|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HỒ CHÍ MINH** |
| **KHOA CÔNG NGHỆ THÔNG TIN**  **'Y&&Y'**  **Logo ĐH Sư Phạm Kỹ Thuật TP Hồ Chí Minh-HCMUTE** |

|  |
| --- |
| **BÁO CÁO ĐỒ ÁN CUỐI KỲ**  **ĐỀ TÀI: XÂY DỰNG KHO DỮ LIỆU CHO KHÁCH SẠN** |
| **Môn học: KHO DỮ LIỆU**  **Mã lớp học phần:**  **DAWH430784\_23\_2\_02**  **GVHD: ThS. Nguyễn Văn Thành** |
| **Nhóm sinh viên thực hiện: Nhóm 6**   |  |  | | --- | --- | | Lại Văn An | 21110363 | | Nguyễn Đức Kha | 21133044 | | Nguyễn Nhật Trường | 21110935 | |

**TP. Hồ Chí Minh, 13 tháng 05 năm 2023**

**DANH SÁCH THÀNH VIÊN THAM GIA**

**THỰC HIỆN ĐỀ TÀI VÀ VIẾT BÁO CÁO**

**Môn: Kho dữ liệu -** *HỌC KÌ II – NĂM HỌC 2022 – 2023*

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **HỌ VÀ TÊN** | **MSSV** | **TỶ LỆ ĐÓNG GÓP** |
| 1 | Nguyễn Đức Kha | 21133044 | 100% |
| 2 | Nguyễn Nhật Trường | 21110935 | 100% |
| 3 | Lại Văn An | 21110363 | 100% |

**Nhận xét của giảng viên:**

**……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………**

*Ngày … tháng 05 năm 2023*

*Giảng viên chấm điểm*

*Ths. Nguyễn Văn Thành*

**LỜI CẢM ƠN**

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến **Trường Đại học Sư phạm Kỹ thuật TPHCM** đã đưa môn học **Kho Dữ Liệu** vào chương trình giảng dạy. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – **Th.S Thầy Nguyễn Văn Thành** đã dạy dỗ, truyền đạt những kiến thức quý báu cho chúng em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học Kho Dữ Liệu của Thầy, nhóm em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để chúng em có thể vững bước sau này. Bộ môn Kho Dữ Liệu là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài đồ án này khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong Thầy xem xét và góp ý để bài đồ án của chúng em được hoàn thiện hơn.

Nhóm chúng em xin chân thành cảm ơn!

**PHÂN CÔNG NHIỆM VỤ**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhiệm vụ** | Nhật Trường | Đức Kha | Văn An |
| Tìm kiếm tập dữ liệu | x | x | x |
| **Mô hình hóa dữ liệu vào workbook** | **x** | **x** | **x** |
| +Dim Date  + Dim Hotel  + Dim Customers |  |  | x |
| + Fact Canceled |  | x |  |
| +Fact Booking | x |  |  |
| **Thực hiện ETL đưa dữ liệu vào các bảng dim và fact dùng SSIS** | **x** | **x** | **x** |
| + Tạo và đưa dữ liệu vào các dim stages và dim |  |  | x |
| + Tạo và đưa dữ liệu vào FactBooking stage và FactBooking | x |  |  |
| + Tạo và đưa dữ liệu vào FactCanceledstage và FactCanceled |  | x |  |
| Nhập dữ liệu vào SSAS, tạo data source, data source view và cube |  | x |  |
| **Tạo các phân cấp hierarchy cho các bảng dim và các attribute relationships.** | **x** | **x** | **x** |
| + Hierarchy cho DimDate | x |  |  |
| + Hierarchy cho DimGeography, DimProduct |  |  | x |
| + Hierarchy cho DimInventory, DimEmployee |  | x |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Thực hiện trả lời các câu hỏi đã đặt ra cho business process: Canceled Analysis** | **x** | **x** | **x** |
| Số lượng huỷ phòng của Resort Hotel và City Hotel ở các tháng theo từng năm. | x |  |  |
| Lượng huỷ phòng theo quốc gia của từng năm. |  | x |  |
| Cho biết số lượng huỷ phòng và giá trung bình ADR qua các tháng theo tường năm của Resort Hotel và City Hotel. |  |  | x |
| **Thực hiện trả lời các câu hỏi đã đặt ra cho business process: Booking Analysis** | **x** |  | **x** |
| Số lượng đặt phòng qua các công ty lữ hành | x |  |  |
| Thống kế số lượng đặt phong ứng với các phương thức đặt phòng. |  |  | x |
| + Trả lời câu hỏi dùng SSAS |  | x |  |
| + Trả lời câu hỏi dùng Power BI Desktop |  | x |  |
| + Trả lời câu hỏi dùng Excel Pivot | x |  |  |

# CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

## Lý do chọn đề tài

## Lý do chọn đề tài kho dữ liệu cho khách sạn xuất phát từ nhu cầu cấp thiết của ngành công nghiệp này trong việc quản lý và khai thác hiệu quả nguồn dữ liệu phong phú. Kho dữ liệu không chỉ giúp tập trung và tổ chức thông tin từ nhiều nguồn như đặt phòng, dịch vụ, và phản hồi khách hàng, mà còn hỗ trợ phân tích xu hướng và dự báo nhu cầu tương lai. Bằng cách tối ưu hóa hoạt động kinh doanh và cá nhân hóa trải nghiệm khách hàng, kho dữ liệu giúp nâng cao hiệu suất và khả năng cạnh tranh của khách sạn. Đồng thời, nó đảm bảo việc tuân thủ các quy định pháp lý về bảo mật thông tin, tạo cơ sở vững chắc cho sự đổi mới và áp dụng công nghệ tiên tiến như trí tuệ nhân tạo và học máy. Với những lợi ích vượt trội này, việc nghiên cứu và triển khai kho dữ liệu trở thành một yếu tố then chốt để thúc đẩy sự phát triển bền vững và thành công của ngành khách sạn

## Tổng quan về tập dữ liệu

### Nguồn dữ liệu

* Nhóm sử dụng Tập dữ liệu Hotel Bookings được lấy từ trang web Kaggle (https://www.kaggle.com/), Kaggle là một nền tảng trực tuyến cho cộng đồng Machine Learning (ML) và Khoa học dữ liệu. Kaggle cho phép người dùng chia sẻ, tìm kiếm các bộ dữ liệu; tìm hiểu và xây dựng models, tương tác với những nhà khoa học và kỹ sư ML trên toàn thế giới; tham gia các cuộc thi để có cơ hội chiến thắng những giải thưởng giá trị.
* Đường dẫn tải tập dữ liệu: <https://www.kaggle.com/code/ahmedsalouh/hotel-bookings-analysis>

### Mô tả chi tiết tập dữ liệu

* Tập dữ liệu ‘dataset.xlsx’ chứa thông tin về việc đặt phòng cho hai khách sạn ở Bồ Đào Nha (một khách sạn nghỉ dưỡng và một khách sạn trong thành phố) dự kiến ​​đến trong khoảng thời gian từ ngày 1 tháng 7 năm 2015 đến ngày 31 tháng 8 năm 2017 và đối với cả hai khách sạn, thông tin giống nhau đã được thu thập như quan sát và đặt phòng khách sạn , bao gồm cả những cái đã bị hủy. Dữ liệu bao gồm các 37 cột với mô tả một số cột chính như sau:

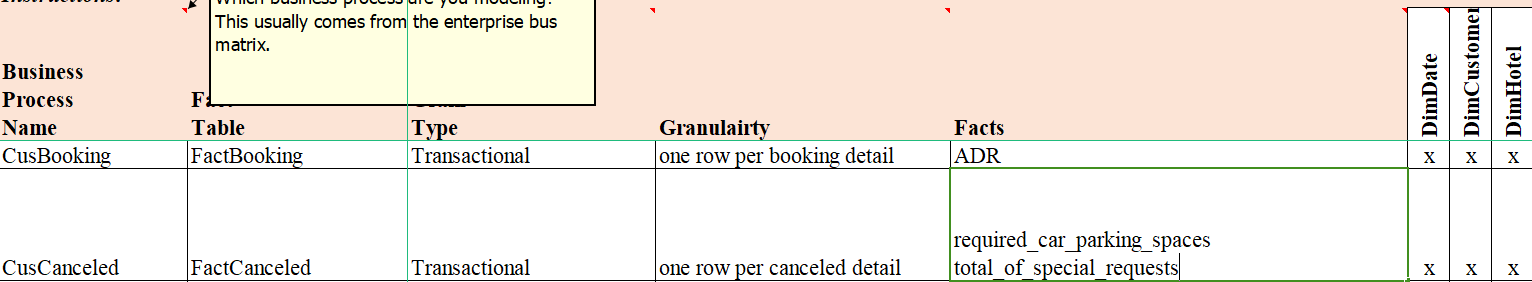
| **Tên cột** | **Mô tả** |
| --- | --- |
| **hotel** | Một trong các khách sạn là khách sạn nghỉ dưỡng và một khách sạn là khách sạn thành phố. |
| **is\_canceled** | Giá trị cho biết liệu đặt phòng có bị hủy (1) hay không (0). |
| **lead\_time** | Thời gian từ khi đặt phòng đến khi khách đến. |
| **arrival\_date\_year** | Năm của ngày đến. |
| **arrival\_date\_month** | Tháng của ngày đến với 12 danh mục: “1” đến “12”. |
| **arrival\_date\_week\_number** | Số tuần của ngày đến. |
| **arrival\_date\_day\_of\_month** | Ngày trong tháng của ngày đến. |
| **country** | Quốc gia xuất xứ. |
| **market\_segment** | Phân khúc thị trường. Trong danh mục, thuật ngữ “TA” nghĩa là “Travel Agents” và “TO” nghĩa là “Tour Operators”. |
| **reserved\_room\_type** | Mã loại phòng được đặt. Mã được trình bày thay vì tên để đảm bảo tính ẩn danh. |
| **agent** | ID của đại lý du lịch đã thực hiện đặt phòng. |
| **adr** | Giá Phòng Trung Bình Hàng Ngày (Tính bằng cách chia tổng số giao dịch lưu trú cho tổng số đêm ở). |
| **customer\_type** | Group – khi đặt phòng liên quan đến một nhóm; Transient – khi đặt phòng không phải là một phần của nhóm hoặc hợp đồng, và không liên quan đến đặt phòng chuyển tiếp khác; Transient-party – khi đặt phòng là dạng chuyển tiếp, nhưng liên quan đến ít nhất một đặt phòng chuyển tiếp khác. |

## Giới thiệu các công cụ được sử dụng

* Visual Studio 2019/2022 tích hợp SSIS (SQL Server Integration Services) và SSAS (SQL Server Analysis Services).
* SQL Server Management Studio 2019
* Power BI.

# **CHƯƠNG 2: THIẾT KẾ XÂY DỰNG CƠ SỞ DỮ LIỆU TÁC NGHIỆP(OLAP)**

## **Xác định các Business Process và bảng Fact**

Xây dựng Detailed Bus Matrix xác định các Business Process, bảng Fact, bảng Dim cần thiết

### **2.1. Business Process: Canceled Analysis**

- Các câu hỏi đặt ra :

**+** Số lượng huỷ phòng của Resort Hotel và City Hotel ở các tháng theo từng năm.

+ Lượng huỷ phòng theo quốc gia của từng năm.

+ Cho biết số lượng huỷ phòng và giá trung bình ADR qua các tháng theo tường năm của Resort Hotel và City Hotel.

**2.2. Business Process: BookingAnalysis**

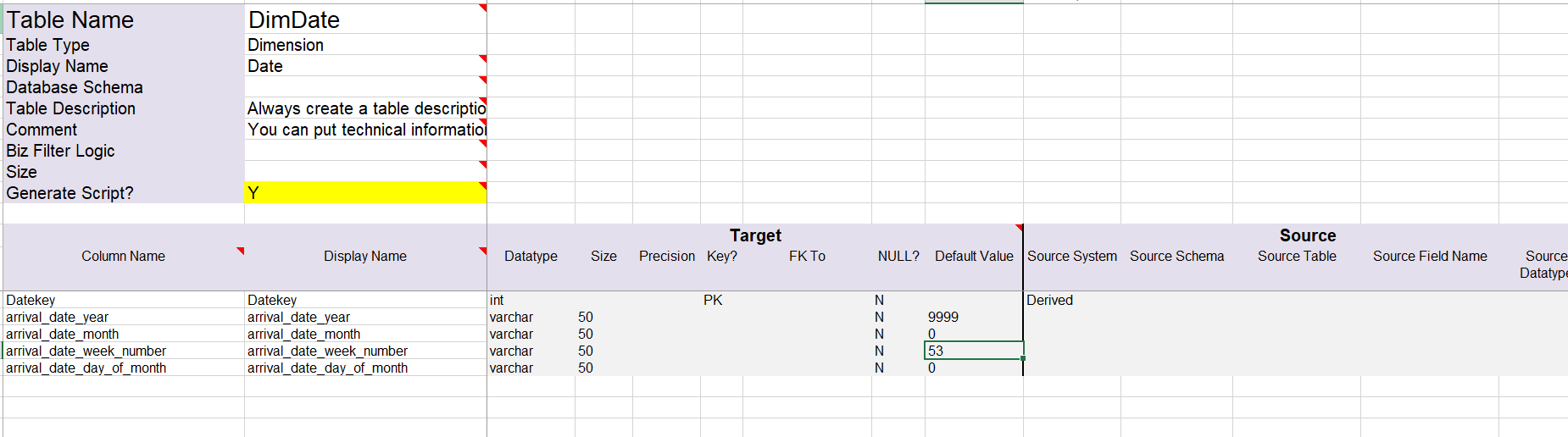
- Các câu hỏi đặt ra :

+ Thống kế số lượng đặt phong ứng với các phương thức đặt phòng.

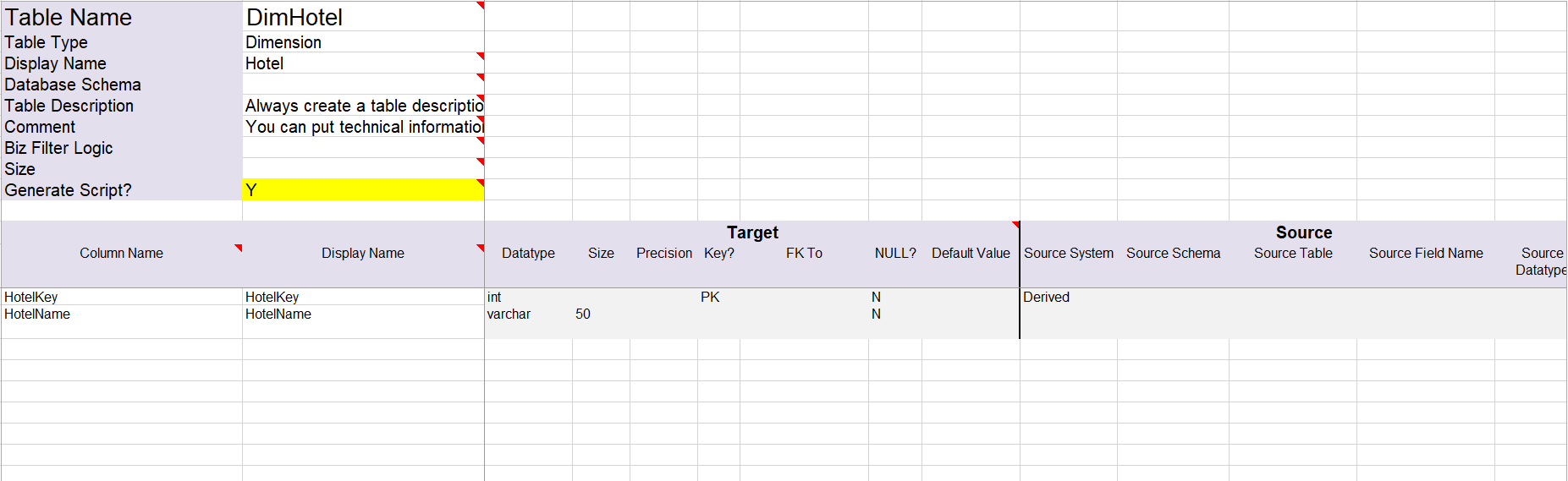
+ Số lượng đặt phòng qua các công ty lữ hành.

**2.3. Xây dựng các bảng Dimension**

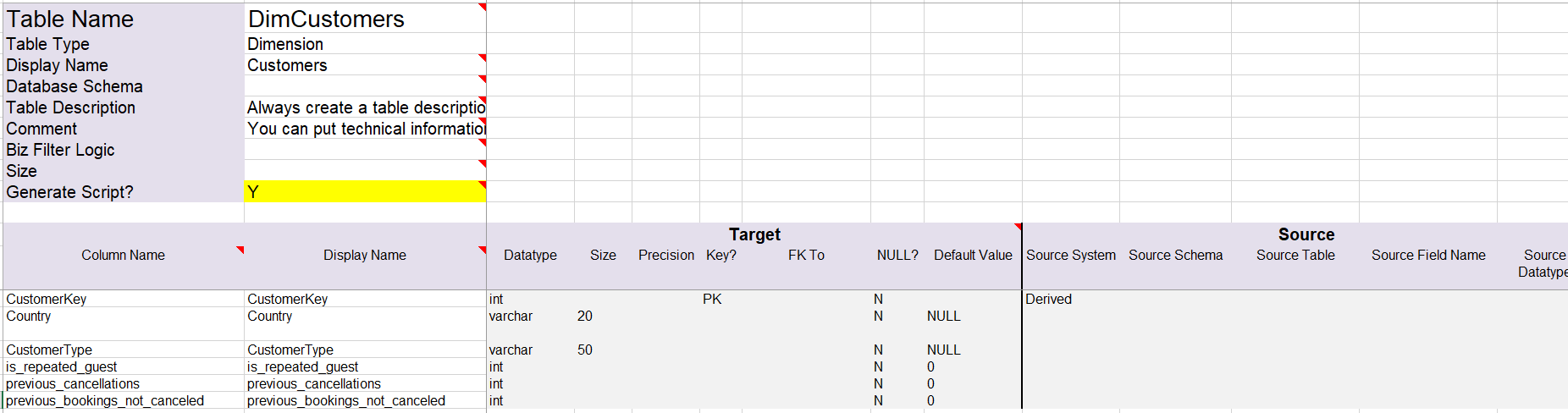
**2.3.1. DimDate**



**2.3.2. DimHotel.**

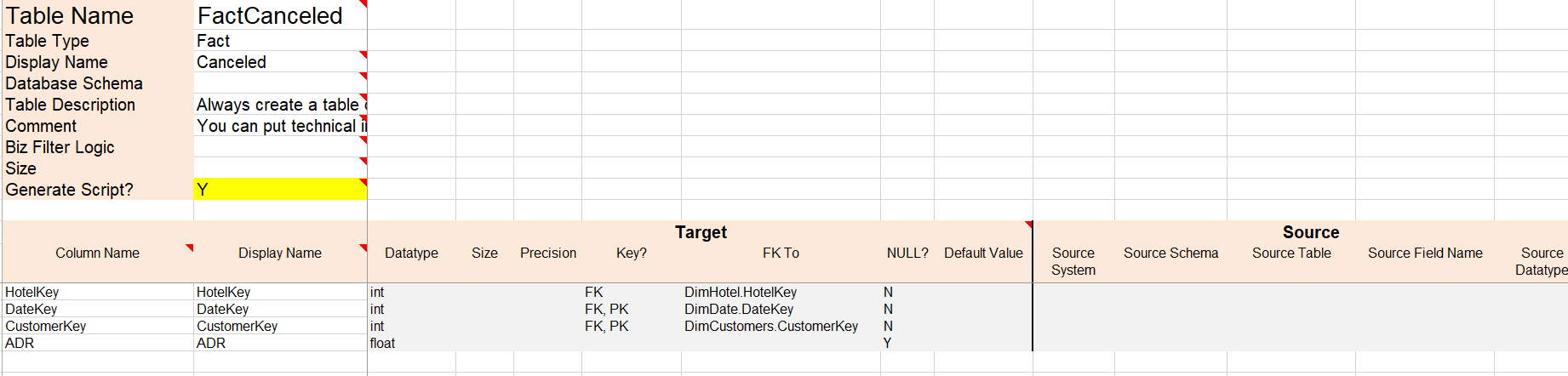


**2.3.3. DimCustomers.**

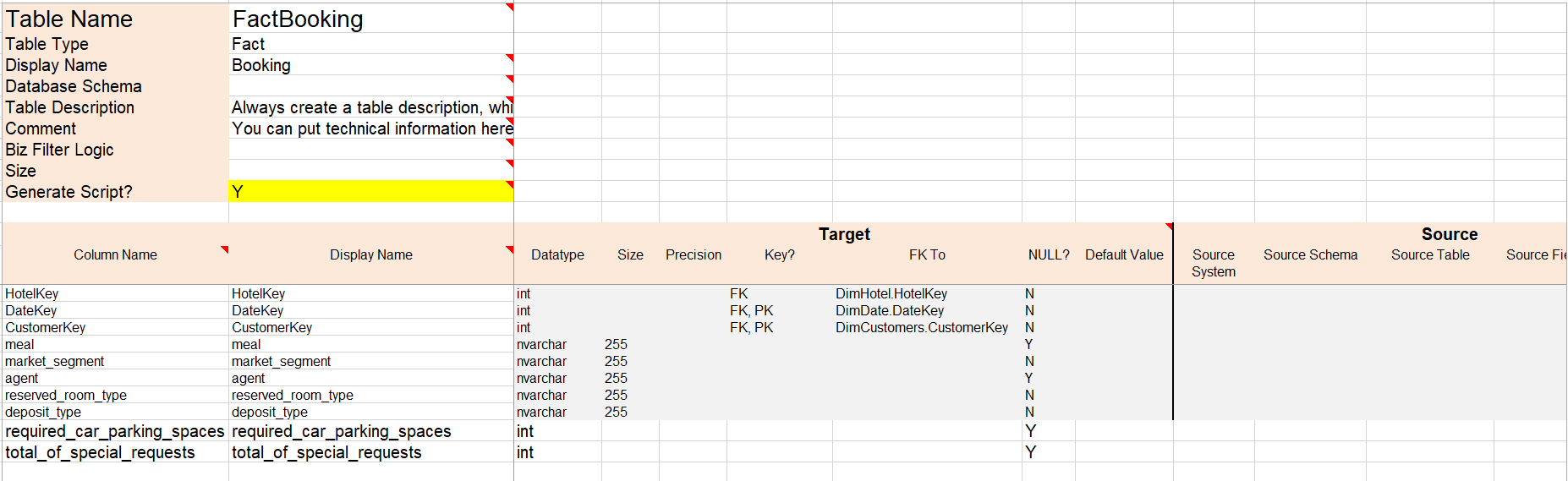


**2.4. Xây dựng các bảng Fact**

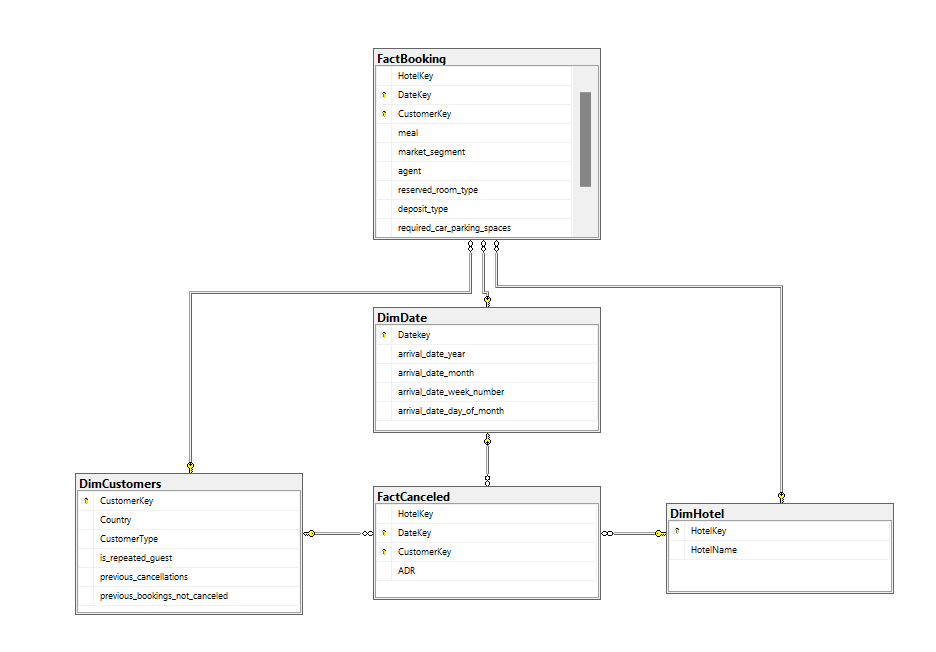
**2.4.1. FactCanceled.**



**2.4.2. FactBooking.**



**2.5. Star Schema (Lược đồ hình sao)**



# **CHƯƠNG 3: TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)**

## **3.1. Tạo ETL project và solution**

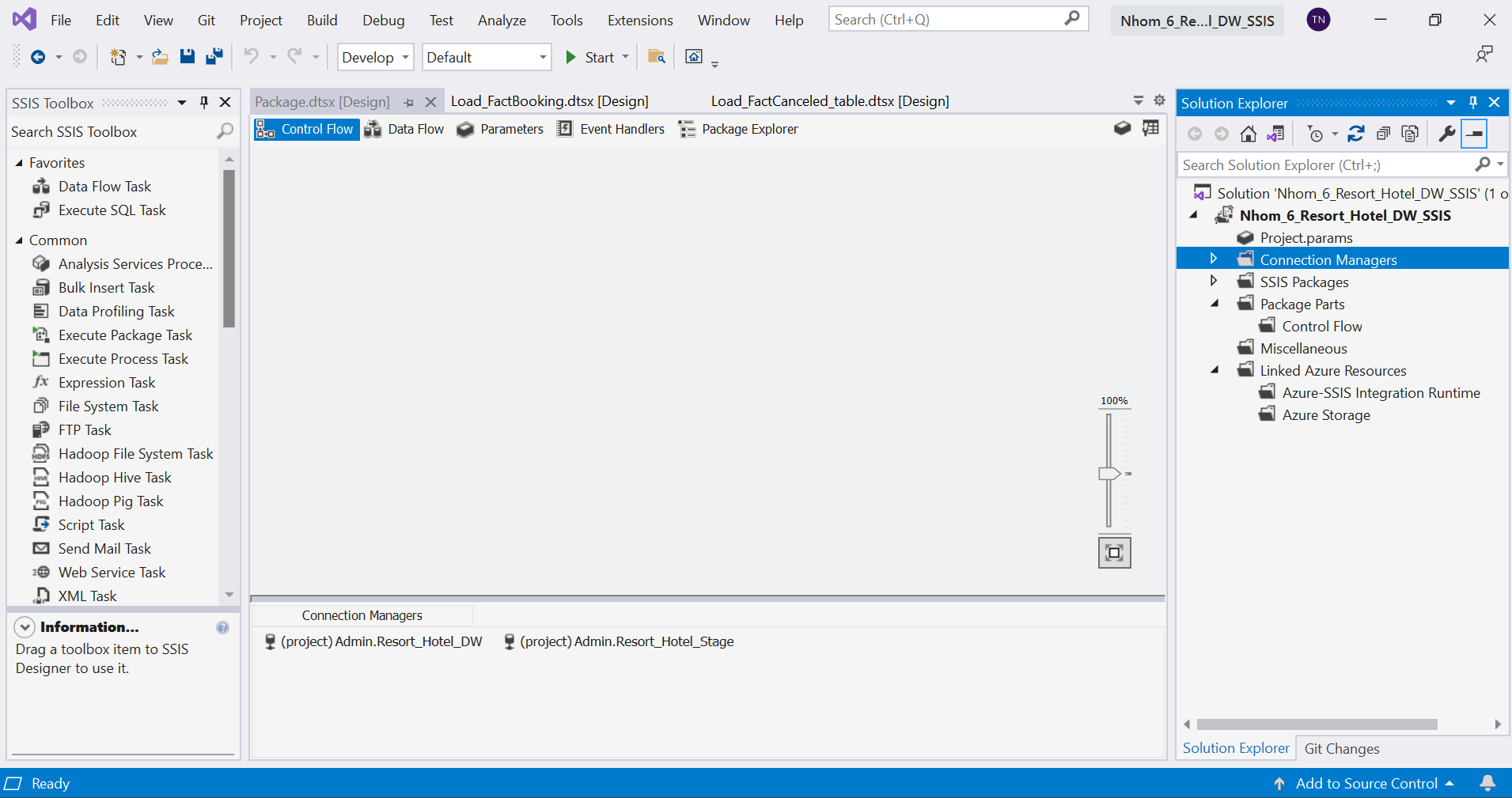
1. Từ **Visual Studio** menu, chọn **File** -> **New** -> **Project**.

2. Từ **dialog**:

a. Chọn **Integration Services Project**

b. Đặt tên: **Nhom\_6\_Resort\_Hotel\_SSIS**

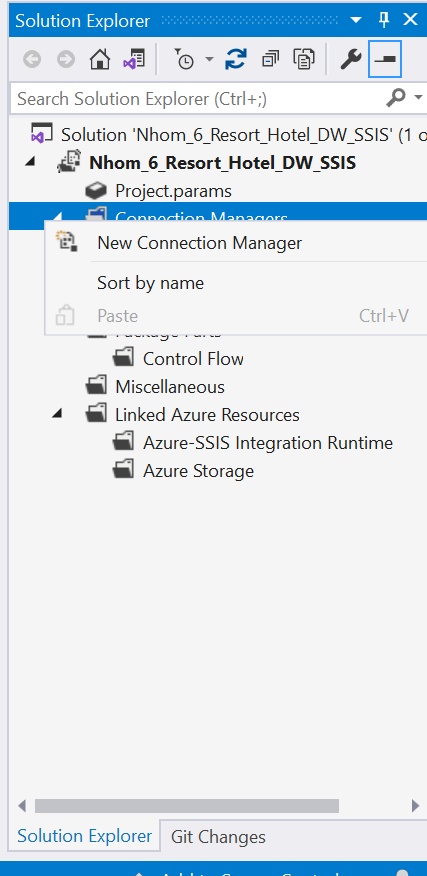
3. Click **OK**



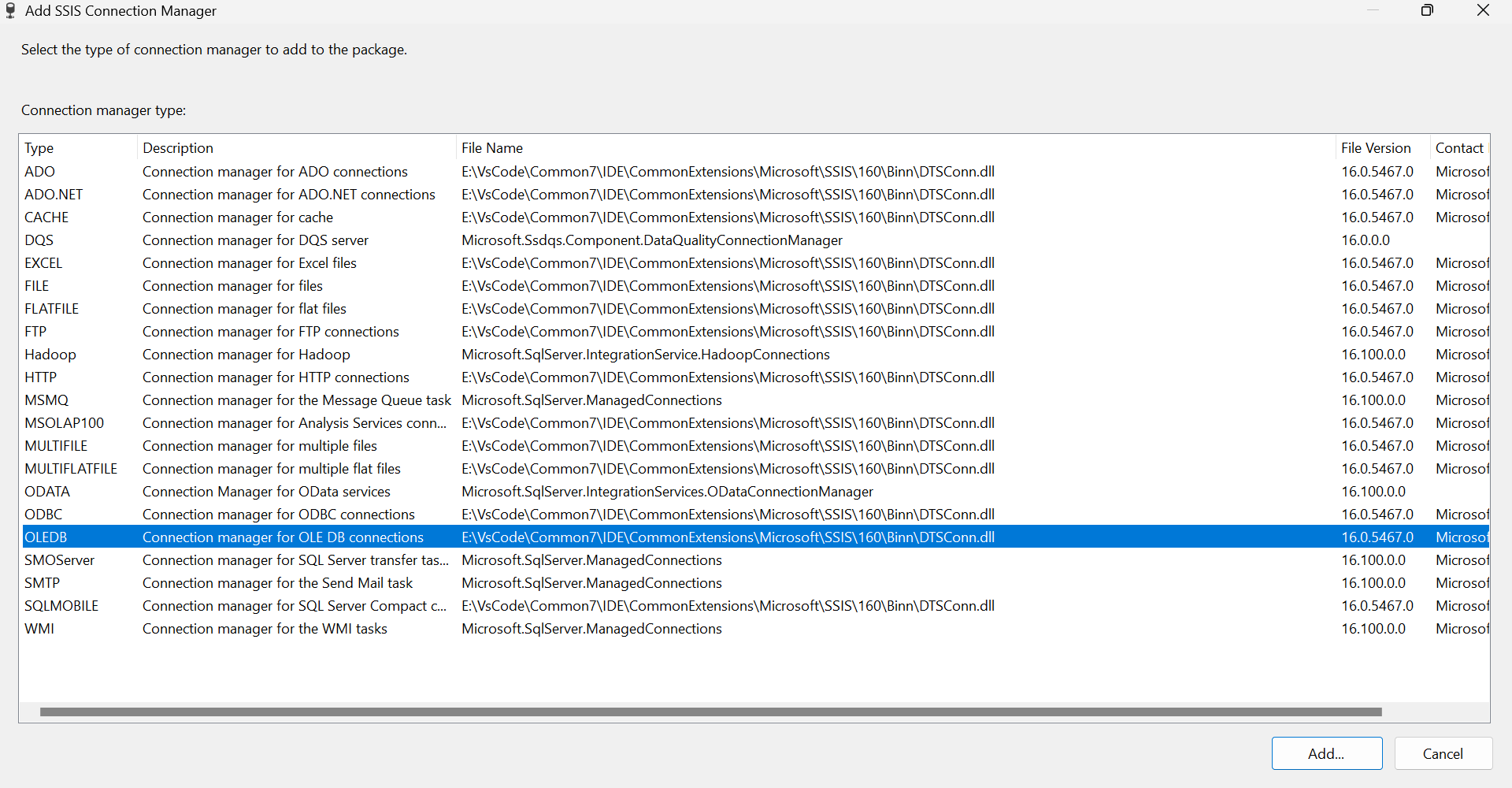
## **3.2. Thiết lập các Connection**

1. Tại **solution explorer**, right-click vào **Connection Managers,** chọn **New**

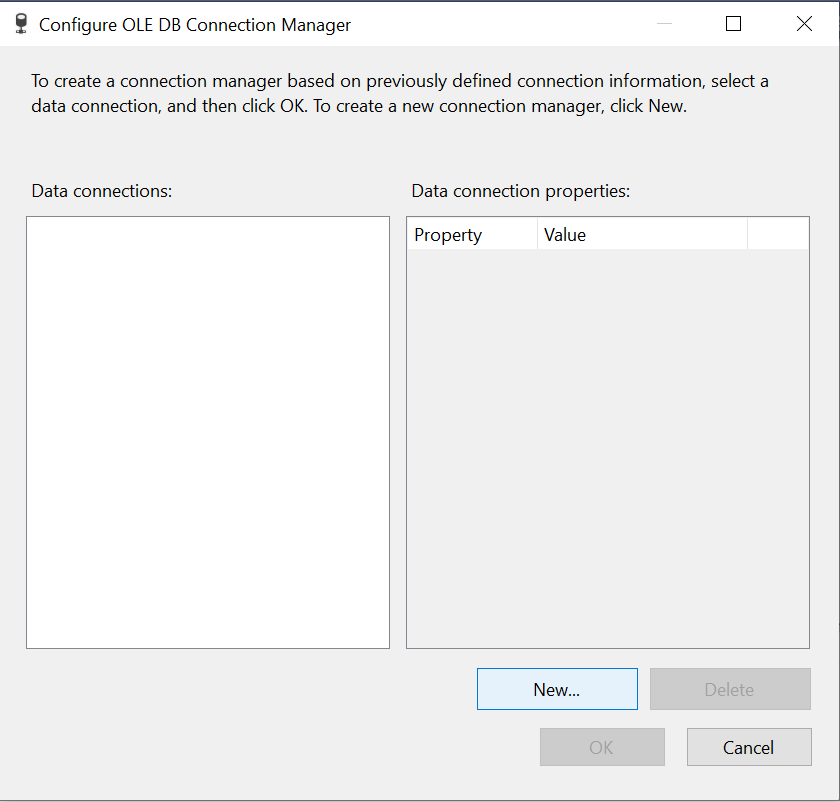
**Connection Manager**.



2. Chọn **OLEDB từ Add SSIS Connection Manager** dialog và click **Add…**



3. **Configure OLE DB Connection Manager** dialog xuất hiện, click **New…**



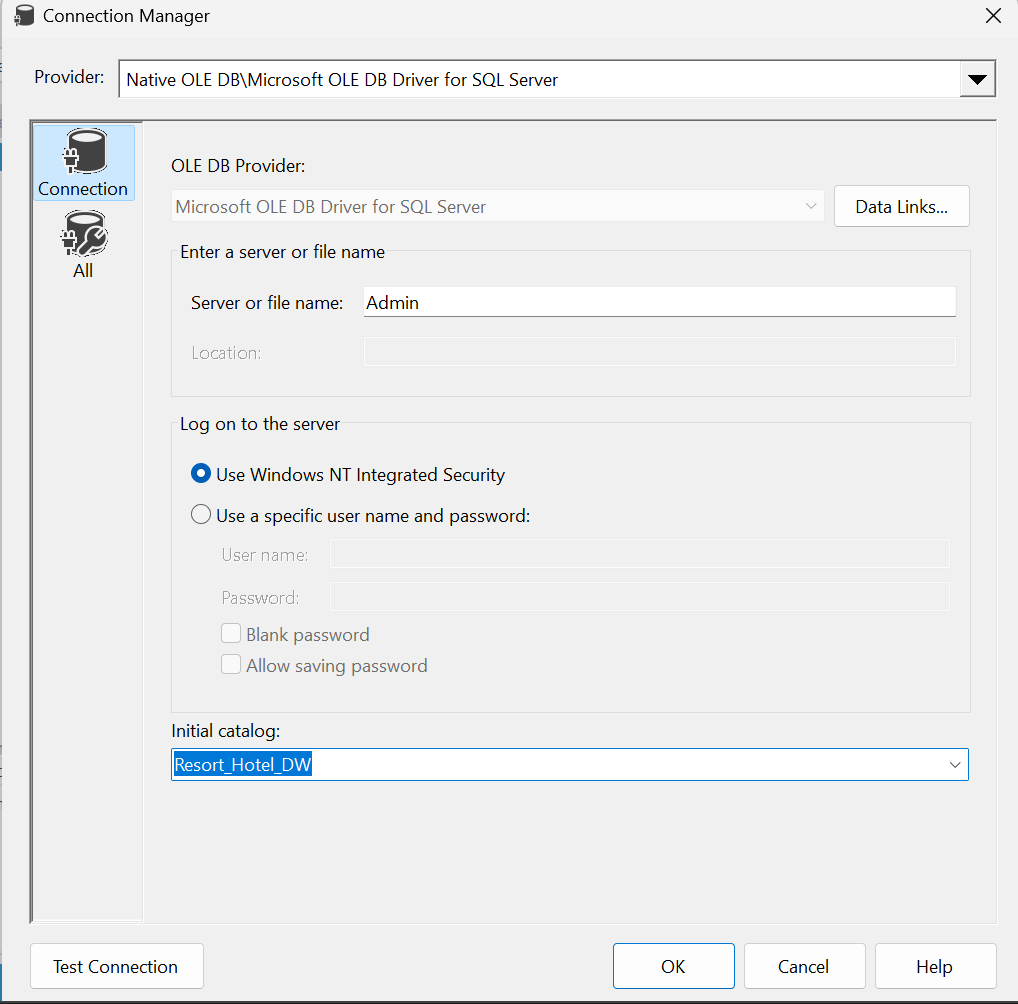
1. Đặt tên cho các thông tin như bên dưới:

Provider: **Microsoft OLE DB Driver for SQL Server**

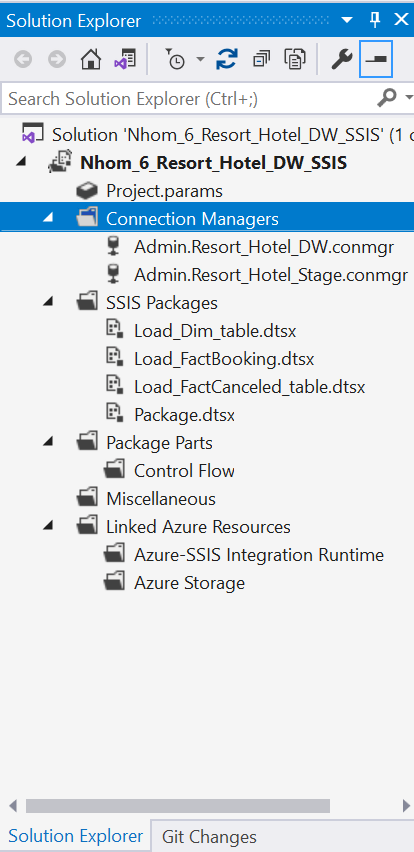
Server Name: **Admin**

Log on to the Server: **Use Windows Authentication**

Connect to a database: **Resort\_Hotel\_DW**

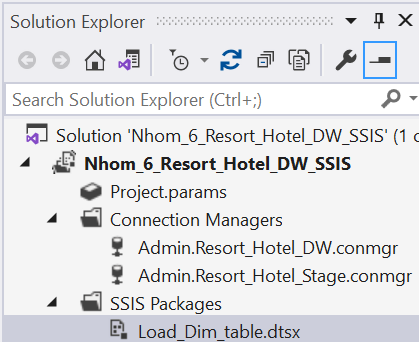


5. Lặp lại các bước trên cho **Resort\_Hotel\_Stage** source database. Sau khi tạo xong **Connection Managers** folder sẽ như bên dưới.

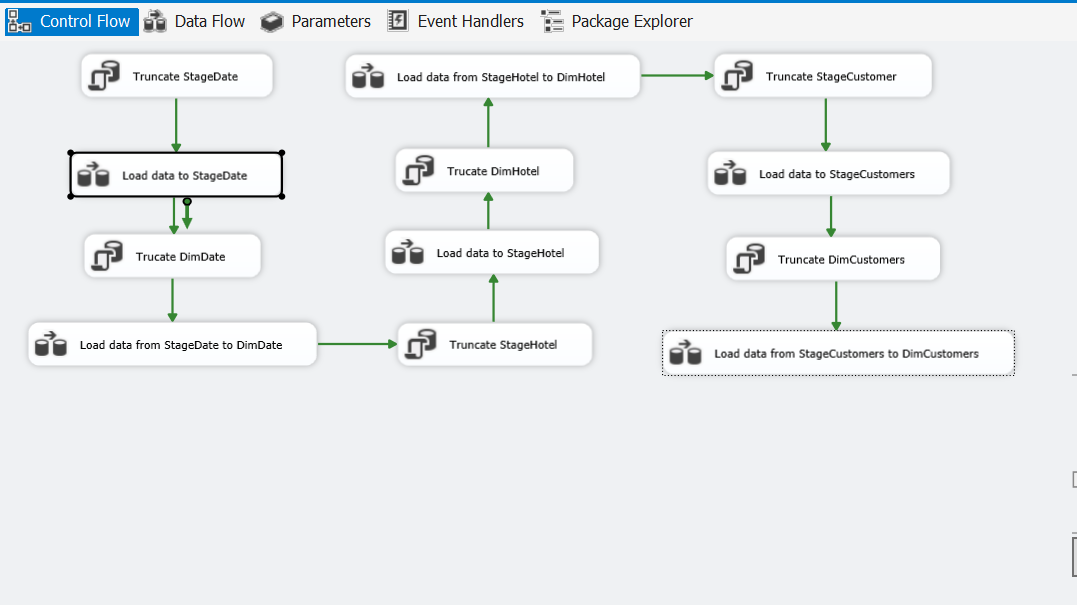


## **3.3. Import Dữ liệu vào các bảng dimension:**

### 3.3.1. **Tạo Package**



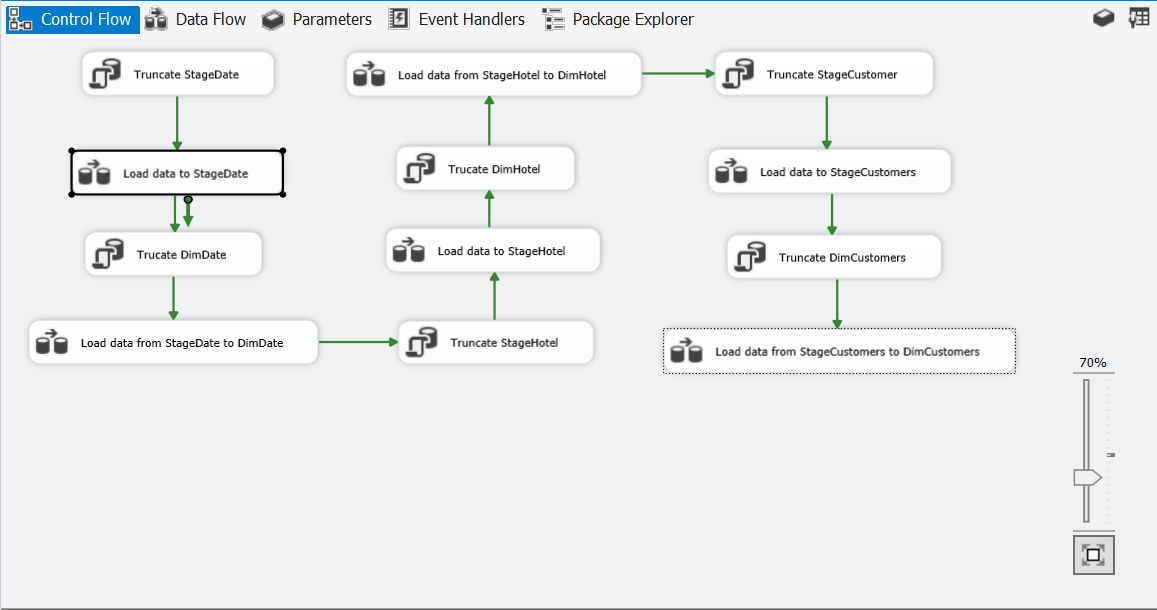
### 3.3.2. Thiết lập Control Flow



Mô tả luồng dữ liệu:

Đầu tiên cần load dữ liệu vào các bảng Stage trước thay vì load dữ liệu trước tiếp vào bảng Dim Mục đích sử lí dữ liệu thô từ excel trước ở các bảng Stage để tránh gặp phải lỗi khi thực hiện load dữ liệu vào bảng Dim.

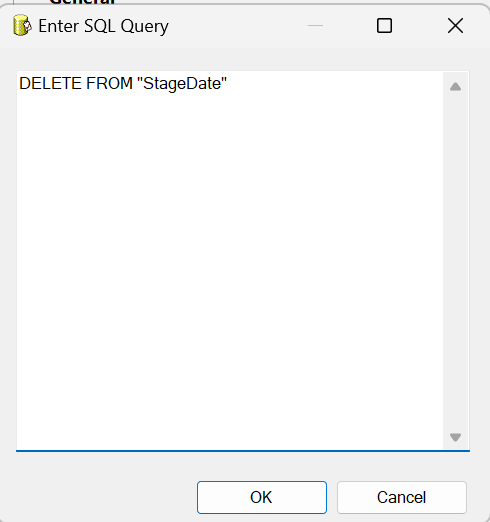
1. Trucate table cho bảng StageDate.
2. Load dữ liệu từ excel vào bảng Stage.
3. Truncate table cho bảng DimDate.
4. Load dữ liệu từ bảng StageDate voà Bảng DimDate.
5. Tiếp đến là lần lược thực hiện các bước 1-4 cho việc load dữ liệu vào DimHotel và DimCustomers.

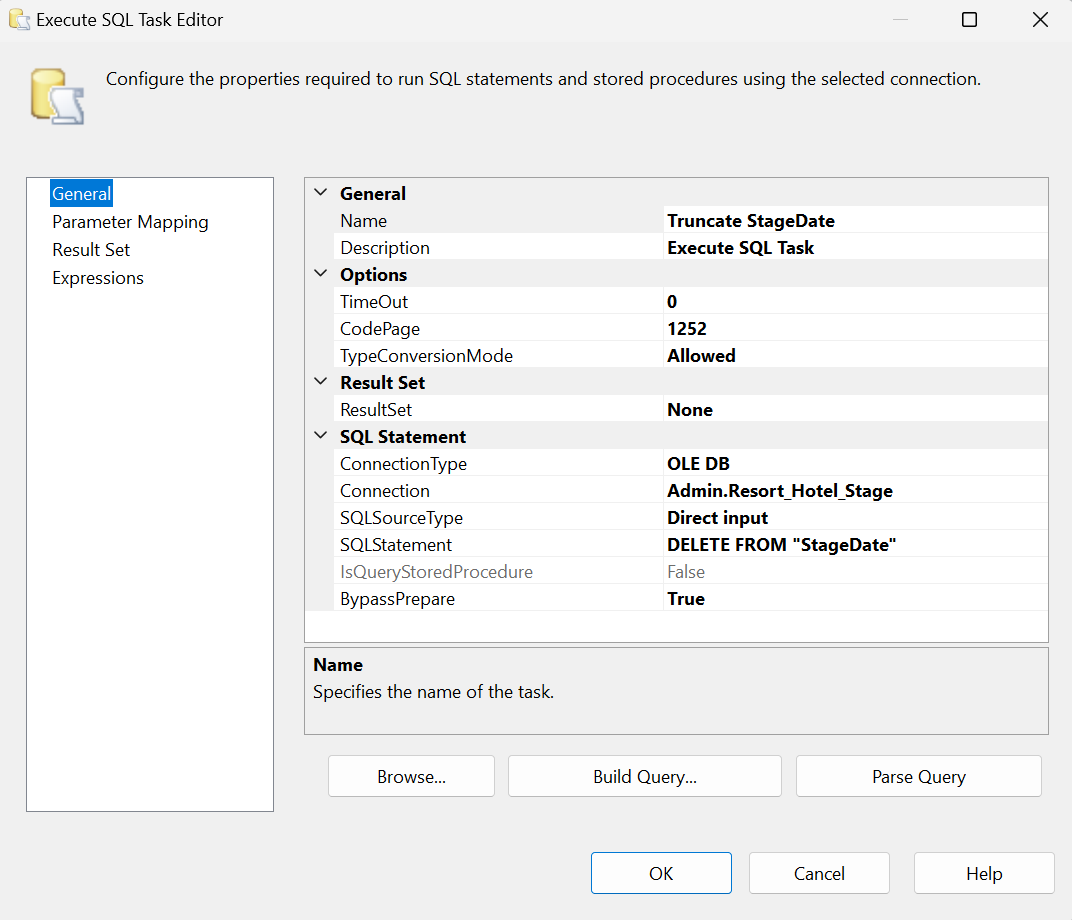


### 3.3.3. Thiết lập Truncate table cho các bảng

1. Thiết lập kết nối đến bảng cơ sở dữ liệu chứa bảng cần thực hiện. Ở đây cần thực hiện Truncate trên bảng StageDate trong cơ sở dữ liệu Resort\_Hotel\_Stage.

2. Thiết lập SQLStagement, cần thực hiện trên bảng nào thì nhập tên bảng đó vào.

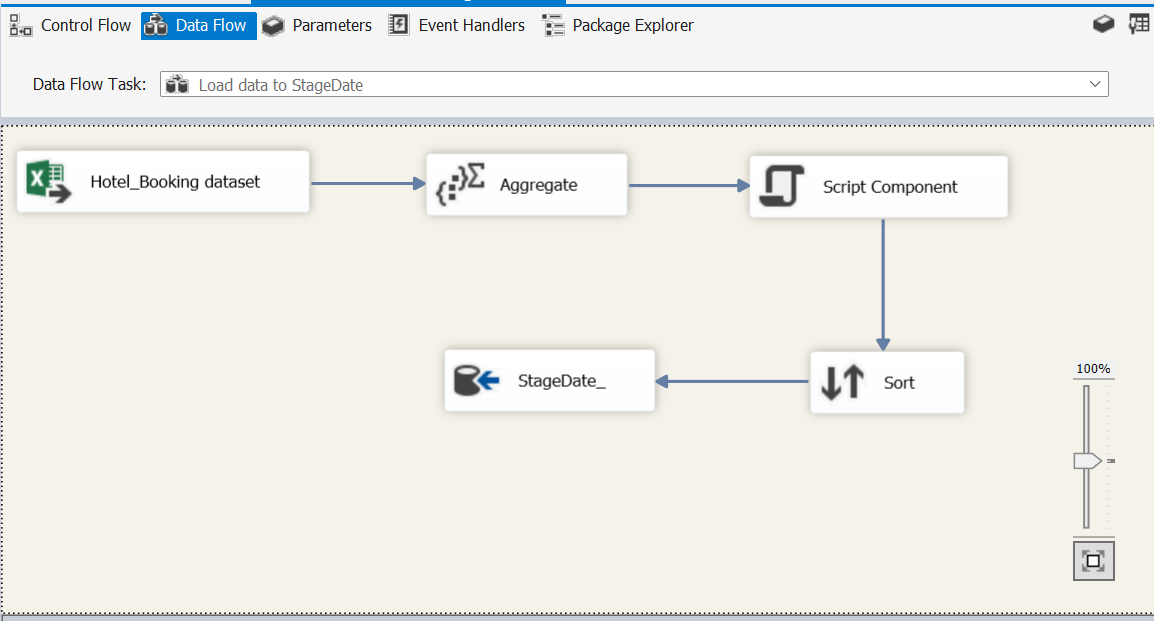




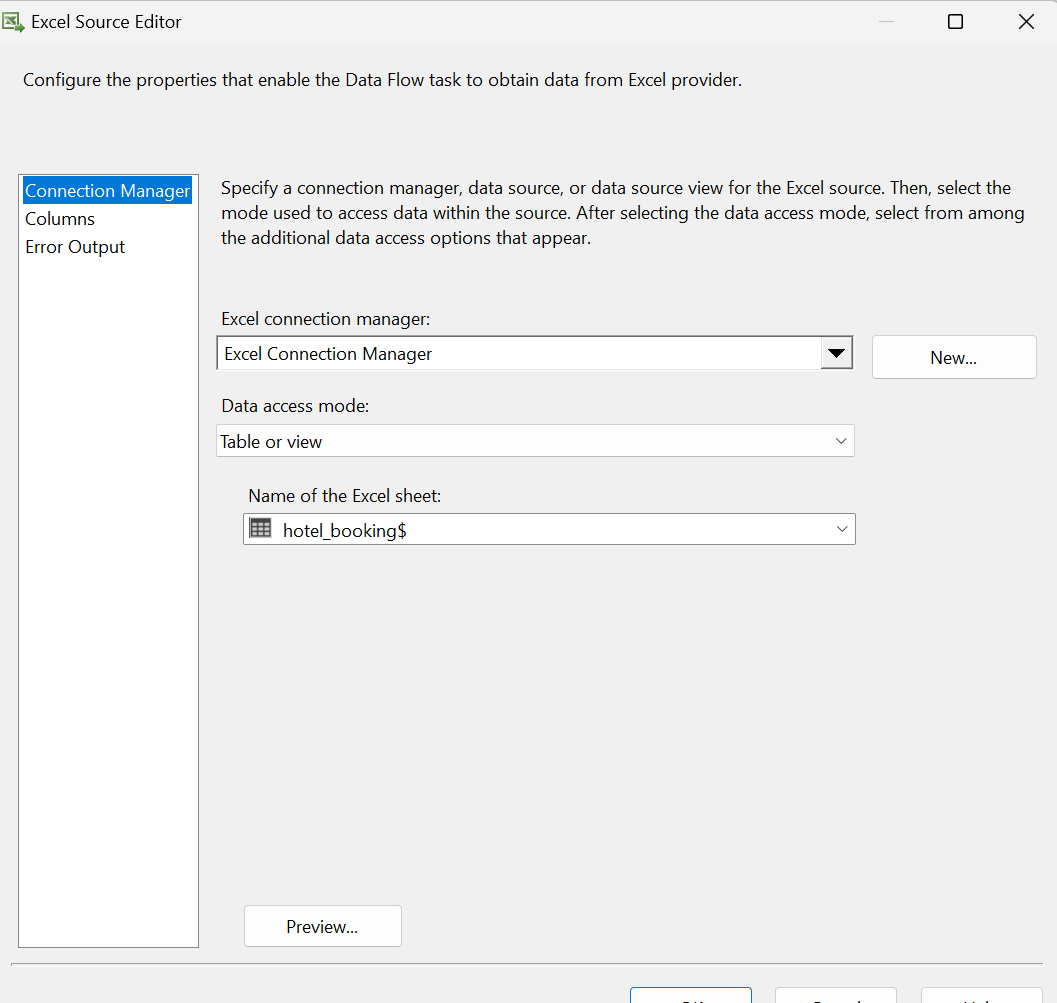
Thực hiện thiết lập cho các bảng tương tự.

3.3.4. Load dữ liệu từ Source vào Stage.

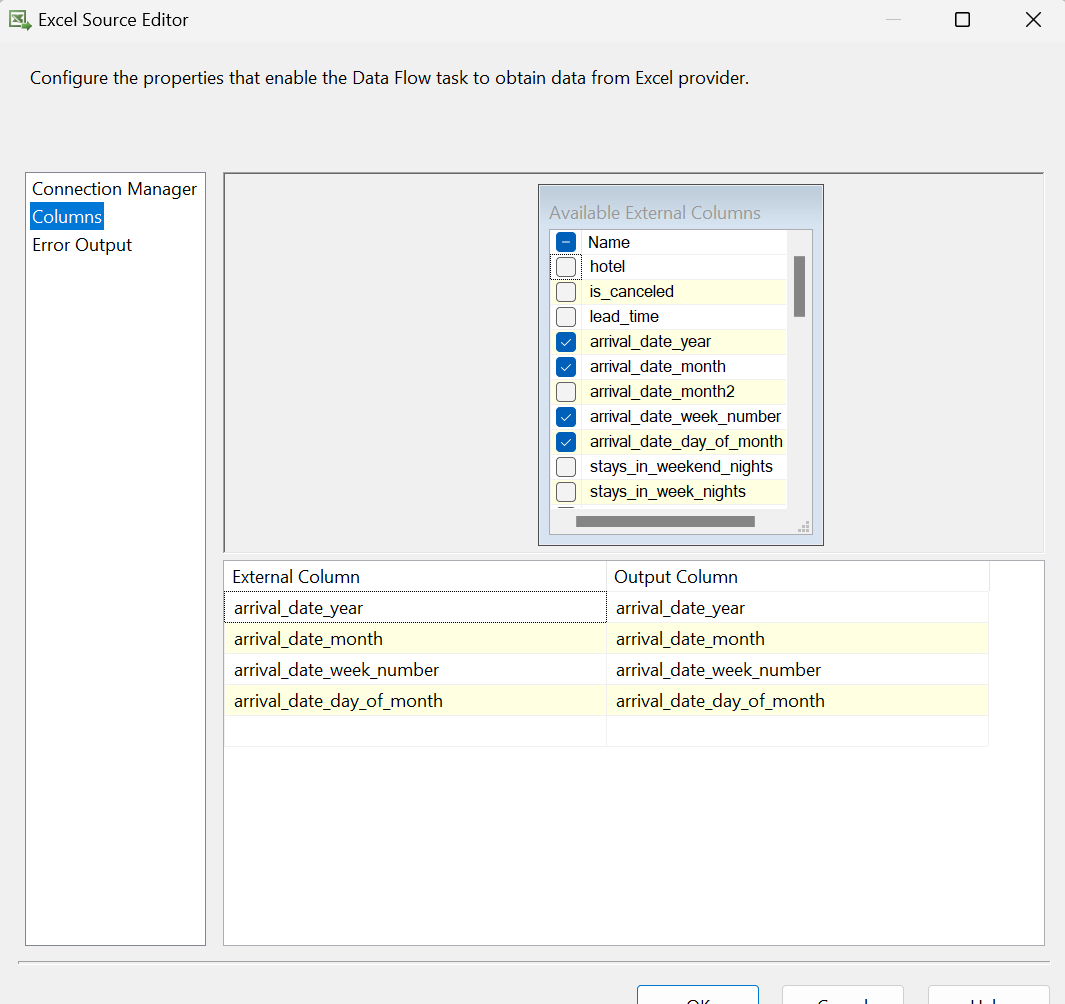
**StageDate:**



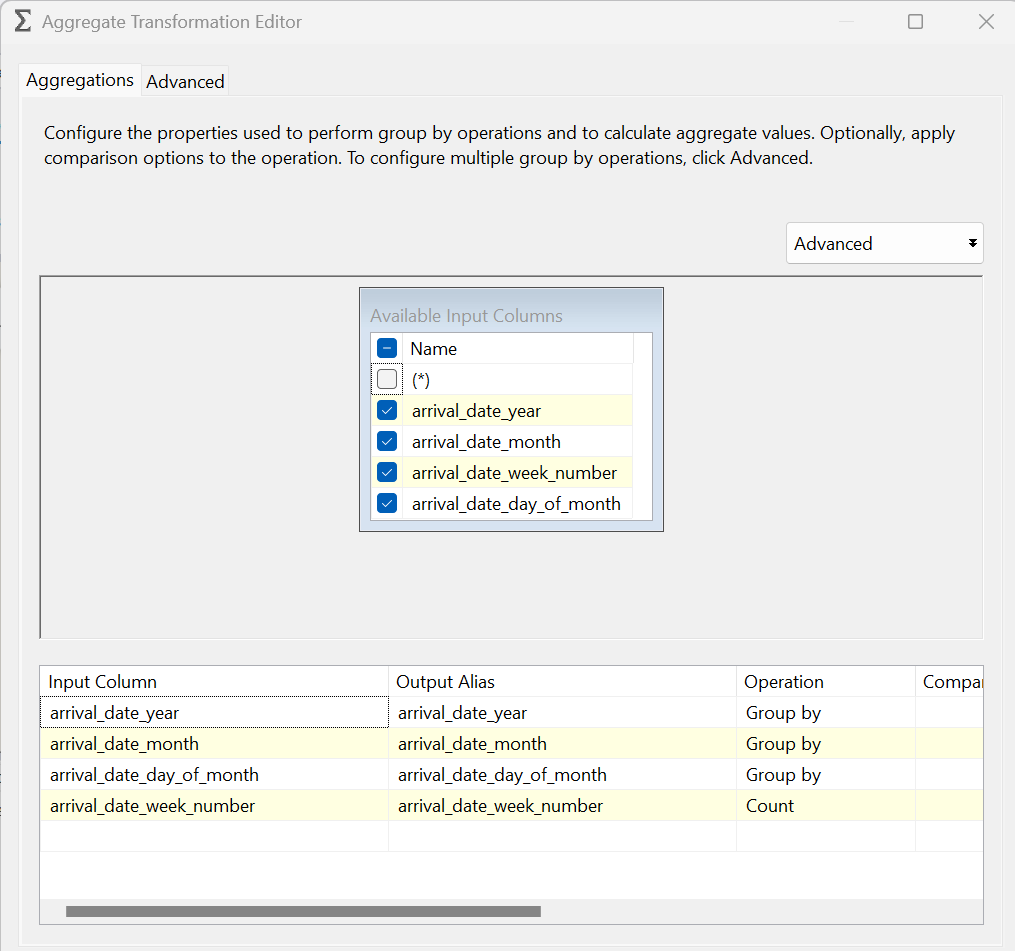
1. Chọn Data Source:



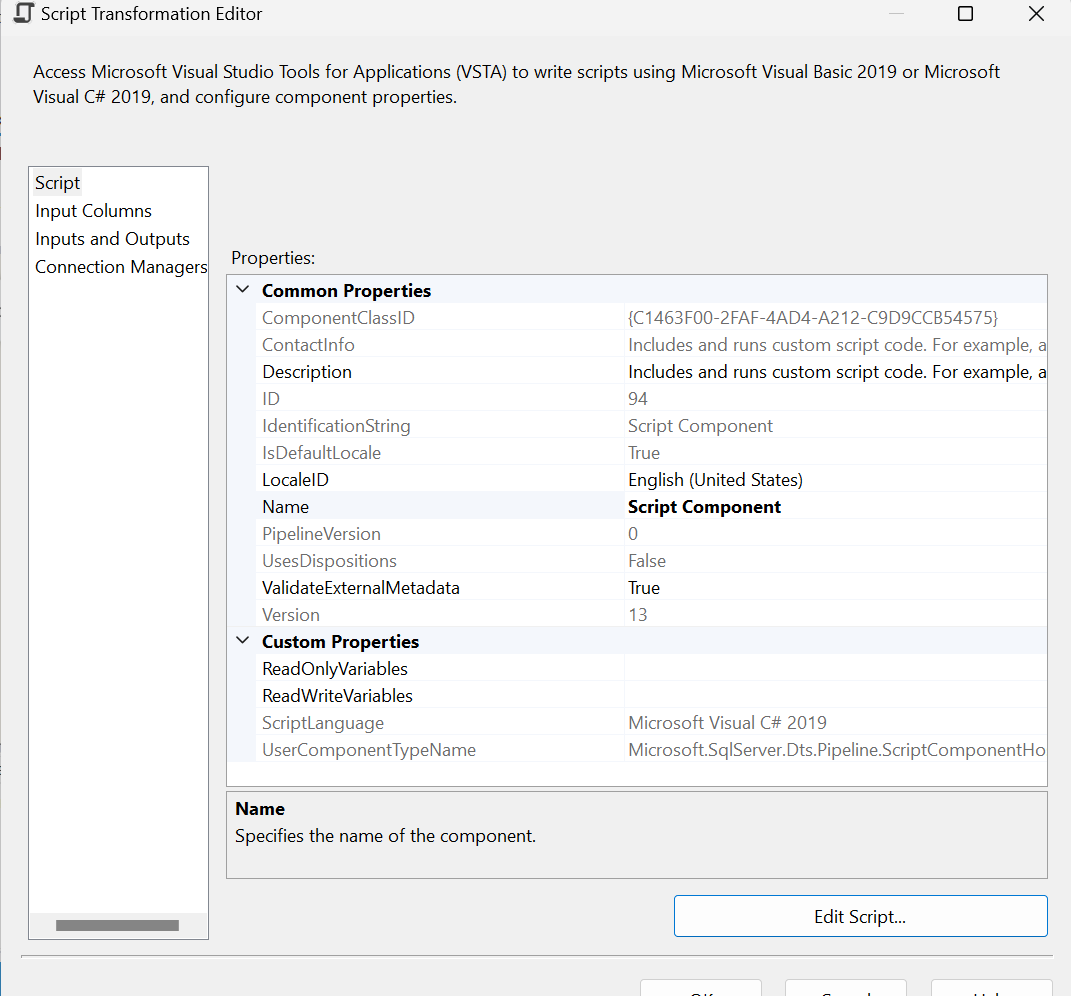
1. Chọn các trường dữ liệu cần cho bảng StageDate;



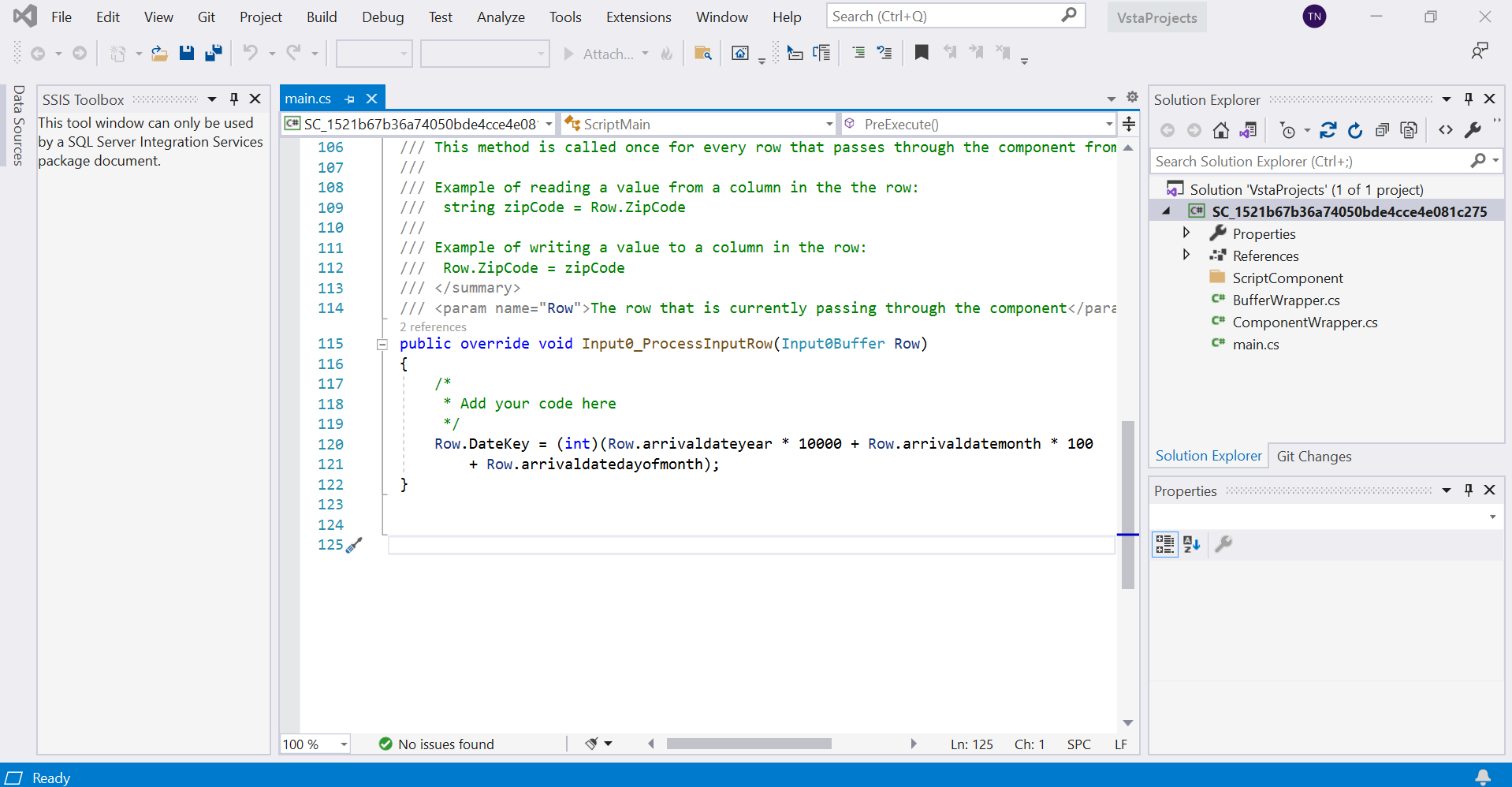
1. Thực hiện Group dữ liệu để tránh dữ liệu trùng lập.



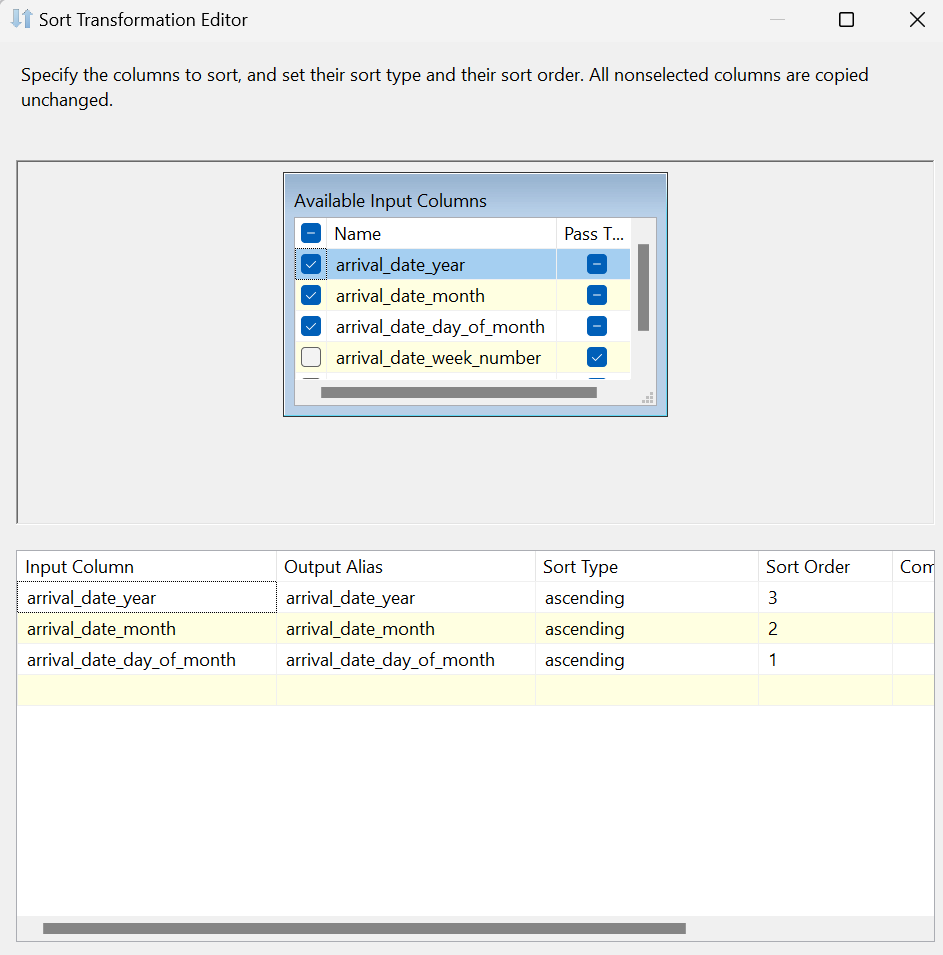
1. Tạo BusinessKey là DateKey:



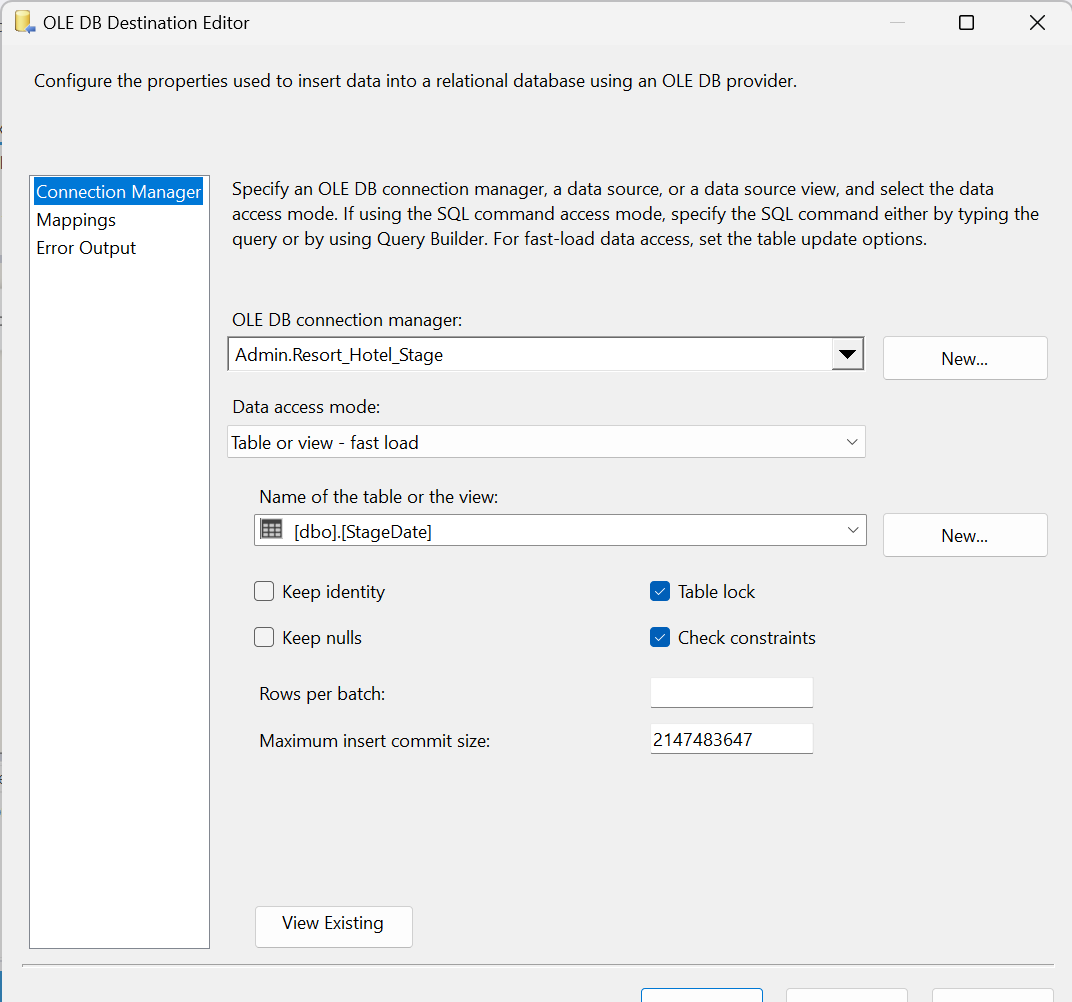
* 1. Chọn Edit Scrip… để thực hiện viết câu lệnh tạo DataKey.

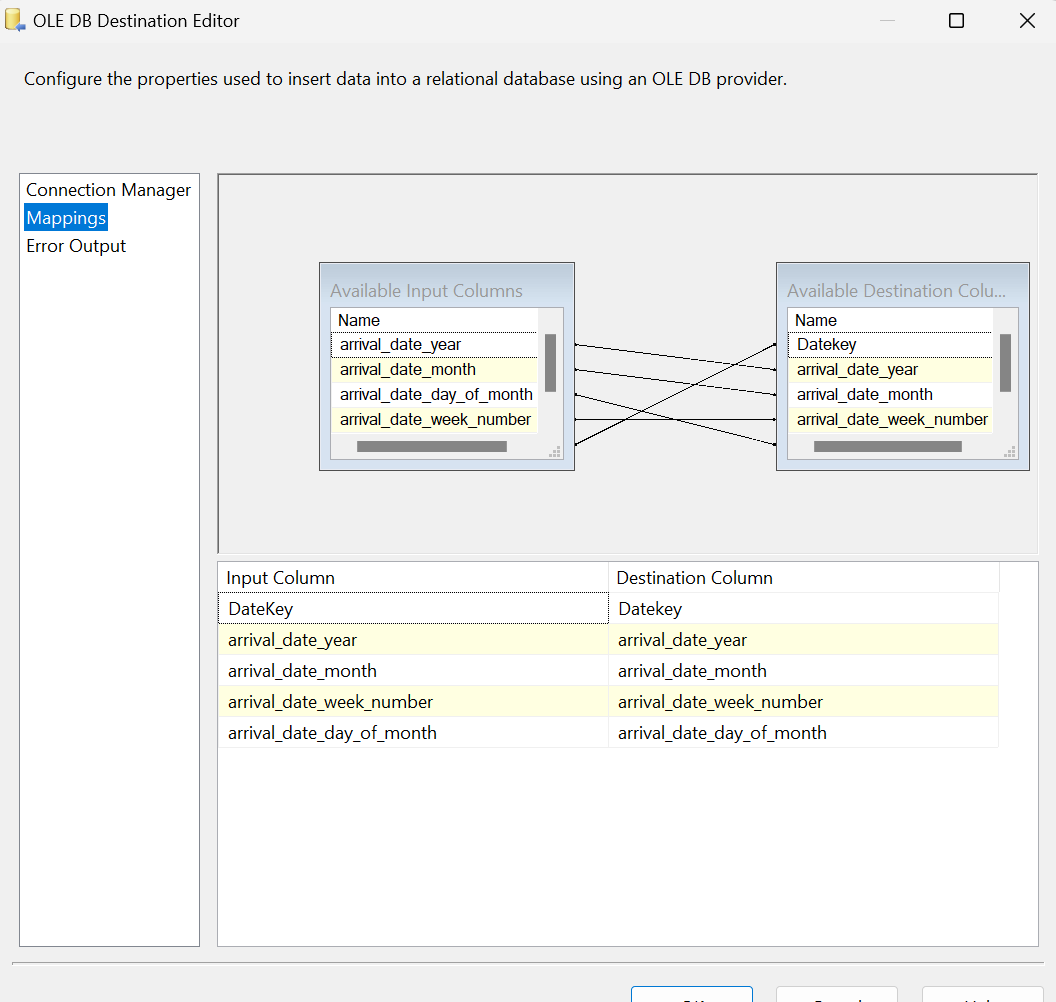


1. Sắp xếp dữ liệu.

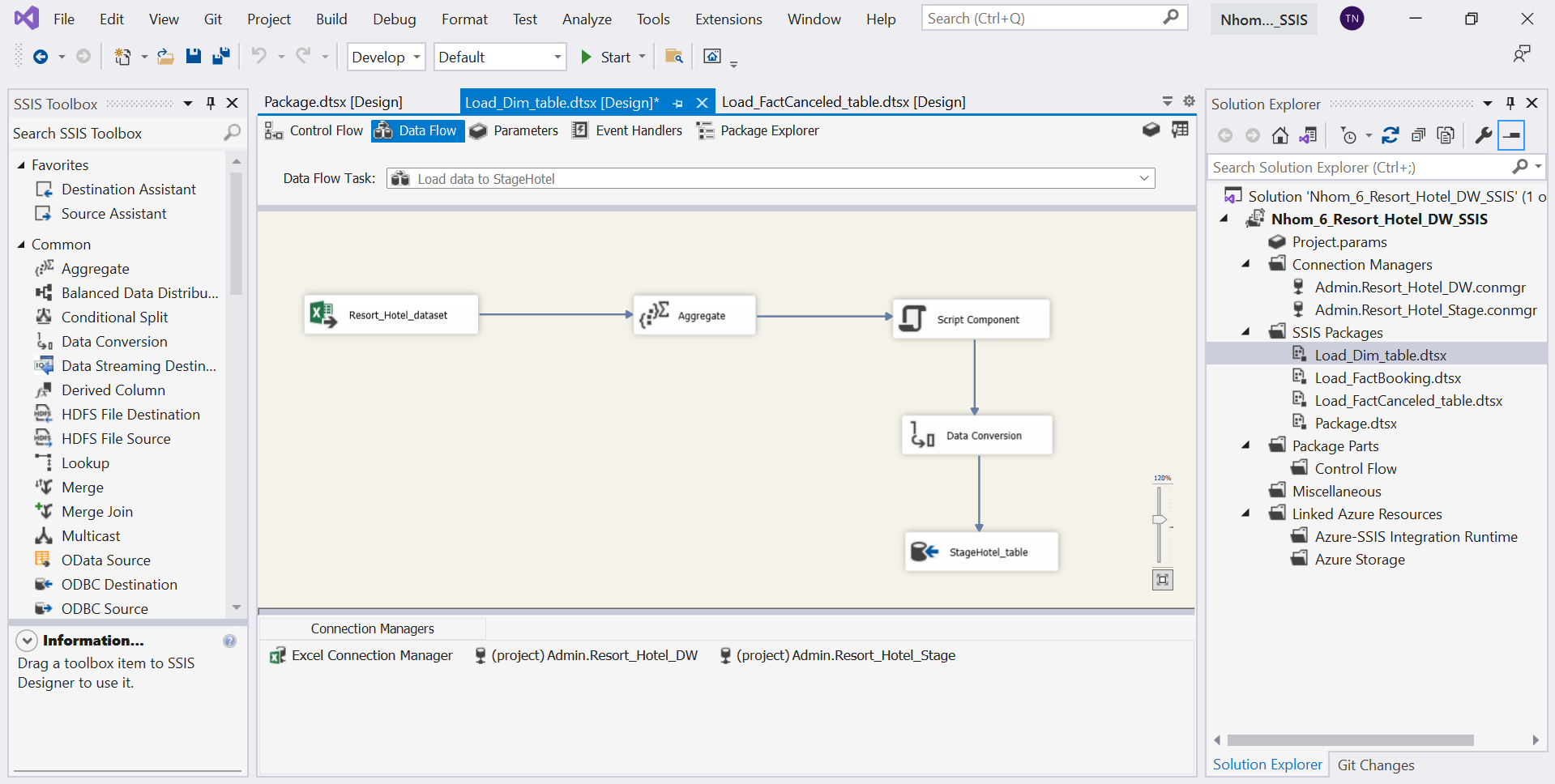


1. Load dữ liệu vào StageDate.

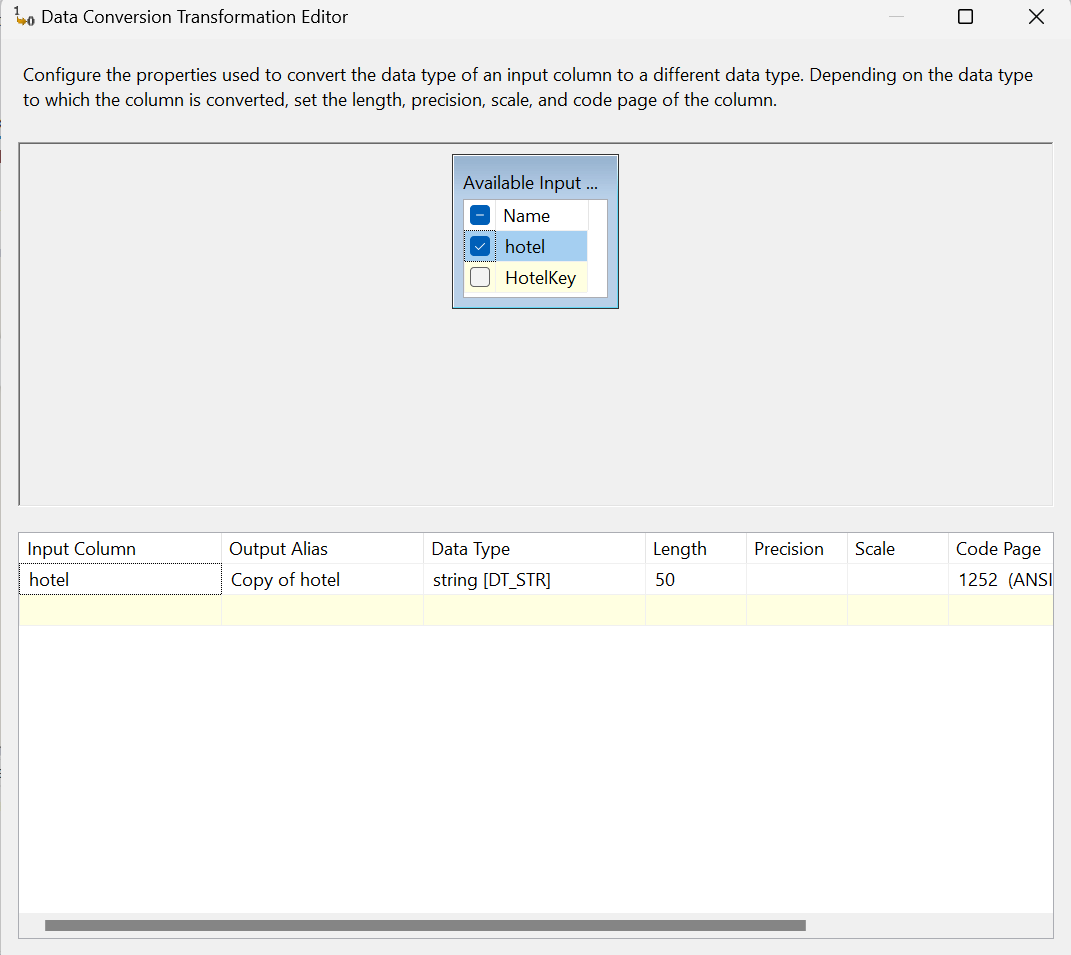




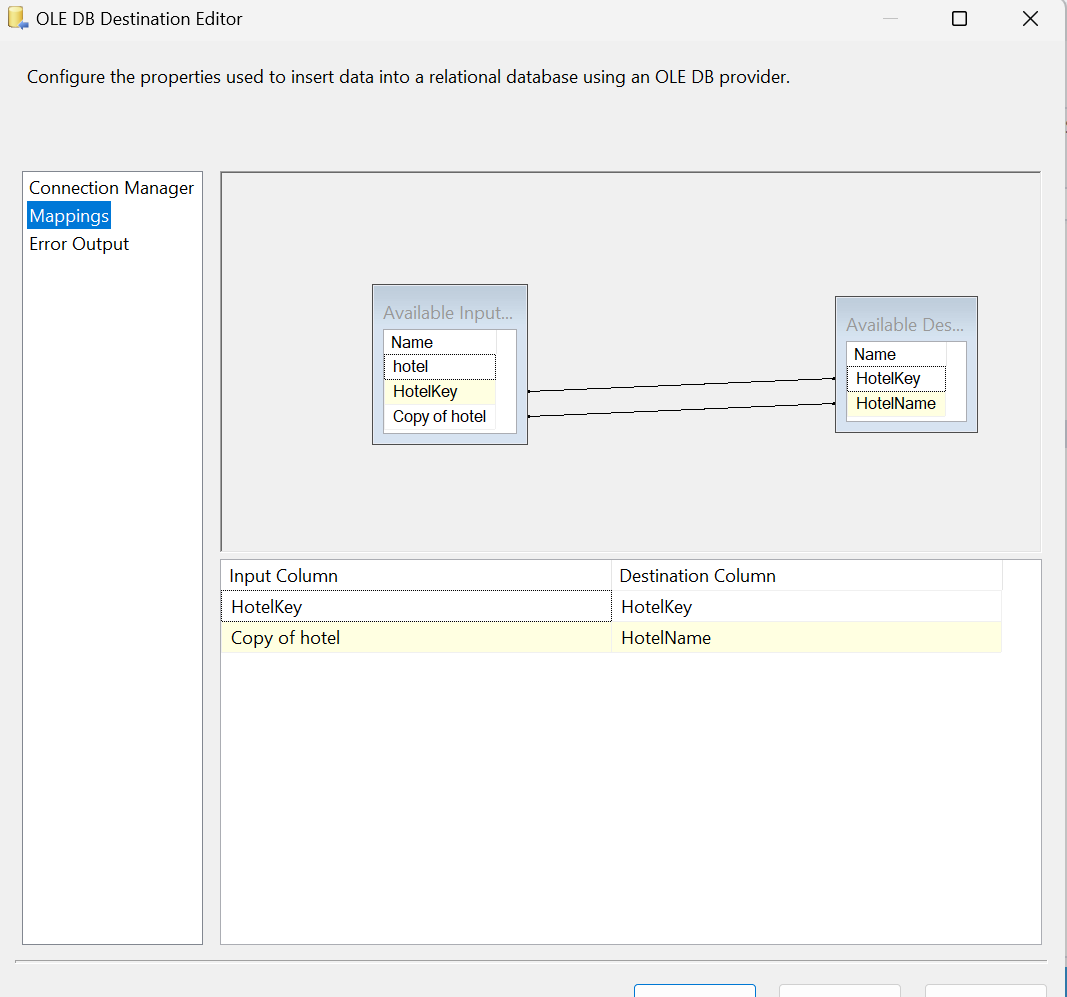
**StageHotel:**



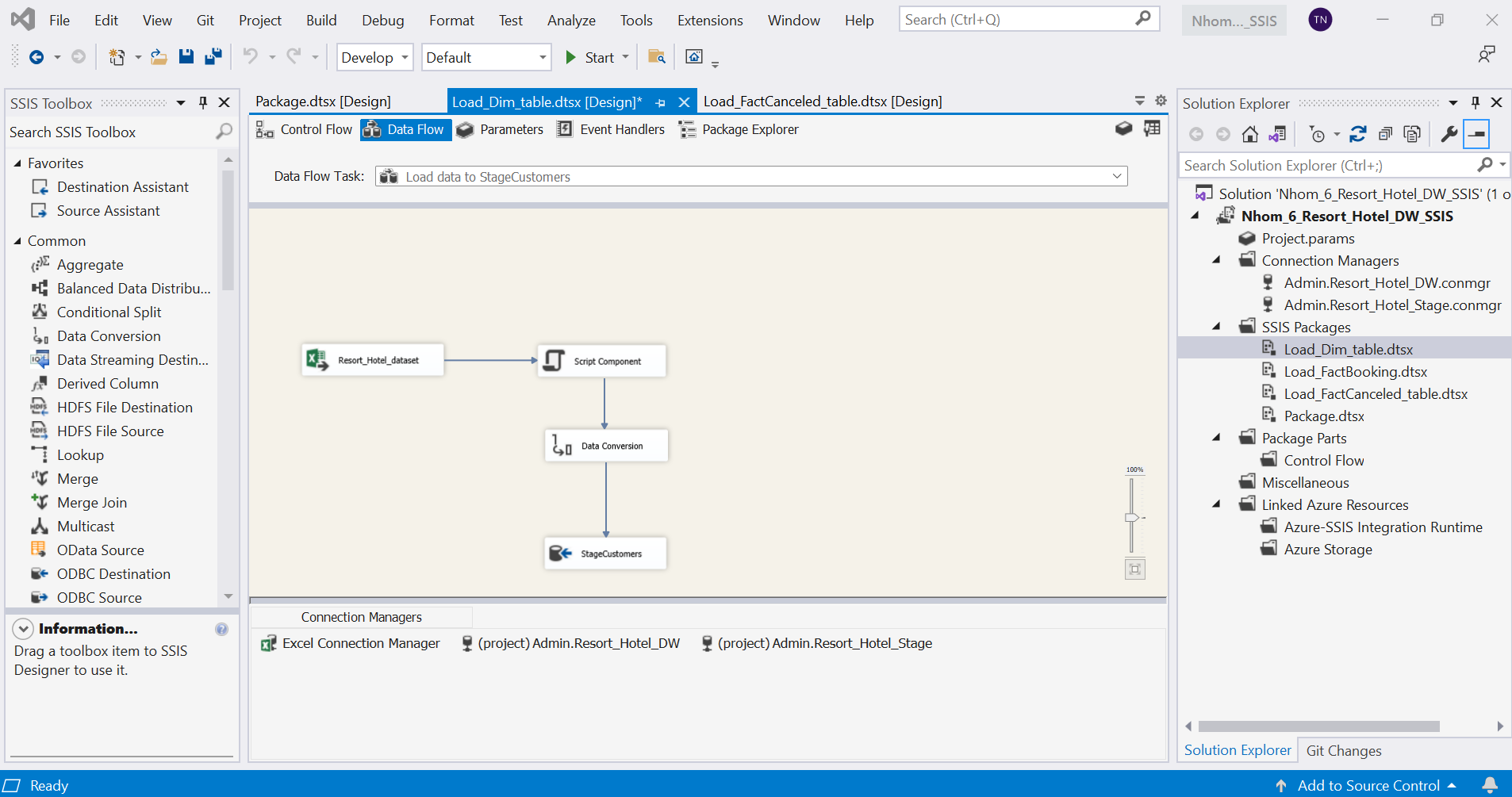
Các bược thực hiện tưng tự như StageDate, tuy nhiên ở StageHotel cần ép dữ liệu của HotelName của excel từ nchar sang String cho phù hợp với kiểu dữ liệu của HotelName trong StageHotel.



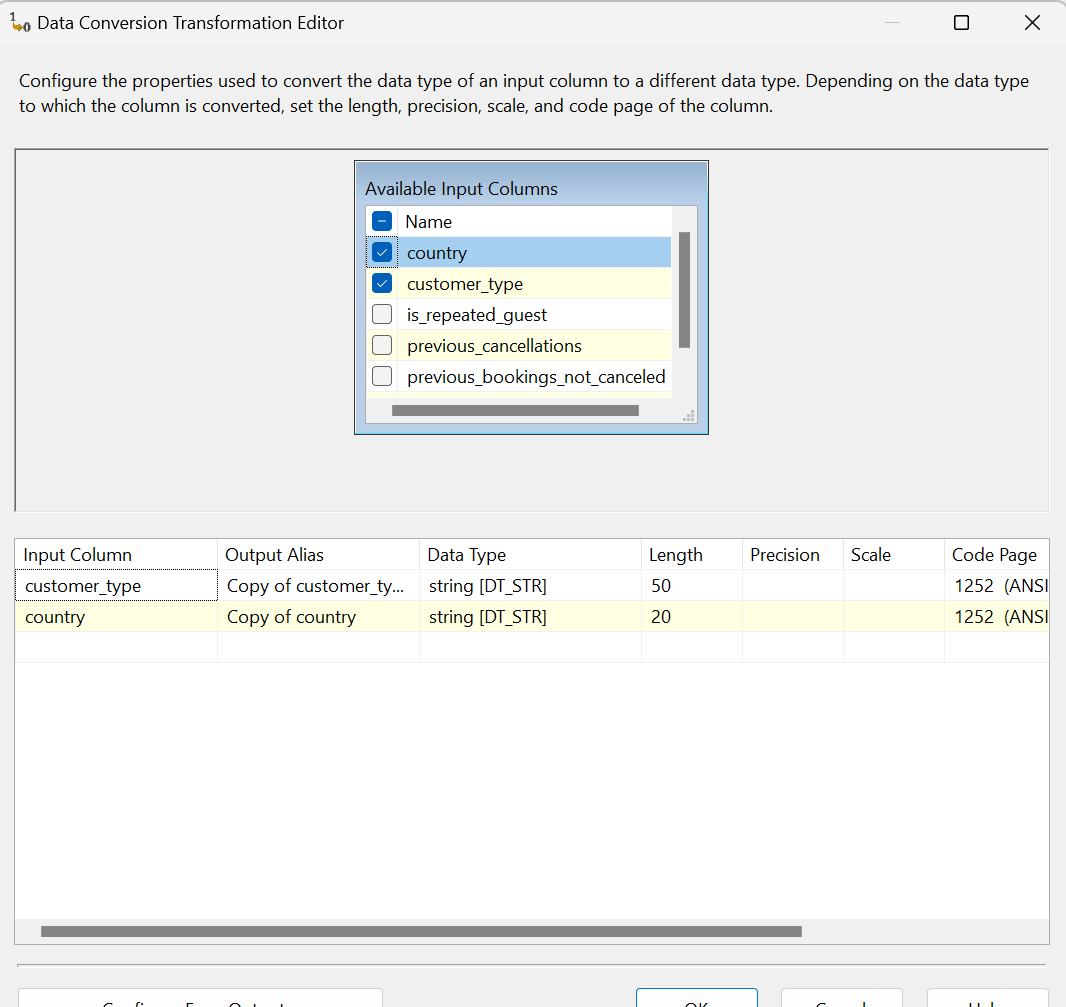
Load dữ liệu vào StageHotel.



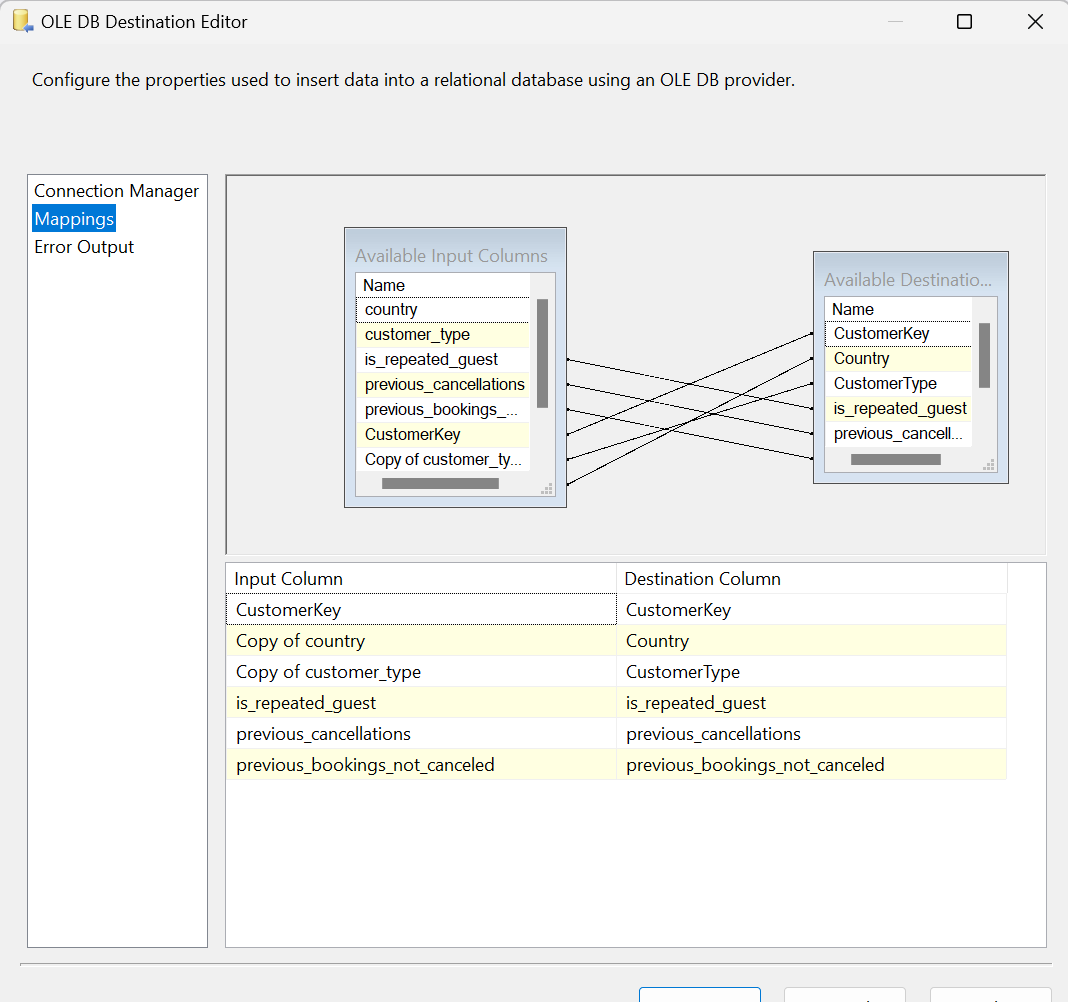
**StageCustomers:**



Ép dữ liệu .

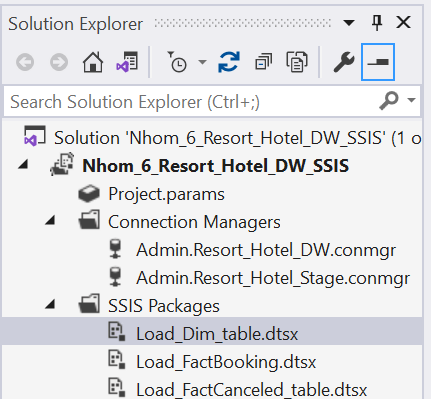


Load dữ liệu vào StageCustomers.



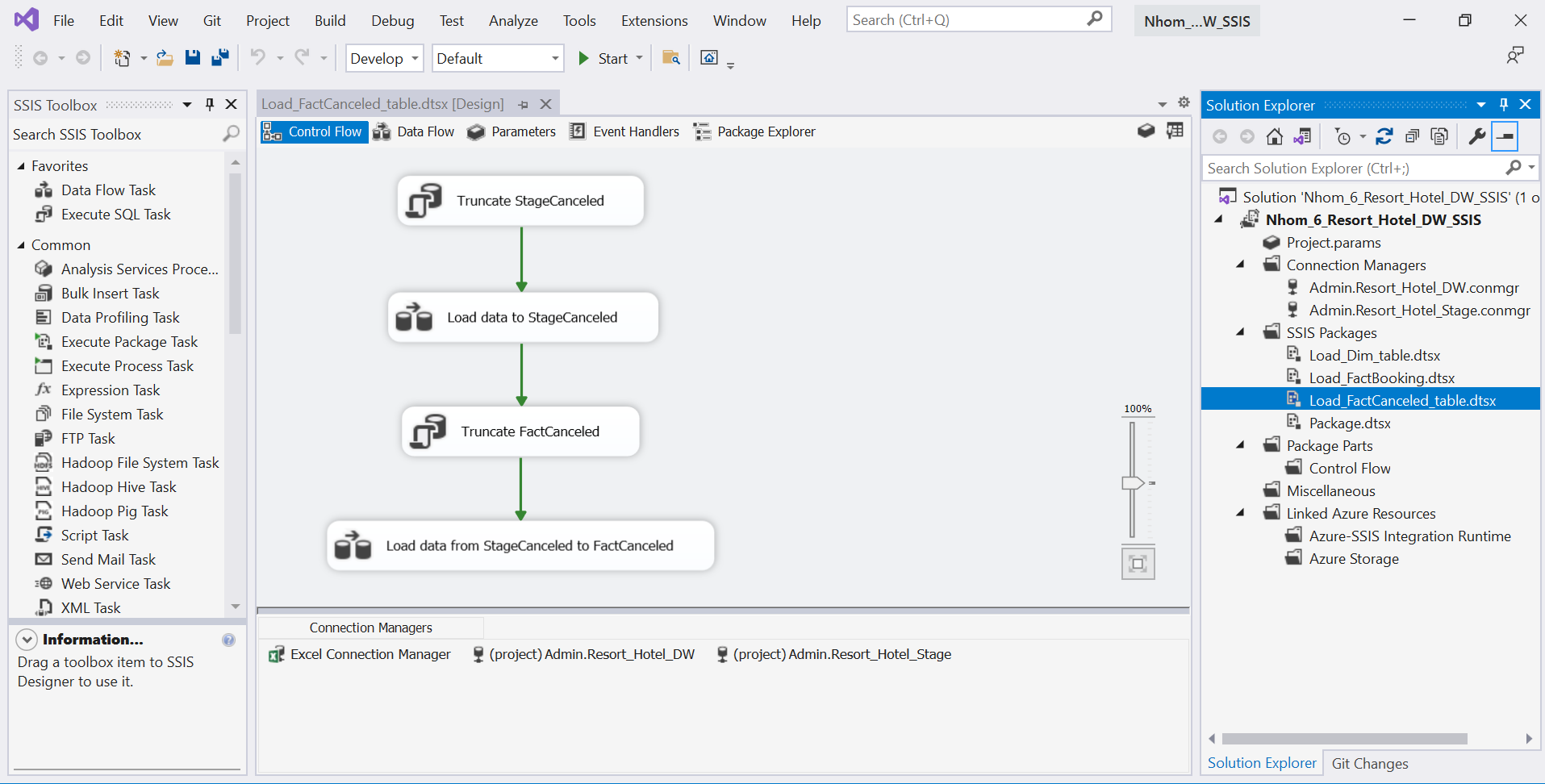
## 3.4. Import dữ liệu vào bảng các bảng fact

### 3.4.1. **Tạo Package**

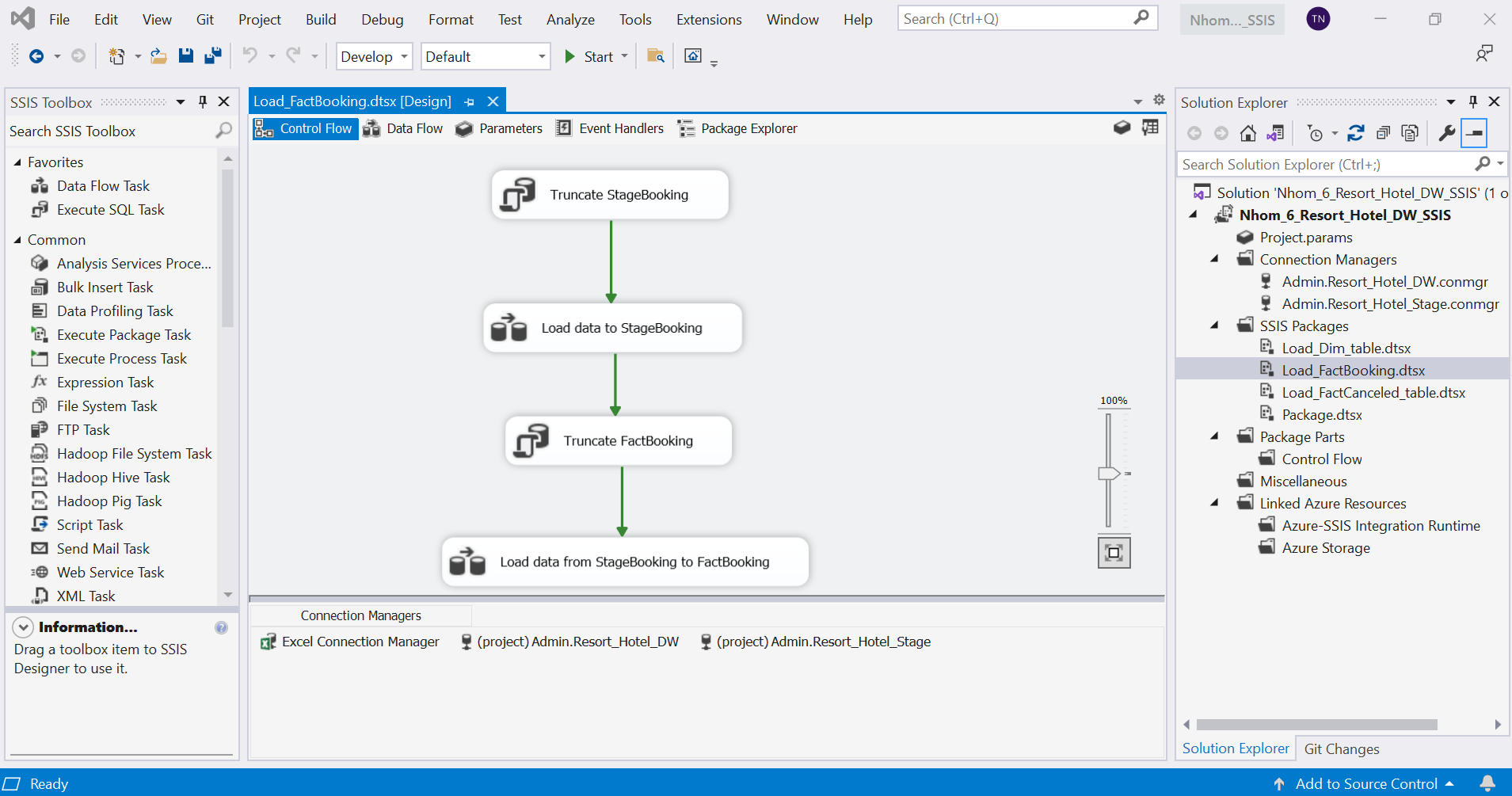


### 3.4.2. Thiết lập Control Flow

**FactCanceled.**



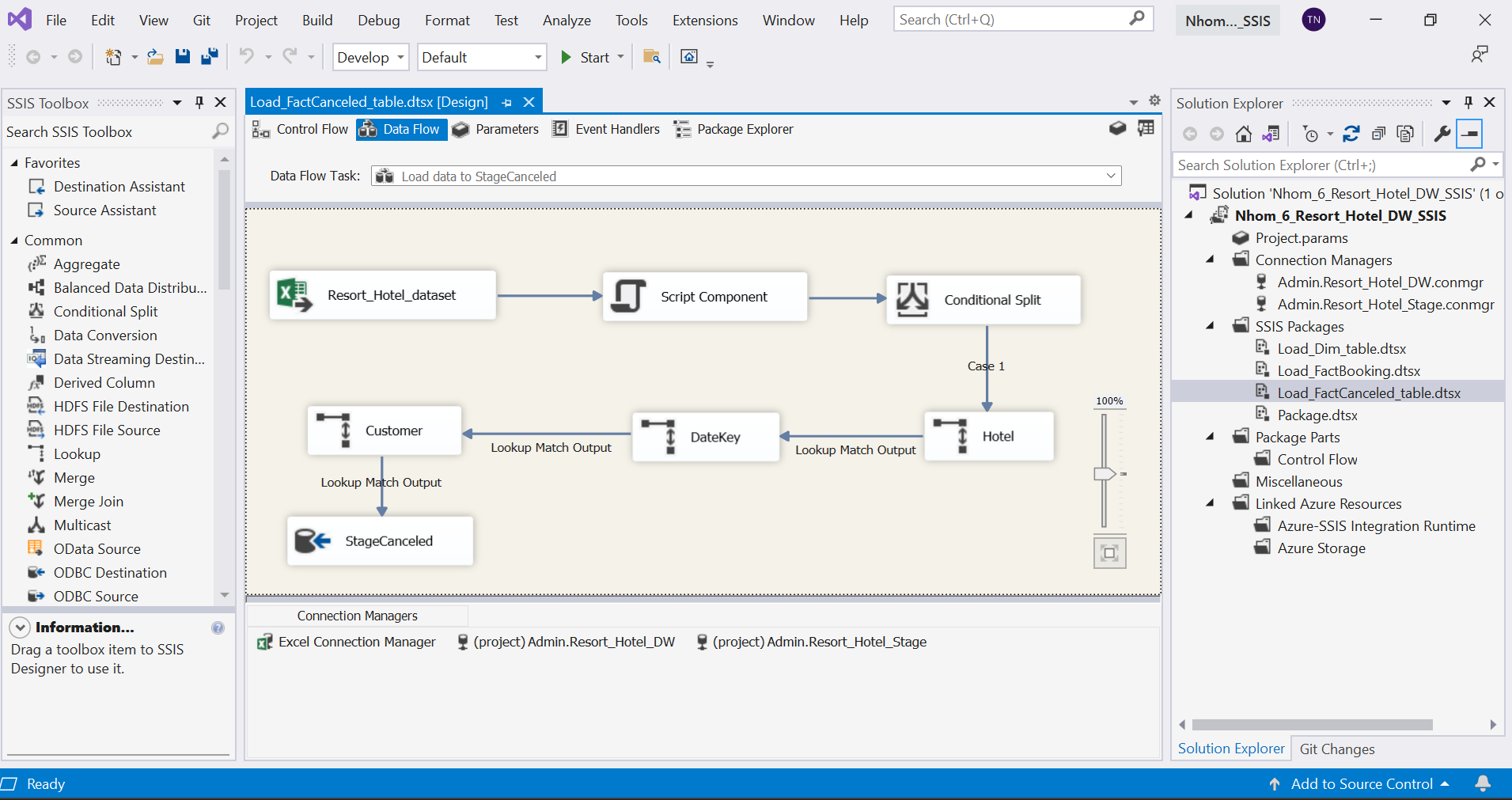
**FactBooking.**



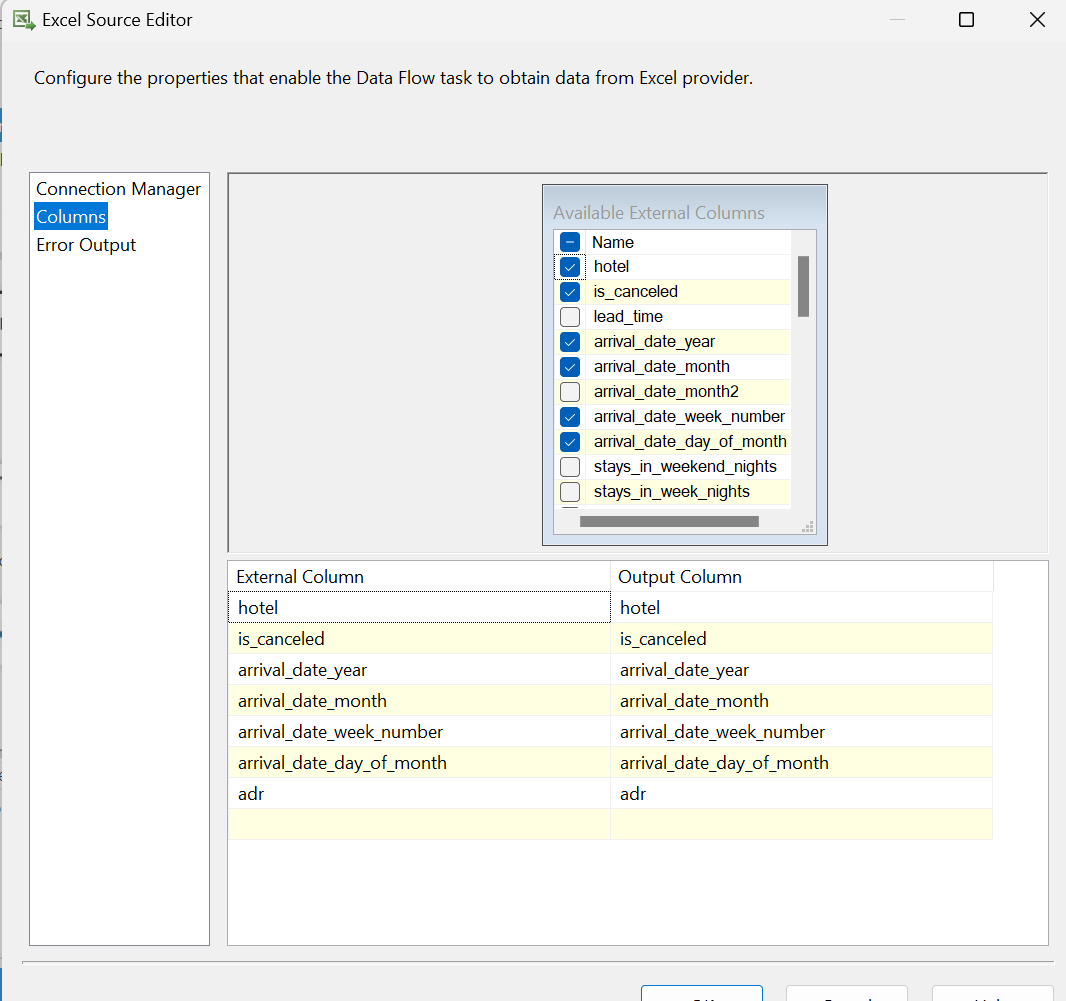
**3.4.3.Load dữ liệu tự excel vào Stage và từ Stage vào Fact**

**FactCanceled.**

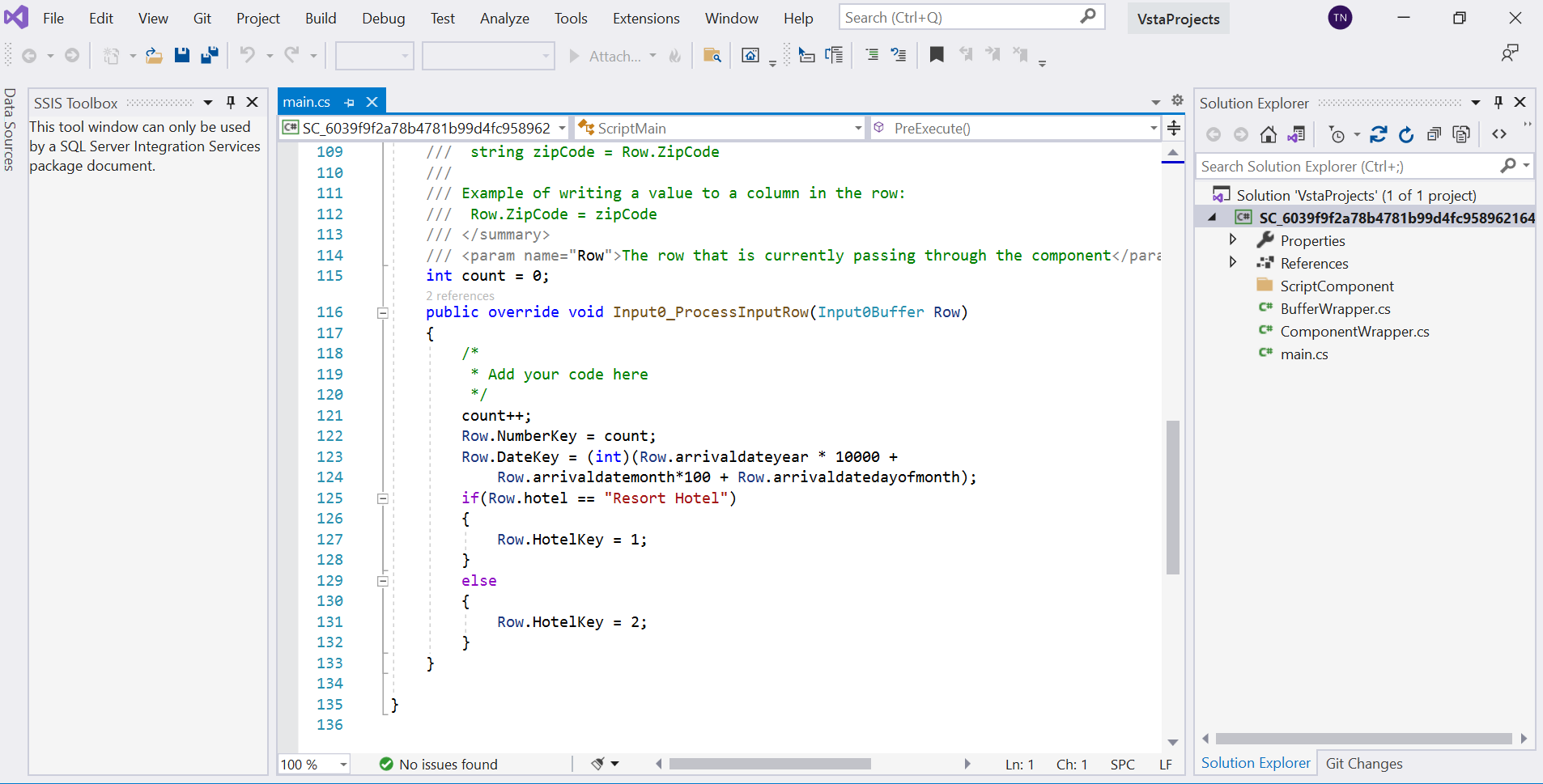
Load dữ liệu từ excel vào StageCanceled.



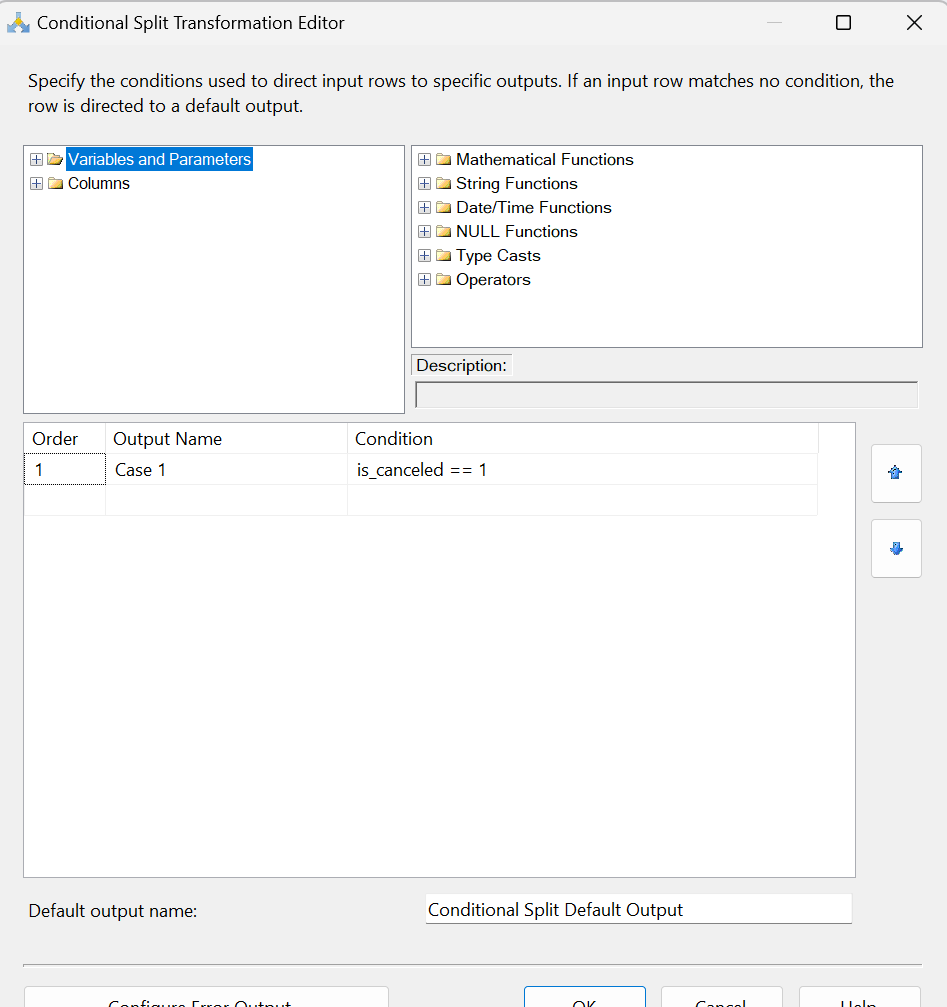
Chọn trường dữ liệu cần thiết.



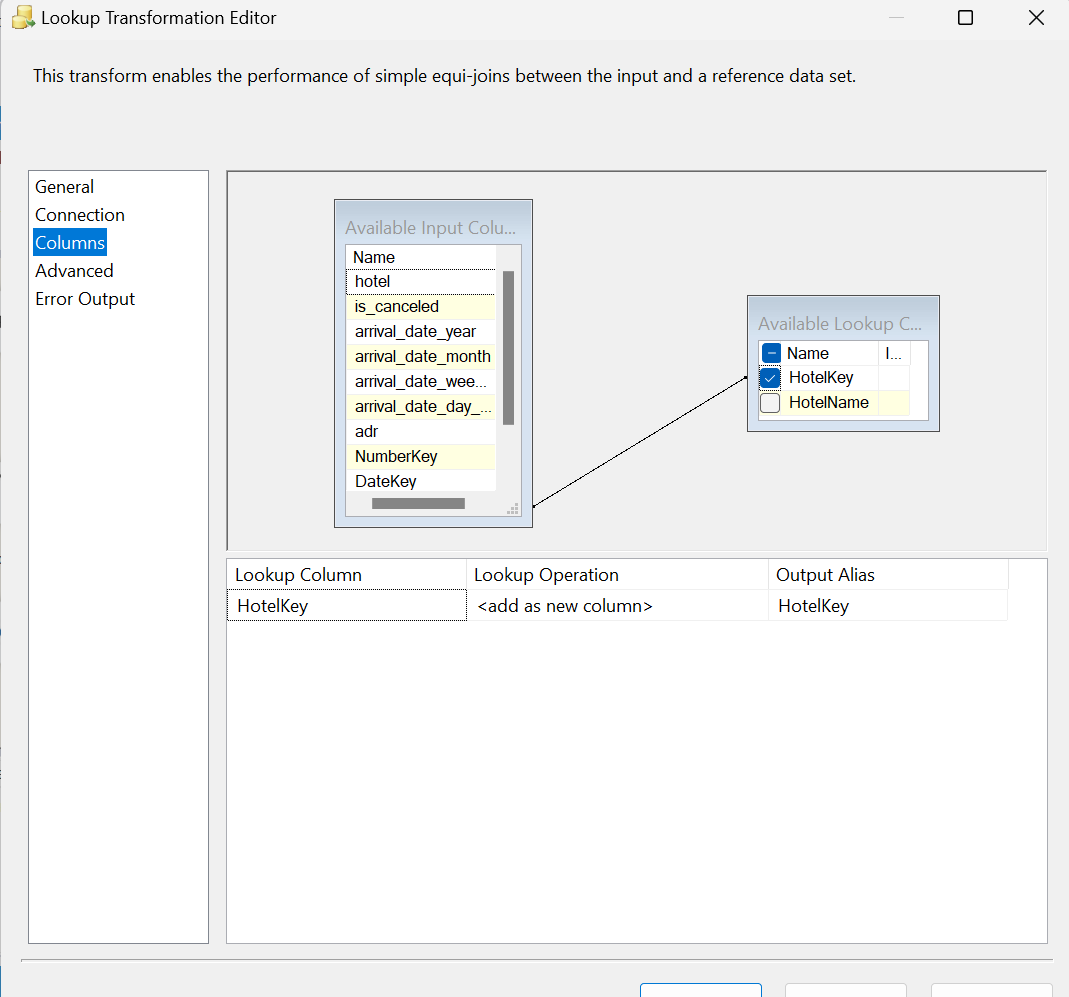
Thực hiện tạo businessKey.

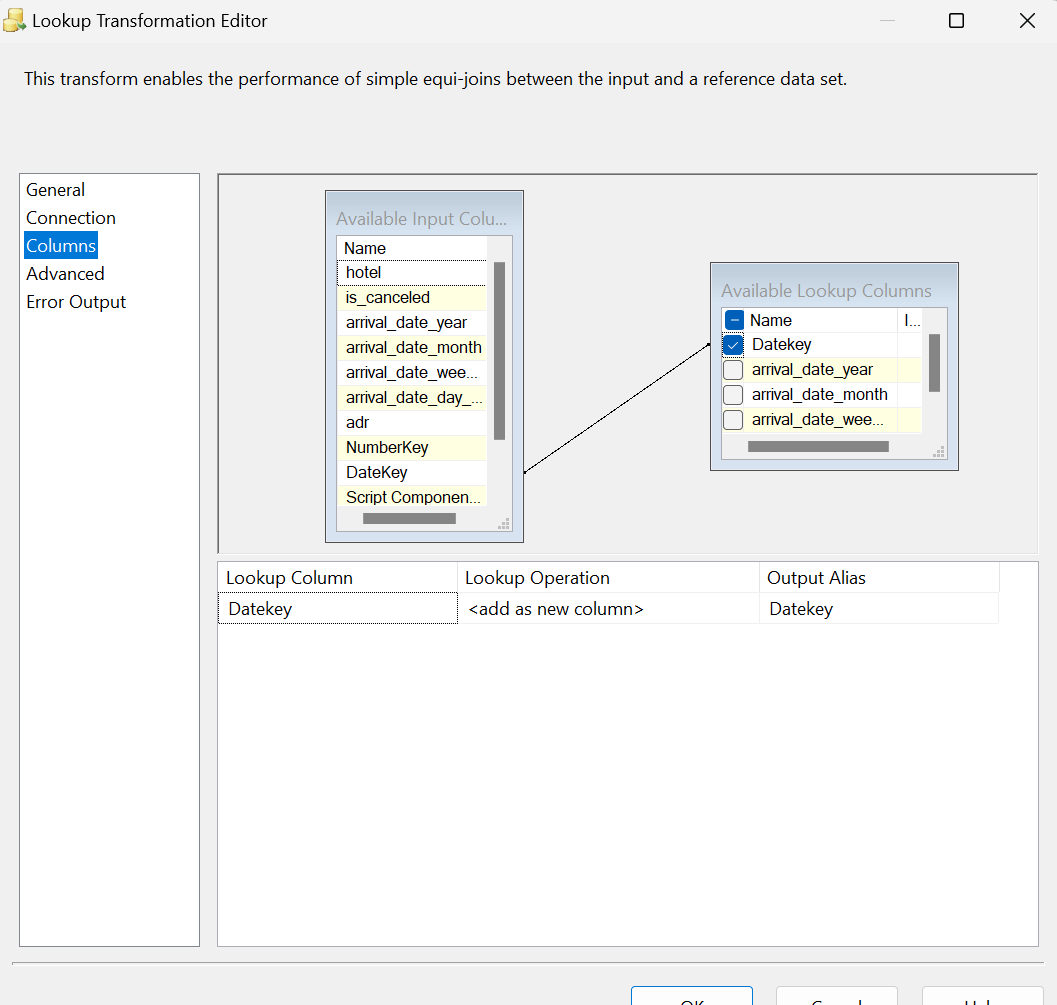


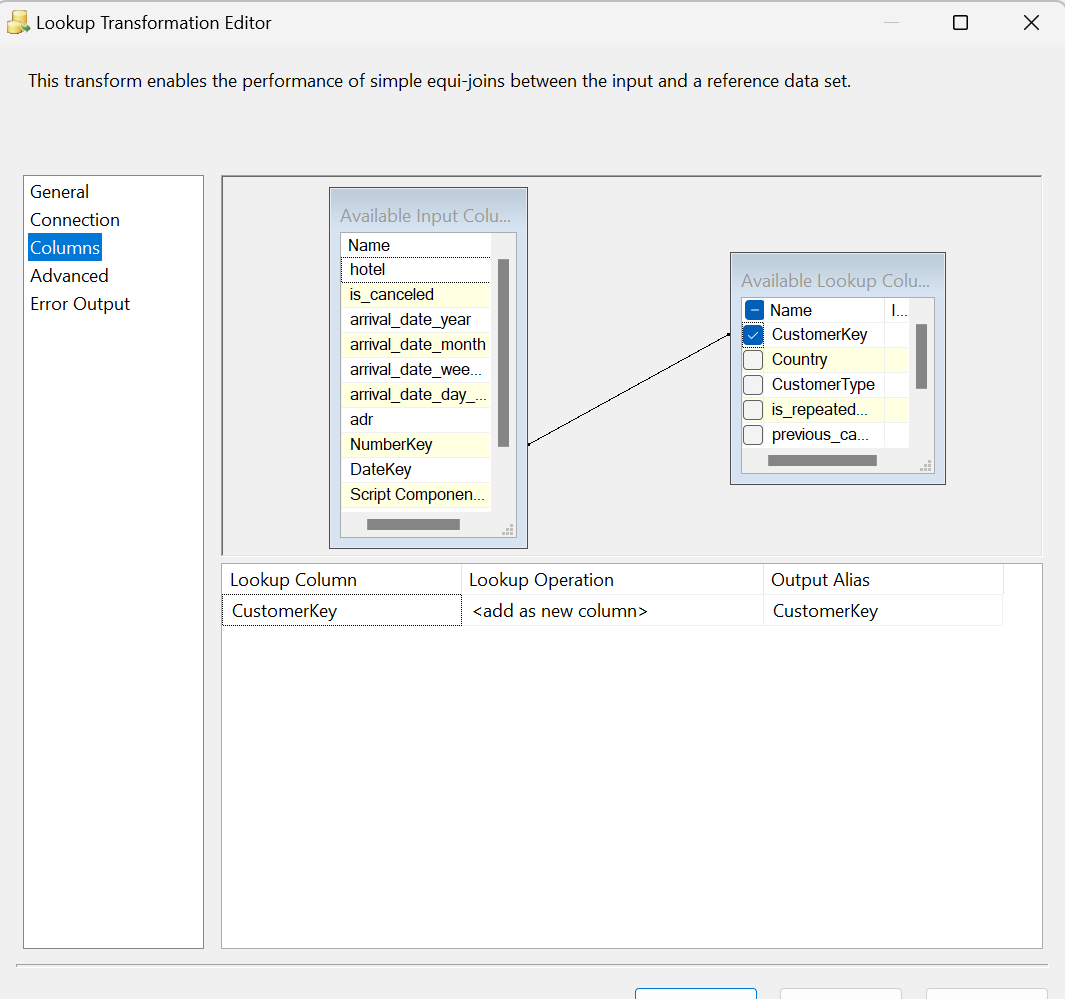
Lọc dữ kiệu theo diều kiện is-canceled = 1.



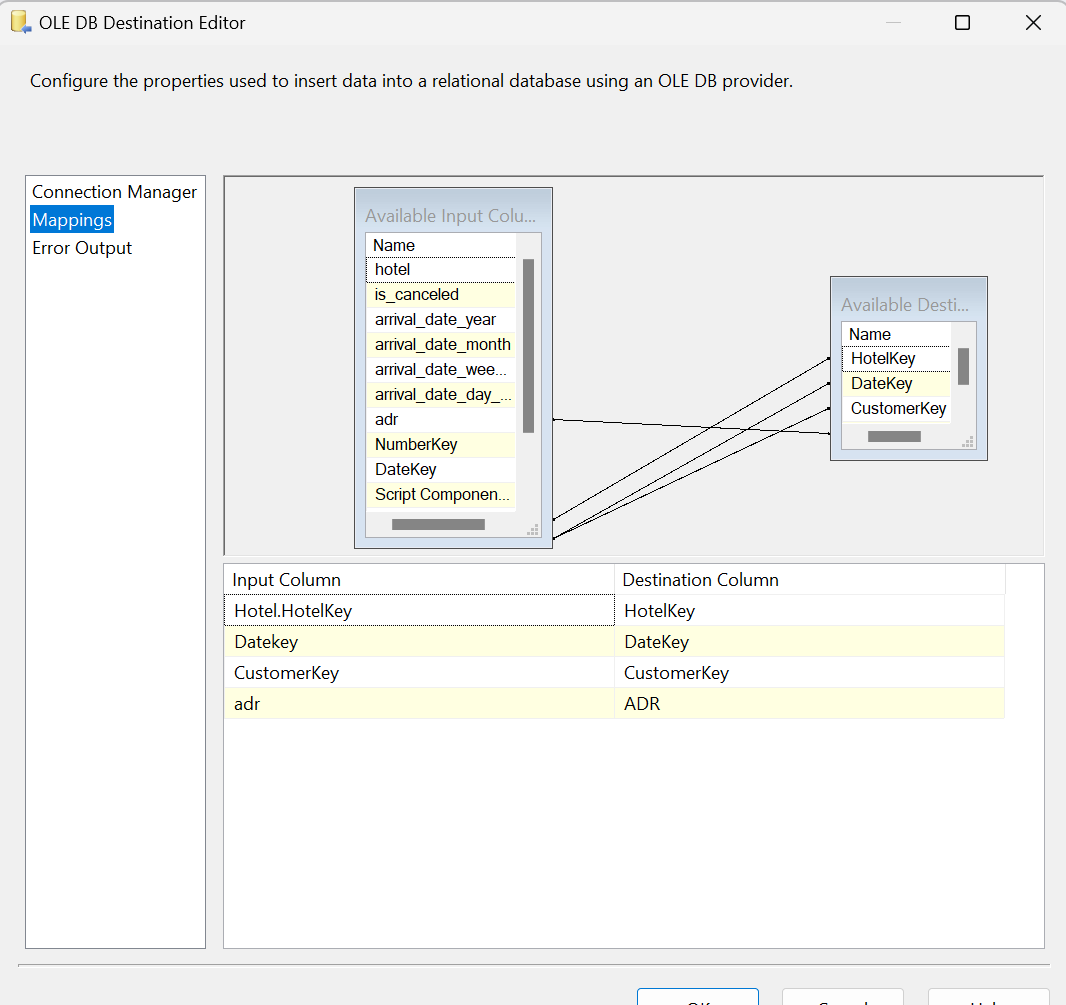
Lookup với các bảng StageHotel, StageDatekey, StageCustomers để tạo liên kết thông qua các businessKey.



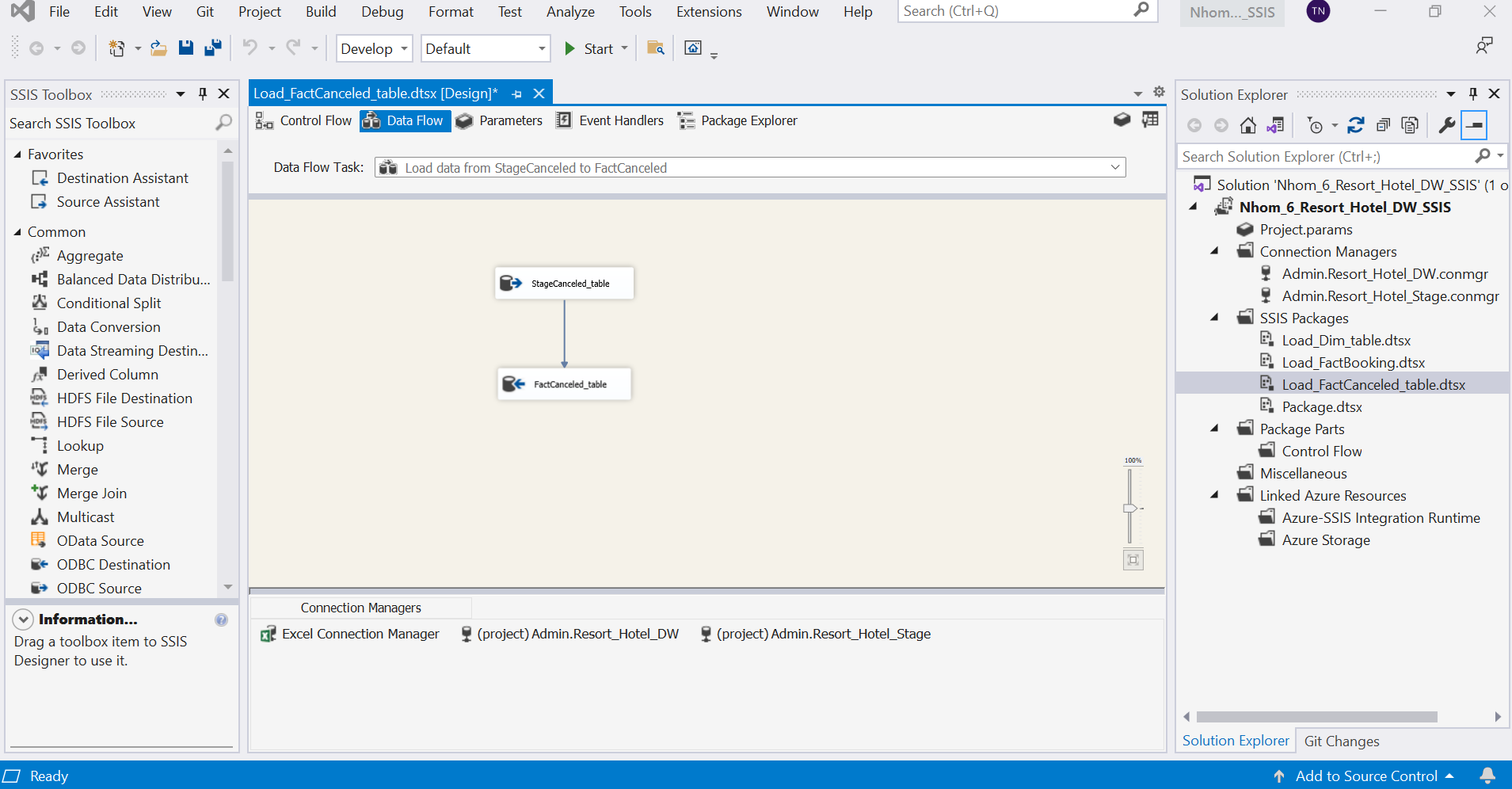




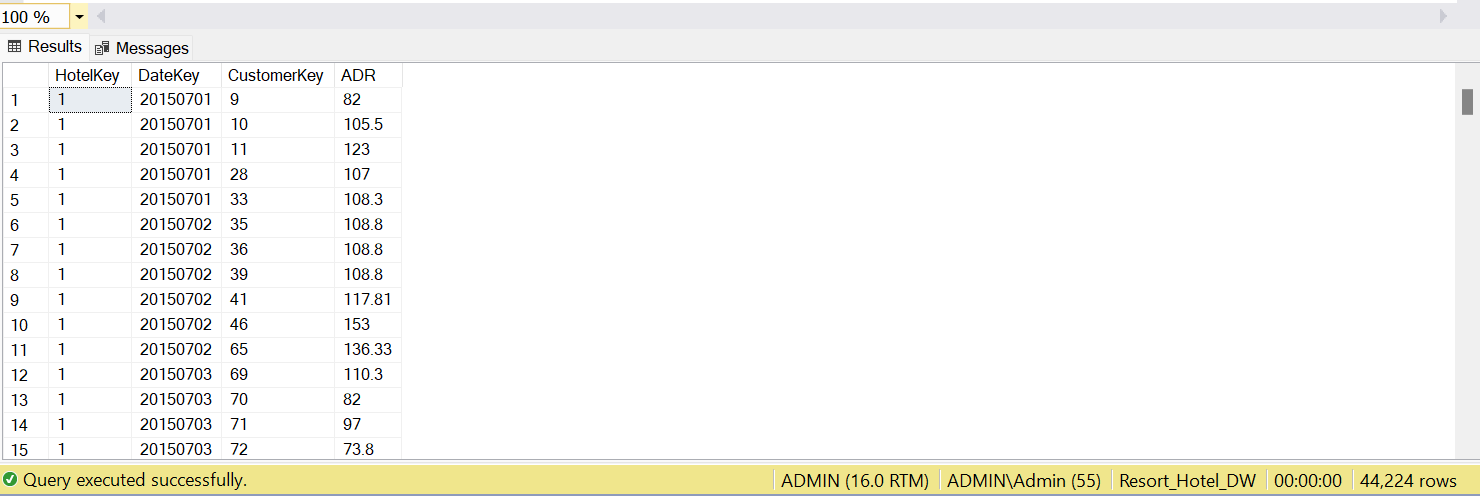
Load dữ liệu vào StageCanceled.



Load dữ liệu vào FactCanceled.



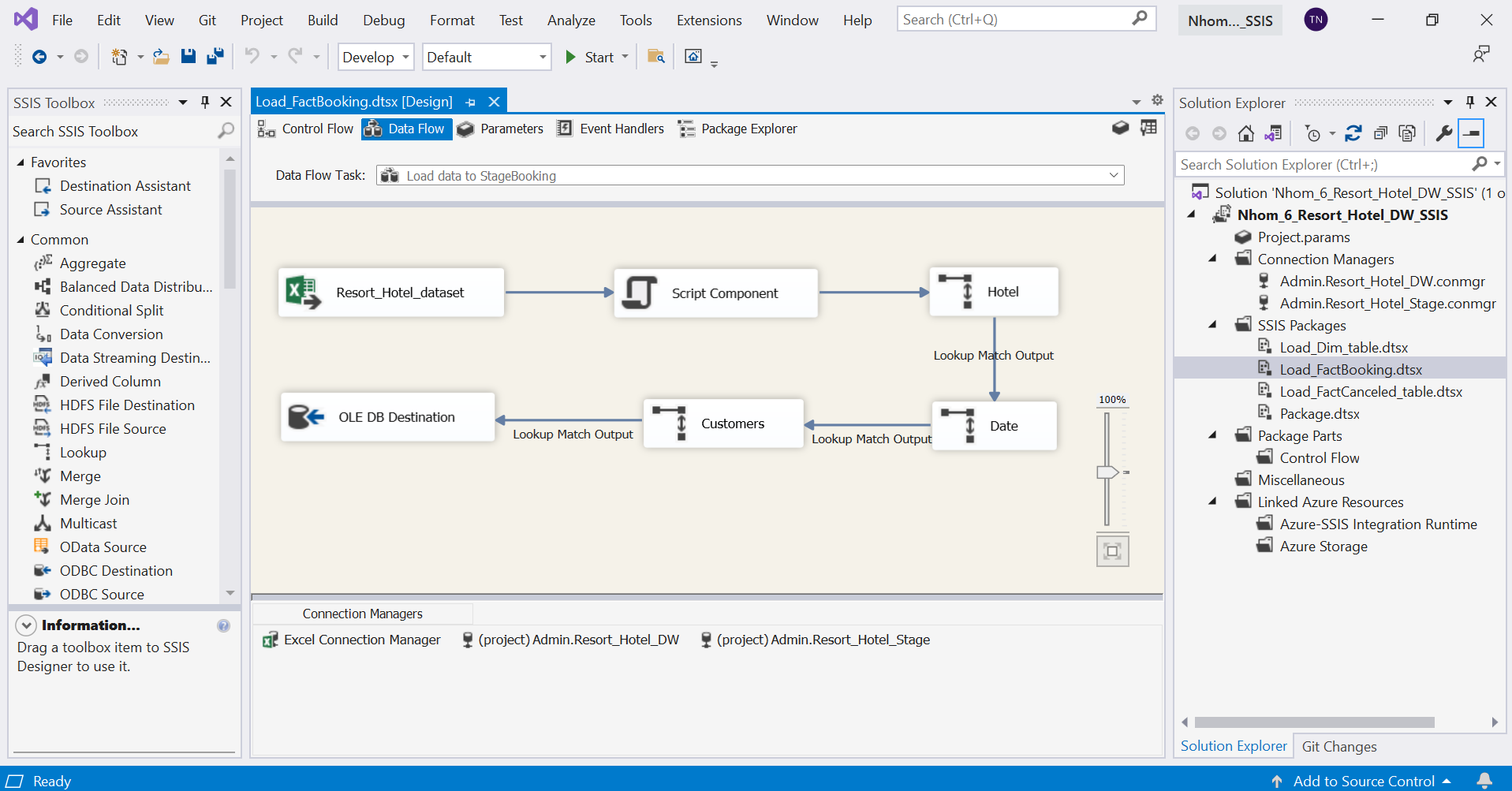
Kêt quả:



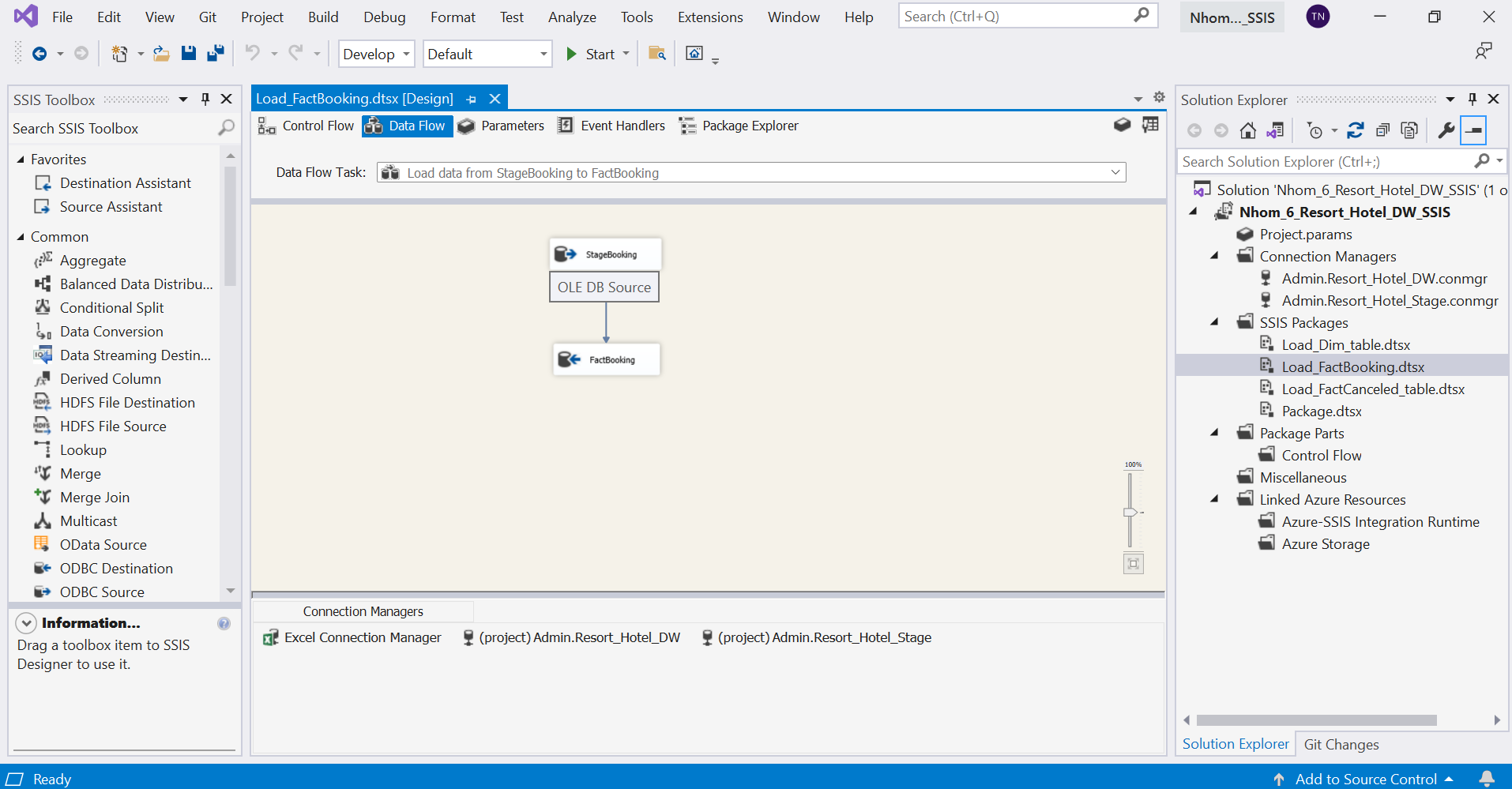
**FactBooking.**

Tương tự như FactCanceled.

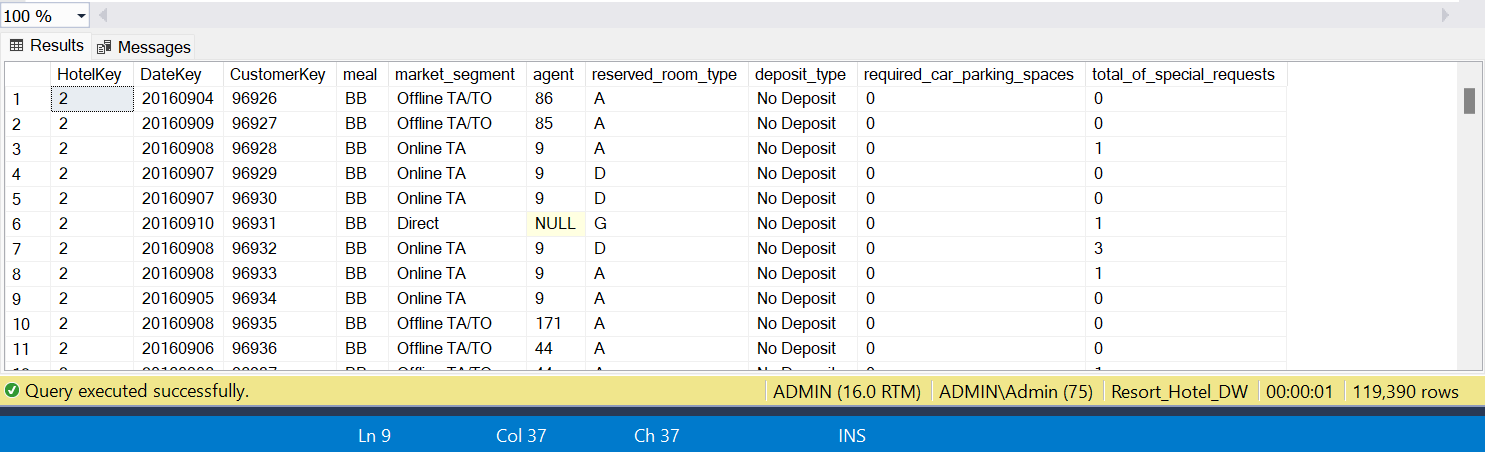
Load dữ liệu từ excel vào StageBooking



Load dữ liệu từ StageBooking vào DimBooking



Kết quả:



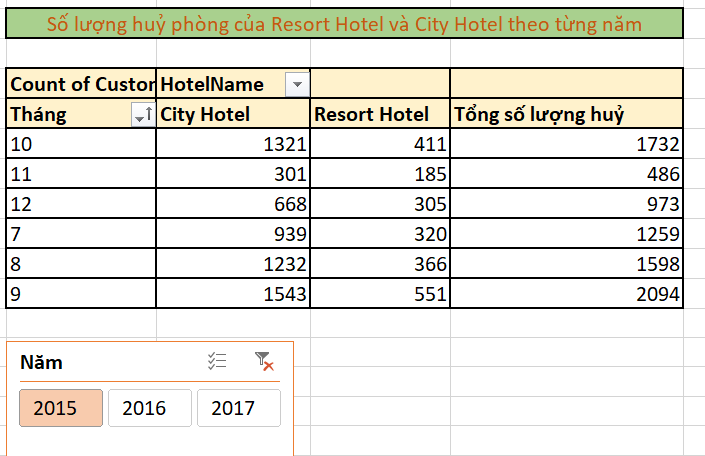
# CHƯƠNG 4: Phân tích dữ liệu

**4.4. Thực hiện tích dữ liệu.**

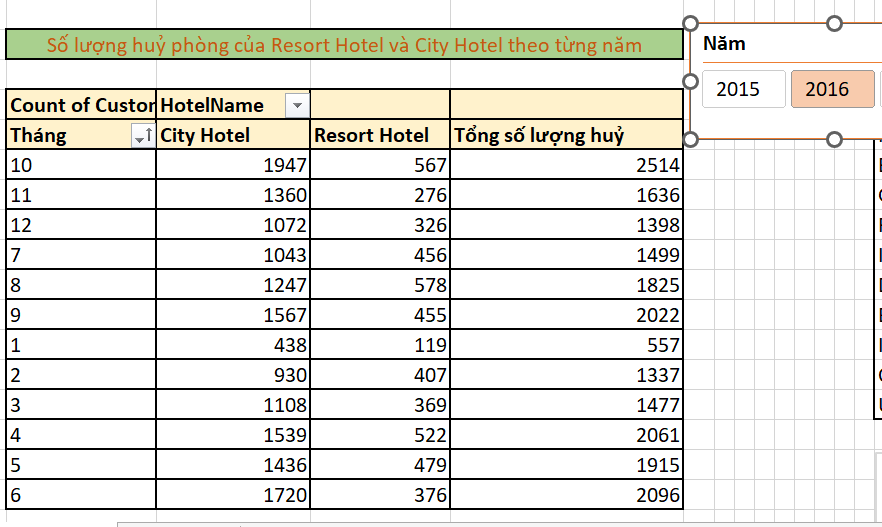
**4.4.1. Câu hỏi: Số lượng huỷ phòng của Resort Hotel và City Hotel ở các tháng theo từng năm.**

##### **4.4.1.2. Sử dụng Pivot Table trong Excel**

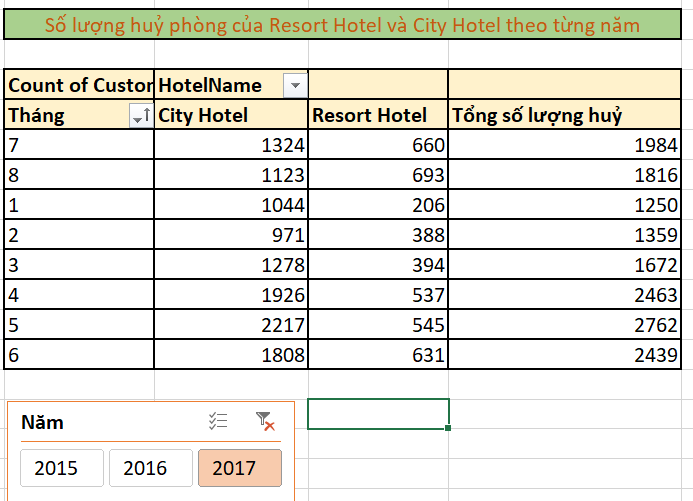
Năm 2015:



Năm 2016:



Năm 2017:



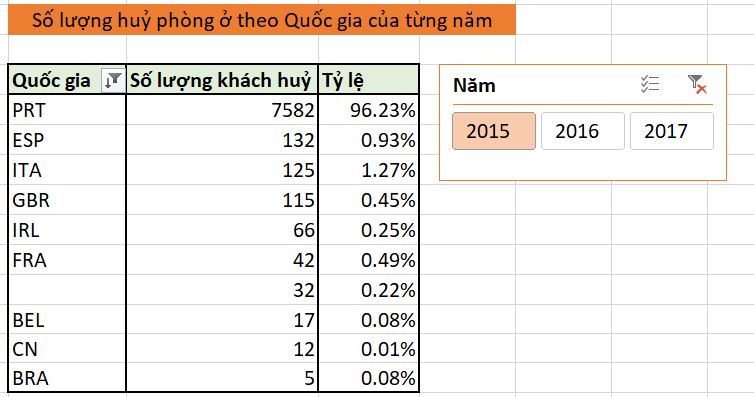
4.4.1.3 Sử dung Power BI

Với các số liệu từ Pivot và Power Bi ta thấy được vào khoảng thời gian từ tháng 4 đến tháng 10 thì số lượng huỷ phòng có sư hướng tăng ở City Hotel. Còn Resort Hotel số lượng huỷ không có sự biến động lớn.Số lượng huỷ phòng ở Resort Hotel thường thấp hơn so với City Hotel.

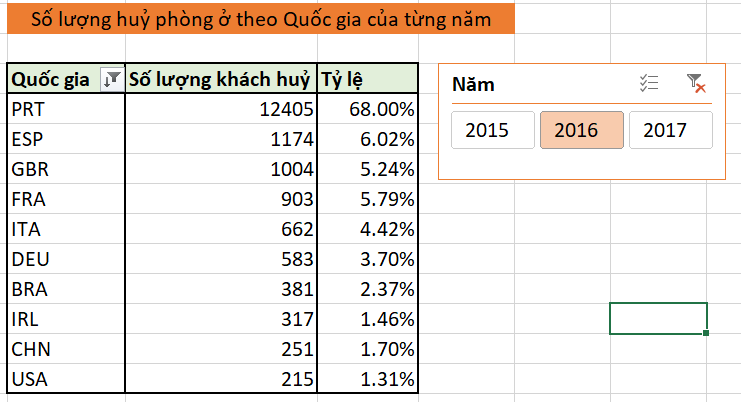
**4.4.2. Câu hỏi: Lượng huỷ phòng theo quốc gia của từng năm.**

**4.4.2.2. Sử dụng Pivot Table trong Excel**

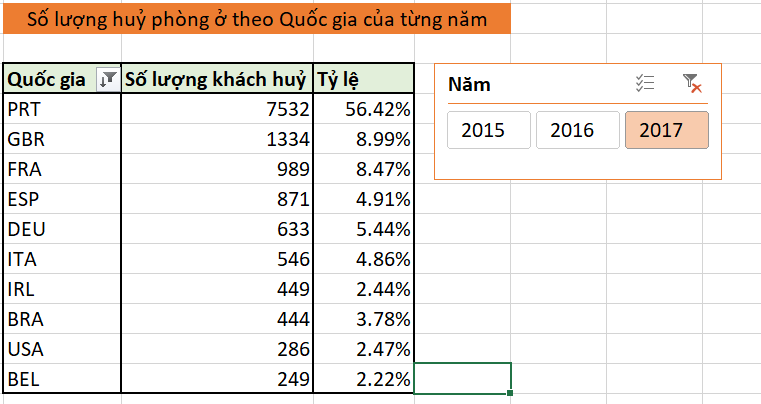
Năm 2015:



Năm 2016:



Năm 2017:



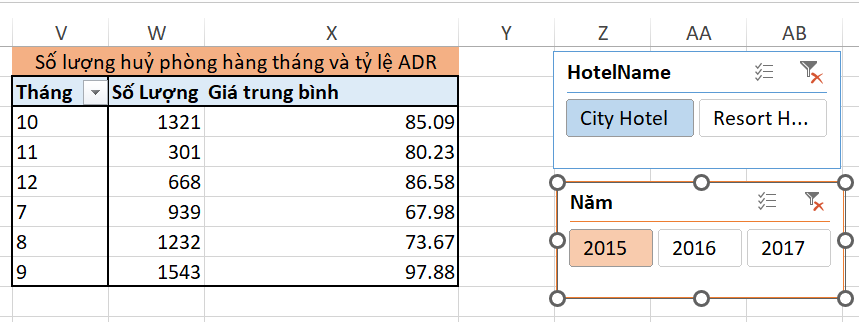
Từ số liệu và biểu đồ thì quốc giá PRT là số lượng huỷ phòng lớn. Cụ thể vào năm 2015 số lượng khách huỷ phòng lên đến 96,23% việc này có ảnh hưởng rất lớn đến việc kinh doanh của khách sạn. Tuy nhiên sau đó các năm 2016, 2017 thì số lượng khách ở PRT huỷ phòng có xu hướng giảm nhớ có biện pháp xử lý kịp thời. Tiếp đến có thấy lượng khách ở các quốc gia ESP, FRA, GBR,. chiếm phân lớn số lượng huỷ phòng. Từ đây còn có biện pháp giảm thiệu việt huỷ phòng bằng cách có chính sách kinh doanh phù hợp với khách hàng ở từng quốc gia.

**4.4.3. Câu hỏi: Cho biết số lượng huỷ phòng và giá trung bình ADR qua các tháng theo tường năm của Resort Hotel và City Hotel.**

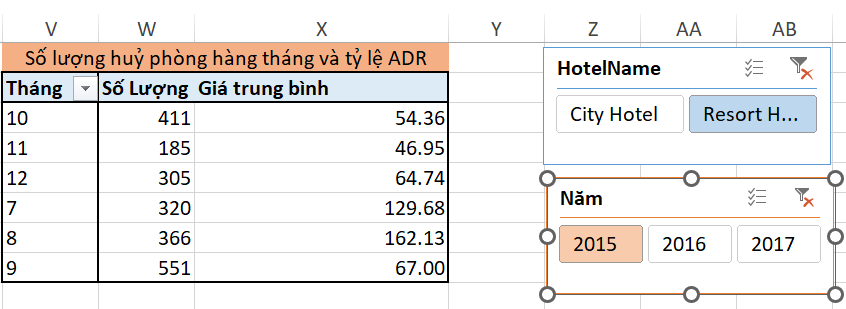
**4.4.3.2. Sử dụng Pivot Table trong Excel**

**Năm 2015:**

**City Hotel.**

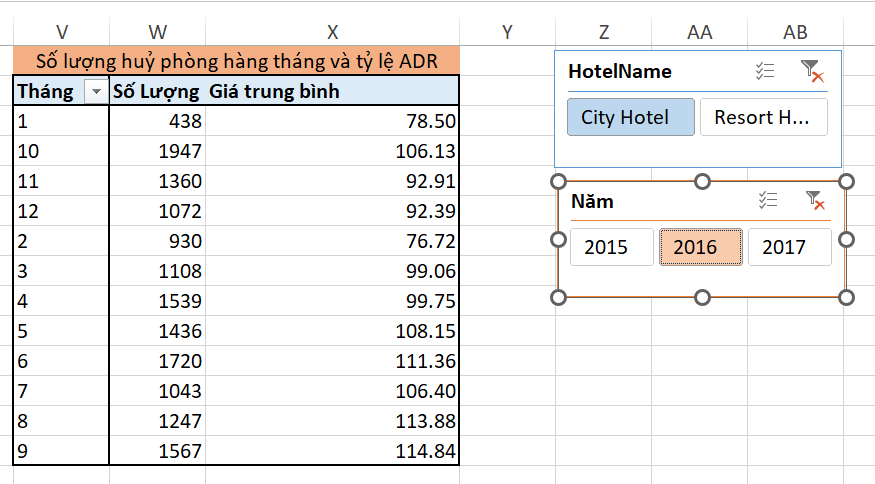


**Resort Hotel.**

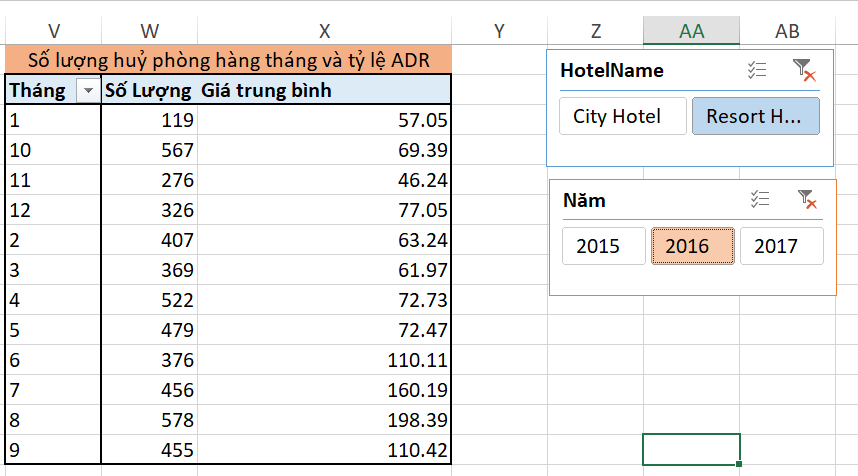


**Năm 2016:**

**City Hotel.**

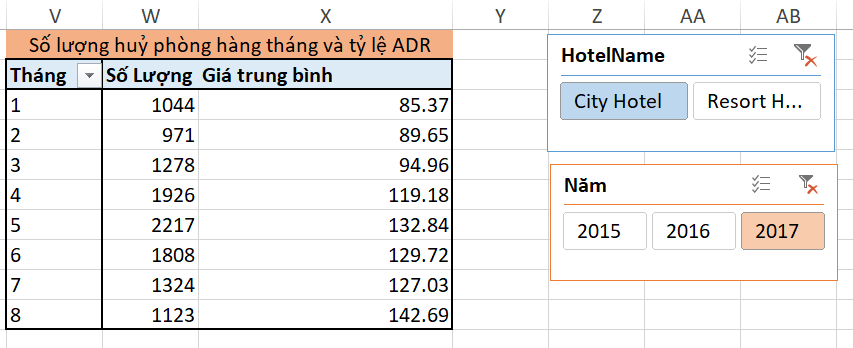


**Resort Hotel.**

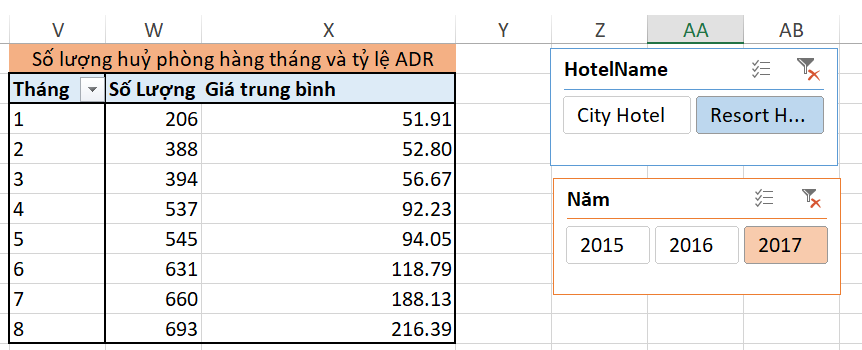


**Năm 2017:**

**City Hotel.**



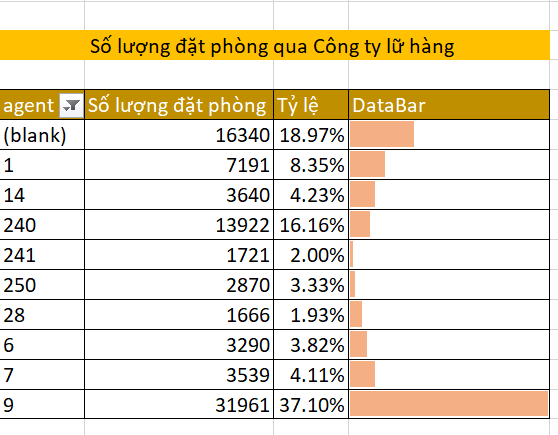
**Resort**



Với ADR trung bình từng tháng và đi kèm số lượng huỷ phòng. Có thể thấy lý do huỷ phòng cũng có bị ảnh hưởng một phần do ADR. Những tháng có lượng huỷ phòng cao thì cũng có phần đi kèm do ADR cao. Và giá ADR có vẻ là tăng theo các lên từ tháng 1 đến tháng 12.

**4.4.4. Câu hỏi: Số lượng đặt phòng qua các công ty lữ hành**

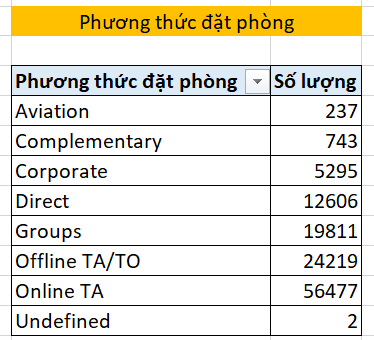
**4.4.4.2 . Sử dụng Privot table**



Đây là dữ liệu top 10 agen có số lượng đạt phòng lớn. Với những lượt đặt phòng không qua agen thì là blank. Với Agen 9 số lượng đặt phòng lớn chiến 37.10% và Agen 240 là 16.16%. Qua đây tao sẽ có các chính sách hợp tác với các công ty lữ hành để toạ mối hợp tác tốt và có các khuyến mãi giúp đẩy kinh doanh của cả hai bên.

**4.4.5. Câu hỏi: Thống kế số lượng đặt phong ứng với các phương thức đặt phòng.**

**4.4.5.2. Sử dụng Pivot Table trong Excel**



Với phương thức Online TA thì có số lượng rất lớn, với phương thức Aviation thì rất ít khách hàng sử dụng đặt phường bằng phương thức này.

# CHƯƠNG 5: KẾT LUẬN

## 5.1. Kết quả đạt được

Thực hiện xây dựng *‘KHO DỮ LIỆU CHO KHÁCH SẠN*’ giúp chúng em hiểu sâu hơn về kiến thức môn học Kho Dữ liệu và thành thạo hơn trong việc sử dụng các công cụ hỗ trợ quá trình ETL đưa dữ liệu vào kho dữ liệu như công cụ SSIS, SSAS, các công cụ để trực quan hóa dữ liệu như Excel Pivot, Power BI Desktop. Nâng cao kỹ năng nhìn nhận và tiếp thu, xử lý thông tin đầu vào, đặt những câu hỏi tiền đề và thảo luận hướng đi để giải quyết và trả lời câu hỏi.

## 5.2. Những hạn chế

Trong quá trình thực hiện xây dựng kho dữ liệu, nhóm chúng em không tránh khỏi gặp một số những khó khăn vì còn hạn chế về mặt kiến thức. Tập dữ liệu còn nhiều dữ kiện chưa được khai thác hết. Những thực hiện của chúng em chưa đạt đến mức nâng cao trong tổng thể quá trình xây dựng một kho dữ liệu hoàn chỉnh.