**SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY**

Data Warehousing and Business Intelligence

**Assignment: 01**



**ID No: IT20024918**

**Name: Mudunkotuwa N.K**

**Batch: DS weekend**

**Table of Contents**

(01)Data Set selection ……………………………………………………………………………………………………………… 3

ER Diagram…………………………………………………………………………………………………………………….. 4

(02)Preparation of Data Sources……………………………………………………………………………………………….. 5

(03)Solution Architechture………………………………………………………………………………….........................8

(04)Data Warehouse design and development ……………………………………………………………………… 10

(05)ETL Development……………………………………………………………………………………………………………….12

Data Extraction…………………………………………………………………………………………………………………12

Extract Hotel Data to Staging ………………………………………………………………………………………… 12

Extract Owner Data to Staging………………………………………………………………………………………..14

Extract Customer Data to Staging……………………………………………………………………………………15

Extract MenuType Data to Staging………………………………………………………………………………….16

Extract CustomerType Data to Staging…………………………………………………………………………….17

Extract RoomType Data to Staging………………………………………………………………………………….18

Extract CustomerAddress Data to Staging……………………………………………………………………….19

Extract Booking Data to Staging………………………………………………………………………………………20

Overall control flow………………………………………………………………………………………………………..21

Data Profiling………………………………………………………………………………………………………………….22

Data Transforming and Loading…………………………………………………………………………………………23

Load Owner Data from Staging to Data Warehouse………………………………………………………..24

Load Hotel Data from Staging to Data Warehouse………………………………………………………….25

Load MealType Data from Staging to Data Warehouse…………………………………………………..27

Load RoomType Data from Staging to Data Warehouse………………………………………………….28

Load CustomerType Data from Staging to Data Warehouse……………………………………………29

Load Customer Data from Staging to Data Warehouse…………………………………………………..30

Dim Date creation…………………………………………………………………………………………………………..32

Load Fact Table……………………………………………………………………………………………………………….35

Overall ETL to data warehouse………………………………………………………………………………………..36

(06)ETL Development – Accumulating fact tables……………………………………………………………………37

(1)Data Set Selection

# This data set contains the details of hotel booking demand for different hotels by different customers during year 2015 to 2017.

# Each of the hotel has a different owner.

# When the customer places a booking ,according to that they can select a room and a meal they prefer to have.

# Necessary modifications has been done for the data set in order to meet the requirements.

# 

# 

# Data set : Cleaned hotel booking demand

# Source :Kaggle

# 

**Link to the source**

<https://www.kaggle.com/datasets/rpereiracruz/cleaned-hotel-bookings>

# 

# ER Diagram

# 

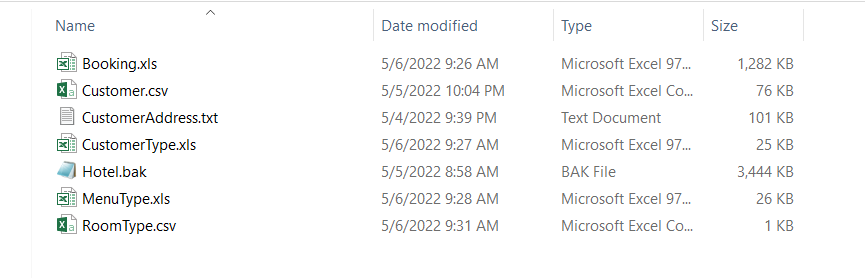
(2)Preparation of Data Sources

The original data set was in one CSV file. Data in the file has been separated into 7 different file types as Excel, CSV, bak and text.

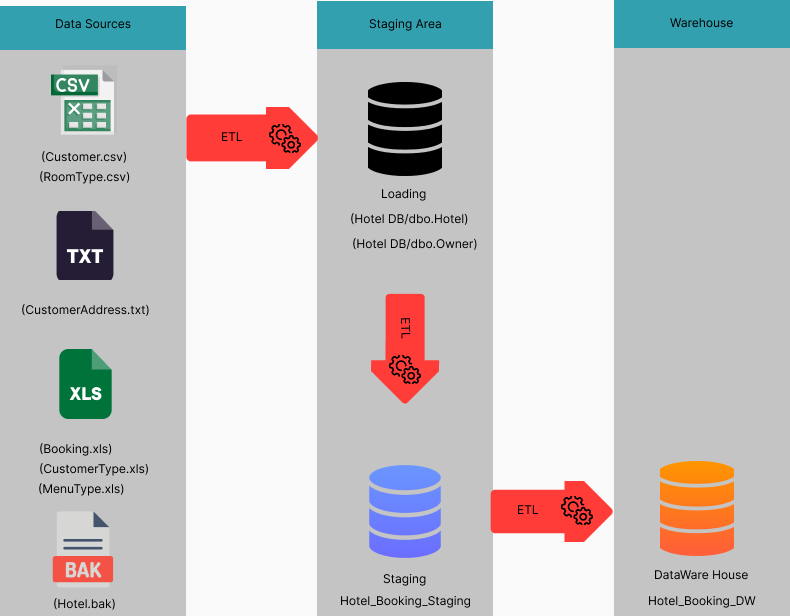
|  |  |
| --- | --- |
| **Table** | **File Type** |
| Booking | Excel file (.xls) |
| Customer | CSV file (.csv) |
| Customer Address | Text file (.txt) |
| CustomerType | Excel file (.xls) |
| MenuType | Excel file (.xls) |
| RoomType | CSV file (.csv) |
| Hotel | Bak file(.bak) |

Hotel.bak file is consists with the Hotel.csv and Owner.csv files. Hotel.bak file was imported to the Hotel database.

After the making files using different formats you can see the multiple data sources as below.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Source Type** | **Source Name** | **Column Name** | **Data Type** | **Description** |
| Excel file | MenuType.xls | MealId | int | Unique ID |
|  |  | MealValue | nvarchar(255) | There are five type of meals as below.FB,HB,SC,BB,  Undefined. |
|  |  | MealTypeDescription | nvarchar(255) | FB – Full Board  HB – Half Board  SC - Self Catering  BB – Bed and Breakfast  Undefined – No meal |
|  | CustomerType.xls | CustomerTypeId | int | Unique ID |
|  |  | CustomerType | nvarchar(255) | There are 4 type of customers.They are   |  | | --- | | Transient, Contract, | | Transient-Party, Group | |  | |  | |
|  | Booking.xls | BookingId | int | Unique ID |
|  |  | HotelId | int | Foreign Key |
|  |  | CustomerId | nvarchar(255) | Foreign Key |
|  |  | ArrivalDate | datetime | This contains the customer arrival date ,year and month |
|  |  | ArrivalDateWeekNumber | int | This contains the customer arrival week number |
|  |  | StaysInWeekendNights | int | No of nights customers going to stay in the week end |
|  |  | StaysInWeekdayNights | int | No of nights customers going to stay in the weekdays |
|  |  | MealId | int | Foreign key |
|  |  | PricePerNight | float | This contains the price that a customer needs to pay per night |
|  |  | Discount | float | This contains the discount that a customer gets |
|  |  | RoomTypeId | int | Foreign key |
| CSV file | Customer.csv | CustomerId | nvarchar(255) | Unique ID |
|  |  | CustomerName | nvarchar(50) | Name of the customer |
|  |  | CustomerOccupation | nvarchar(50) | Occupation of the customer |
|  |  | CustomerPhone | nvarchar(50) | Contact number of the phone |
|  |  | CustomerTypeId | int | Foreign key |
|  | RoomType.csv | RoomtypeId | int | Unique ID |
|  |  | RoomTypeValue | nvarchar(50) | From A –I there different types of rooms that a customer can book |
|  | CustomerAddres.txt | CustomerId | nvarchar(255) | Foreign key |
| Text file |  | CustomerHomeAddress | nvarchar(50) | Home address of the customer |
|  |  | CustomerCity | nvarchar(50) | Customer’s living city |
|  |  | CustomerState | nvarchar(50) | Customer’s living state |
|  |  | CustomerCountry | nvarchar(50) | Customer’s living Country |
|  |  | CustomerZipCode | nvarchar(50) | Customer’s city ZipCode |
| Database File | dbo.Hotel | HotelId | int | Unique Id |
|  |  | HotelName | nvarchar(100) | Name of the hotel |
|  |  | latitude | float | Hotel location - latitude |
|  |  | longitude | float | Hotel location - longitude |
|  |  | OwnerId | int | Foreign key |
|  | dbo.Owner | OwnerId | int | Unique key |
|  |  | Owner\_name | nvarchar(50) | Name of the owner |
|  |  | Gender | nvarchar(50) | Name of the gender |
|  |  | Phone | nvarchar(50) | Owner’s contact number |
|  |  | Owner\_since | date |  |

(3)Solution Architechture

The above architecture represents the High-Level BI solution to the warehouse design

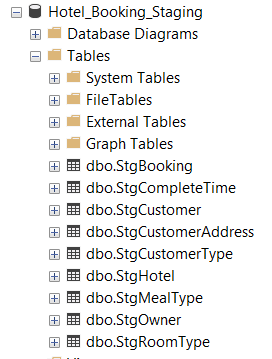
**Data Sources**

Data Sources In the above diagram, ‘.txt’ component represents Text files, ‘.xls’ component is used to represent Excel files, ‘.csv’ component is used to represent Comma Separated files, and ‘.bak’ component represents Database files.

**Staging Area**

Loading DB component represents the process of the creating database tables. Owner and Hotel csv files were imported to the database and the relevant database tables were created. These tables were used as the DB source data. Staging DB component represents the creation of staging level tables through the ‘Extract’ process.

After extracting data into the staging the staging database looks as below.

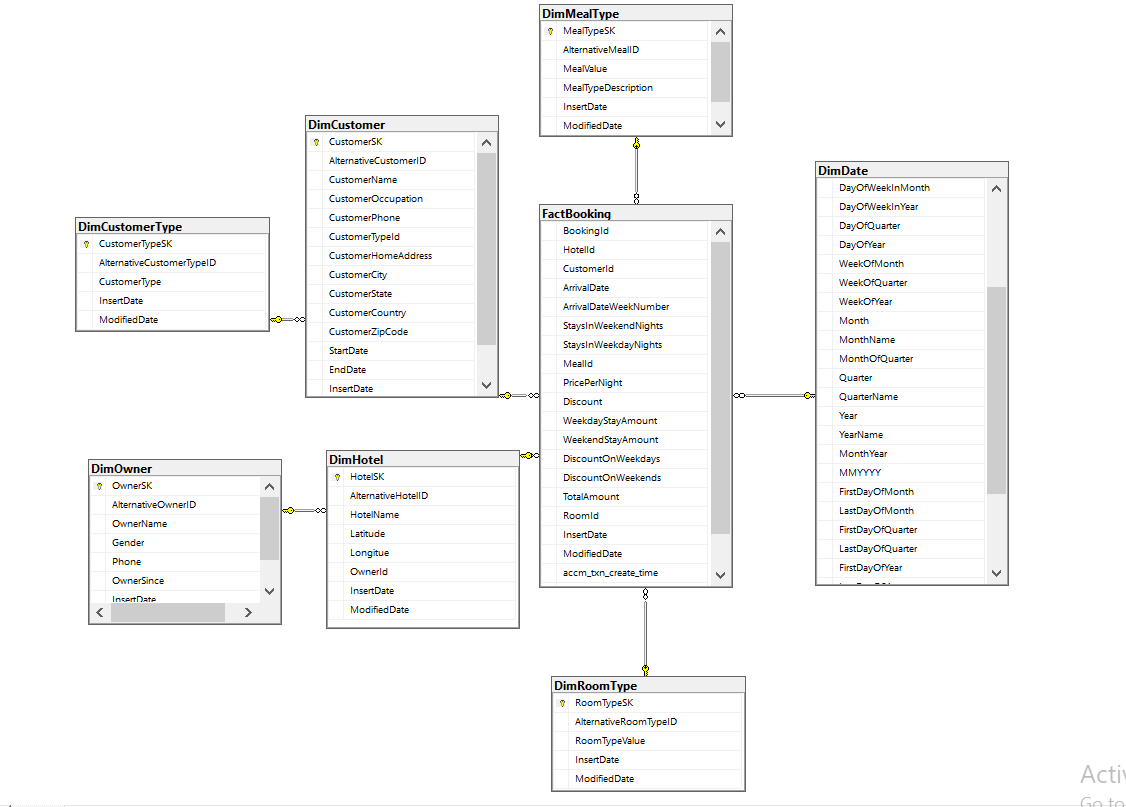


**Data Warehouse**

The tables at the staging are then profiled and after performing a rich set of ETL tasks, data is loaded to the data warehouse where from that several reporting tools and analysing tools can use data for reporting mining and analyzing.

(4)Data Warehouse design and development

The following diagram shows how the fact table and dimension tables were combined in a rational manner to implement the data warehouse.



**Schema Type**

**Snowflake** schema type was used for this scenario since the tables need to be **normalized.**

**Dimension Types**

• Hierarchical Dimensions

* Date – all the hierarchies in date
* Customer– Country → State → City → Zip → HomeAddress
* DimCustomerType is taken as a hierarchy of DimCustomer since one customer might use only one CustomerType but single CustomerType is used by several customers.
* DimOwner is taken as a hierarchy of DimHotel since one Hotel might have only one Owner but single Ownercan have many hotels.

• Slowly Changing Dimension

➢ Customer

➢ Following columns were set as changing attributes.

* CustomerCity
* CustomerHomeAddress
* CustomerZipcode
* CustomerState

• Fact Table

➢ Numbers – Stays in weekday nights ,Stays in weekend nights, Price per night,Weekday stay amount,Weekend stay amount,Discount

➢ FKs – Hotel ID, Customer ID, Meal ID,Room ID

**Calculations**

1. WeekdayStayAmount in [dbo].[FactBooking] table is calculated by,

([NoOfWeekdayNights]\*[PricePerNight])

1. WeekendStayAmount in [dbo].[FactBooking] table is calculated by,

([NoOfWeekendNights]\*[PricePerNight])

1. DiscountOnweekdays in [dbo].[FactBooking] table is calculated by,

([NoOfWeekdayNights]\*[PricePerNight]\*[Discount])

1. DiscountOnweekends in [dbo].[FactBooking] table is calculated by,

([NoOfWeekendNights]\*[PricePerNight]\*[Discount])

1. TotalAmount in [dbo].[FactBooking] table is calculated by,

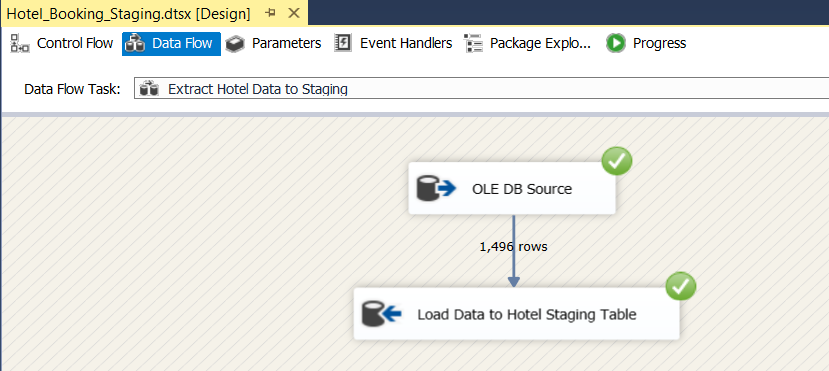
[WeekdayStayAmount]+[WeekendStayAmount]–[DiscountOnweekdays] – [DiscountOnweekends]

(5)ETL development

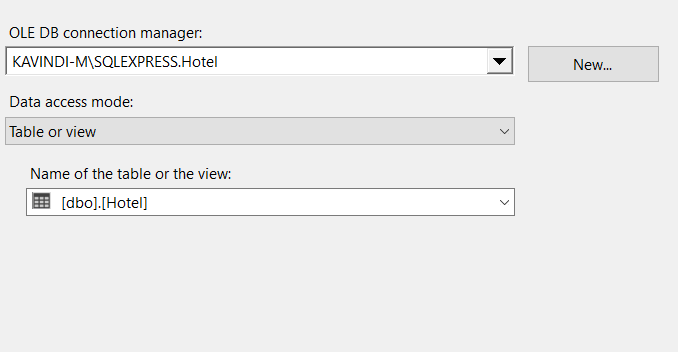
As the first step data has been extracted from sources to staging area. Data flow task has been used for every extraction.

**DATA EXTRACTION**

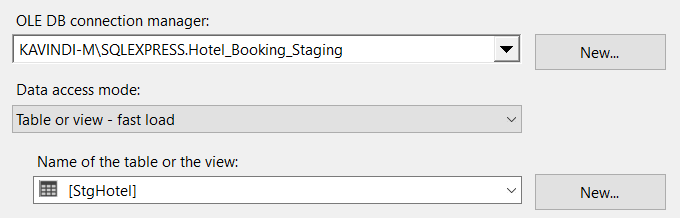
**Extract Hotel Data to Staging area**



Hotel data in Hotel database table has been extracted and loaded to Hotel Staging table

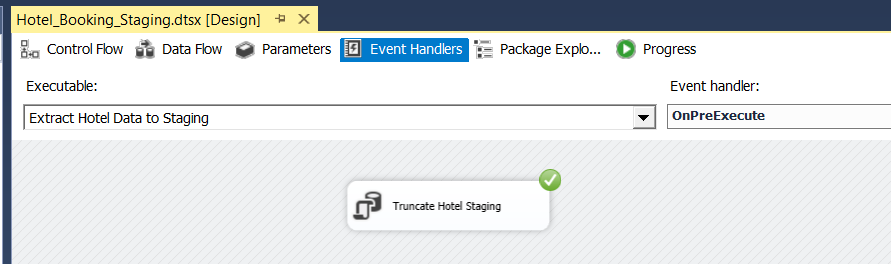
Used OLE DB Source as dbo.Hotel in the Hotel database

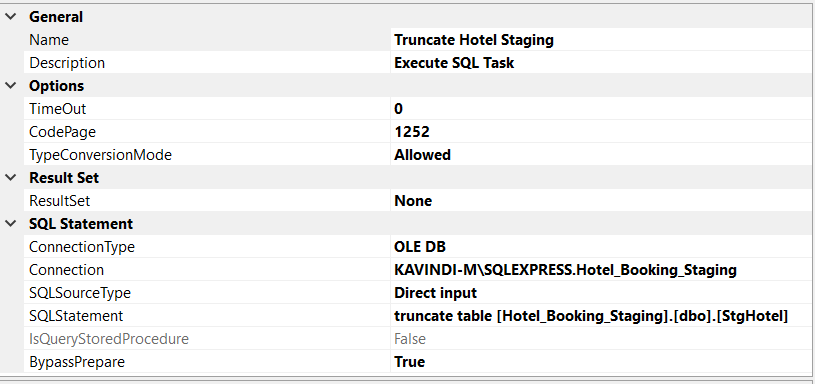
OLE DB Destination for create new table as StgHotel in the Hotel\_Booking\_Staging database.



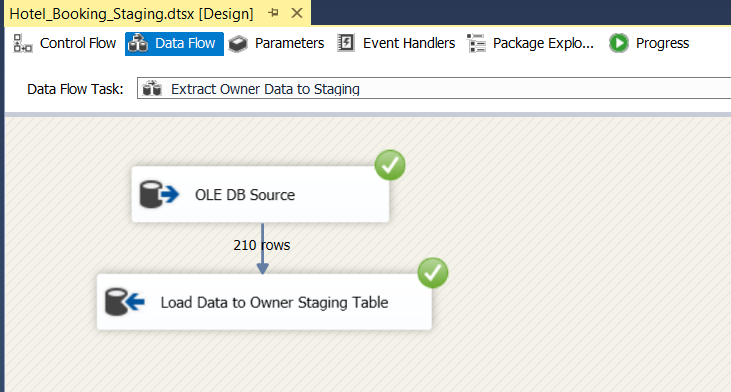
**Event handler was used to do the truncate**

Used Execute SQL Task SSIS tool to Truncate table for SQL command as truncate table dbo.StgHotel



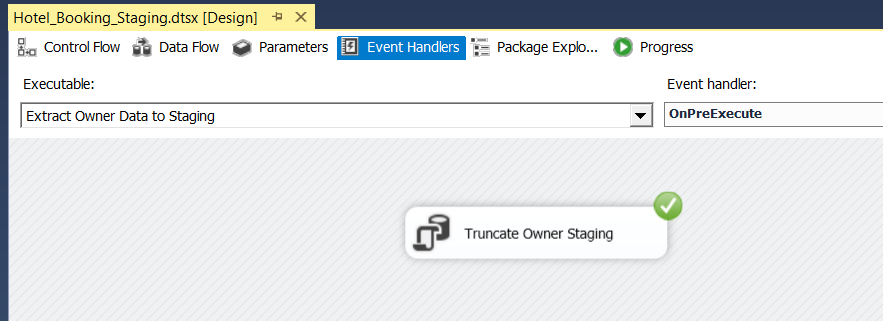


**Extract Owner Data to Staging area**

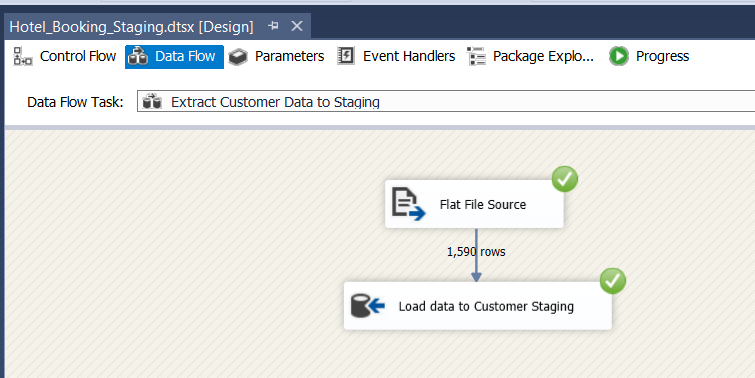
****

Owner data in Hotel database table has been extracted and loaded to Owner Staging table

**Event handler was used to do the truncate**

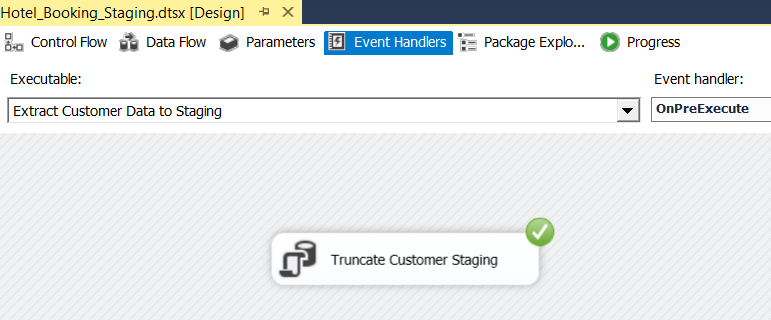
****

**Extract Customer Data to Staging area**

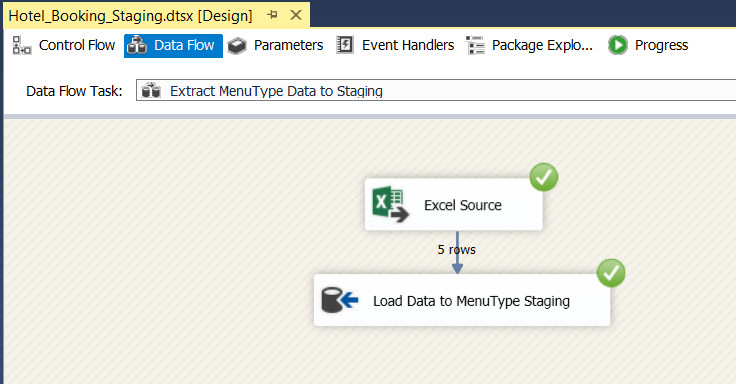
****

Customer data in Customer.CSV has been extracted and loaded to Customer Staging table

**Event handler was used to do the truncate**

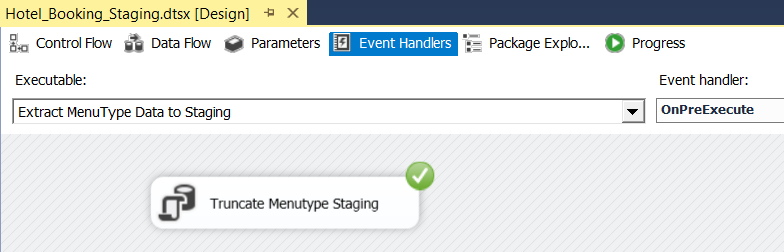


**Extract MenuType Data to Staging area**

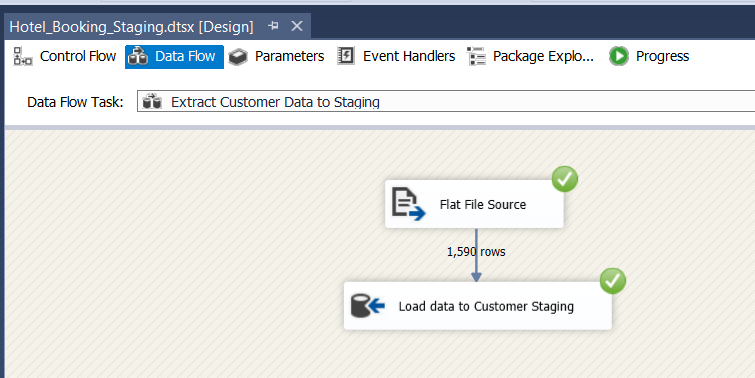
****

MenuType data in MenuType.xls has been extracted and loaded to MenuType Staging table

**Event handler was used to do the truncate**

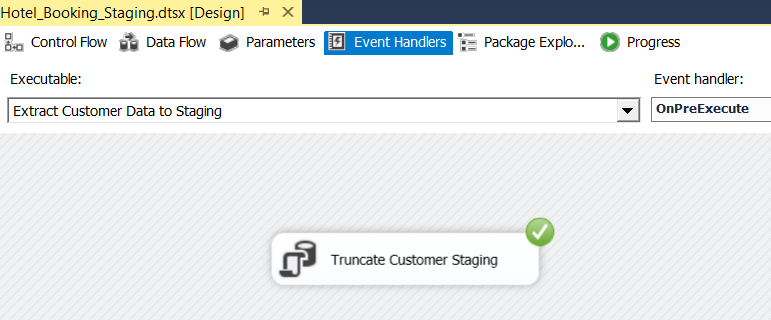


**Extract CustomerType Data to Staging area**

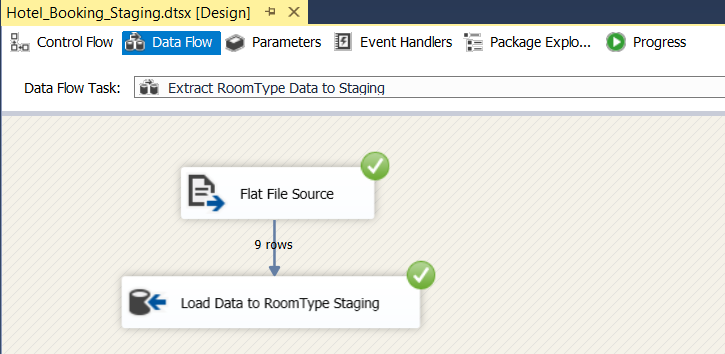


CustomrType in CustomerType.csvhas been extracted and loaded to CustomerType Staging table

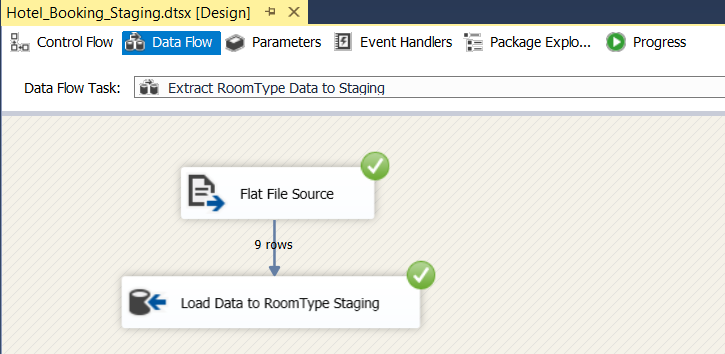
**Event handler was used to do the truncate**

****

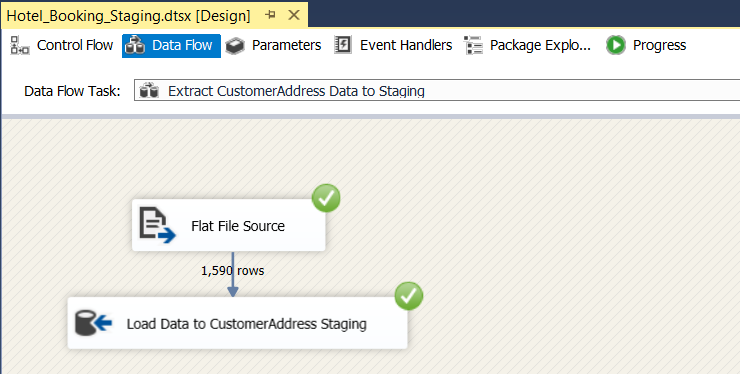
**Extract RoomType Data to Staging area**

****

RoomType data in RoomType.csv has been extracted and loaded to RoomType Staging table

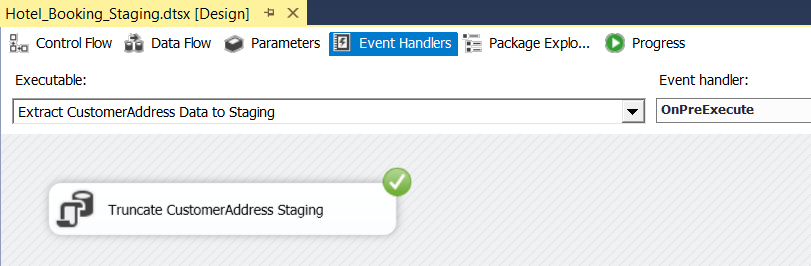
**Event handler was used to do the truncate**

**Extract CustomerAddress Data to Staging area**

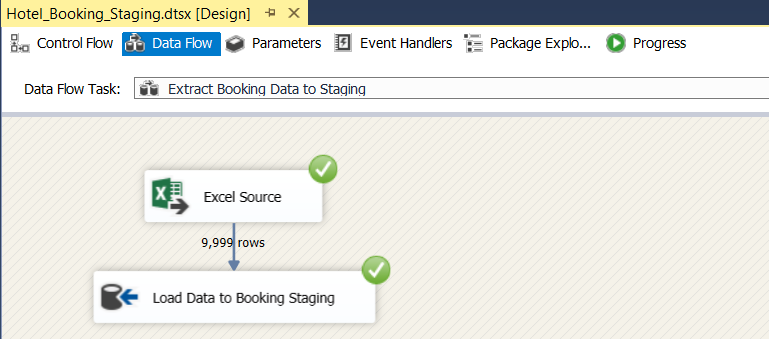
****

CustomerAddress data in CustomerAddress.txt has been extracted and loaded to CustomerAddress Staging table

**Event handler was used to do the truncate**

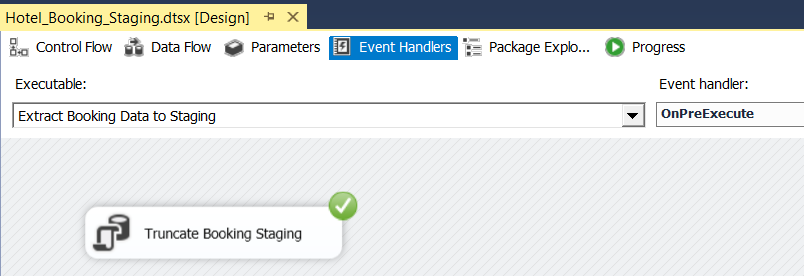
****

**Extract Booking Data to Staging area**

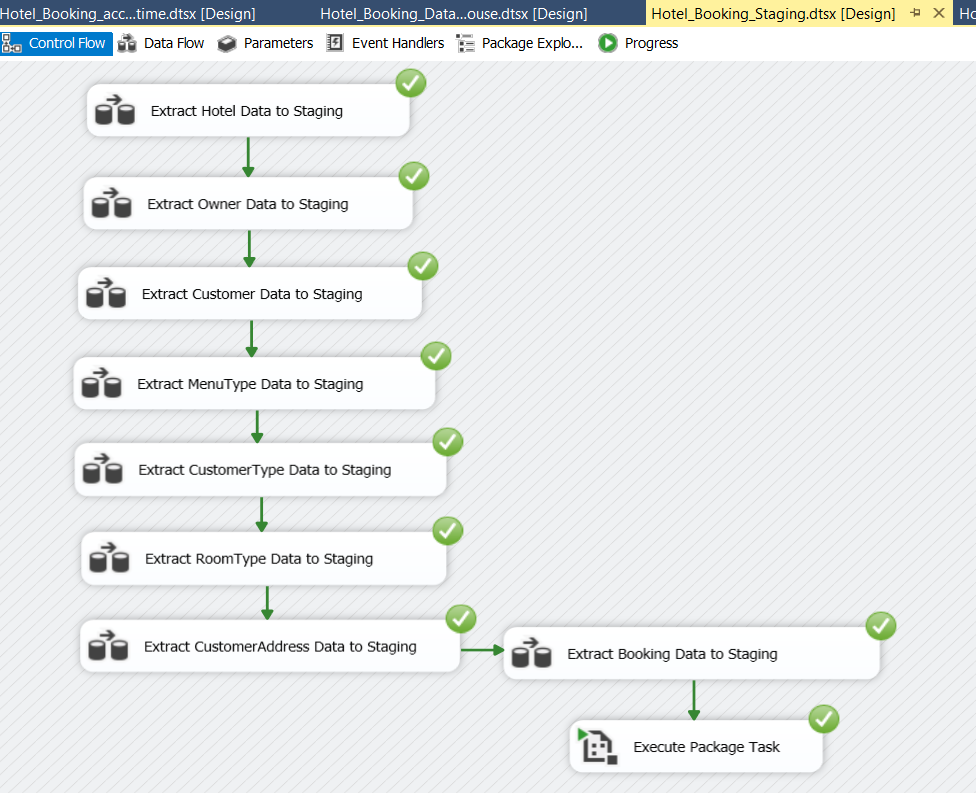
****

Booking data in Booking.xls has been extracted and loaded to BookingStaging table

**Event handler was used to do the truncate**

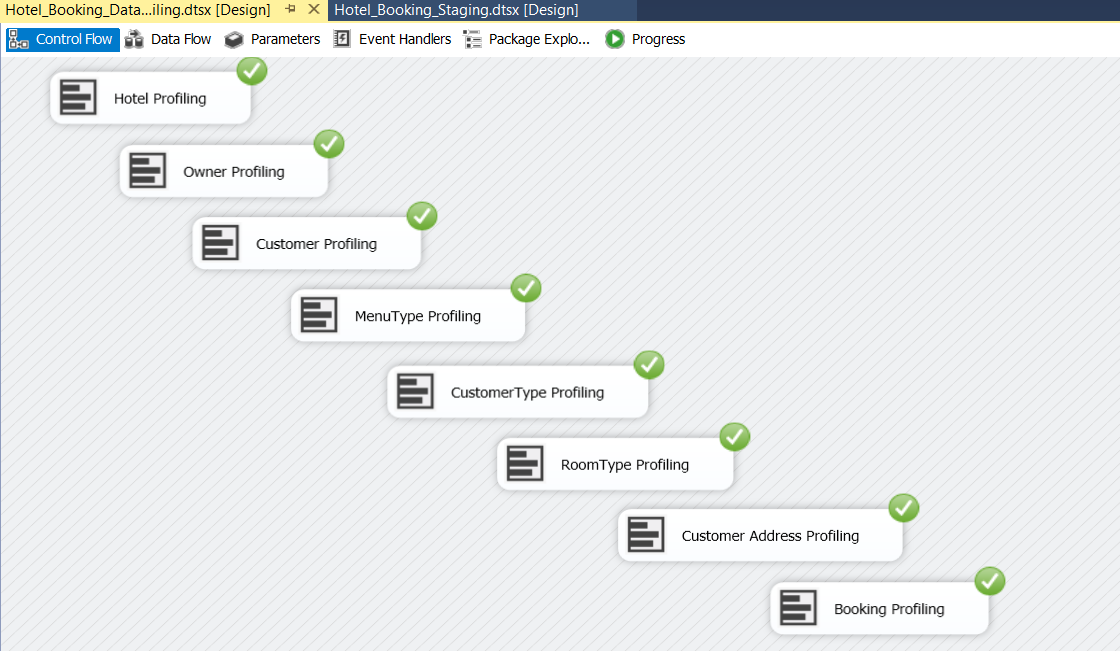


**Overall control flow**

****

**Data Profiling**

Before loading the created staging tables to the data warehouse each and every staging table has been profiled.

****

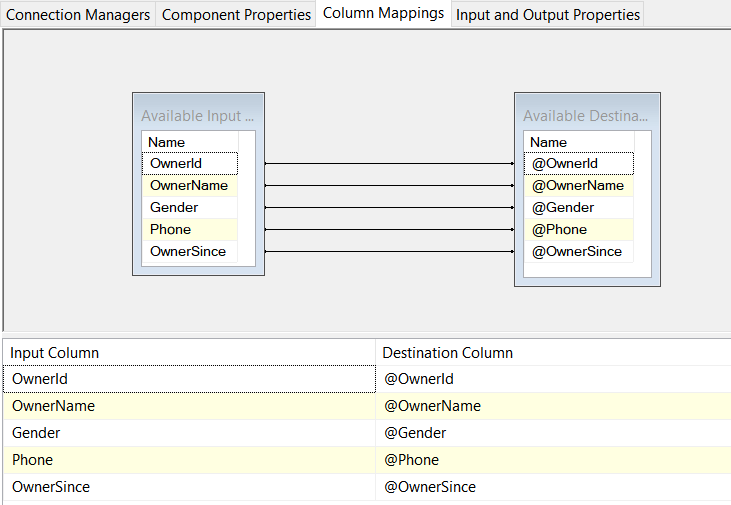
Data Transforming and Loading

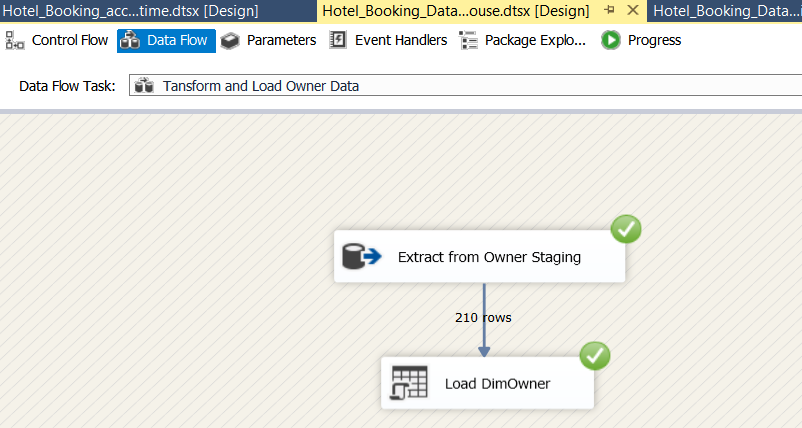
(a)Load Hierarchical Dimensions

This design is consisting with two hierarchical dimensions. They are ,

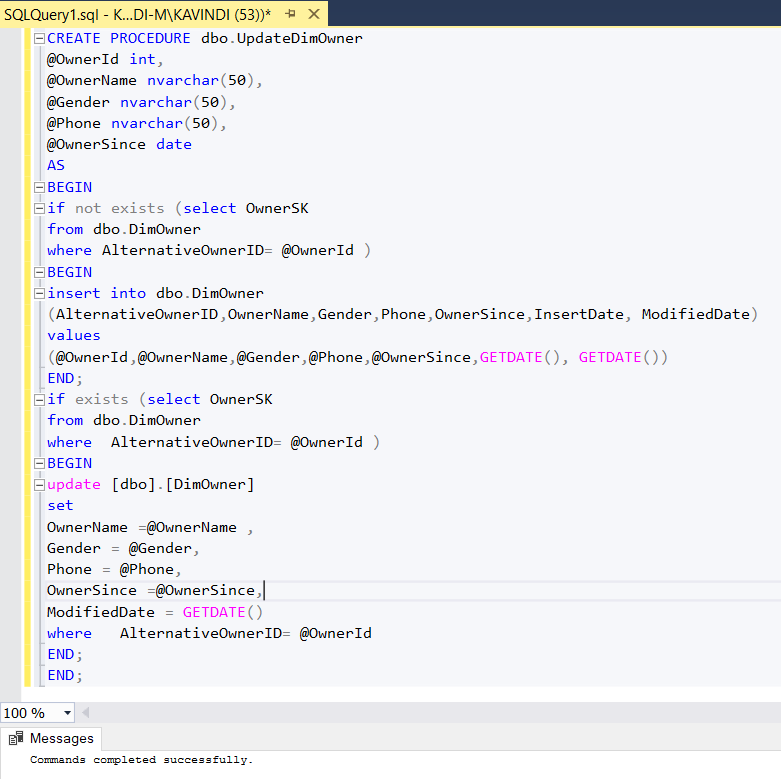
(1)Owner Dimension –Sub category of Hotel Dimension

(2)CutomerType Dimension – Sub category of Customer dimension

**Owner Data from Staging to Data Warehouse**

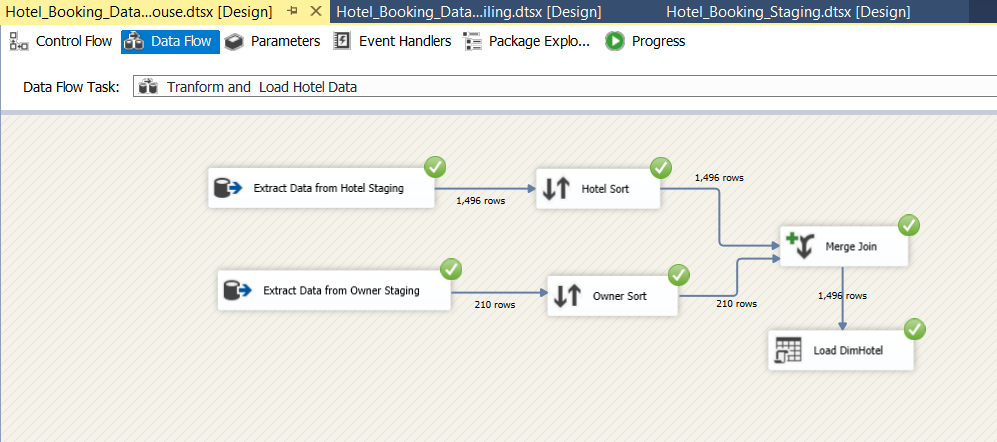


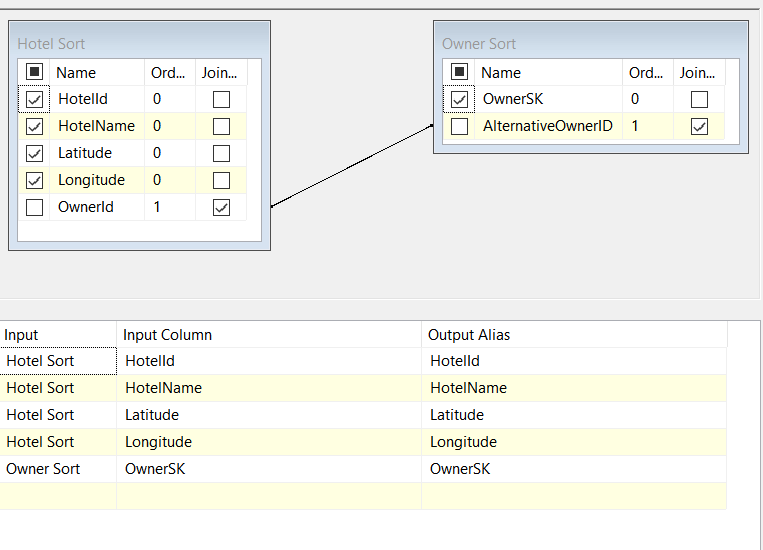
Owner data has been loaded to DimOwner table in data warehouse.The following procedure is used to load data into the DimOwner table.



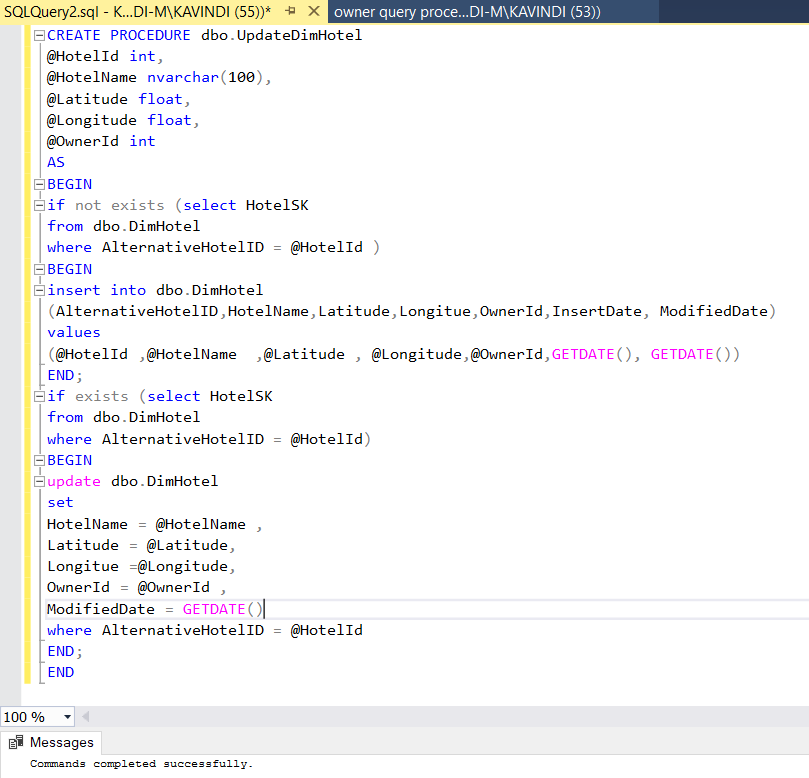
**Hotel Data from Staging to Data Warehouse**

As mentioned in above Owner is a hierarchical dimension of Hotel.To get the details of the surrogate key of the DimOwner merge has been used as below.

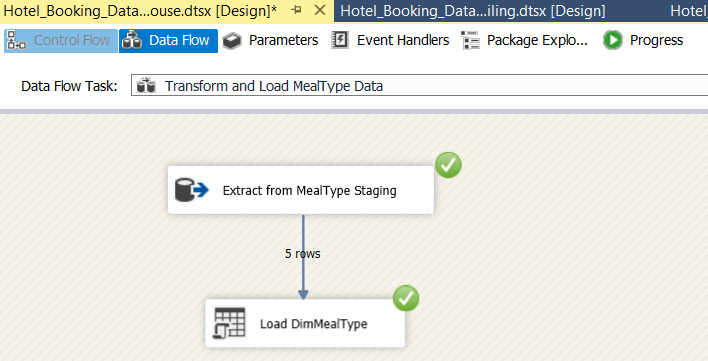
****

****

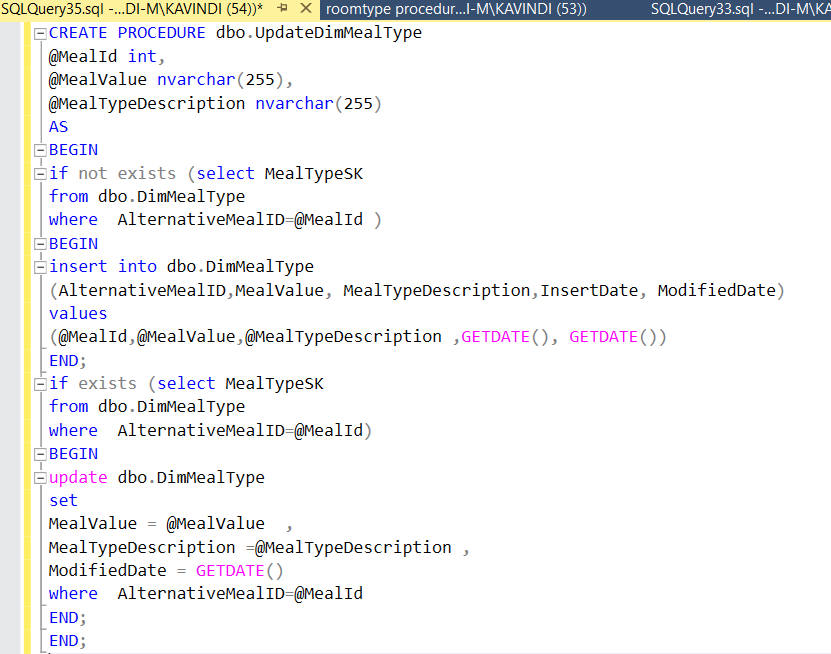
Hotel data has been loaded to DimHotel table in data warehouse. The following procedure is used to load data into the DimHotel table.



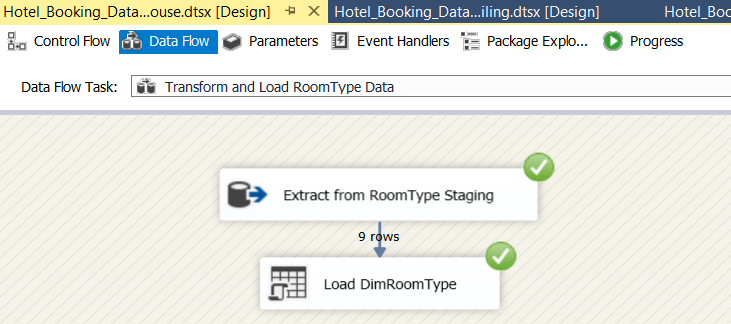
**MealType Data from Staging to Data Warehouse**

****

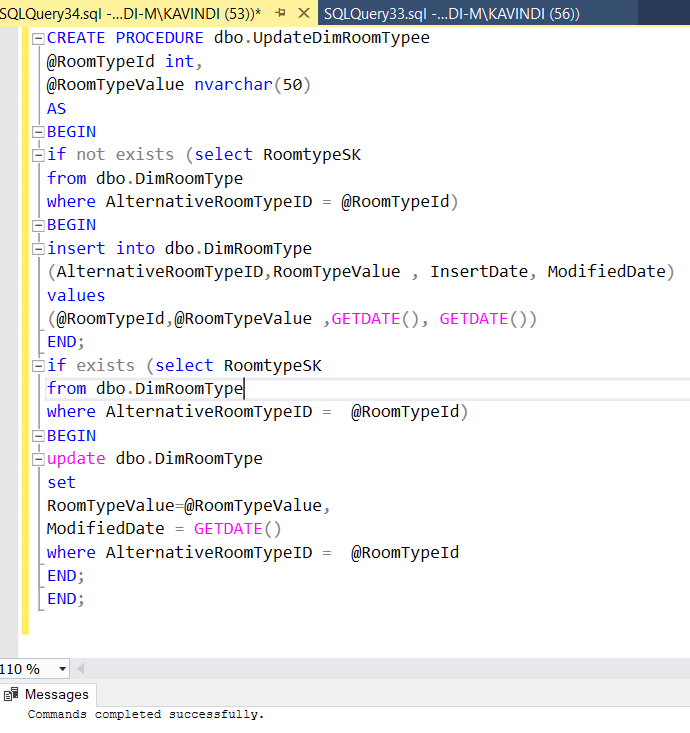
MealType data has been loaded to DimMealType table in data warehouse. The following procedure is used to load data into the DimMealType table.



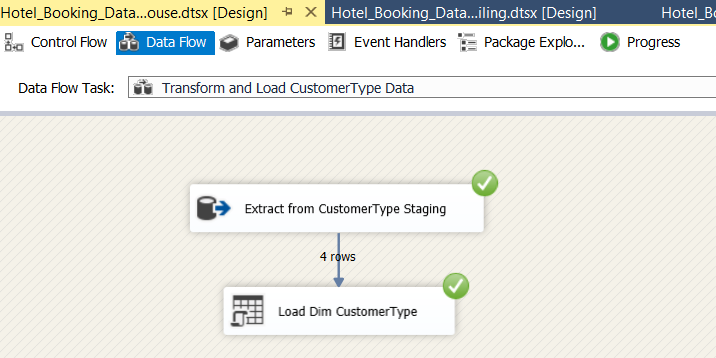
**RoomType Data from Staging to Data Warehouse**

****

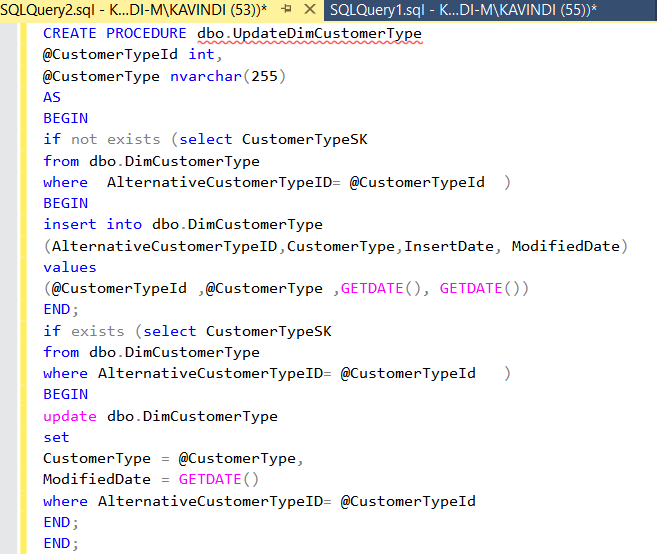
RoomType data has been loaded to DimRoomType table in data warehouse. The following procedure is used to load data into the DimRoomType table.



**CustomerType Data from Staging to Data Warehouse**

****

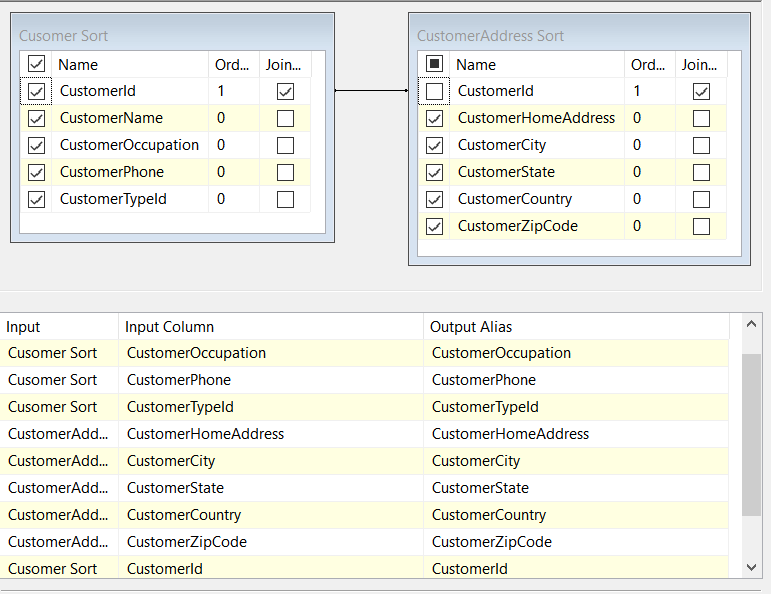
CustomerType data has been loaded to DimCustomerType table in data warehouse. The following procedure is used to load data into the DimCustomerType table.



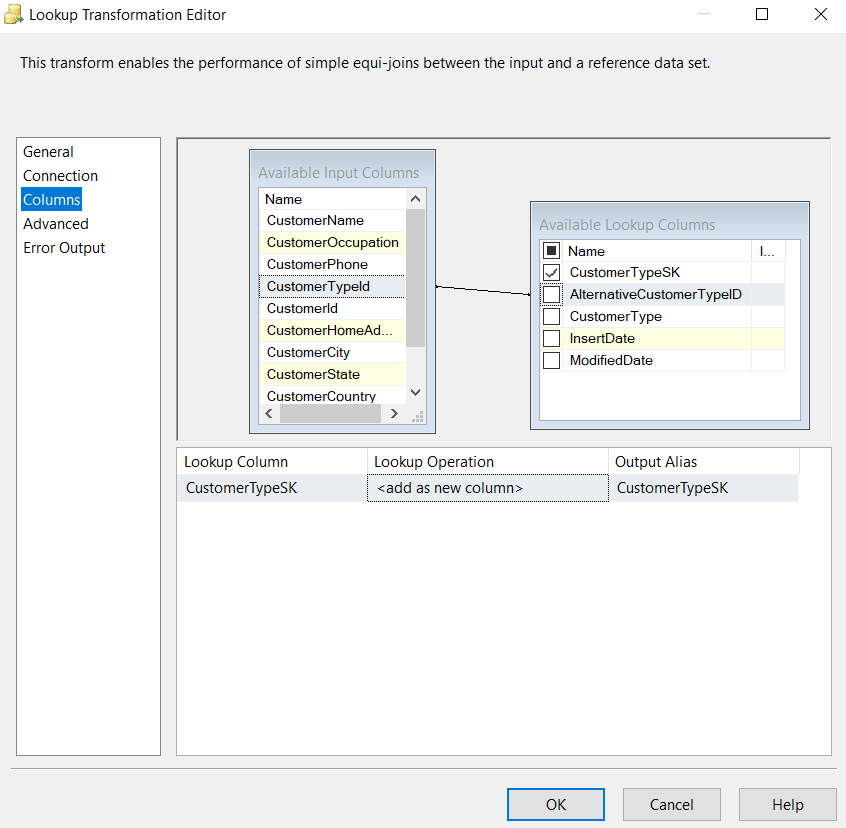
(b)Load Slowly changing dimensions

**Customer Data from Staging to Data Warehouse**

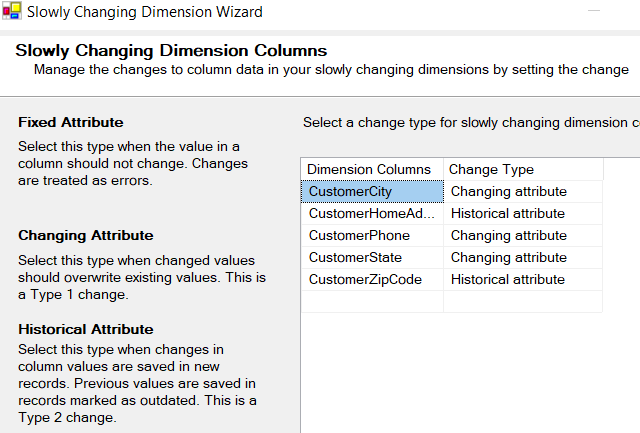
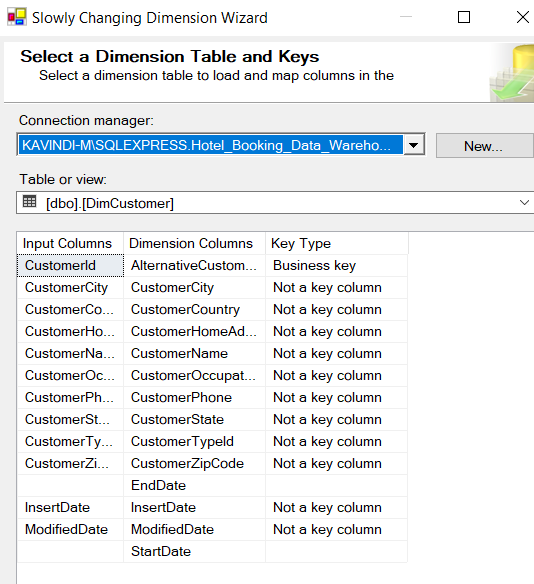
CustomerAddress staging table has been merged into Customer table. Merge has been done by sorting the common attribute of both the tables which is CustomerId.Surrogate key of the hierarchical dimension of Customer, which is DimCustomerType has been obtained through a lookup.



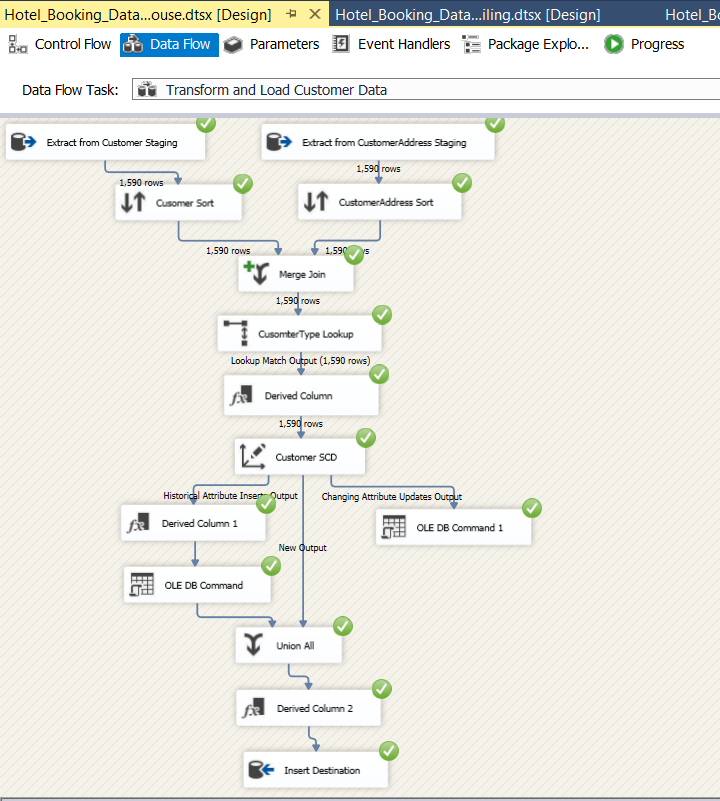
Merge Join



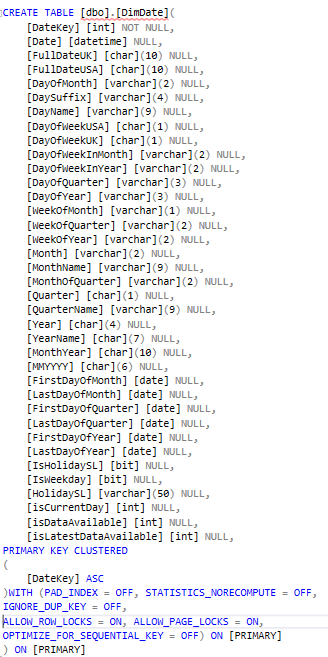
* As explained above, DimCustomer dimension has been identified as a slowly changing dimension. Hence necessary steps has been followed to make DimCustomer a slowly changing dimension



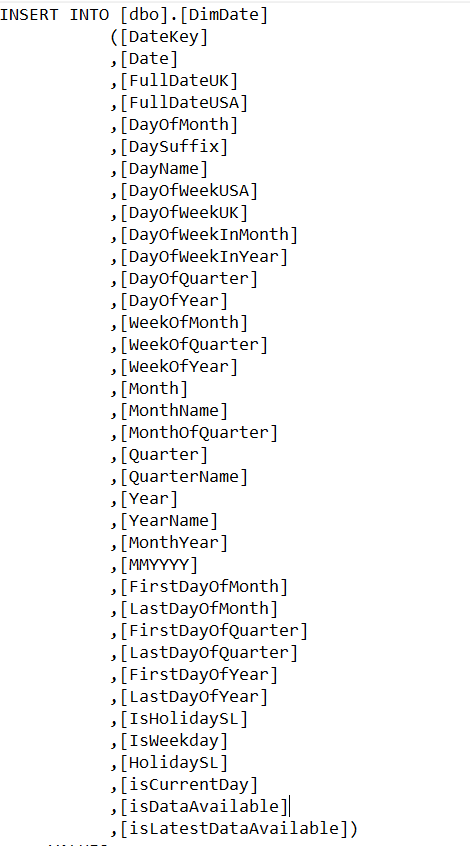
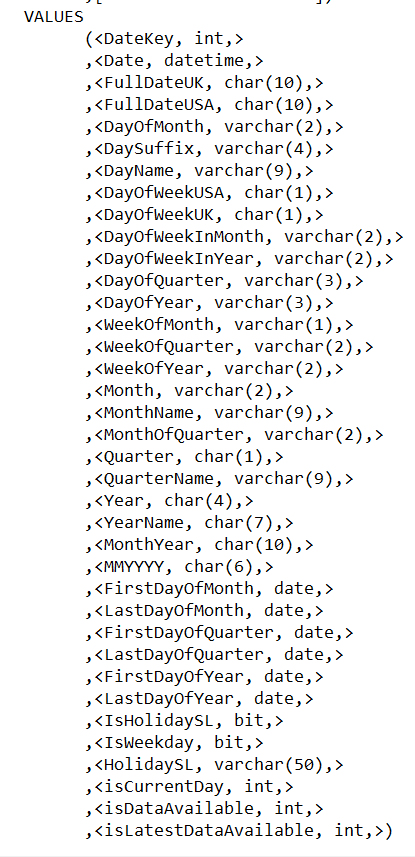
Following all the steps ,finally Customer data has been Loaded to DimCustomer table in data warehouse.

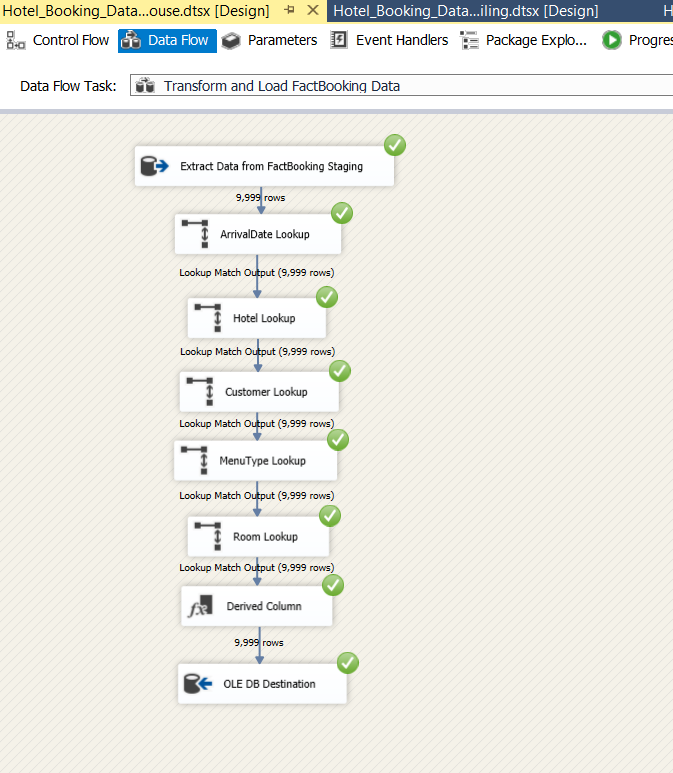


Dim Date Creation



DimDate Data inserting

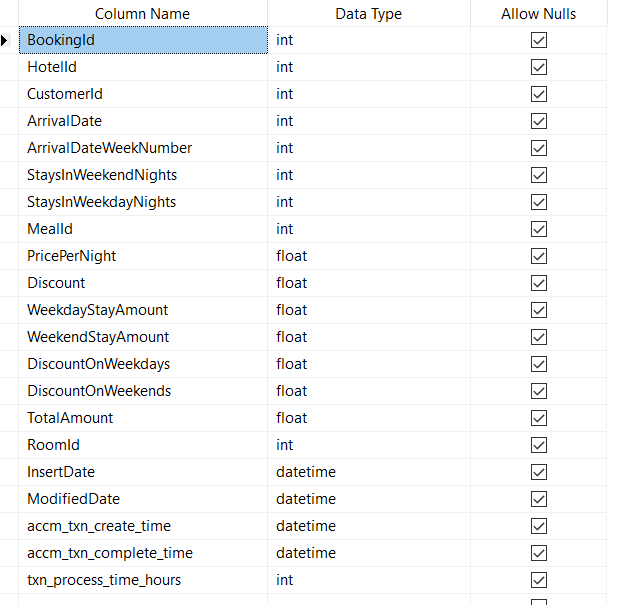
(C)Load Fact Table



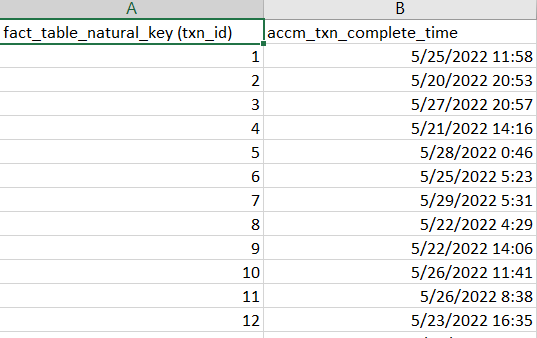
Overall ETL to data warehouse



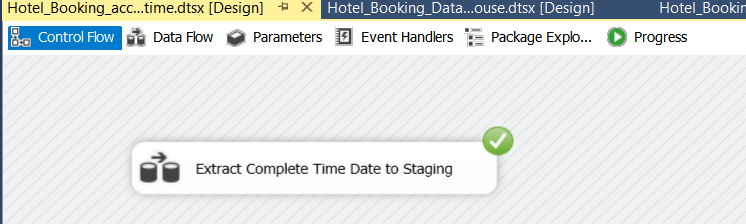
(6)ETL development – Accumilating fact tables

The fact table was extended by adding the last 3 rows as shown below.

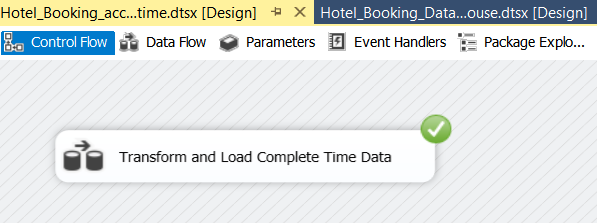
Then a separate data source (csv file) named CompletionTimw.csv was created and the structure of this file is shown below



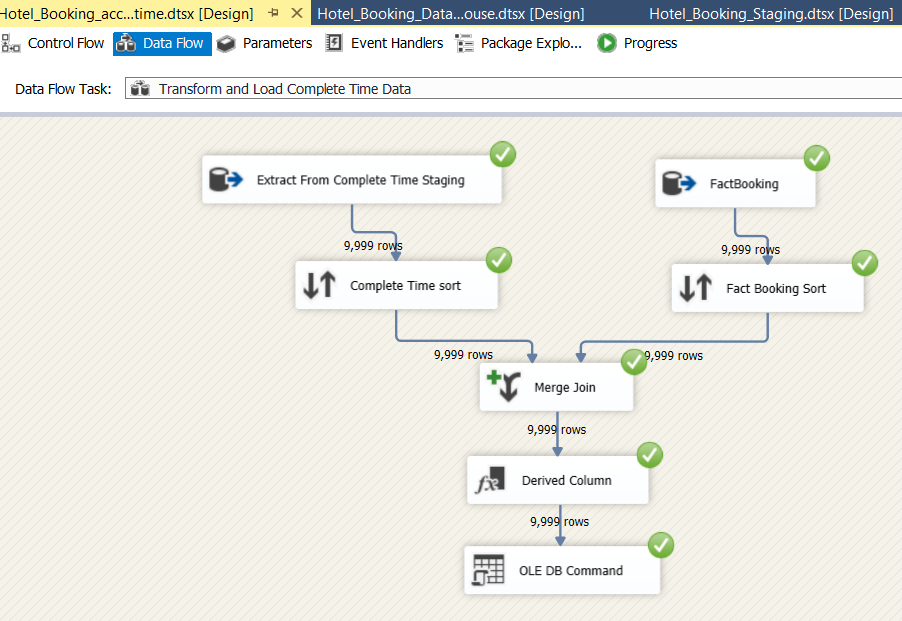
**Extract CompleteTime Data to Staging area**



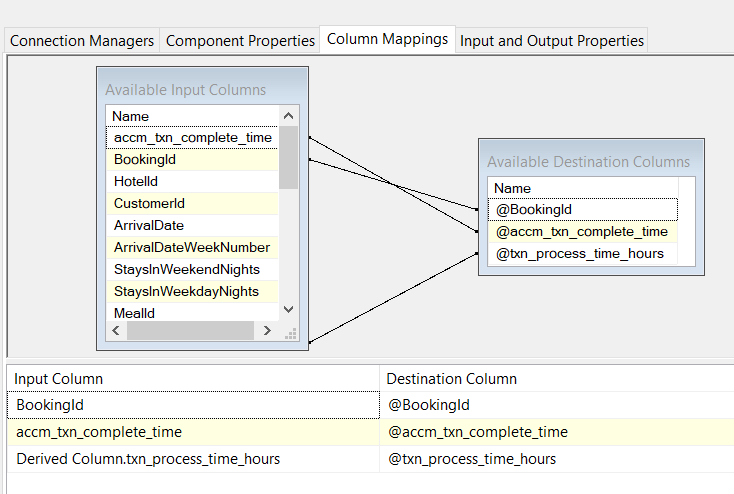
**Hotel Data from Staging to Data Warehouse**



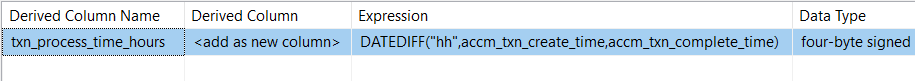
Data from the Complete Time staging table and FactBooking fact table were extracted and merged. Merge has been performed by sorting both the tables using the fields fact\_table\_natural\_key (txn\_id) & BookingId respectively. Finally, the merged data was loaded to the FactBooking fact table in the data warehouse.

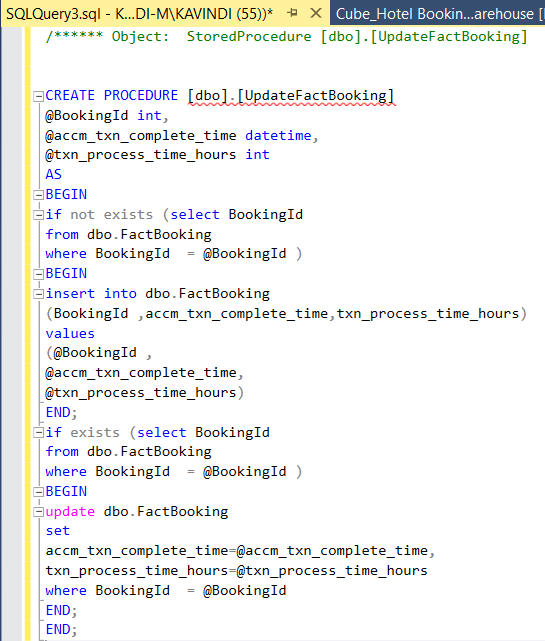


Relevant column mappings is shown below

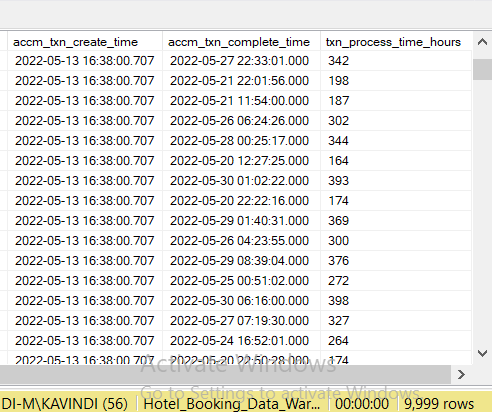


A Derived Columns task has been used to derive the values for txn\_process\_time\_hours column by getting the date difference of accn\_txn\_complete\_time & accn\_txn\_create\_time



The following procedure is used in order to load the data

A screenshot of the FactBookings fact table of the data warehouse after accumulating (Completing Step 6) is shown below.



*Thank You!*