest].[Country].&[Switzerland]);
```

➤ Kết quả:

```
SELECT  
    { [Measures].[Total Amount] } ON COLUMNS,  
    FILTER(  
        NONEMPTY(  
            [Dim Guest].[Guest ID].[Guest ID].MEMBERS * [Dim Guest].[First Name].[First Name].MEMBERS * [Dim Guest].[Last Name].[Last Name].MEMBERS,  
            [Measures].[Total Amount]  
        ),  
        [Measures].[Total Amount] > 1000  
    ) ON ROWS  
FROM [Hotel]  
WHERE  
    ([Dim Guest].[Country].&[Switzerland]);
```

Guest ID	First Name	Last Name	Total Amount
315	Aaron	House	1120
328	Lisa	Bradford	1115
336	Miguel	Rivera	1035
604	Holly	Moore	1095
2862	Jeffrey	Morales	1175
3015	Ashley	Hale	1120
4250	Christine	Parsons	1190
6473	Jacob	Alexander	1035
28722	Sarah	Cole	1220
29511	Rebecca	Brown	1085

3.2.8. Số lượng khách nữ từ Thụy Sĩ (Switzerland) với tổng số tiền chi tiêu lớn hơn 1000

- Sử dụng Dice
- Câu lệnh truy vấn MDX:

```
SELECT  
    { [Measures].[Total Amount] } ON COLUMNS,  
    FILTER(  
        NONEMPTY(  
            [Dim Guest].[Guest ID].[Guest ID].MEMBERS * [Dim  
Guest].[First Name].[First Name].MEMBERS * [Dim Guest].[Last  
Name].[Last Name].MEMBERS,  
            [Measures].[Total Amount]  
        ),  
        [Measures].[Total Amount] > 1000  
    ) ON ROWS  
FROM [Hotel]
```

WHERE ([Dim Guest].[Country].&[Switzerland], [Dim Guest].[Gender].&[Female]);

➤ Kết quả:

```
SELECT
{ [Measures].[Total Amount] } ON COLUMNS,
FILTER(
NONEMPTY(
[Dim Guest].[Guest ID].[Guest ID].MEMBERS * [Dim Guest].[First Name].[First Name].MEMBERS * [Dim Guest].[Last Name].[Last Name].MEMBERS,
[Measures].[Total Amount]
),
[Measures].[Total Amount] > 1000
) ON ROWS
FROM [Hotel]
WHERE
([Dim Guest].[Country].&[Switzerland], [Dim Guest].[Gender].&[Female]);
```

Guest ID	First Name	Last Name	Total Amount
315	Aaron	House	1120
328	Lisa	Bradford	1115
336	Miguel	Rivera	1035
604	Holly	Moore	1095
2862	Jeffrey	Morales	1175
4250	Christine	Parsons	1190
28722	Sarah	Cole	1220
29511	Rebecca	Brown	1085

3.2.9. Số lượng đặt phòng qua các nguồn đặt phòng qua các năm

➤ Sử dụng Roll up

➤ Câu lệnh truy vấn MDX:

```
SELECT NON EMPTY [Measures].[Fact Count] ON COLUMNS,
NON EMPTY CROSSJOIN ([Dim Reservation].[Reservation
Source].children, [Checkin Date].[Nam].children) ON ROWS
FROM [Hotel];
```

➤ Kết quả:

```
SELECT NON EMPTY [Measures].[Fact Count] ON COLUMNS,  
NON EMPTY CROSSJOIN ([Dim Reservation].[Reservation Source].children, [Checkin Date].[Nam].children) ON ROWS  
FROM [Hotel];
```

Reservation Source	Nam	Fact Count
Booking Platform	2022	1104
Booking Platform	2023	453
Booking Platform	2024	22
Phone	2022	1108
Phone	2023	483
Phone	2024	28
Walk-in	2022	1091
Walk-in	2023	457
Walk-in	2024	29
Website	2022	1153
Website	2023	506
Website	2024	22

3.2.10. Thông tin khách hàng nam ở phòng loại ‘Deluxe’, chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 740

- Sử dụng Dice
- Câu lệnh truy vấn MDX:

```
SELECT  
  { [Measures].[Total Amount] } ON COLUMNS,  
  FILTER(  
    NONEMPTY(  
      [Dim Guest].[Guest ID].[Guest ID].MEMBERS * [Dim  
Guest].[First Name].[First Name].MEMBERS * [Dim Guest].[Last  
Name].[Last Name].MEMBERS,  
      [Measures].[Total Amount]  
    ),  
    [Measures].[Total Amount] > 740  
  ) ON ROWS  
FROM [Hotel]  
WHERE ([Dim Room].[Room Type].&[Deluxe], [Dim  
Guest].[Gender].&[Male]);
```

- Kết quả:

```
SELECT
  { [Measures].[Total Amount] } ON COLUMNS,
  FILTER(
    NONEMPTY(
      [Dim Guest].[Guest ID].[Guest ID].MEMBERS * [Dim Guest].[First Name].[First Name].MEMBERS * [Dim Guest].[Last Name].[Last Name].MEMBERS,
      [Measures].[Total Amount]
    ),
    [Measures].[Total Amount] > 740
  ) ON ROWS
FROM [Hotel]
WHERE
  ([Dim Room].[Room Type].&[Deluxe], [Dim Guest].[Gender].&[Male]);
```

Guest ID	First Name	Last Name	Total Amount
850	Alexander	Solis	745
1545	Kyle	Lewis	750
2395	Brittany	Price	750
4812	Victoria	White	745
4823	Randall	Juarez	750
4912	Susan	Harris	745
5608	Mark	Golden	750
6671	Katelyn	Mcgee	750
29441	William	Sullivan	745

3.2.11. Tìm các khách hàng đặt phòng có số lượng đêm ở lớn hơn 4 trong quý 4/2023



- Tạo Name set **[SoDemLonHon4]**: Lọc ra các Guest ID có số lượng đêm ở lớn hơn 4



- Câu lệnh truy vấn MDX:

```
SELECT NON EMPTY { [Measures].[Total Nights] } ON COLUMNS,
  NON EMPTY { ([Dim Guest].[Guest ID].[Guest ID].ALLMEMBERS * [Dim
Guest].[First Name].[First Name].ALLMEMBERS * [Dim Guest].[Last
Name].[Last Name].ALLMEMBERS ) } ON ROWS
FROM ( SELECT ( { [Checkin Date].[date].[Nam].&[2023], [Checkin
Date].[date].[Quy].&[4]&[2023] } ) ON COLUMNS
FROM ( SELECT ( [SoDemLonHon4] ) ON COLUMNS FROM [Hotel]))
WHERE ( [Checkin Date].[date].CurrentMember );
```

- Kết quả:

Dimension	Hierarchy	Operator	Filter Expression
Dim Guest	 Guest ID	In	SoDemLonHon4
Checkin Date	 Checkin Date.date	Equal	{ 2023, 4 }
<Select dimension>			

Guest ID	First Name	Last Name	Total Nights
108	Gabriel	Gonzalez	4
161	Nancy	Campbell	5
197	Melissa	Martinez	5
203	Kelly	Thompson	4
215	James	Miller	4
270	Jamie	Morales	4
309	Melissa	Jones	5
468	Michelle	Gonzalez	4
534	Tiffany	Fletcher	4
687	Lawrence	Ewing	5
1140	Amber	Hoffman	4
1151	Michelle	Warner	4
1185	Laura	Bullock	4
1413	Kelly	Davis	5
1436	Jonathan	Lopez	4
1579	Steven	Bright	4
1711	Aaron	Sosa	5
1739	Leslie	Hodges	5

3.2.12. Mỗi quý trong năm 2023, cho biết tháng nào có tổng doanh thu cao nhất

➤ Câu lệnh truy vấn MDX:

```
SELECT [Measures].[Total Amount] ON COLUMNS,
NON EMPTY GENERATE(
    [Checkin Date].[Quy].children,
    TOPCOUNT([Checkin Date].[Quy].CURRENTMEMBER * [Checkin
Date].[Thang].children,1,[Measures].[Total Amount])
) ON ROWS
FROM [Hotel]
WHERE [Checkin Date].[Nam].&[2023];
```

➤ Kết quả:


```
SELECT [Measures].[Total Amount] ON COLUMNS,
NON EMPTY GENERATE(
[Checkin Date].[Quy].children,
TOPCOUNT([Checkin Date].[Quy].CURRENTMEMBER * [Checkin Date].[Thang].children, 1, [Measures].[Total Amount])
) ON ROWS
FROM [Hotel]
WHERE [Checkin Date].[Nam].&[2023];
```

Quy	Thang	Total Amount
1	1	126790
2	6	65936
3	8	62667
4	10	70660

3.2.13. Doanh thu của loại phòng ‘Deluxe’ qua từng năm

➤ Câu lệnh truy vấn MDX:

```
SELECT NON EMPTY [Checkin Date].[Nam].children ON ROWS,
[Measures].[Total Amount] ON COLUMNS
FROM [Hotel]
WHERE [Dim Room].[Room Type].&[Deluxe];
```

➤ Kết quả:

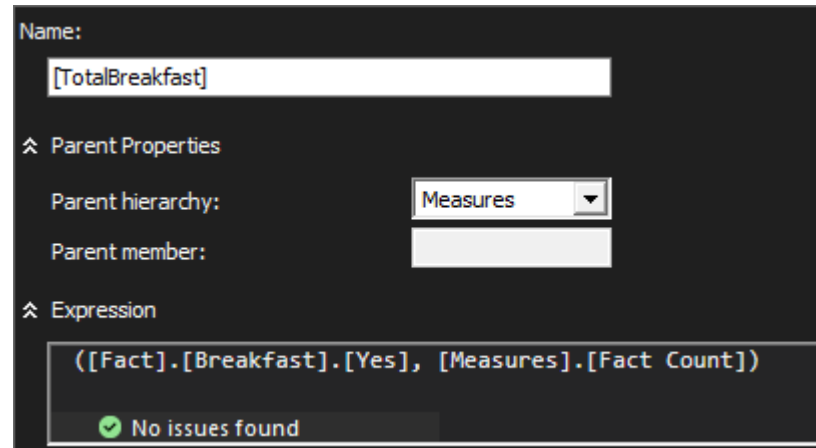
```
SELECT NON EMPTY [Checkin Date].[Nam].children ON ROWS, [Measures].[Total Amount] ON COLUMNS
FROM [Hotel]
WHERE [Dim Room].[Room Type].&[Deluxe];
```

Nam	Total Amount
2022	588244
2023	245827
2024	14360

3.2.14. Phần trăm số lượng đặt phòng có sử dụng bữa sáng

➤ Tạo Calculate member:

- ✓ **[TotalBreakfast]:** Tính tổng số lượng đơn đặt phòng có sử dụng bữa sáng



Name: [TotalBreakfast]

Parent Properties

Parent hierarchy: Measures

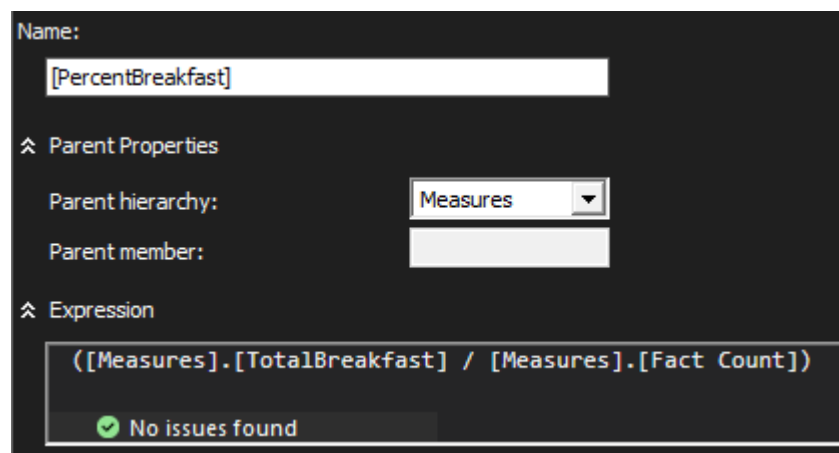
Parent member:

Expression

`([Fact].[Breakfast].[Yes], [Measures].[Fact Count])`

✓ No issues found

- ✓ **[PercentBreakfast]**: Tính phần trăm đơn đặt phòng có sử dụng bữa sáng



Name: [PercentBreakfast]

Parent Properties

Parent hierarchy: Measures

Parent member:

Expression

`([Measures].[TotalBreakfast] / [Measures].[Fact Count])`

✓ No issues found

- Câu lệnh truy vấn MDX:

```
WITH
MEMBER [Measures].[TotalBreakfast] AS
    ([Fact].[Breakfast].[Yes], [Measures].[Fact Count])

MEMBER [Measures].[PercentBreakfast] AS
    ([Measures].[TotalBreakfast] / [Measures].[Fact Count]) * 100,
    FORMAT_STRING = "0.00"

SELECT
    {[Measures].[TotalBreakfast], [Measures].[PercentBreakfast]}
ON COLUMNS
FROM [Hotel];
```

- Kết quả:

```
WITH
MEMBER [Measures].[TotalBreakfast] AS
    ([Fact].[Breakfast].[Yes], [Measures].[Fact Count])

MEMBER [Measures].[PercentBreakfast] AS
    ([Measures].[TotalBreakfast] / [Measures].[Fact Count]) * 100,
    FORMAT_STRING = "0.00"

SELECT
    {[Measures].[TotalBreakfast], [Measures].[PercentBreakfast]} ON COLUMNS
FROM
    [Hotel];
```

TotalBreakfast	PercentBreakfast
3222	49.907063197026

3.2.15. Phần trăm doanh thu từ mỗi nguồn đặt phòng năm 2023

➤ Tạo Calculate member:

✓ **[TotalRevenueByReservationSource]**: Tổng doanh thu một nguồn đặt phòng

Name:

Parent Properties

Parent hierarchy:

Parent member:

Change

Expression

✓ No issues found

✓ **[TotalRevenueAllReservationSources]**: Tổng doanh thu tất cả các nguồn đặt phòng

Name:

Parent Properties

Parent hierarchy:

Parent member:

Change

Expression

✓ No issues found

- ✓ **[PercentRevenueByReservationSource]:** Tính phần trăm doanh thu nguồn đặt phòng, lấy tổng doanh thu một nguồn chia cho tổng doanh thu của tất cả các nguồn đặt phòng

The screenshot shows the SSAS cube editor interface. The 'Name' field is set to '[PercentRevenueByReservationSource]'. Under 'Parent Properties', the 'Parent hierarchy' is set to 'Measures' and the 'Parent member' is empty, with a 'Change' button next to it. The 'Expression' field contains the MDX formula:
$$([Measures].[TotalRevenueByReservationSource] / [Measures].[TotalRevenueAllReservationSources])$$
. At the bottom, a green checkmark and the text 'No issues found' indicate the formula is valid.

- Câu lệnh truy vấn MDX:

```
WITH
MEMBER [Measures].[TotalRevenueByReservationSource] AS
    ([Dim Reservation].[Reservation Source].CurrentMember,
    [Measures].[Total Amount])

MEMBER [Measures].[TotalRevenueAllReservationSources] AS
    SUM(
        [Dim Reservation].[Reservation Source].[Reservation
Source].MEMBERS,
        [Measures].[Total Amount]
    )

MEMBER [Measures].[PercentRevenueByReservationSource] AS
    ([Measures].[TotalRevenueByReservationSource] /
    [Measures].[TotalRevenueAllReservationSources]) * 100,
    FORMAT_STRING = "0.00"

SELECT
    {[Measures].[TotalRevenueByReservationSource],
    [Measures].[PercentRevenueByReservationSource]} ON COLUMNS,
    NON EMPTY [Dim Reservation].[Reservation Source].children ON
ROWS
FROM [Hotel]
WHERE [Checkin Date].[Nam].&[2023];
```

- Kết quả:

```

    [Measures].[Total Amount]
)

MEMBER [Measures].[PercentRevenueByReservationSource] AS
    ([Measures].[TotalRevenueByReservationSource] / [Measures].[TotalRevenueAllReservationSources]) * 100,
    FORMAT_STRING = "0.00"

SELECT
    {[Measures].[TotalRevenueByReservationSource], [Measures].[PercentRevenueByReservationSource]} ON COLUMNS,
    NON EMPTY [Dim Reservation].[Reservation Source].children ON ROWS
FROM
    [Hotel]
WHERE [Checkin Date].[Nam].&[2023];

```

Reservation Source	TotalRevenueByReservationSource	PercentRevenueByReservationSource
Booking Platform	190866	23.7887274440448
Phone	199429	24.855983388547
Walk-in	192047	23.9359222671742
Website	219996	27.4193669002341

3.2.16. Top 3 doanh thu theo quốc gia từ các đơn đặt phòng có sử dụng dịch vụ đưa đón sân bay

- Tạo Calculate member **[RevenueByAirport]**: Tính doanh thu các đơn đặt phòng có sử dụng dịch vụ đưa đón sân bay

Name:

Parent Properties

Parent hierarchy:

Parent member:

Expression

☒ No issues found

- Câu lệnh truy vấn MDX:

```

WITH
MEMBER [Measures].[RevenueByAirport] AS
    ([Dim Guest].[Country].CurrentMember, [Fact].[Airport].[Yes],
    [Measures].[Total Amount])

SELECT
    {[Measures].[RevenueByAirport]} ON COLUMNS,
    TOPCOUNT(
        NONEMPTY([Dim Guest].[Country].Children,
        [Measures].[RevenueByAirport])),
        3,

```

```
        [Measures].[RevenueByAirport]
    ) ON ROWS
FROM [Hotel]
WHERE ([Checkin Date].[Nam].&[2023] );
```

➤ Kết quả:

```
WITH
MEMBER [Measures].[RevenueByAirport] AS
    ([Dim Guest].[Country].CurrentMember, [Fact].[Airport].[Yes], [Measures].[Total Amount])

SELECT
    {[Measures].[RevenueByAirport]} ON COLUMNS,
    TOPCOUNT(
        NONEMPTY([Dim Guest].[Country].Children, [Measures].[RevenueByAirport]),
        3,
        [Measures].[RevenueByAirport]
    ) ON ROWS
FROM
    [Hotel]
WHERE ([Checkin Date].[Nam].&[2023] );
```

Country	RevenueByAirport
United States	30253
Bulgaria	21568
Romania	17891

3.2.17. Số lượng đặt phòng của từng loại phòng theo mỗi nguồn đặt phòng

➤ Câu lệnh truy vấn MDX:

```
SELECT
    NON EMPTY {[Measures].[Fact Count]} ON COLUMNS,
    NON EMPTY CROSSJOIN(
        [Dim Room].[Room Type].[Room Type].Members,
        [Dim Reservation].[Reservation Source].[Reservation
Source].Members
    ) ON ROWS
FROM [Hotel];
```

➤ Kết quả:

```
SELECT
    NON EMPTY {[Measures].[Fact Count]} ON COLUMNS,
    NON EMPTY CROSSJOIN(
        [Dim Room].[Room Type].[Room Type].Members,
        [Dim Reservation].[Reservation Source].[Reservation Source].Members
    ) ON ROWS
FROM
    [Hotel];
```

Room Type	Reservation Source	Fact Count
Deluxe	Booking Platform	553
Deluxe	Phone	529
Deluxe	Walk-in	557
Deluxe	Website	550
Standard	Booking Platform	511
Standard	Phone	553
Standard	Walk-in	504
Standard	Website	558
Suite	Booking Platform	515
Suite	Phone	537
Suite	Walk-in	516
Suite	Website	573

3.2.18. Tổng số lượng đặt phòng theo từng loại phòng và tầng

- Câu lệnh truy vấn MDX:

```
SELECT
    {[Measures].[Fact Count]} ON COLUMNS,
    NON EMPTY CROSSJOIN(
        [Dim Room].[Room Type].[Room Type].Members,
        [Dim Room].[Floor Number].[Floor Number].Members
    ) ON ROWS
FROM [Hotel];
```

- Kết quả:

```
SELECT
    {[Measures].[Fact Count]} ON COLUMNS,
    NON EMPTY CROSSJOIN(
        [Dim Room].[Room Type].[Room Type].Members,
        [Dim Room].[Floor Number].[Floor Number].Members
    ) ON ROWS
FROM
    [Hotel];
```

Room Type	Floor Number	Fact Count
Deluxe	1	584
Deluxe	2	517
Deluxe	3	508
Deluxe	4	580
Standard	1	513
Standard	2	597
Standard	3	497
Standard	4	519
Suite	1	521
Suite	2	501
Suite	3	585
Suite	4	534

3.2.19. Trung bình số đêm ở theo từng tháng trong năm 2023

- Tạo Calculate member **[AverageNightsStay]**: Tính trung bình số đêm ở

Name:

Parent Properties

Parent hierarchy:

Parent member:

Expression

```
IIF(
    [Measures].[Fact Count] = 0,
    NULL,
    [Measures].[Total Nights] / [Measures].[Fact Count]
```

✓ No issues found

- Câu lệnh truy vấn MDX:

```
WITH
MEMBER [Measures].[AverageNightsStay] AS
    IIF(
        [Measures].[Fact Count] = 0,
        NULL,
        [Measures].[Total Nights] / [Measures].[Fact Count]
```



```
),  
FORMAT_STRING = "0.00"
```

```
SELECT  
    {[Measures].[AverageNightsStay]} ON COLUMNS,  
    NON EMPTY [Checkin Date].[Thang].[Thang].Members ON ROWS  
FROM [Hotel]  
WHERE ([Checkin Date].[Nam].&[2023]);
```

➤ Kết quả:

```
FORMAT_STRING = "0.00"  
SELECT  
    {[Measures].[AverageNightsStay]} ON COLUMNS,  
    NON EMPTY [Checkin Date].[Thang].[Thang].Members ON ROWS  
FROM  
    [Hotel]  
WHERE  
    ([Checkin Date].[Nam].&[2023]);
```

Thang	AverageNightsStay
1	2.68085106382979
10	2.98148148148148
11	2.88387096774194
12	3.08176100628931
2	3.09174311926605
3	2.90344827586207
4	3.01481481481481
5	2.97945205479452
6	3.00684931506849
7	3.00735294117647
8	3.02721088435374
9	2.83076923076923

3.2.20. Tổng doanh thu theo loại phòng

➤ Câu lệnh truy vấn MDX:

```
SELECT  
    {[Measures].[Total Amount]} ON COLUMNS,  
    NON EMPTY [Dim Room].[Room Type].[Room Type].Members ON ROWS  
FROM [Hotel];
```

➤ Kết quả:

```
SELECT
    {[Measures].[Total Amount]} ON COLUMNS,
    NON EMPTY [Dim Room].[Room Type].[Room Type].Members ON ROWS
FROM
    [Hotel];
```

Room Type	Total Amount
Deluxe	848431
Standard	456379
Suite	1379228

3.2.21. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng

- Câu lệnh truy vấn MDX:

```
SELECT [Measures].[Fact Count] ON COLUMNS,
NON EMPTY Generate(
    [Dim Reservation].[Reservation Source].children,
    TopCount(
        [Dim Reservation].[Reservation Source].currentmember *
        [Dim Room].[Room Type].children,
        2,
        [Measures].[Fact Count]
    )
) ON ROWS
FROM [Hotel];
```

- Kết quả:

```
SELECT [Measures].[Fact Count] ON COLUMNS,
NON EMPTY Generate(
    [Dim Reservation].[Reservation Source].children,
    TopCount(
        [Dim Reservation].[Reservation Source].currentmember * [Dim Room].[Room Type].children,
        2,
        [Measures].[Fact Count]
    )
) ON ROWS
FROM [Hotel];
```

Reservation Source	Room Type	Fact Count
Booking Platform	Deluxe	553
Booking Platform	Suite	515
Phone	Standard	553
Phone	Suite	537
Walk-in	Deluxe	557
Walk-in	Suite	516
Website	Suite	573
Website	Standard	558

3.2.22. Liệt kê top 2 loại phòng có doanh thu nhiều nhất trong cả hai năm 2022 và năm 2023

➤ Câu lệnh truy vấn MDX:

✓ **Cách 1:** Dùng hàm Intersect

```
SELECT {
    ([Measures].[Total Amount],[Checkin Date].[Nam].&[2022]),
    ([Measures].[Total Amount],[Checkin Date].[Nam].&[2023])
} ON COLUMNS,
INTERSECT(
    {TOPCOUNT([Dim Room].[Room Type].children,
                2,
                ([Measures].[Total Amount],[Checkin
Date].[Nam].&[2022])
            )},
    {TOPCOUNT([Dim Room].[Room Type].children,
                2,
                ([Measures].[Total Amount],[Checkin
Date].[Nam].&[2023])
            )}
) ON ROWS
FROM [Hotel];
```

✓ **Cách 2:** Tạo các Member mới và dùng hàm Intersect

```
WITH
-- Tính tổng doanh thu cho năm 2022
MEMBER [Measures].[Total Amount 2022] AS
    ([Measures].[Total Amount], [Checkin Date].[Nam].&[2022])

-- Tính tổng doanh thu cho năm 2023
MEMBER [Measures].[Total Amount 2023] AS
    ([Measures].[Total Amount], [Checkin Date].[Nam].&[2023])

-- Tạo tập hợp các loại phòng có doanh thu cao nhất trong cả hai
năm
SET [Top Rooms 2022] AS
    TOPCOUNT (
        [Dim Room].[Room Type].children,
        2,
        [Measures].[Total Amount 2022]
    )

SET [Top Rooms 2023] AS
    TOPCOUNT (
        [Dim Room].[Room Type].children,
```

```
2,  
[Measures].[Total Amount 2023]  
)
```

```
SET [Intersected Top Rooms] AS  
Intersect([Top Rooms 2022], [Top Rooms 2023])
```

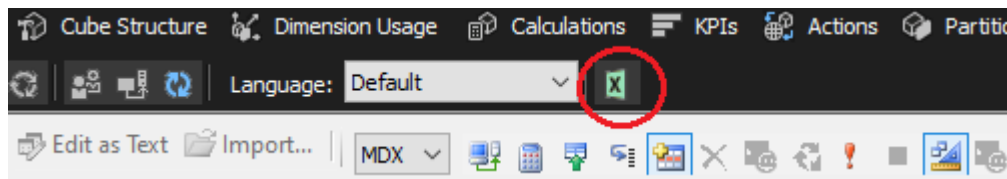
```
SELECT  
    {[Measures].[Total Amount 2022], [Measures].[Total Amount  
2023]} ON COLUMNS,  
    [Intersected Top Rooms] ON ROWS  
FROM [Hotel];
```

➤ Kết quả:

Room Type	Total Amount 2022	Total Amount 2023
Suite	943462	413713
Deluxe	588244	245827

3.3. Thực hiện trên Excel

Chọn Excel để phân tích Pivot trong Excel



3.3.1. Tổng doanh thu theo loại phòng qua các năm (group by)

- Kéo [Total Amount] trong 'Measures' vào Values
- Kéo [Room Type] trong 'Dim Room' vào Rows
- Kéo [Nam] trong 'Checkin Date' vào Columns

Total Amount	Column Labels				
Row Labels	2022	2023	2024	Grand Total	
Deluxe	588244	245827	14360	848431	
Standard	307262	142798	6319	456379	
Suite	943462	413713	22053	1379228	
Grand Total	1838968	802338	42732	2684038	

PivotTable Fields

Choose fields to add to report:

Search

More Fields

- ☐ Checkin Date.Date
- ☒ Nam
- ☐ Checkin Date.Ngày
- ☐ Checkin Date.Quý

Drag fields between areas below:

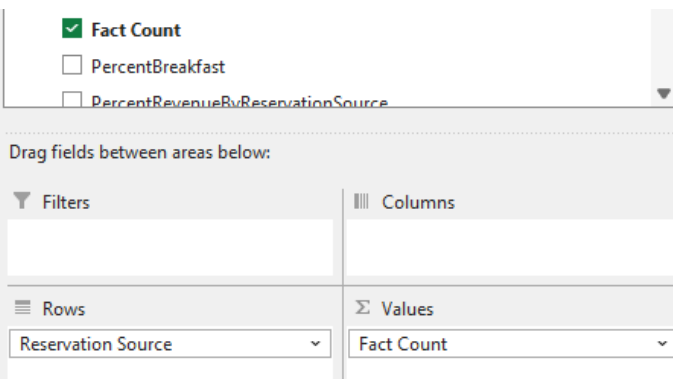
Filters	Columns
	Nam
Rows	Values
Room Type	Total Amount

3.3.2. Tổng số đơn đặt phòng theo nguồn đặt phòng (group by)

- Kéo [Fact Count] trong 'Measures' vào Values

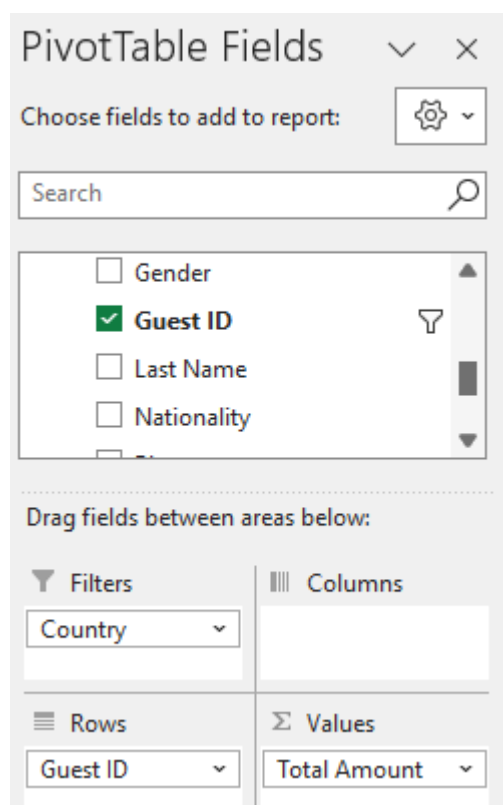
- Kéo [Reservation Source] trong ‘Dim Reservation’ vào Rows

Row Labels	Fact Count
Booking Platform	1579
Phone	1619
Walk-in	1577
Website	1681
Grand Total	6456

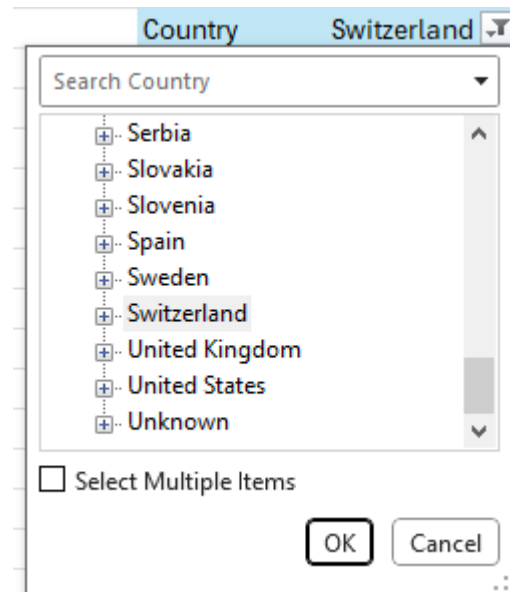


3.3.3. Số lượng khách từ Thụy Sĩ (Switzerland) , chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 1000 (table)

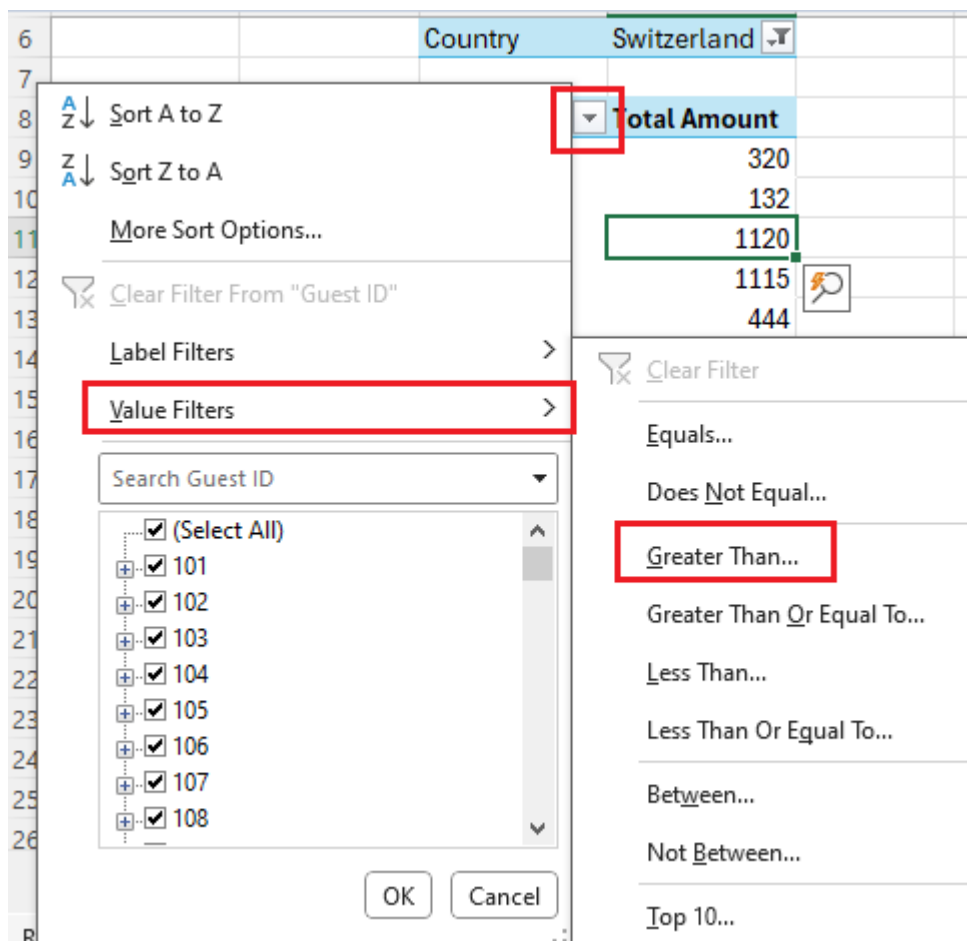
- Kéo [Guest ID] từ Dimension ‘Dim Guest’ vào Rows
- Kéo [Total Amount] từ Measures vào Values
- Kéo [Country] từ Dimension ‘Dim Guest’ vào Filters



- Lọc để chỉ hiển thị khách hàng từ Thụy Sĩ bằng cách chọn Switzerland trong Filter Country



- Click vào mũi tên thả xuống bên cạnh Row Labels (Guest ID)
- Chọn Value Filters > Greater Than....



- Nhập 400 để lọc các giá trị Total Amount lớn hơn 1000 và “OK”

Value Filter (Guest ID) ? X

Show items for which

Total Amount is greater than 1000

OK Cancel

➤ Kết quả:

Country	Switzerland
Row Labels	Total Amount
315	1120
328	1115
336	1035
604	1095
2862	1175
3015	1120
4250	1190
6473	1035
28722	1220
29511	1085
Grand Total	11190

3.3.4. Tạo ma trận hiển thị số lượng khách hàng theo loại phòng và nguồn đặt phòng (matrix)

- Kéo [Room Type] trong Dim Room vào Columns
- Kéo [Reservation Source] trong Dim Reservation vào Rows
- Kéo [Fact Count] trong Measures vào Values

Fact Count	Column Labels			
Row Labels	Deluxe	Standard	Suite	Grand Total
Booking Platform	553	511	515	1579
Phone	529	553	537	1619
Walk-in	557	504	516	1577
Website	550	558	573	1681
Grand Total	2189	2126	2141	6456

PivotTable Fields

Choose fields to add to report:

Search

☐ Payment Status

☒ Dim Reservation

☒ Reservation Source

☒ Dim Room

Drag fields between areas below:

Filters

Columns

Room Type

Rows

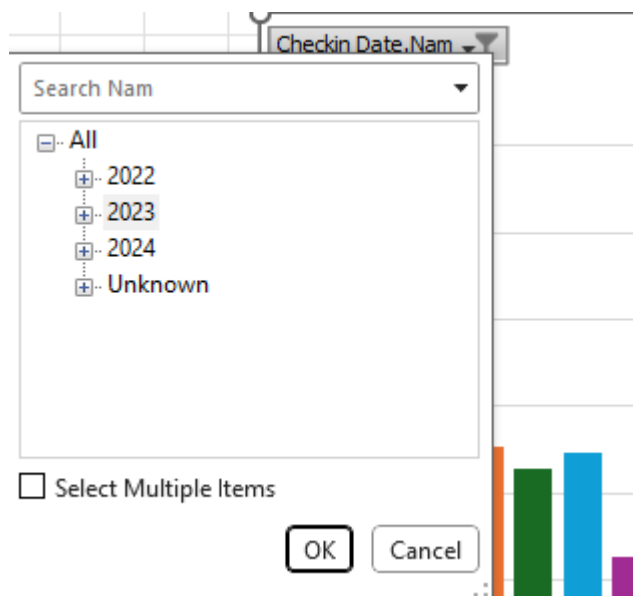
Reservation Source

Values

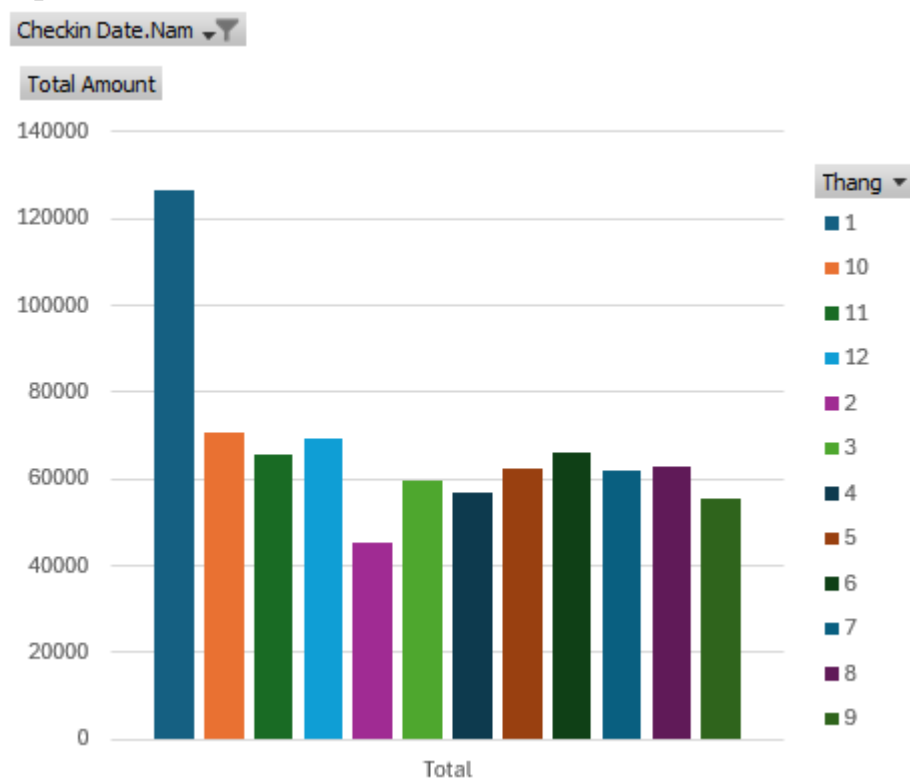
Fact Count

3.3.5. Tạo biểu đồ hiển thị tổng doanh thu theo tháng năm 2023 (chart)

- Kéo [Nam] trong 'Checkin Date' vào Filters
- Kéo [Thang] trong 'Checkin Date' vào Columns
- Kéo [Total Amount] trong 'Measures' vào Values
- Chọn 2023 để lọc doanh thu vào năm 2023

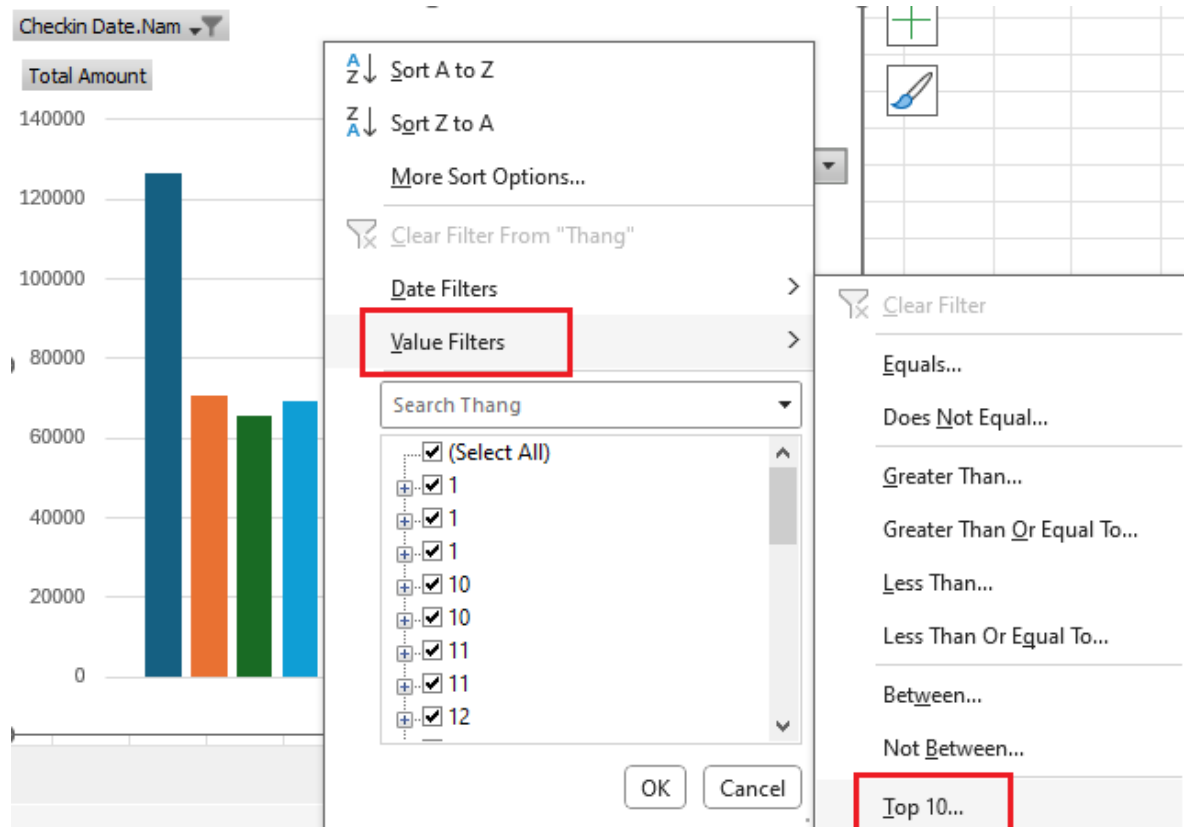


- Kết quả:



3.3.6. Tạo biểu đồ hiển thị top 2 doanh thu theo tháng năm 2023 (chart)

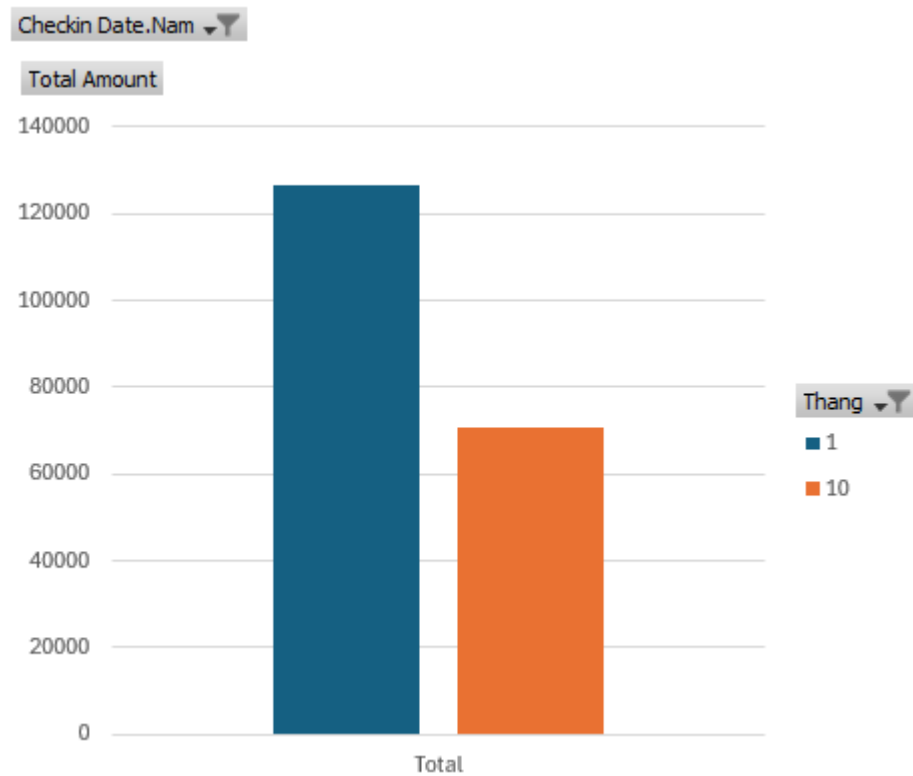
- Trong mục [Thang] → chọn Value Filters → chọn Top 10



- Chỉnh sửa thành top 2 theo doanh thu (Total Amount)

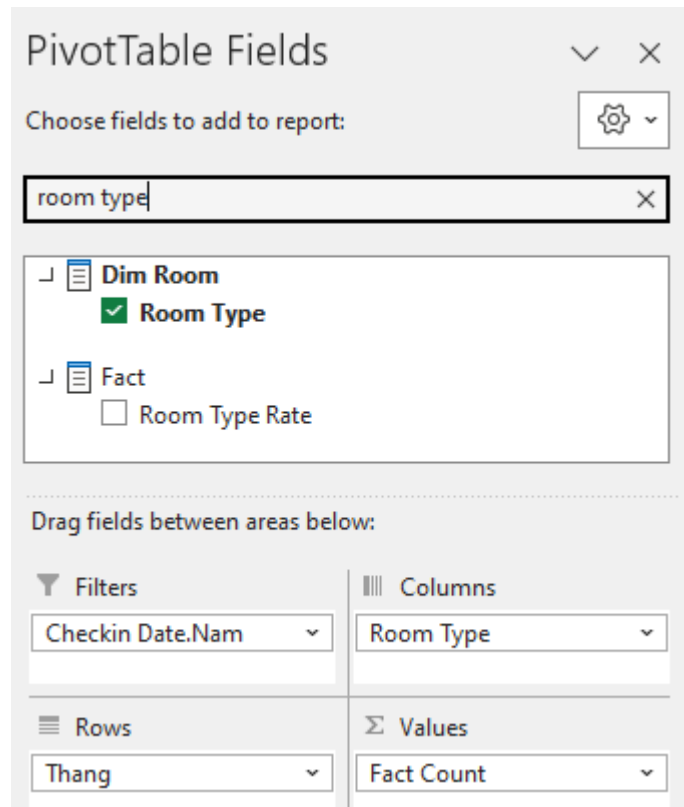
The screenshot shows a 'Top 10 Filter (Thang)' dialog box. It has a 'Show' section with a dropdown menu set to 'Top', a text input field containing '2', a dropdown menu set to 'Items', and a dropdown menu set to 'Total Amount'. The 'OK' button is highlighted with a blue border.

- Kết quả:

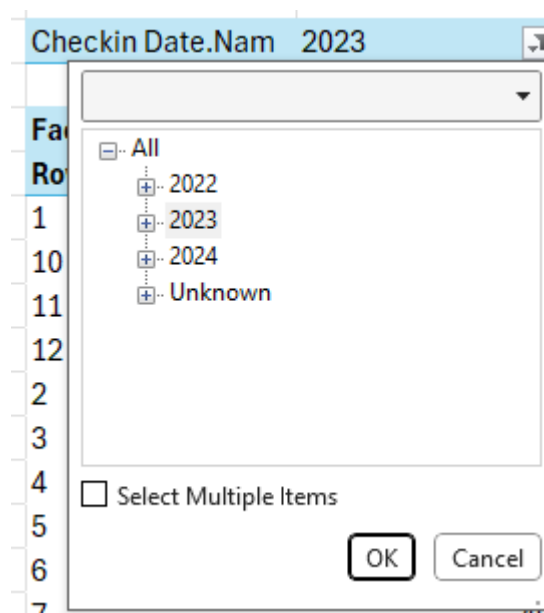


3.3.7. Số lượng đặt phòng theo từng tháng và từng loại phòng trong năm 2023

- Kéo [Room Type] trong 'Dim Room' vào Columns
- Kéo [Thang] trong 'Checkin Date' vào Rows
- Kéo [Fact Count] trong Measures vào Values
- Kéo [Nam] trong 'Checkin Date' vào Filters



- Lọc dữ liệu của Nam = 2023

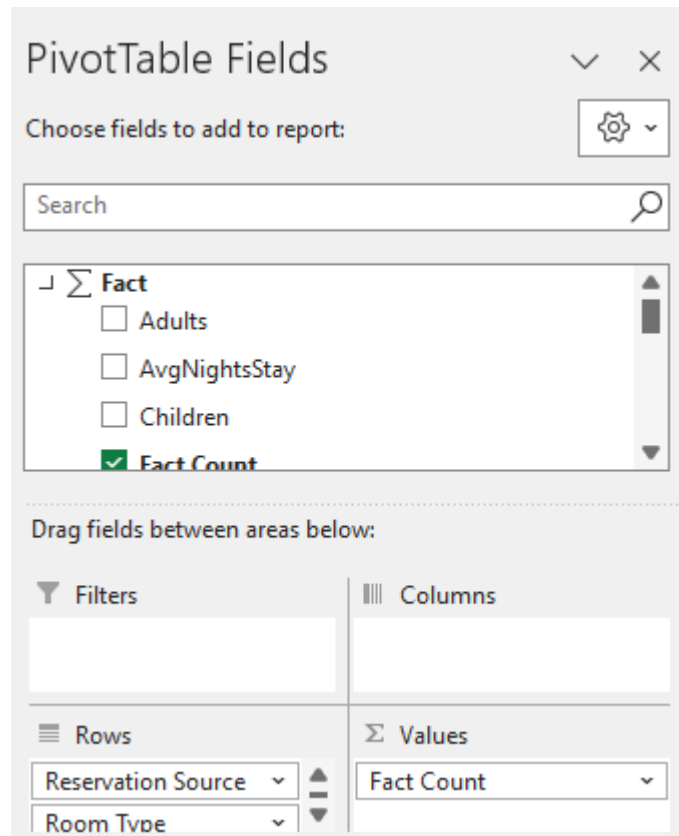


- Kết quả:

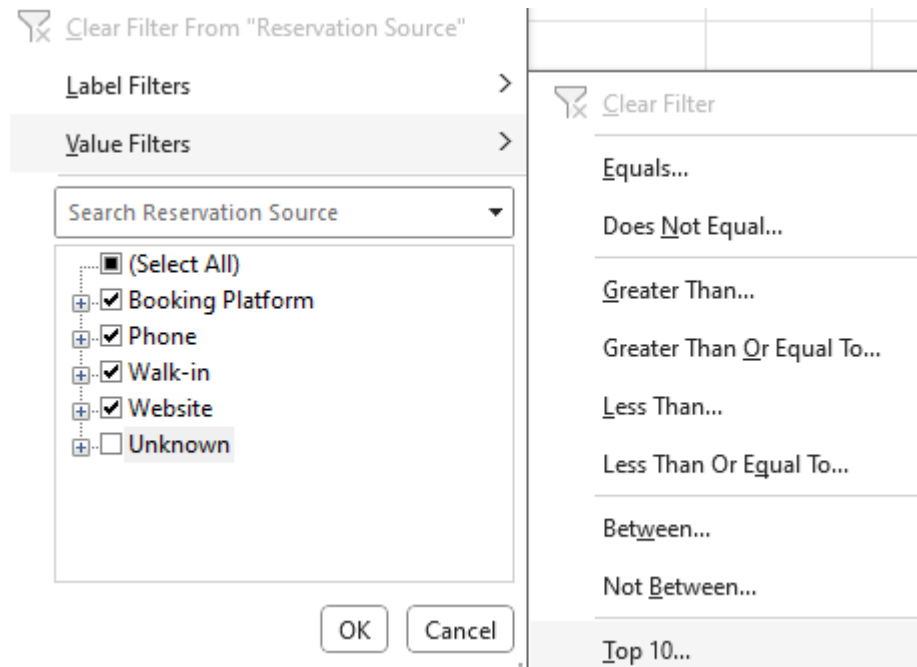
Checkin Date.Nam	2023				
Fact Count	Column Labels				
Row Labels	Deluxe	Standard	Suite	Grand Total	
1	113	107	109	329	
10	48	56	58	162	
11	60	50	45	155	
12	56	51	52	159	
2	33	42	34	109	
3	54	46	45	145	
4	42	51	42	135	
5	49	49	48	146	
6	50	42	54	146	
7	40	43	53	136	
8	46	55	46	147	
9	40	41	49	130	
Grand Total	631	633	635	1899	

3.3.8. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng

- Kéo [Room Type] trong 'Dim Room' vào Columns
- Kéo [Reservation Source] trong 'Dim Reservation' vào Rows
- Kéo [Fact Count] trong Measures vào Values



- Ở Column labels, chọn Sort → Value Filters → Top 10



- Chọn Top 2 theo số lượt thuê để liệt kê top 2 loại phòng

Top 10 Filter (Room Type) ? X

Show

Top 2 Items by Fact Count

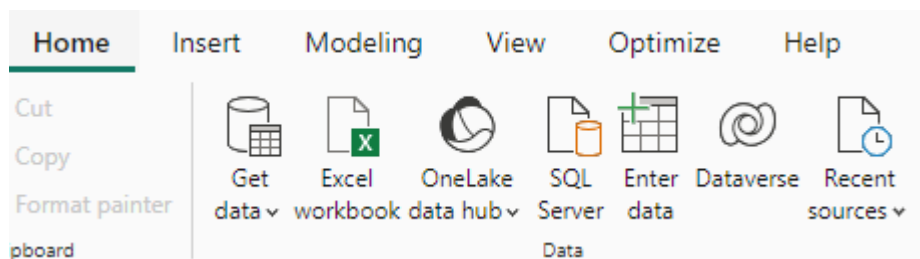
OK Cancel

➤ Kết quả:

Row Labels	Fact Count
Booking Platform	
Deluxe	553
Suite	515
Phone	
Standard	553
Suite	537
Walk-in	
Deluxe	557
Suite	516
Website	
Standard	558
Suite	573
Grand Total	4362

3.4. Thực hiện trên Power BI

Bước 1: Đưa dữ liệu vào Power BI: Trong thẻ Home → chọn Get data → Analysis Services



Bước 2: Điền tên Server chứa Database

SQL Server Analysis Services database

Server ⓘ

LAPTOP-IK43O6K9\TPA

Database (optional)

☐ Import

☒ Connect live

▸ MDX or DAX query (optional)

OK

Cancel

Bước 3: Chọn Cube chứa bảng Dim và Fact dùng để trực quan dữ liệu

Navigator

A tree view showing the structure of the SQL Server Analysis Services (SSAS) instance. The root node is 'LAPTOP-IK43O6K9\TPA [5]'. Under this root, there are several folders: 'CK1_2017_2018', 'example', 'hotel_ssas [1]', 'olap', and 'SSAS_Hotel [1]'. The 'SSAS_Hotel [1]' folder is expanded, showing a sub-folder named 'Hotel' which is selected and highlighted.

Hotel

Last Modified: 05/26/2024 11:10:29

This model contains the following dimensions and measures

Booking Date, Checkin Date, Checkout Date, Dim Guest, Dim Payment, Dim Reservation, Dim Room, Adults, Children, Total Nights, Total Amount, Fact Count

OK

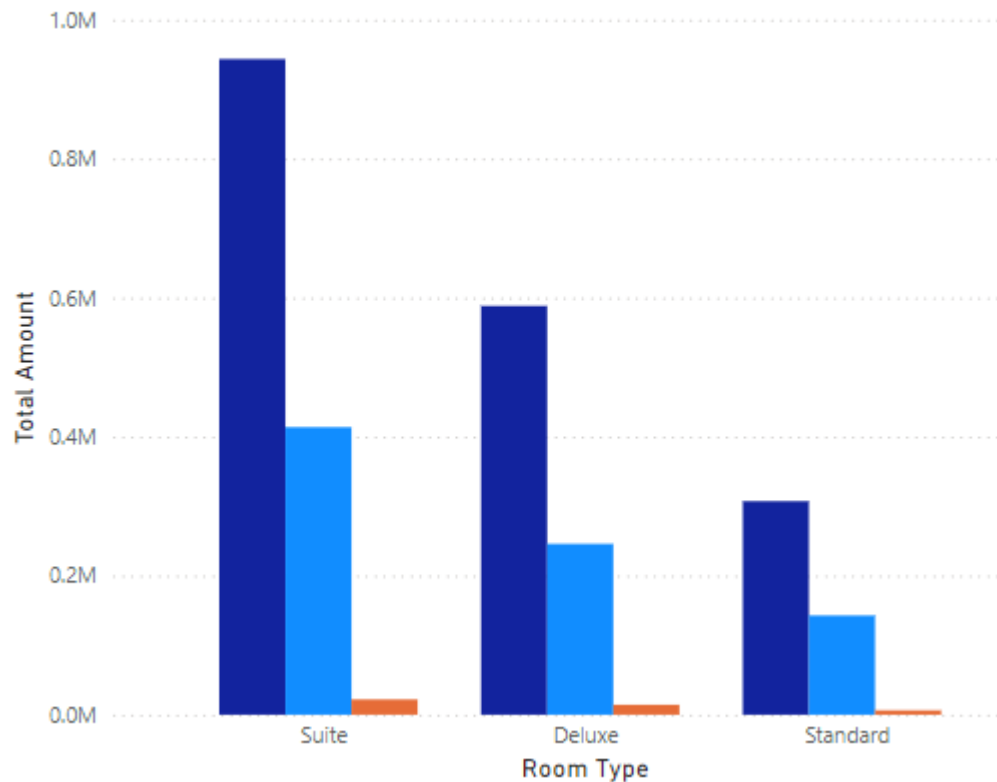
Cancel

3.4.1. Tổng doanh thu theo loại phòng qua các năm

- Column Chart

Total Amount by Room Type and Nam

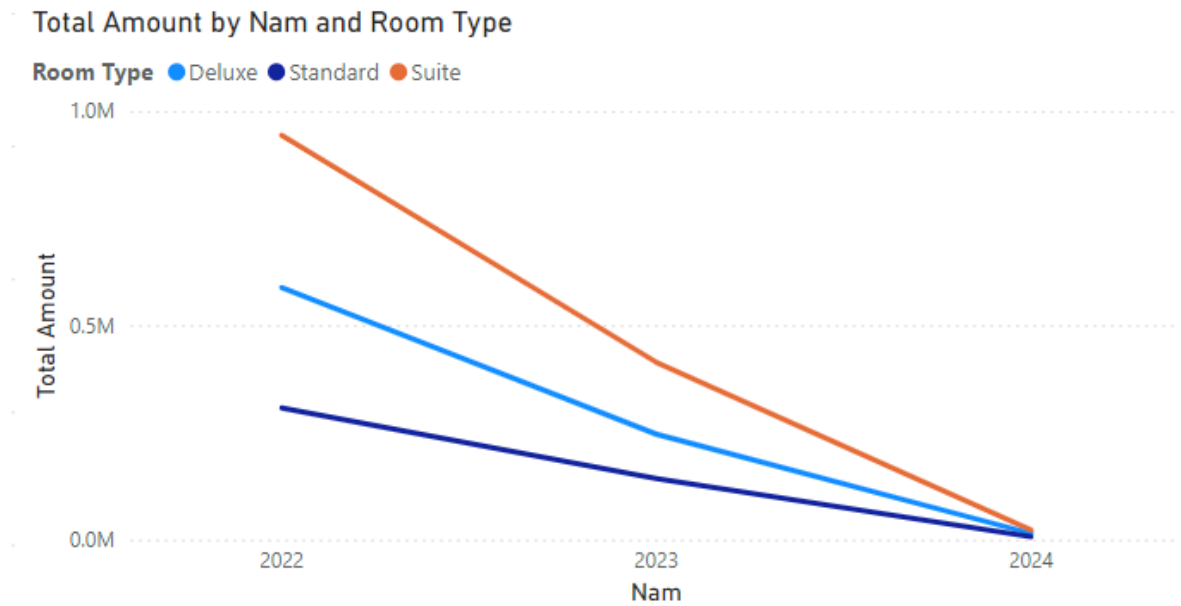
Nam ● 2022 ● 2023 ● 2024



- Table

Room Type	2022	2023	2024	Total
Deluxe	588,244.00	245,827.00	14,360.00	848,431.00
Standard	307,262.00	142,798.00	6,319.00	456,379.00
Suite	943,462.00	413,713.00	22,053.00	1,379,228.00
Total	1,838,968.00	802,338.00	42,732.00	2,684,038.00

- Line chart



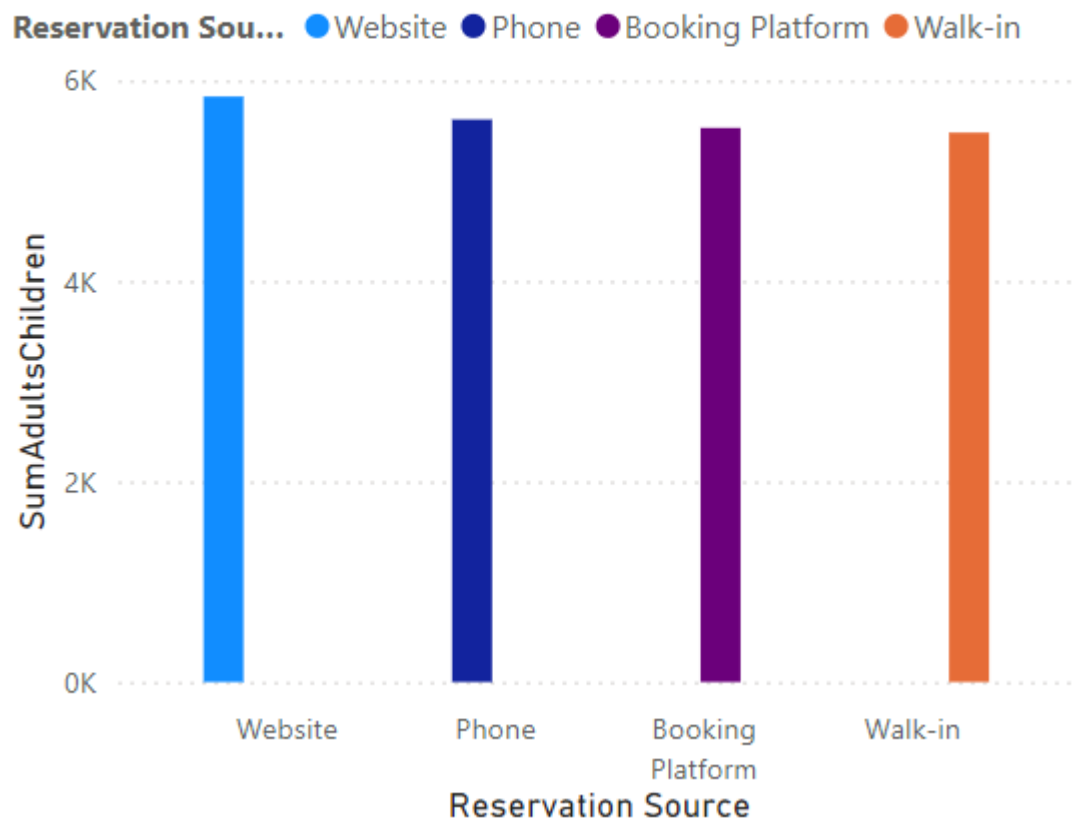
Từ kết quả trên, ta có:

- Cả 3 năm (2022, 2023, 2024) thì loại phòng Suite có doanh thu cao nhất trong các loại phòng
- Loại phòng Standard có doanh thu thấp nhất ở cả 3 năm

3.4.2. Tổng số lượng khách hàng theo nguồn đặt phòng

- Sử dụng Calculate member [SumAdultsChildren] đã tạo trước trong SSAS
- **Column chart**

SumAdultsChildren by Reservation Source and Reservation Source

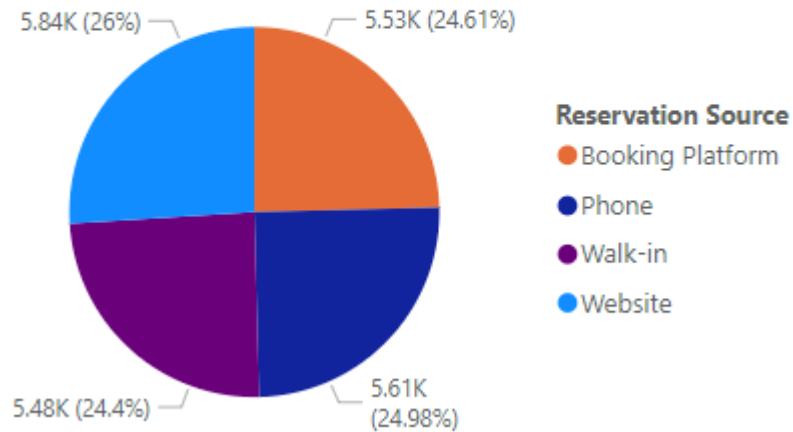


- **Table**

Reservation Source	SumAdultsChildren
Booking Platform	5,528.00
Phone	5,612.00
Walk-in	5,481.00
Website	5,841.00
Total	22,462.00

- **Pie chart**

SumAdultsChildren by Reservation Source



3.4.3. Số lượng khách từ Thụy Sĩ (Switzerland), chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 1000

- Filter Total Amount > 1000

Show items when the value

is greater than ▼

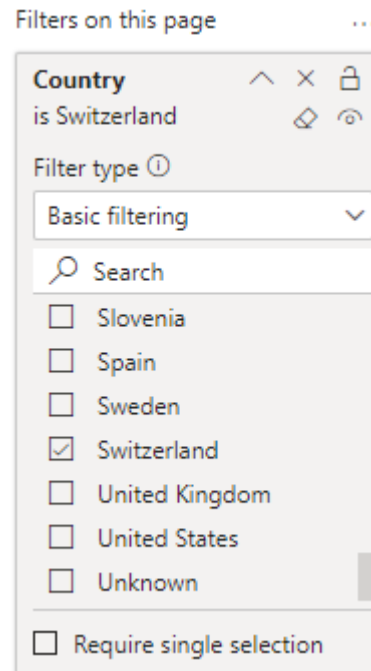
1000

☒ And ☐ Or

▼

Apply filter

- Filter Country = 'Switzerland'



➤ Kết quả:

Guest ID	First Name	Last Name	Total Amount
315.00	Aaron	House	1,120.00
328.00	Lisa	Bradford	1,115.00
336.00	Miguel	Rivera	1,035.00
604.00	Holly	Moore	1,095.00
2,862.00	Jeffrey	Morales	1,175.00
3,015.00	Ashley	Hale	1,120.00
4,250.00	Christine	Parsons	1,190.00
6,473.00	Jacob	Alexander	1,035.00
28,722.00	Sarah	Cole	1,220.00
29,511.00	Rebecca	Brown	1,085.00
Total			11,190.00

3.4.4. Tạo ma trận hiển thị số lượng đặt phòng theo loại phòng và nguồn đặt phòng

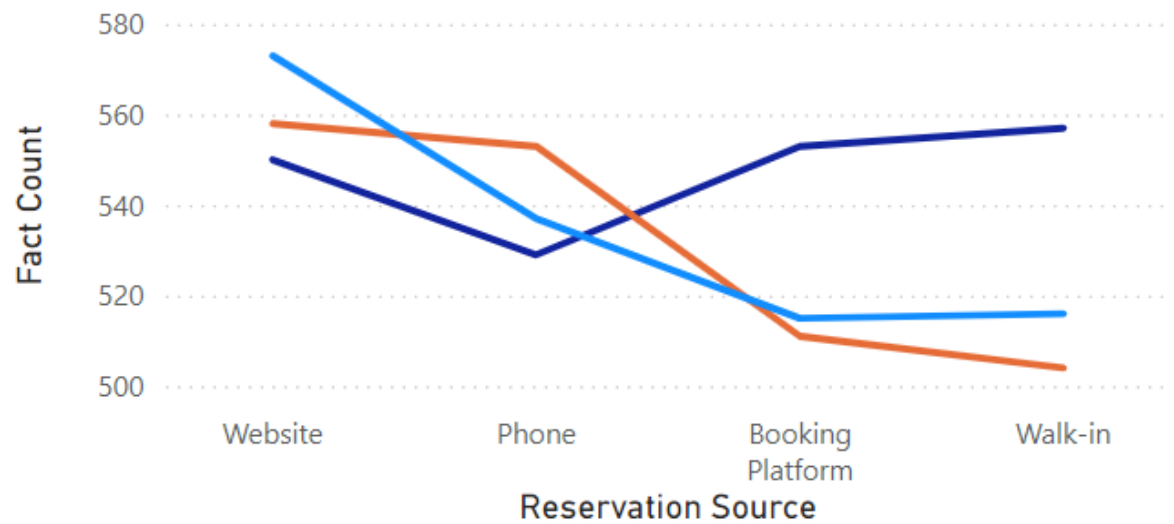
- Matrix

Reservation Source	Deluxe	Standard	Suite	Total
Booking Platform	553	511	515	1579
Phone	529	553	537	1619
Walk-in	557	504	516	1577
Website	550	558	573	1681
Total	2189	2126	2141	6456

- **Line chart**

Fact Count by Reservation Source and Room Type

Room Type ● Deluxe ● Standard ● Suite



3.4.5. Tạo biểu đồ hiển thị tổng doanh thu theo tháng năm 2023

➤ Filter [Nam] = 2023

Nam ^ x

is 2023

Filter type ⓘ

Basic filtering v

Search

☒ Select all

☐ 2022

☒ 2023

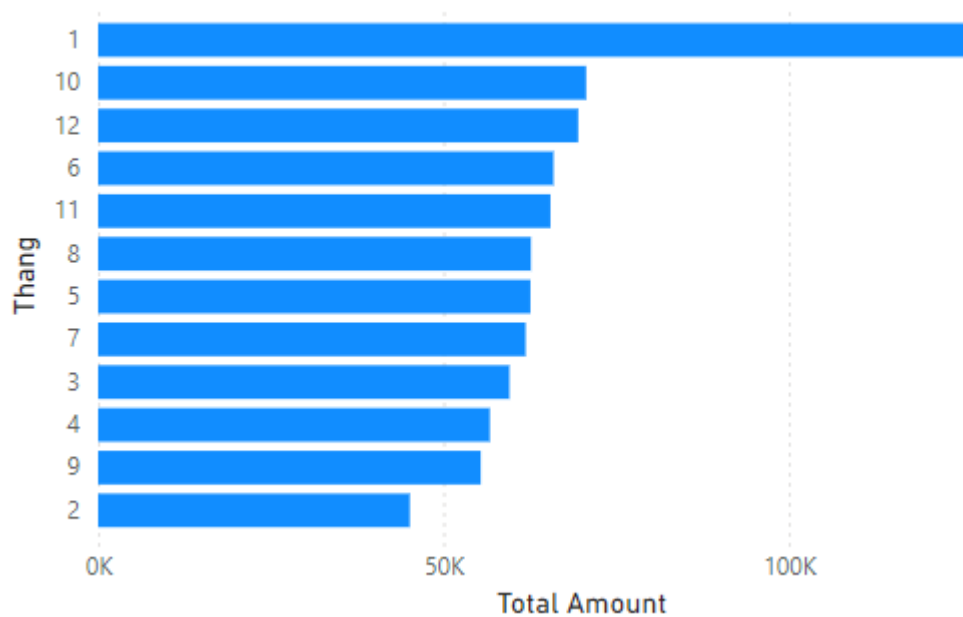
☐ 2024

☐ Unknown

☐ Require single selection

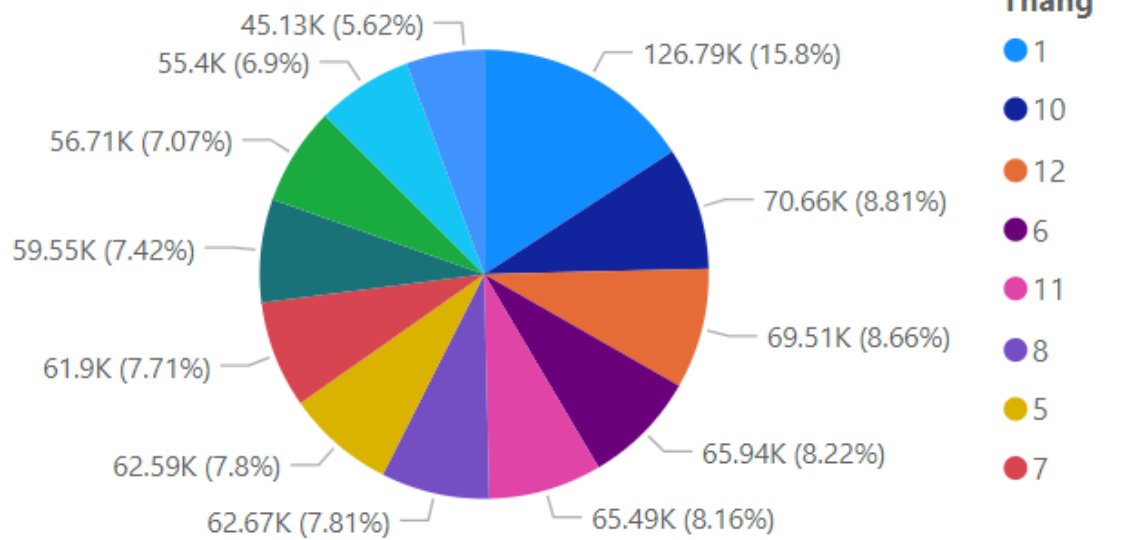
- **Bar chart**

Total Amount by Thang



- **Pie chart**

Total Amount by Tháng



- Table

Tháng	Total Amount
1	126,790.00
10	70,660.00
12	69,510.00
6	65,936.00
11	65,491.00
8	62,667.00
5	62,592.00
7	61,899.00
3	59,553.00
4	56,714.00
9	55,399.00
2	45,127.00
Total	802,338.00

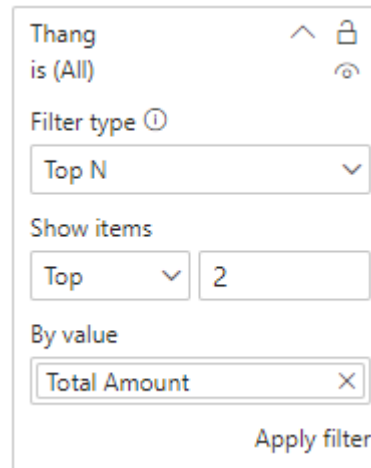
Từ kết quả trên, ta có:

- Tháng 1/2023 có doanh thu cao nhất tất cả các tháng trong năm

- Tháng 2 có doanh thu thấp nhất năm 2023

3.4.6. Tạo biểu đồ hiển thị top 2 doanh thu theo tháng năm 2023

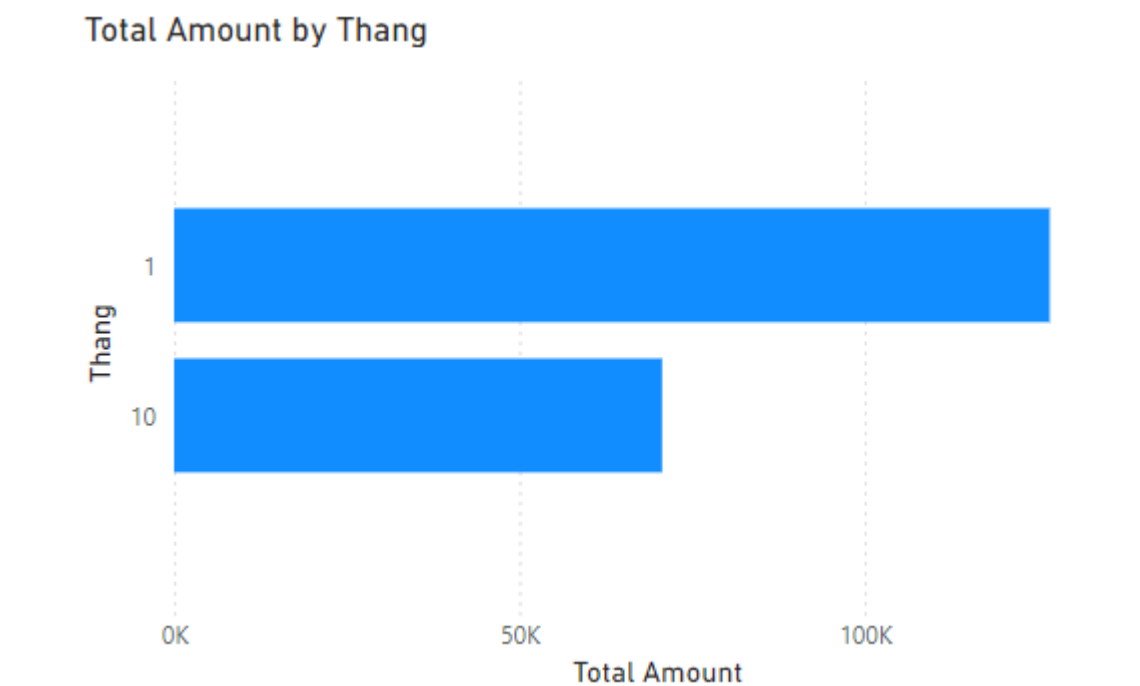
- Filter Top 2 doanh thu (Total Amount) theo tháng → Apply filter



- **Table**

Thang	Total Amount
1	126,790.00
10	70,660.00
Total	197,450.00

- **Bar chart**



3.4.7. Số lượng đặt phòng theo từng tháng và từng loại phòng trong năm 2023

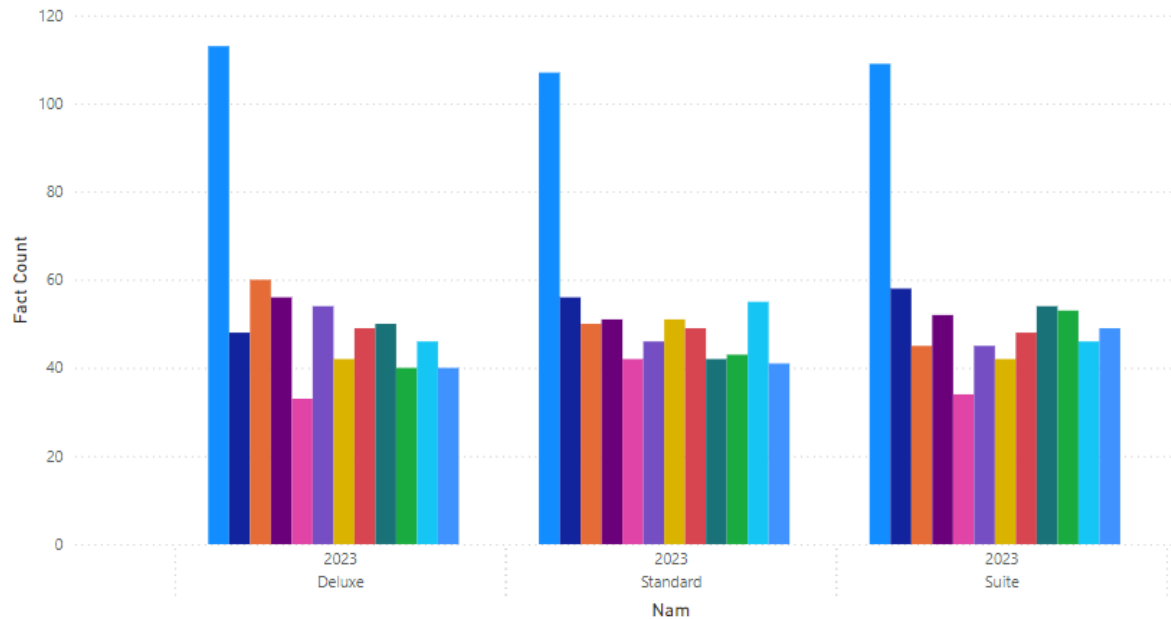
- Matrix

Room Type	1	10	11	12	2	3	4	5	6	7	8	9	Total
⊕ Deluxe	113	48	60	56	33	54	42	49	50	40	46	40	631
⊕ Standard	107	56	50	51	42	46	51	49	42	43	55	41	633
⊕ Suite	109	58	45	52	34	45	42	48	54	53	46	49	635
Total	329	162	155	159	109	145	135	146	146	136	147	130	1899

- Column chart

Fact Count by Room Type, Nam and Thang

Thang 1 10 11 12 2 3 4 5 6 7 8 9



3.4.8. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng

- Lọc ra top 2 loại phòng theo số lượt thuê

Room Type
top 2 by Fact Count

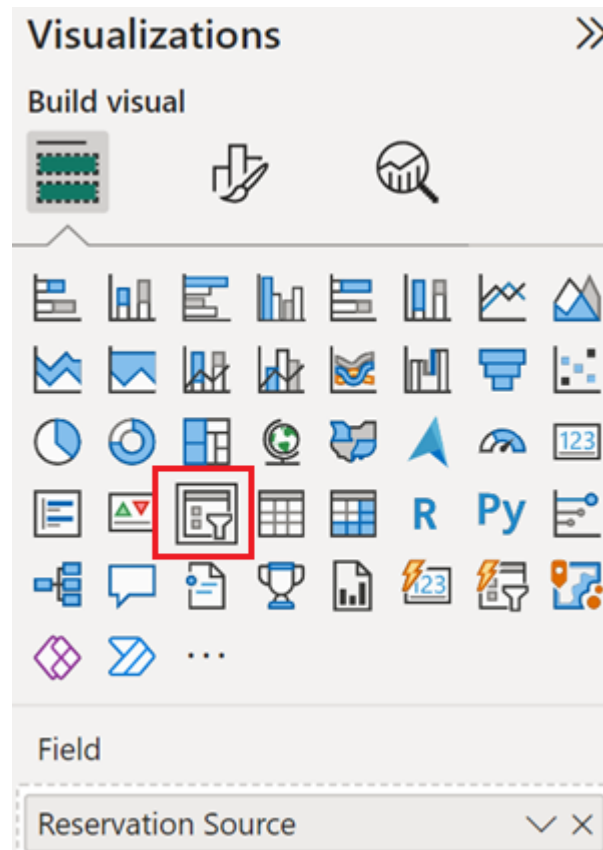
Filter type ⓘ
Top N

Show items
Top 2

By value
Fact Count

Apply filter

- Chọn Slicer trong Visualizations để tạo bộ lọc dữ liệu cần lọc (theo Reservation Source)



- **Chọn Reservation Source là Booking Platform**

Reservation Source	Room Type	Fact Count
Booking Platform	Deluxe	553
Booking Platform	Suite	515
Total		1068

Reservation Source

☒ Booking Platform

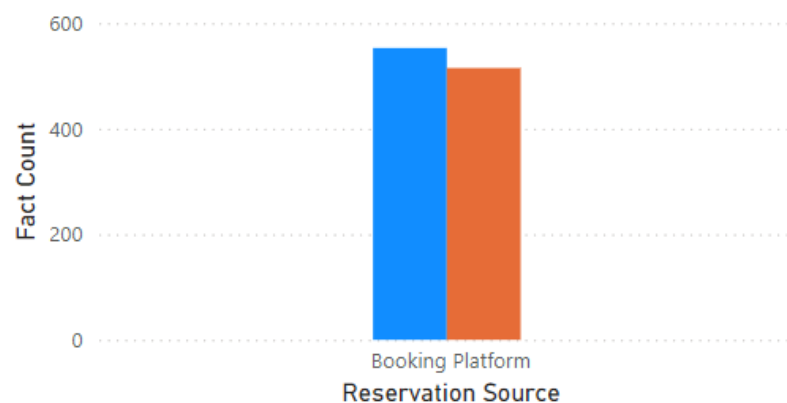
☐ Phone

☐ Walk-in

☐ Website

Fact Count by Reservation Source and Room Type

Room Type ● Deluxe ● Suite



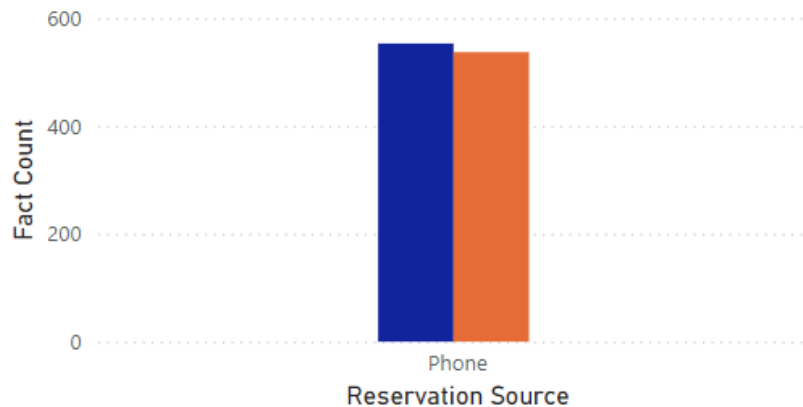
- **Chọn Reservation Source là Phone**

Reservation Source	Room Type	Fact Count
Phone	Standard	553
Phone	Suite	537
Total		1090

- Reservation Source
- ☐ Booking Platform
 - ☒ Phone
 - ☐ Walk-in
 - ☐ Website

Fact Count by Reservation Source and Room Type

Room Type ● Standard ● Suite



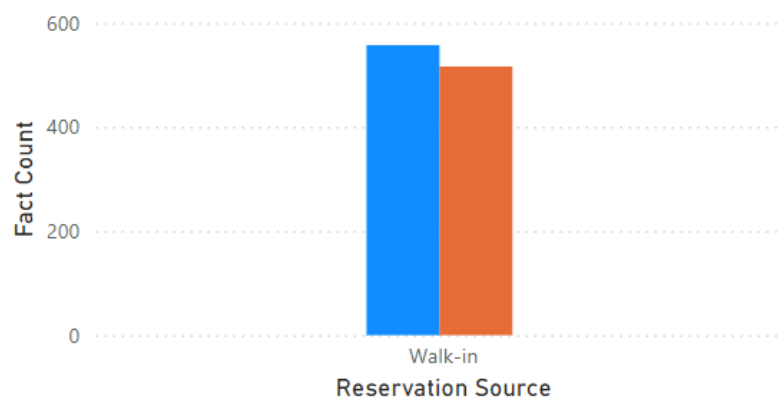
- **Chọn Reservation Source là Walk-in**

Reservation Source	Room Type	Fact Count
Walk-in	Deluxe	557
Walk-in	Suite	516
Total		1073

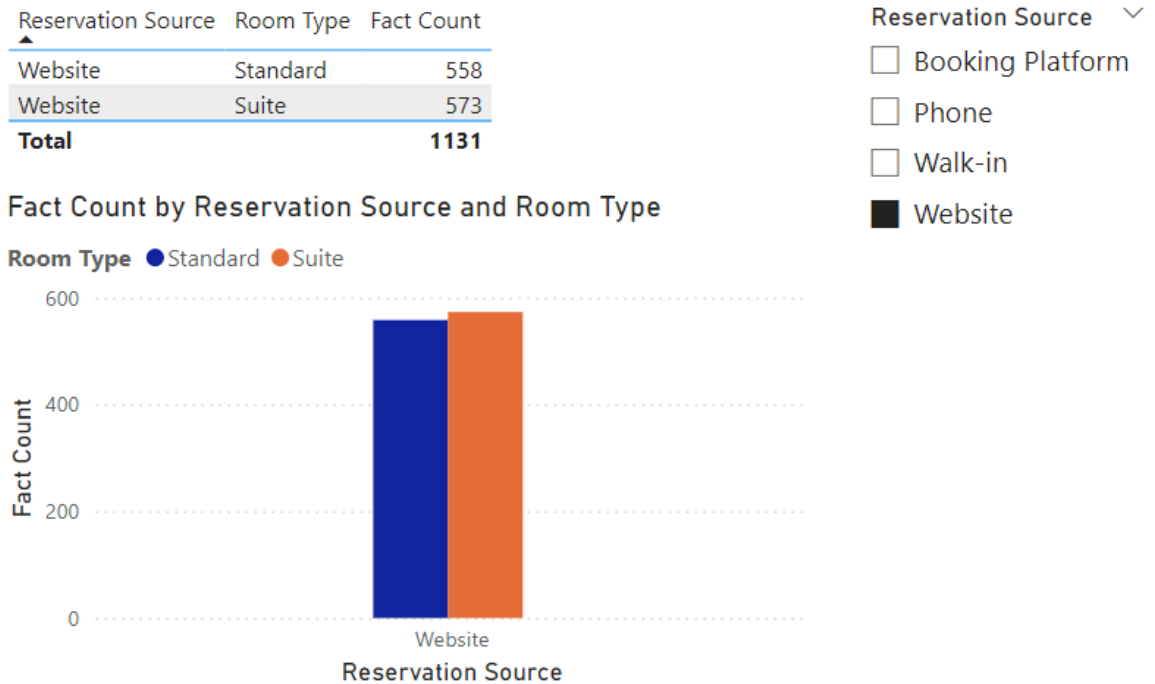
- Reservation Source
- ☐ Booking Platform
 - ☐ Phone
 - ☒ Walk-in
 - ☐ Website

Fact Count by Reservation Source and Room Type

Room Type ● Deluxe ● Suite



- **Chọn Reservation Source là Website**



Từ kết quả trên, ta có:

- Reservation Source là Booking Platform và Walk-in sẽ có loại phòng *Deluxe* và *Suite* được thuê nhiều nhất, trong đó cả hai nguồn đặt chỗ đều có xu hướng đặt loại phòng *Deluxe* nhiều hơn so với *Suite*
- Reservation Source là Phone và Website sẽ có loại phòng *Standard* và *Suite* được thuê nhiều nhất, trong đó:
 - + Với nguồn đặt chỗ Phone, số lượng thuê phòng loại *Standard* nhiều hơn so với *Suite*
 - + Với nguồn đặt chỗ Website, số lượng thuê phòng loại *Suite* nhiều hơn so với *Standard*

CHƯƠNG 4. QUÁ TRÌNH DATA MINING

4.1. Tổng quan đề tài

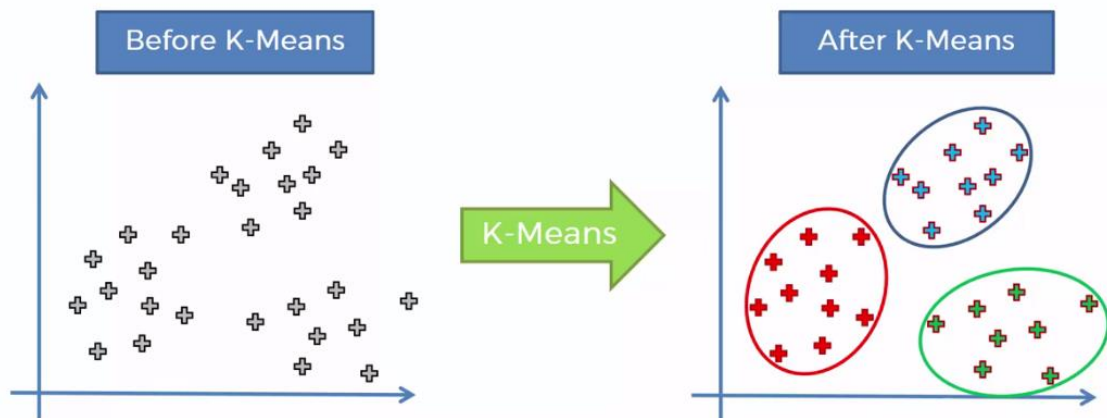
Đề tài: Phân tích và phân cụm khách hàng dựa trên dữ liệu đặt phòng khách sạn.

Mục tiêu của dự án là áp dụng phân tích RFM (Recency, Frequency và Monetary) và các phương pháp phân cụm để hiểu và phân loại khách hàng từ các quốc gia khác nhau dựa trên hành vi đặt phòng khách sạn. Qua đó, xác định các nhóm khách hàng tiềm năng từ từng quốc gia và phát triển chiến lược tiếp thị hiệu quả hơn để tối đa hóa giá trị từ từng nhóm khách hàng.

4.2. Lý thuyết mô hình phân cụm

4.2.1. K-Means

K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát (tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.



Thuật toán phân cụm k-means thường được sử dụng trong các ứng dụng cổ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu,...

Thuật toán k-means có thể được chia thành các bước như sau:

- **Bước 1: Tạo các trung tâm ngẫu nhiên**

$$\mathbb{C}^{(0)} = \{m_1^{(0)}, m_2^{(0)}, \dots, m_k^{(0)}\}$$

- **Bước 2: Gán các điểm dữ liệu vào các cụm**

$$\mathbb{S}_i^{(t)} = \left\{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2\right\}, \forall j, 1 \leq j \leq k$$

Với mỗi điểm dữ liệu, ta sẽ tính khoảng cách của nó tới các trung tâm và sẽ gán chúng vào trung tâm gần nhất. Tập hợp các điểm được gán vào cùng 1 trung tâm sẽ tạo thành cụm.

- **Bước 3: Cập nhật trung tâm**

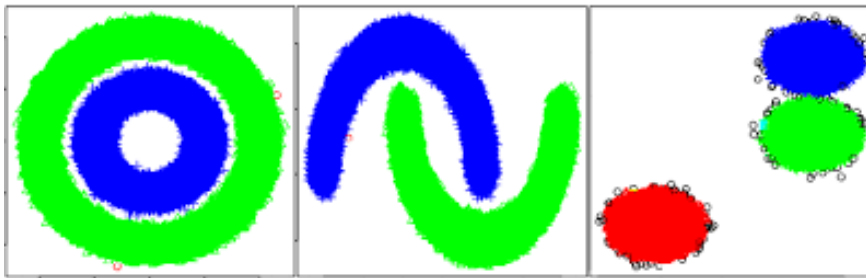
Với mỗi cụm đã tìm được ở bước 2, trung tâm mới sẽ là trung bình cộng của các điểm dữ liệu trong cụm đó.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x \in S_i^{(t)}} x_j$$

⇒ Thuật toán sẽ lặp lại các bước trên cho tới khi đạt được kết quả chấp nhận được.

4.2.2. DBSCAN (Density-Based Clustering)

Giải thuật DBSCAN (Density Based Spatial Clustering of Application with Noise) được Ester, Kriegel và Sander đề xuất năm 1996 khi nghiên cứu các thuật toán gom cụm dữ liệu không gian dựa trên định nghĩa cụm là tập tối đa các điểm liên thông về mật độ. Giải thuật DBSCAN phát hiện các cụm có hình dạng tùy ý, khả năng phát hiện nhiễu tốt. DBSCAN thực hiện tốt trên không gian nhiều chiều; thích hợp với cơ sở dữ liệu có mật độ phân bố dày đặc kể cả có phân tử nhiễu.

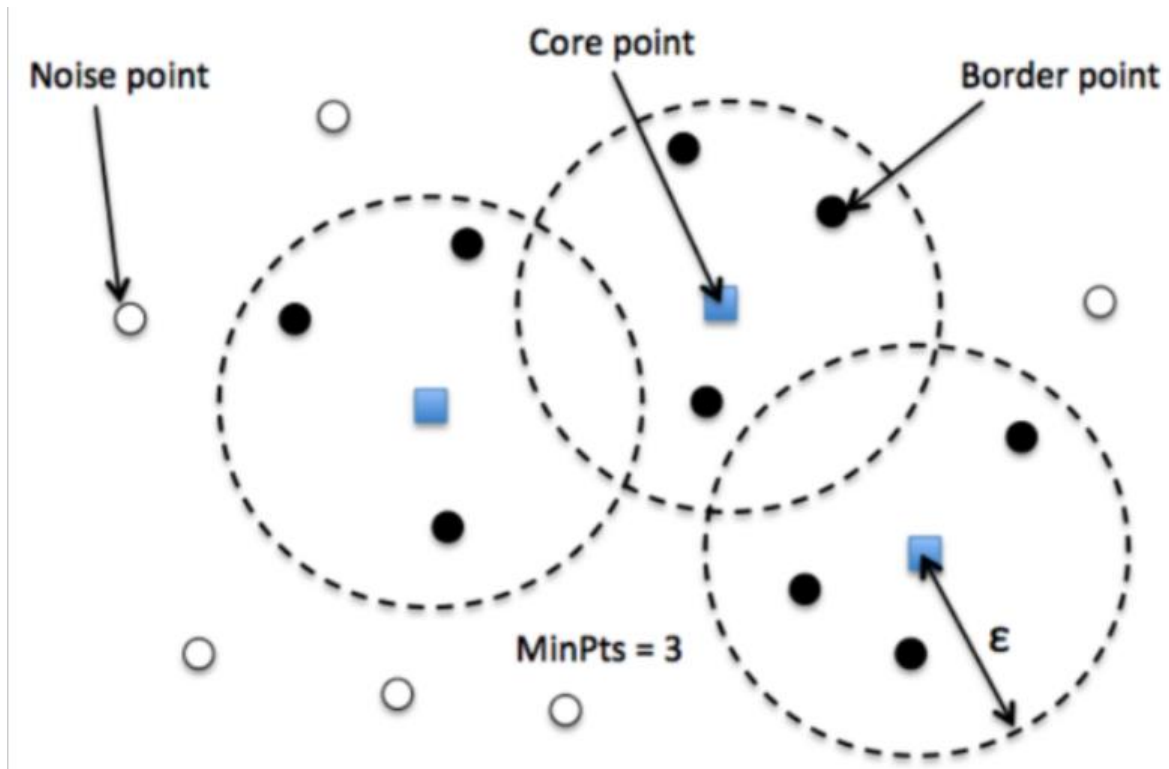


Nguyên lý hoạt động của DBSCAN:

DBSCAN dựa trên ý tưởng rằng một cụm dữ liệu là một vùng mật độ cao được tách biệt bởi các vùng mật độ thấp. Cụ thể, nó xác định các cụm dựa trên hai tham số chính:

- Epsilon (ϵ): Bán kính của vùng lân cận xung quanh một điểm dữ liệu.
- MinPts: Số lượng điểm tối thiểu cần thiết trong vùng lân cận của một điểm để xác định nó là một điểm lõi.

Phân loại dạng điểm trong DBSCAN:



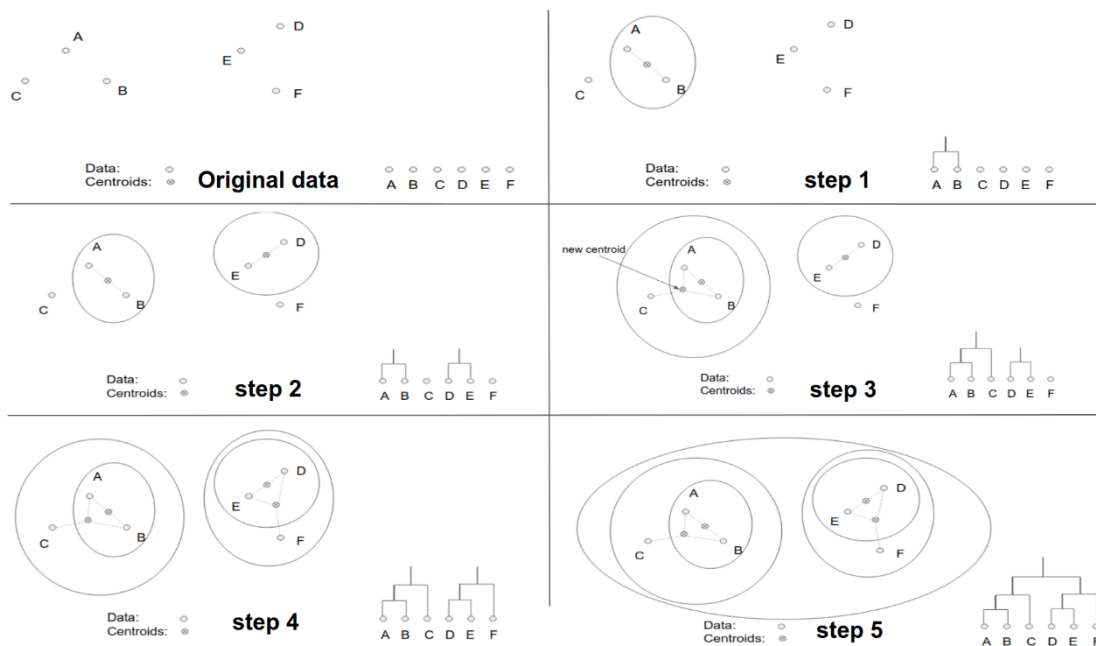
Căn cứ vào vị trí của các điểm dữ liệu so với cụm chúng ta có thể chia chúng thành ba loại:

- Điểm lõi (Core Point): Một điểm được gọi là điểm lõi nếu trong vùng lân cận ϵ của nó có ít nhất MinPts điểm.
- Điểm biên (Border Point): Một điểm được gọi là điểm biên nếu nó nằm trong vùng lân cận của một điểm lõi nhưng không phải là một điểm lõi.
- Điểm nhiễu (Noise Point): Một điểm được gọi là điểm nhiễu nếu nó không phải là điểm lõi và không nằm trong vùng lân cận của bất kỳ điểm lõi nào.

4.2.3. Agglomerative Clustering

Agglomerative Clustering, hay còn gọi là phân cụm tập hợp, là một thuật toán phân cụm phân cấp từ dưới lên. Thuật toán này bắt đầu với mỗi điểm dữ liệu là một cụm riêng lẻ, sau đó lặp lại quá trình hợp nhất các cụm nhỏ thành các cụm lớn hơn cho đến khi đạt được một tiêu chí dừng cụ thể (ví dụ: số lượng cụm mong muốn).

Quá trình của Agglomerative Clustering bao gồm các bước:



- Bắt đầu với n cụm, mỗi cụm chỉ chứa một điểm dữ liệu.
- Tính toán ma trận khoảng cách giữa các cụm.
- Tìm cặp cụm gần nhau nhất (có khoảng cách nhỏ nhất) và hợp nhất chúng thành một cụm mới.
- Cập nhật ma trận khoảng cách bằng cách tính khoảng cách giữa cụm mới với các cụm còn lại.
- Lặp lại bước 3-4 cho đến khi đạt được số lượng cụm mong muốn hoặc một tiêu chí dừng khác.

Các định nghĩa khoảng cách giữa các cụm, bao gồm:

- Khoảng cách liên kết tối thiểu (single linkage): Khoảng cách giữa hai cụm là khoảng cách nhỏ nhất giữa các điểm trong hai cụm.
- Khoảng cách liên kết tối đa (complete linkage): Khoảng cách giữa hai cụm là khoảng cách lớn nhất giữa các điểm trong hai cụm.
- Khoảng cách liên kết trung bình (average linkage): Khoảng cách giữa hai cụm là trung bình khoảng cách giữa các điểm trong hai cụm.

Agglomerative Clustering có thể tạo ra các cụm có hình dạng và kích thước khác nhau, phù hợp với nhiều loại dữ liệu. Phân tích dataset gốc

4.2.4. Thống kê mô tả

Tính toán đại lượng thống kê mô tả: Count, Min, Max, Mean, Mode,... trên tập dữ liệu.

- Thêm thư viện và đọc file excel của tập dữ liệu, kiểm tra 5 dòng đầu tiên dữ liệu.

```
data = pd.read_excel('Hotel Reservations Data.xlsx')
data.head()
```

✓ 0.0s

	Reservation ID	Guest ID	First Name	Last Name	Gender	Email	Phone	Nationality	Birthdate	Address	City	Postal Code	Country	Check-in Date	Check-out Date	Room Number	Floor Number	Room Type	Rate
0	1001	101	Laura	Weiss	Male	xconley@example.org	+1-777-290-9299x1874	Sweden	1990-06-20	194 Stewart Squares	233
1	1002	102	Austin	Henderson	Female	willamaustin@example.org	4268908795	Cyprus	1999-07-22	31442 Morris Port Apt. 423	132
2	1003	103	Jamie	Smith	Male	benjaminporter@example.com	+1-563-234-8041x0677	Italy	1978-09-16	851 Ashley Junctions Apt. 370	88
3	1004	104	Brian	Erickson	Male	johnmelton@example.org	+1-377-838-9030x072	Slovakia	1958-02-21	7221 Lewis Burg	227
4	1005	105	Cristian	Taylor	Male	salazarkelly@example.com	5043212352	Norway	1984-11-30	9874 Melanie Ford Suite 715	229

5 rows x 32 columns

Tập dữ liệu đặt phòng khách sạn với các thông tin mã đặt phòng, thông tin khách đặt phòng, thời gian đặt/giao/trả phòng, thông tin dịch vụ đi kèm, thông tin về phòng đã đặt.

- Kiểm tra kiểu dữ liệu.

```
data.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9974 entries, 0 to 9973
Data columns (total 32 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Reservation ID                        9974 non-null   int64
1   Guest ID                             9974 non-null   int64
2   First Name                           9974 non-null   object
3   Last Name                            9974 non-null   object
4   Gender                               9974 non-null   object
5   Email                                9974 non-null   object
6   Phone                                9974 non-null   object
7   Nationality                          9974 non-null   object
8   Birthdate                            9974 non-null   datetime64[ns]
9   Address                              9974 non-null   object
10  City                                 9974 non-null   object
11  Postal Code                          9974 non-null   int64
12  Country                              9974 non-null   object
13  Check-in Date                        9974 non-null   datetime64[ns]
14  Check-out Date                       9974 non-null   datetime64[ns]
15  Room Number                          9974 non-null   int64
16  Floor Number                         9974 non-null   int64
17  Room Type                            9974 non-null   object
18  Adults                               9974 non-null   int64
19  Children                             9974 non-null   int64
...
30  Airport Pickup Included              9974 non-null   object
31  Room Type Rate                       9974 non-null   int64
dtypes: datetime64[ns](4), int64(10), object(18)
memory usage: 2.4+ MB
```

Tập dữ liệu bao gồm các kiểu dữ liệu như: datetime, int và object.

- Kiểm tra số lượng dòng và số lượng thuộc tính của tập dữ liệu cho thấy tập dữ liệu gồm 9974 dòng và 32 thuộc tính.

```
data.shape  
✓ 0.0s
```

(9974, 32)

- Sử dụng phương thức **describe()** để tính toán thống kê mô tả tất cả các cột có kiểu dữ liệu là có dạng numeric trong toàn tập dữ liệu.

```
data.describe().T  
✓ 0.8s
```

	count	mean	min	25%	50%	75%	max	std
Reservation ID	9974.0	18870.937137	1001.0	3999.25	29174.5	32331.75	37036.0	14336.091517
Guest ID	9974.0	17970.937137	101.0	3099.25	28274.5	31431.75	36136.0	14336.091517
Birthdate	9974	1974-08-22 21:41:32.640866272	1943-02-05 00:00:00	1959-02-20 06:00:00	1974-06-19 12:00:00	1990-04-16 00:00:00	2006-01-25 00:00:00	NaN
Postal Code	9974.0	49824.038701	505.0	24745.0	49514.5	75010.5	99938.0	28876.035247
Check-in Date	9974	2023-02-01 00:12:33.639462400	2022-01-06 00:00:00	2022-08-02 06:00:00	2023-02-10 00:00:00	2023-08-03 00:00:00	2024-02-06 00:00:00	NaN
Check-out Date	9974	2023-02-03 20:03:39.450571264	2022-01-10 00:00:00	2022-08-06 00:00:00	2023-02-12 00:00:00	2023-08-06 00:00:00	2024-02-10 00:00:00	NaN
Room Number	9974.0	50.362944	1.0	26.0	50.0	75.0	100.0	28.692588
Floor Number	9974.0	2.497293	1.0	2.0	2.0	3.0	4.0	1.115753
Adults	9974.0	2.496992	1.0	1.0	2.0	4.0	4.0	1.122293
Children	9974.0	0.988169	0.0	0.0	1.0	2.0	2.0	0.814525
Total Nights	9974.0	2.827151	1.0	2.0	3.0	4.0	5.0	1.399747
Total Amount	9974.0	408.760076	60.0	205.0	320.0	572.0	1250.0	282.541411
Booking Date	9974	2023-01-09 23:00:05.053137920	2022-01-01 00:00:00	2022-07-11 00:00:00	2023-01-20 00:00:00	2023-07-11 00:00:00	2023-12-30 00:00:00	NaN
Room Type Rate	9974.0	144.940746	60.0	83.0	135.0	212.0	250.0	62.424222

Qua bảng mô tả dữ liệu:

- Các ngày sinh (Birthdate) trải dài từ năm 1943 đến 2006, điều này cho thấy sự đa dạng về độ tuổi của khách hàng.
- Số người lớn (Adults) trung bình là 2.50, phần lớn các đặt phòng dành cho 2 hoặc 4 người lớn.
- Số trẻ em (Children) đi cùng gia đình với trung bình là 0.99, với nhiều đặt phòng không có trẻ em hoặc chỉ có 1-2 trẻ em.
- Tổng số đêm (Total Nights) trung bình là 2.83, phần lớn các đặt phòng có thời gian lưu trú từ 2-3 đêm.
- Tổng số tiền phải trả (Total Amount) trung bình là 408.76, với số tiền cho mỗi lần chi trả thấp nhất là 60\$, số tiền lớn nhất lên đến 1250\$, cho thấy sự đa dạng trong mức chi tiêu của khách hàng.

4.3. Tiền xử lý dữ liệu

4.3.1. Xử lý giá trị thiếu

- Kiểm tra dữ liệu rỗng trong tập, tiến hành xử lý bằng cách thay thế giá trị trung bình, trung vị,... hoặc loại bỏ hoàn toàn dòng chứa dữ liệu rỗng.

```
#kiểm tra dữ liệu thiếu trong tập dữ liệu
data.isnull().sum()

✓ 0.0s
```

Reservation ID	0
Guest ID	0
First Name	0
Last Name	0
Gender	0
Email	0
Phone	0
Nationality	0
Birthdate	0
Address	0
City	0
Postal Code	0
Country	0
Check-in Date	0
Check-out Date	0
Room Number	0
Floor Number	0
Room Type	0
Adults	0
Children	0
Total Nights	0
Total Amount	0
Payment Status	0
Special Requests	0
Reservation Source	0
...	
Breakfast Included	0
Spa Package Included	0
Airport Pickup Included	0
Room Type Rate	0
dtype: int64	

Tập dữ liệu hoàn toàn đầy đủ, không chứa giá trị nul trong tập.

4.3.2. Tính toán các chỉ số RFM

```
#tính toán RFM
format = '%Y-%m-%d'
#lấy ngày lớn nhất trong reservation date
import datetime
current_date = max(data['Booking Date']) + datetime.timedelta(days=1)

# Tính toán RFM
RFM = data.groupby('Country').agg(
    {
        'Booking Date': lambda x: (current_date - x.max()).days,
        'Reservation ID': 'count',
        'Total Amount': 'sum'
    }
)
RFM.rename(columns={'Booking Date': 'Recency', 'Reservation ID': 'Frequency', 'Total Amount': 'MoneytaryValue'}, inplace=True)
RFM.head(10)

✓ 0.0s
```

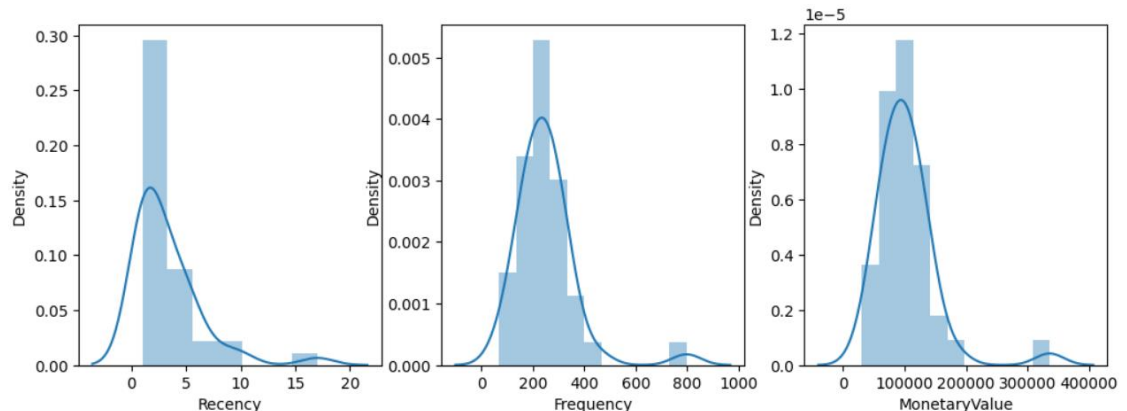
	Recency	Frequency	MoneytaryValue
Country			
Argentina	9	129	54902
Australia	2	88	40672
Austria	1	173	63538
Belgium	5	315	128269
Brazil	1	223	91063
Bulgaria	1	455	182564
Canada	5	158	62089
China	5	134	54652
Croatia	2	292	128315
Cyprus	1	236	96100

Tính toán các giá trị của mô hình phân tích RFM, trong bài thực hiện tính toán RFM theo từng quốc gia.

- **Recency:** khoảng thời gian giữ thời gian đặt phòng khách sạn đến ngày đặt phòng cuối cùng ghi nhận trong tập dữ liệu.
- **Frequency:** Số lần đặt phòng.
- **Monetary:** Tổng số tiền đã chi tiêu.

4.3.3. Khám phá và xử lý dữ liệu RFM

➤ Kiểm tra phân phối dữ liệu

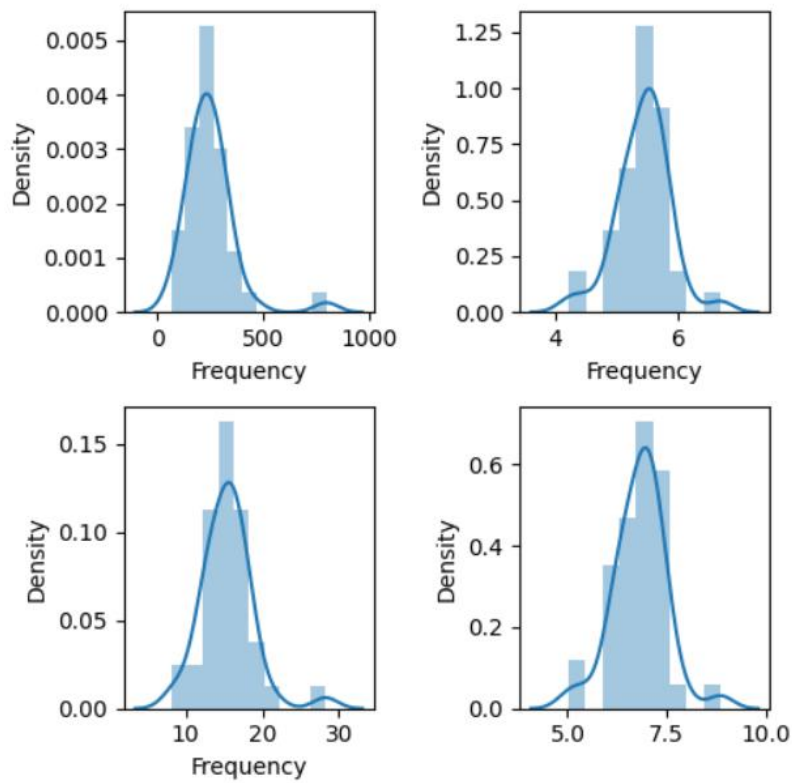


Phân phối dữ liệu lệch phải, nếu sử dụng dữ liệu cho việc chạy model sẽ không mang lại hiệu quả cao ➔ tiến hành chuyển đổi dữ liệu.

➤ Dùng hàm `analyze_skewness()` kiểm tra xem áp dụng phương pháp transform phù hợp với mô hình sau đó thực hiện transform.

```
def analyze_skewness(x):  
    fig, ax = plt.subplots(2,2,figsize=(5,5))  
    sns.distplot(RFM[x], ax=ax[0,0])  
    sns.distplot(np.log(RFM[x]), ax=ax[0,1])  
    sns.distplot(np.sqrt(RFM[x]), ax=ax[1,0])  
    sns.distplot(stats.boxcox(RFM[x])[0], ax=ax[1,1])  
    plt.tight_layout()  
    plt.show()  
  
    print(RFM[x].skew().round(2))  
    print(np.log(RFM[x]).skew().round(2))  
    print(np.sqrt(RFM[x]).skew().round(2))  
    print(pd.Series(stats.boxcox(RFM[x])[0]).skew().round(2))
```

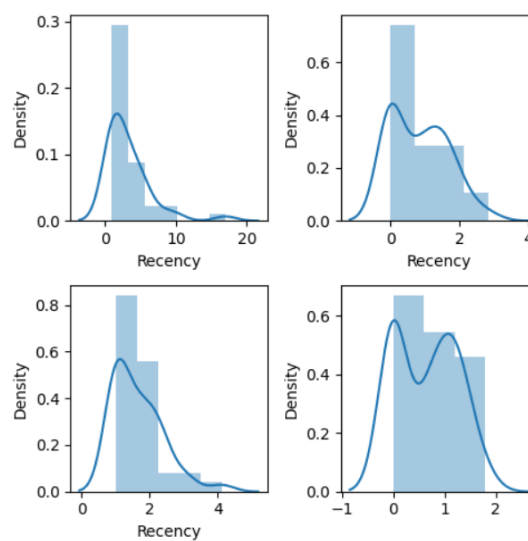
➤ Tiến hành kiểm tra trên cột Frequency cho thấy chuyển đổi dữ liệu theo kiểu boxcox là tốt nhất (do độ lệch thấp nhất là 0.02).



4

2.58
-0.18
1.14
0.02

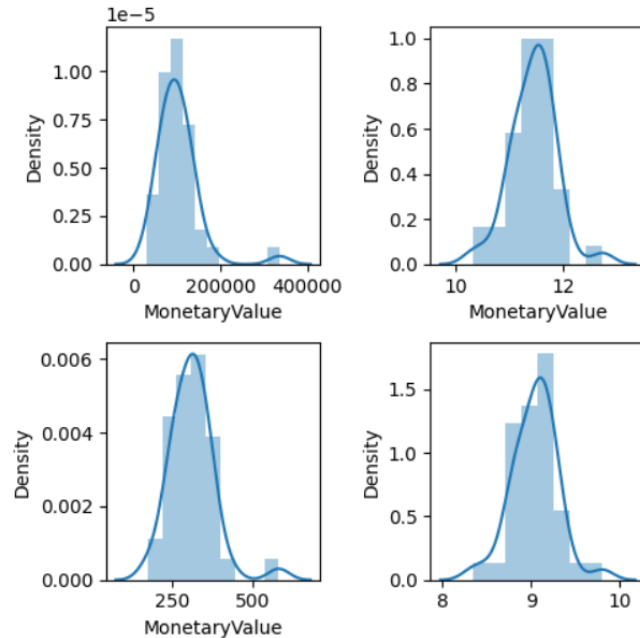
- Tiến hành kiểm tra trên cột Recency cho thấy chuyển đổi dữ liệu theo kiểu boxcox là tốt nhất (do độ lệch thấp nhất là 0.14).



4

2.47
0.47
1.27
0.14

- Tương tự tiến hành kiểm tra trên cột MonetaryValue cho thấy chuyển đổi dữ liệu theo kiểu boxcox là tốt nhất (do độ lệch thấp nhất là -0.01)



2.74
0.09
1.33
-0.01

- Transform dữ liệu theo các kỹ thuật phù hợp.

```
#transform du lieu
RFMT = pd.DataFrame()
RFMT['Recency'] = stats.boxcox(RFM['Recency'])[0]
RFMT['Frequency'] = stats.boxcox(RFM['Frequency'])[0]
RFMT['MoneytaryValue'] = stats.boxcox(RFM['MoneytaryValue'])[0]
RFMT.head()
```

	Recency	Frequency	MoneytaryValue
0	1.530962	5.954460	8.718772
1	0.615195	5.396312	8.529519
2	0.000000	6.394608	8.810051
3	1.229277	7.326524	9.241099
4	0.000000	6.783893	9.032513

4.3.4. Chuẩn hóa dữ liệu

Sử dụng phương pháp StandScaler là một bộ biến đổi (transformer) được sử dụng để chuẩn hóa dữ liệu bằng cách loại bỏ giá trị trung bình và chia cho độ lệch chuẩn của mỗi đặc trưng.

```
scaler = StandardScaler()  
scaler.fit(RFMT)  
RFMT = scaler.transform(RFMT)  
pd.DataFrame(RFMT).head()
```

✓ 0.0s

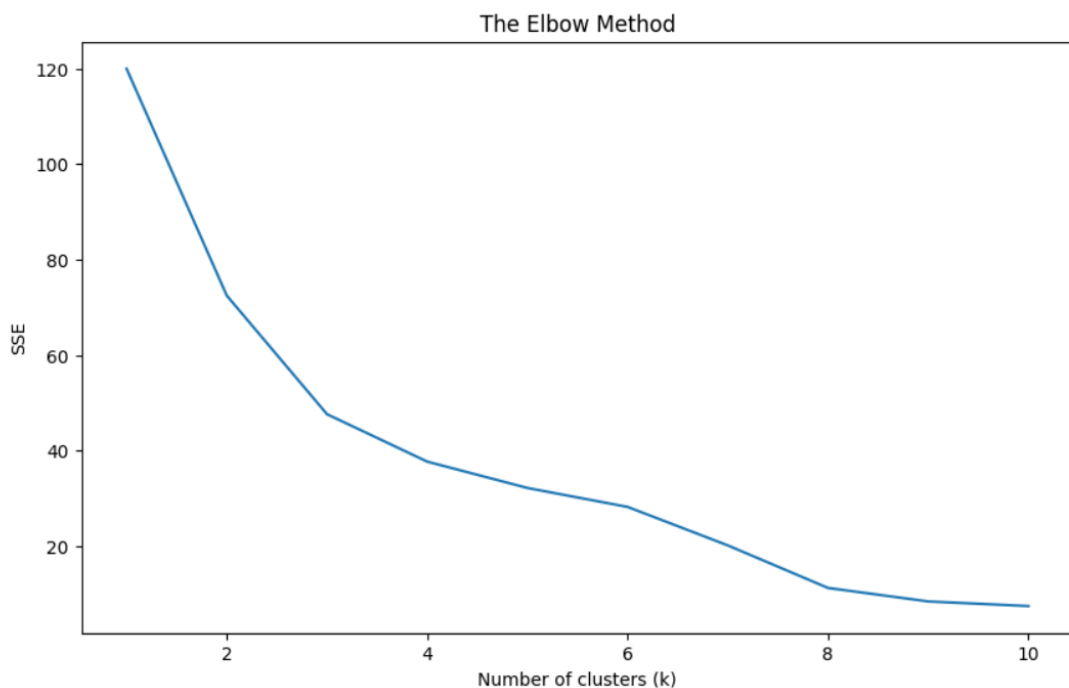
	0	1	2
0	1.522591	-1.332014	-1.280421
1	-0.049365	-2.182302	-2.024205
2	-1.105377	-0.661490	-0.921682
3	1.004735	0.758199	0.772388
4	-1.105377	-0.068449	-0.047381

4.4. Xây dựng mô hình phân cụm

4.4.1. Xác định số lượng cụm (k) tối ưu bằng kỹ thuật Elbow

Kỹ thuật Elbow là một phương pháp phổ biến để xác định số lượng cụm (k) tối ưu trong các thuật toán phân cụm. Cách thức hoạt động của kỹ thuật Elbow:

- Chạy thuật toán phân cụm (ví dụ K-Means) với các giá trị k khác nhau, thường từ 1 đến 10 hoặc nhiều hơn tùy thuộc vào dữ liệu.
- Tính toán một chỉ số đánh giá chất lượng phân cụm, thường là Inertia (tổng bình phương khoảng cách các điểm đến tâm cụm gần nhất).
- Vẽ một đồ thị biểu diễn giá trị Inertia theo số lượng cụm k.
- Tìm "khủy tay" (elbow) trên đồ thị, tức là điểm mà sau đó giá trị Inertia giảm chậm hơn. Điểm này được coi là số lượng cụm tối ưu.



➔ Qua biểu đồ: elbow xuất hiện ở $k = 4$, vì sau đó inertia giảm chậm hơn. Do đó, số lượng cụm tối ưu là 4.

4.4.2. Xây dựng các mô hình

Xây dựng mô hình K-Means, DBSCAN và Agglomerative Clustering. K-Means và Agglomerative Clustering phân chia theo số cụm là 4 như đã xác định ở trên.

```
#K-Means
kmeans = KMeans(n_clusters = 4, random_state=0)
kmeans_labels = kmeans.fit_predict(RFMT)

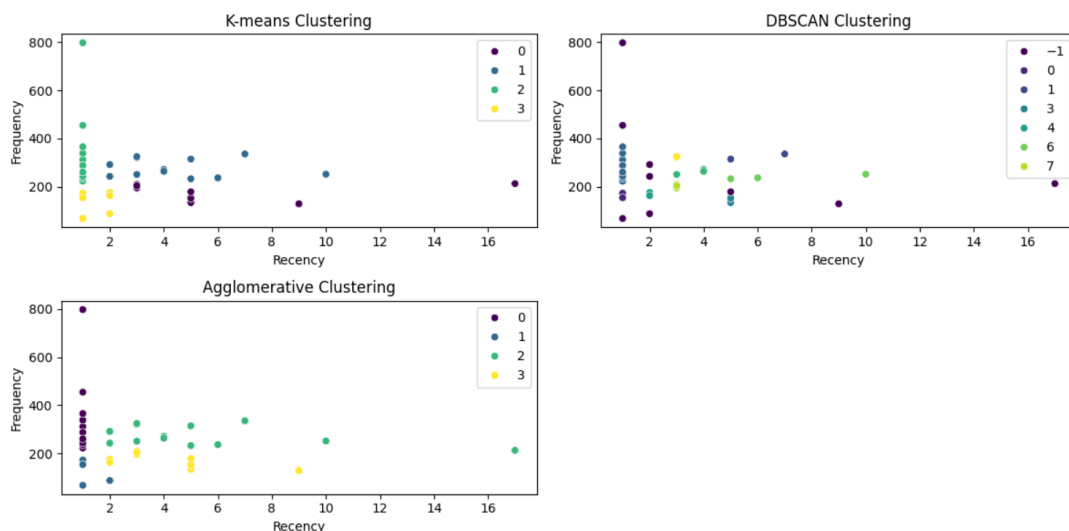
# DBSCAN
dbscan = DBSCAN(eps=0.5, min_samples=2)
dbscan_labels = dbscan.fit_predict(RFMT)

# Agglomerative Clustering
agg_clustering = AgglomerativeClustering(n_clusters=4)
agg_labels = agg_clustering.fit_predict(RFMT)
dbscan_silhouette = silhouette_score(RFMT, dbscan_labels)
```

4.4.3. Đánh giá mô hình và trực quan hóa kết quả

Sử dụng Silhouette Score xác định chất lượng của việc phân cụm. Silhouette Score tính toán độ tương đồng của một điểm dữ liệu với các điểm dữ liệu trong cùng cụm (cohesion) so với các điểm dữ liệu trong các cụm khác (separation).

Silhouette Score for K-means: 0.3742805771100933
Silhouette Score for DBSCAN: 0.31108696839761496
Silhouette Score for Agglomerative Clustering: 0.4022408237937481



- **K-means: 0.37**, các cụm được tạo ra bởi K-means có độ tương đồng trong nội bộ cụm và độ phân biệt giữa các cụm ở mức trung bình.
- **DBSCAN: 0.31**, hơn so với K-means ➔ độ phân biệt giữa các cụm được DBSCAN tạo ra không tốt bằng K-means.
- **Agglomerative Clustering: 0.40**, giá trị cao nhất trong 3 phương pháp.

➔ Dựa trên các giá trị Silhouette Score, Agglomerative Clustering cho kết quả phân cụm tốt nhất, tiếp đến là K-means, và DBSCAN cho kết quả phân cụm kém hơn. Sử dụng thuật toán Agglomerative Clustering tiến hành phân tích phân cụm khách hàng.

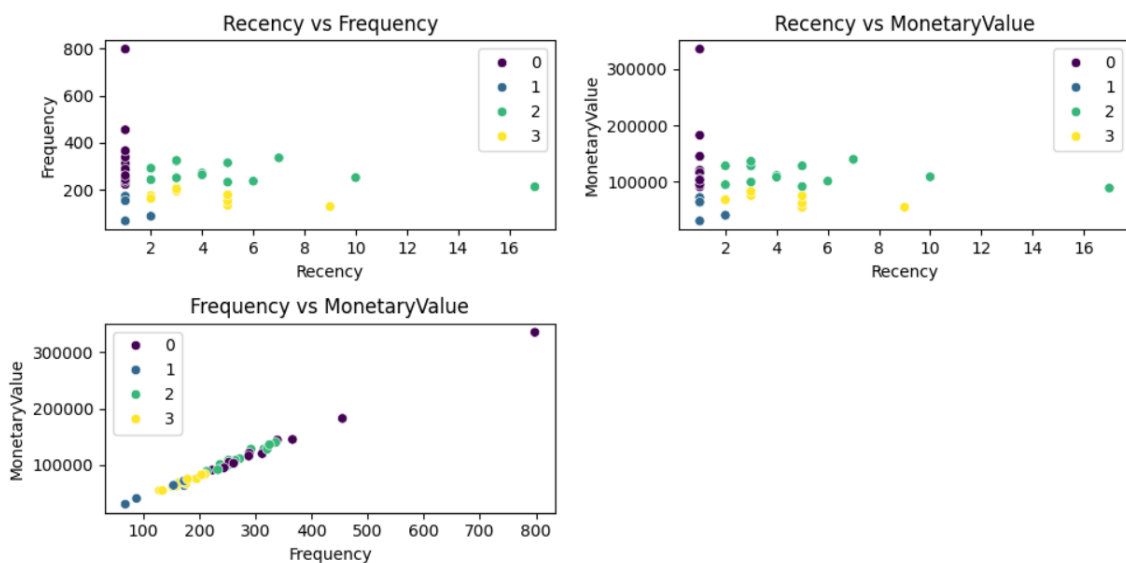
4.4.4. Phân tích & phân cụm khách hàng

- Sử dụng thuật toán Agglomerative Clustering tiến hành gán nhãn phân cụm cho các dòng dữ liệu:

```
RFM['cluster'] = model.labels_  
RFM.head()
```

	Recency	Frequency	MonetaryValue	Cluster
Country				
Argentina	9	129	54902	3
Australia	2	88	40672	1
Austria	1	173	63538	1
Belgium	5	315	128269	2
Brazil	1	223	91063	0

- Trực quan hóa kết quả phân cụm bằng các cặp thuộc tính khác nhau:



- Tính toán giá trị trung bình RFM tìm ra các đặc trưng của từng nhóm:

	Recency	Frequency	MonetaryValue
Cluster			
0	1.00	338.67	137934.17
1	1.20	131.20	54173.40
2	5.46	273.38	112715.77
3	4.20	170.00	68559.10

Nhận xét:

- **Cluster 0:** Khách hàng VIP có giá trị cao nhất với tần suất và giá trị chi tiêu rất cao.
- **Cluster 1:** Khách hàng trung thành với tần suất và giá trị chi tiêu cao.
- **Cluster 2:** Khách hàng có giá trị cao nhưng không hoạt động gần đây.
- **Cluster 3:** Khách hàng thường xuyên với giá trị chi tiêu trung bình.

Đặt tên các phân loại dựa theo đặc điểm chung:

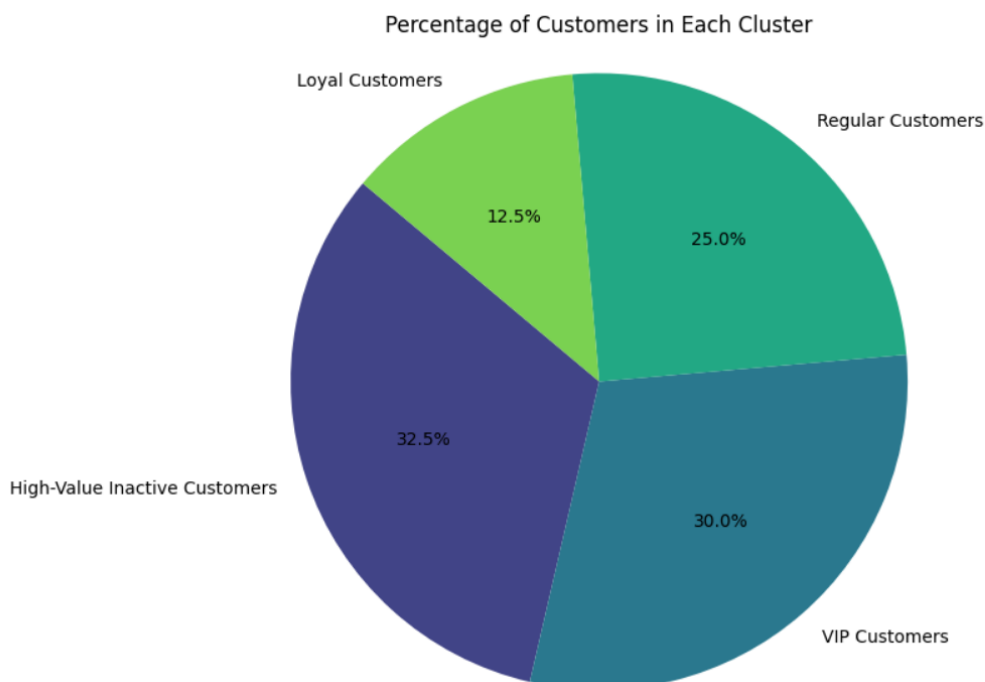
```
RFM['cluster'] = agg_labels
def name_clusters(row):
    if row['cluster'] == 0:
        return 'VIP Customers'
    elif row['cluster'] == 1:
        return 'Loyal Customers'
    elif row['cluster'] == 2:
        return 'High-Value Inactive Customers'
    elif row['cluster'] == 3:
        return 'Regular Customers'
    else:
        return 'Other'

RFM['clusterName'] = RFM.apply(name_clusters, axis=1)
RFM.head()
```

✓ 0.1s

	Recency	Frequency	MonetaryValue	Cluster	ClusterName
Country					
Argentina	9	129	54902	3	Regular Customers
Australia	2	88	40672	1	Loyal Customers
Austria	1	173	63538	1	Loyal Customers
Belgium	5	315	128269	2	High-Value Inactive Customers
Brazil	1	223	91063	0	VIP Customers

- Vẽ biểu đồ cho thấy tỉ lệ phân phối của từng nhóm khách hàng trên tập dữ liệu:



→Hiển thị các quốc gia thuộc từng nhóm:

Cluster 0:
['Brazil', 'Bulgaria', 'Cyprus', 'Finland', 'Germany', 'Italy', 'Lithuania', 'Portugal', 'Romania', 'Slovakia', 'Switzerland', 'United States']

Cluster 1:
['Australia', 'Austria', 'France', 'Netherlands', 'Russia']

Cluster 2:
['Belgium', 'Croatia', 'Czech Republic', 'Greece', 'Hungary', 'Latvia', 'Luxembourg', 'Norway', 'Poland', 'Serbia', 'Slovenia', 'Spain', 'Sweden']

Cluster 3:
['Argentina', 'Canada', 'China', 'Denmark', 'Estonia', 'India', 'Ireland', 'Japan', 'Malta', 'United Kingdom']

➤ Các chiến lược phù hợp cho từng nhóm khách hàng

High-Value Inactive Customers (32%):

- Chiếm tỷ lệ lớn nhất trong các cụm khách hàng. Các quốc gia trong cụm này có giá trị chi tiêu cao nhưng tần suất đặt phòng giảm sút.
- **Mục tiêu:** Kích thích hoạt động trở lại của khách hàng từ các quốc gia này bằng cách: gửi các chiến dịch tiếp thị và ưu đãi đặc biệt, khảo sát để hiểu nguyên nhân giảm sút tần suất đặt phòng và cải thiện dịch vụ.

VIP Customers (30%):

- Chiếm tỷ cao thứ 2 lệ gần như tương đương với nhóm High-Value Inactive. Các quốc gia này có khách hàng có giá trị chi tiêu cao và thường xuyên đặt phòng.
- **Mục tiêu:** cần duy trì và nâng cao trải nghiệm để giữ chân nhóm khách hàng này.
- **Chiến lược:**
 - Cung cấp các chương trình ưu đãi đặc biệt và dịch vụ cao cấp.
 - Tạo các chương trình khách hàng thân thiết và tích điểm.
 - Thường xuyên gửi lời cảm ơn và đánh giá dịch vụ từ khách hàng.

Regular Customers (25%):

- Chiếm tỷ lệ khá đáng kể. Các quốc gia này có khách hàng thường xuyên đặt phòng nhưng với giá trị chi tiêu trung bình.
- **Mục tiêu:** Tăng tần suất đặt phòng và giá trị chi tiêu.
- **Chiến lược:**
 - Cung cấp các chương trình khuyến mãi định kỳ.
 - Tăng cường chăm sóc khách hàng để nâng cao mức độ hài lòng.
 - Tạo các gói dịch vụ hấp dẫn để khách hàng chi tiêu nhiều hơn.

Loyal Customers (13%):

- Chiếm tỷ lệ nhỏ nhất trong các cụm khách hàng. Nhóm khách hàng trung thành, thường xuyên đặt phòng với giá trị chi tiêu cao.
- **Mục tiêu:** Khuyến khích tiếp tục đặt phòng thường xuyên và biến họ thành VIP Customers.
- **Chiến lược:** khuyến khích họ tiếp tục đặt phòng bằng các chương trình tích điểm hoặc giảm giá, tăng cường chăm sóc khách hàng và cá nhân hóa trải nghiệm.