**Sri Lanka Institute of Information Technology**



**Data Warehousing & Business Intelligence**

**Assignment 01**

**IT Number – IT20083182**

**Submitted by – Senadeera N. A. J. N.**

**Batch – Year 03 Semester 01 (Y3S1.5.1 (DS))**

Contents

[01. Data set selection 3](#_Toc105344460)

[1.1 Data set name 3](#_Toc105344461)

[1.2 About Dataset 3](#_Toc105344462)

[1.3 ER Diagram 3](#_Toc105344463)

[02. Preparation of data 4](#_Toc105344464)

[03. Solution Architecture 4](#_Toc105344465)

[04. Data Warehouse Design & Development 6](#_Toc105344466)

[05. ETL Development 7](#_Toc105344467)

[5.1 ETL – Source to Staging 7](#_Toc105344468)

[5.1.1 Load data User to staging 8](#_Toc105344469)

[5.1.2 Load data Review to staging 8](#_Toc105344470)

[5.1.3 Load data Location to staging (.txt file) 9](#_Toc105344471)

[5.1.4 Load data Hotel to staging 9](#_Toc105344472)

[5.1.5 Load data Hotel Category to staging 9](#_Toc105344473)

[5.1.6 Load data accm\_txn\_complete\_time to staging 10](#_Toc105344474)

[06. Staging to DW 10](#_Toc105344475)

[6.1 ETL System to Datawarehouse 10](#_Toc105344476)

[6.1.1 Transfer and Load DimUser Data from staging 11](#_Toc105344477)

[6.1.2 Transfer and Load DimReview Data from staging 11](#_Toc105344478)

[6.1.3 Transfer and Load DimHotelCategory Data from staging 11](#_Toc105344479)

[6.1.4 Transfer and Load DimLocation Data from staging (Slowly changing dimension) 12](#_Toc105344480)

[6.1.5 Load FactHotel Data from staging 13](#_Toc105344481)

[07. Datawarehouse Updating 13](#_Toc105344482)

[7.1 Datawarehouse updating 13](#_Toc105344483)

[7.1.1 Update FactHotel accm\_txn\_complete\_time 14](#_Toc105344484)

[7.1.2 Update FactHotel txn\_process\_time\_hours 14](#_Toc105344485)

[7.2 Accumulated Fact Table (FactHotel) 14](#_Toc105344486)

# Data set selection

* 1. Data set name**:** **Hotel Reviews**

Provided by: kaggle.com

Source link: <https://www.kaggle.com/datasets/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews_Jun19.csv>

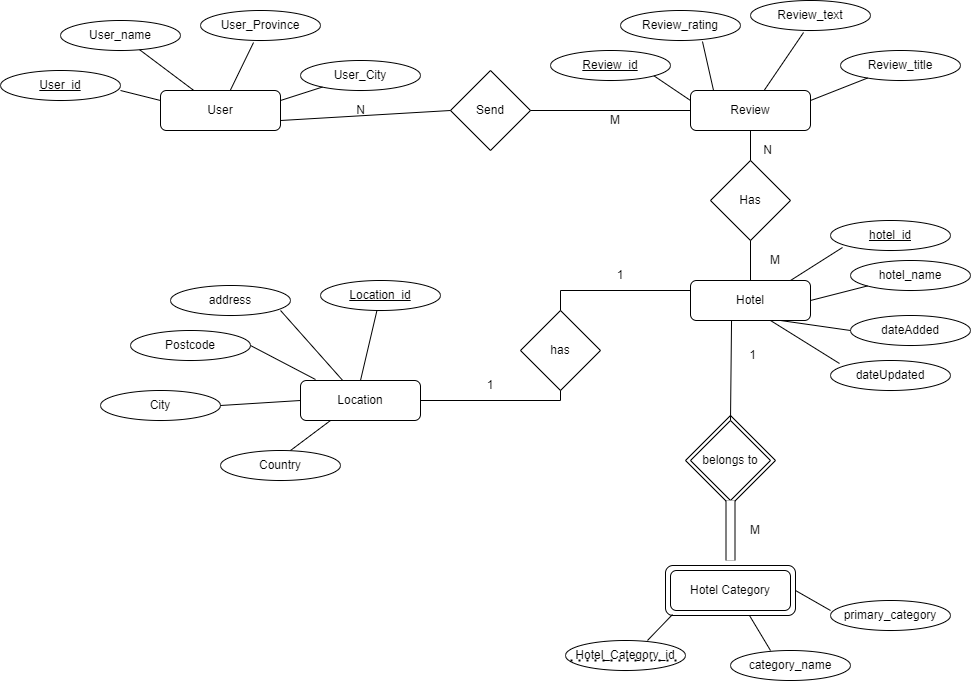
* 1. About Dataset:

This is a list of 1,000 hotels and their reviews provided by [Datafiniti's Business Database](https://datafiniti.co/products/business-data/). The dataset includes hotel location, name, rating, review data, title, username, and more.

You can use this data to [compare hotel reviews on a state-by-state basis](https://datafiniti.co/state-state-comparison-hotel-reviews/); experiment with sentiment scoring and other natural language processing techniques. The review data lets you correlate keywords in the review text with ratings. E.g.:

* What are the bottom and top states for hotel reviews by average rating?
* What is the correlation between a state’s population and their number of hotel reviews?
* What is the correlation between a state’s tourism budget and their number of hotel reviews?

## ER Diagram



# Preparation of data

All the data sources are provided in csv format by the web site. In preparation of data sources, some changes have done for the source format (some columns were added, separated into another table) of the given files as converting into text files and importing csv files into a source database.

Final State of Preparation of the source data formats before Transforming data =>

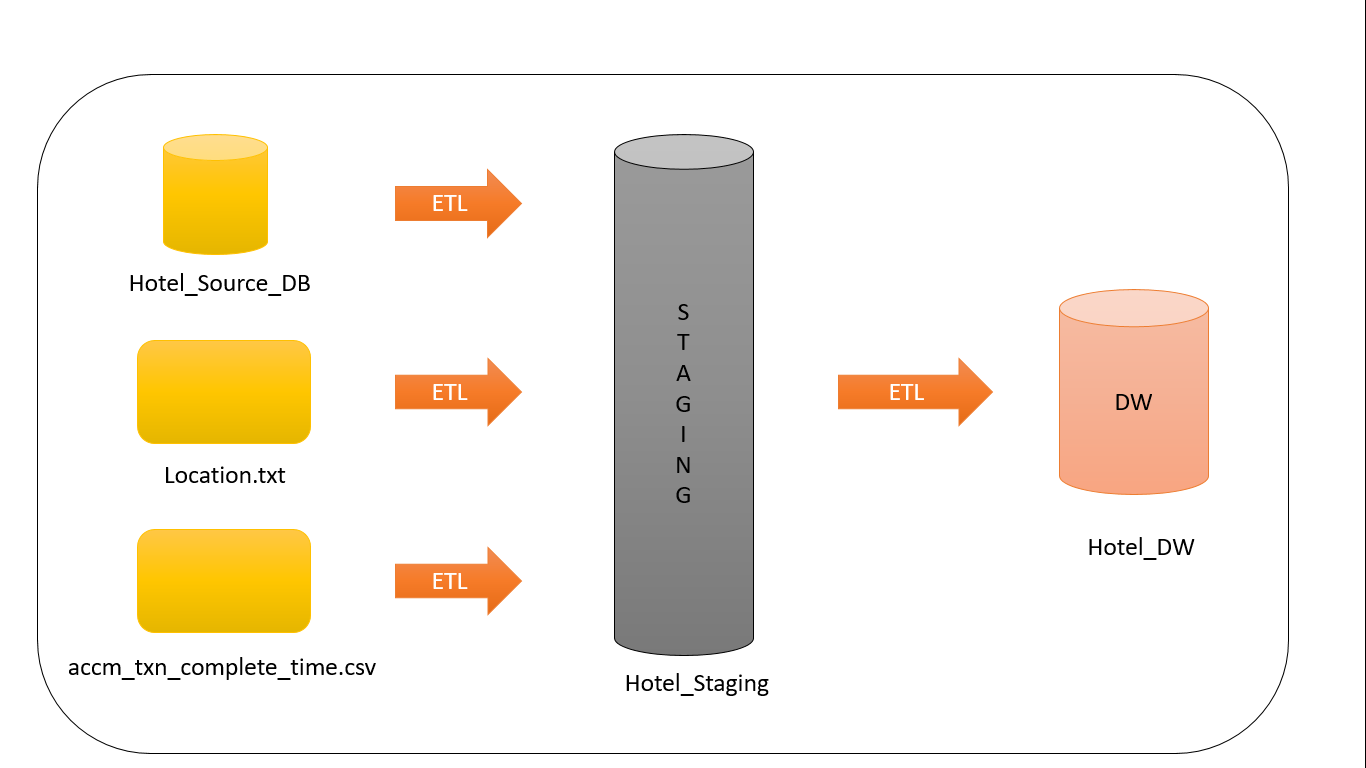
1. CSV files (.csv)
   * Hotel CSV File.
   * Hotel Category CSV File.
   * User CSV File.
   * Review CSV File.
   * accm\_txn\_complete\_time CSV

These csv files are imported into SSMS, database created as **Hotel\_Source\_DB** database.

1. Text file (.txt)

* Location text file

# Solution Architecture



**Hotel\_Staging.**

* accm\_txn\_complete\_time
* stgHotel
* stgHotalCategory
* stgReview
* StgUser
* stgLocation

**Hotel\_DW**

* DimDate
* DimHotelCategory
* DimLocation
* DimReview
* DimUser
* FactHotel

**Architecture Components.**

* Data Sources.

Operational System (**Accumulating**).

External Sources.

* Extract, Transform and Load.

Extract – reading data from source systems.

Transform – Combine data from multiple sources, De-duplicating.

* Data Warehouse

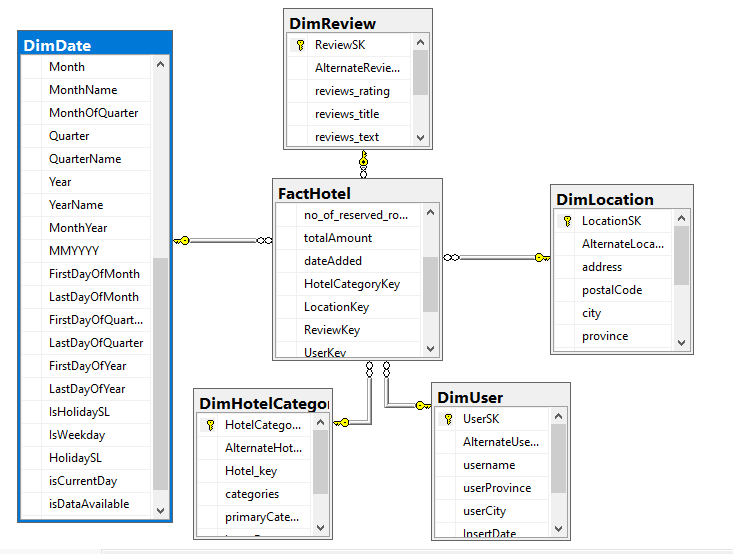
EDW and Data Mart.

Dimensional Modeling- Facts and Dimensions.

Many schemas – In here I use star schema.

# Data Warehouse Design & Development

**Relational Diagram – Star Schema**



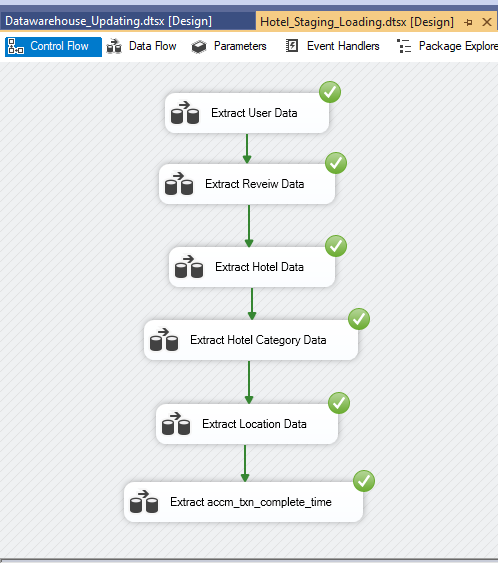
**DimLocation** is **slowly changing dimensions**. Address and city may be changed in future. Therefore, I get it as slowly changing attribute.

**Address -> PostalCode -> City -> Province ->Country**

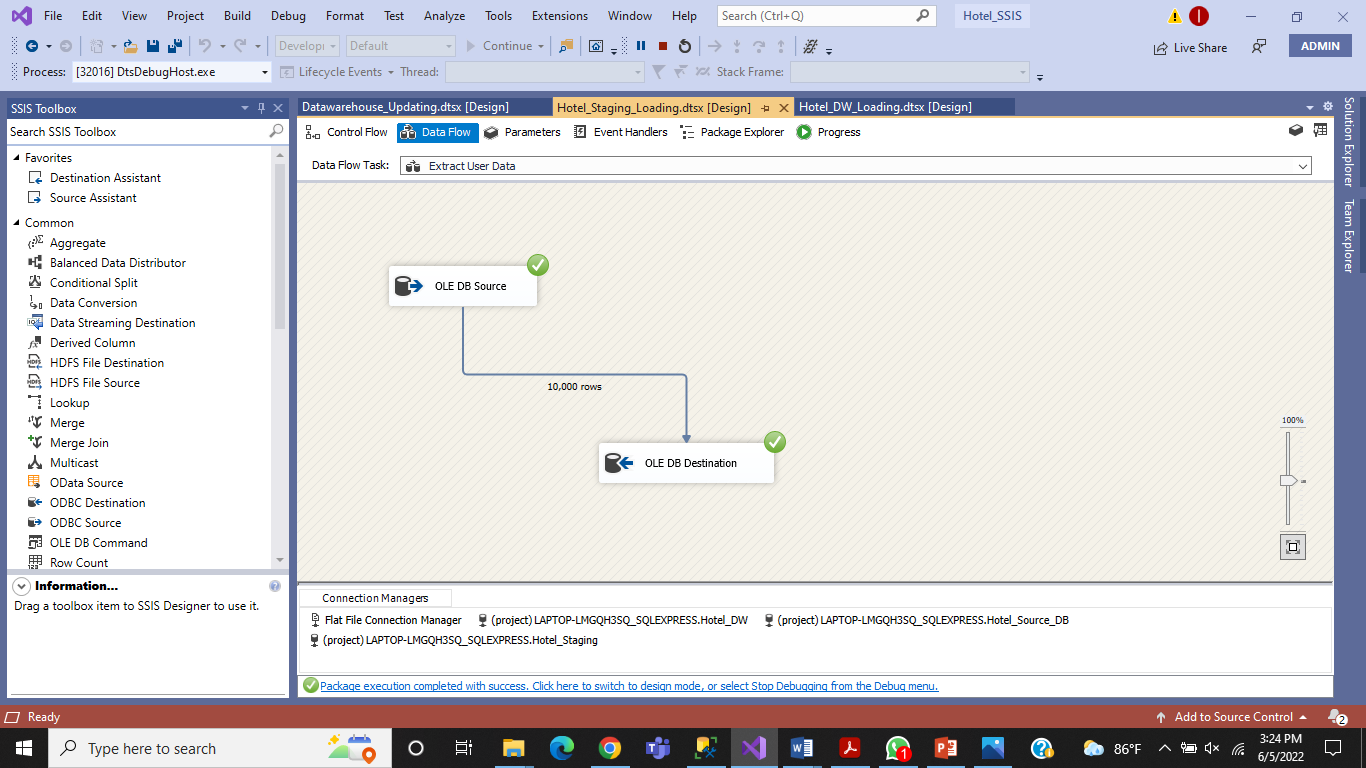
**This is the Hierarchies (Location table.)**

# ETL Development

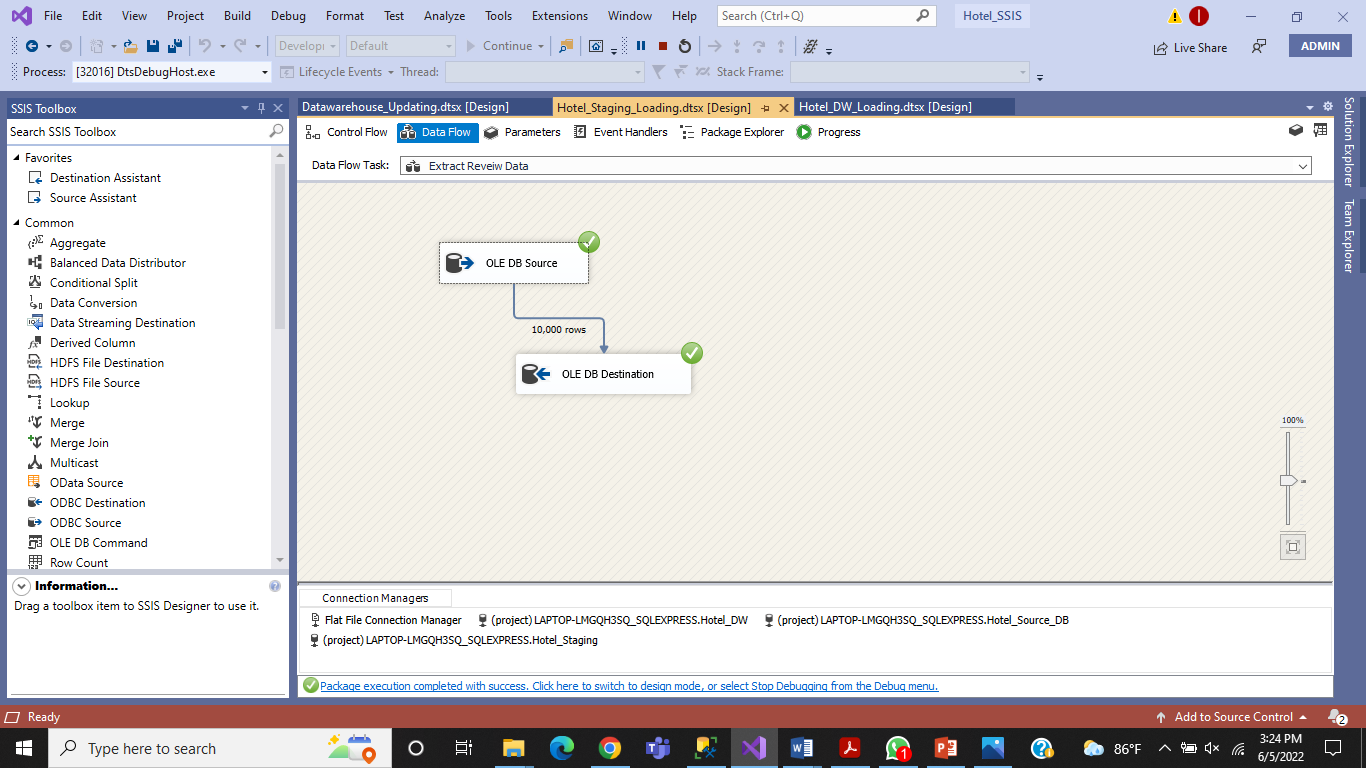
## 5.1 ETL – Source to Staging

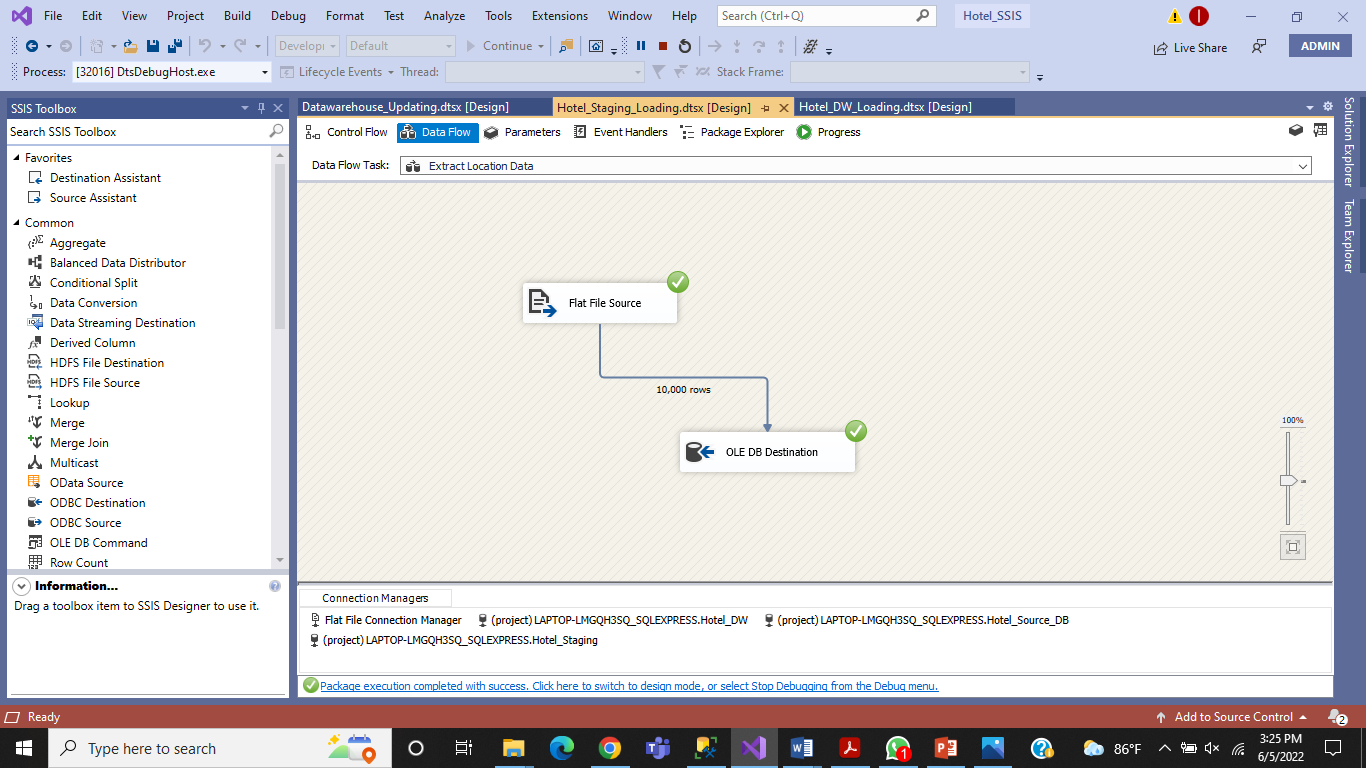


### 5.1.1 Load data User to staging

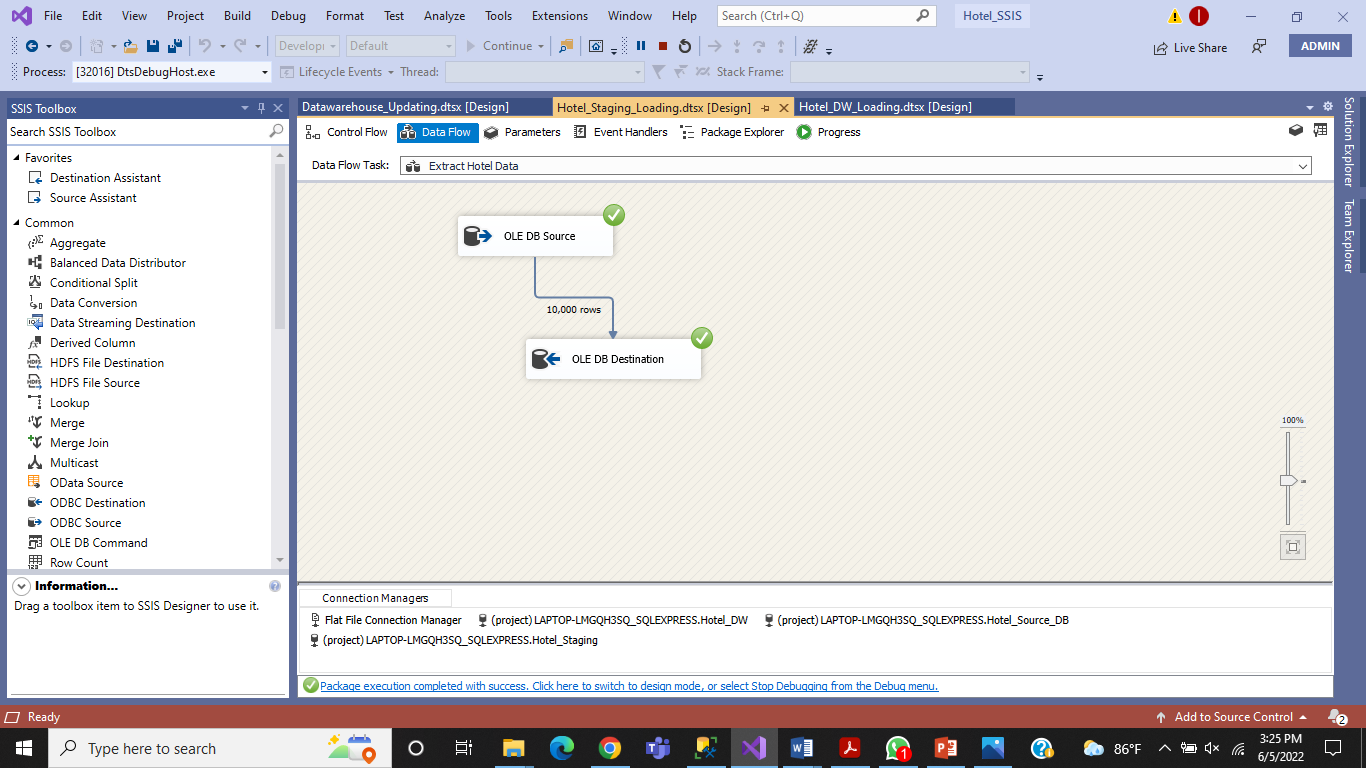


### 5.1.2 Load data Review to staging

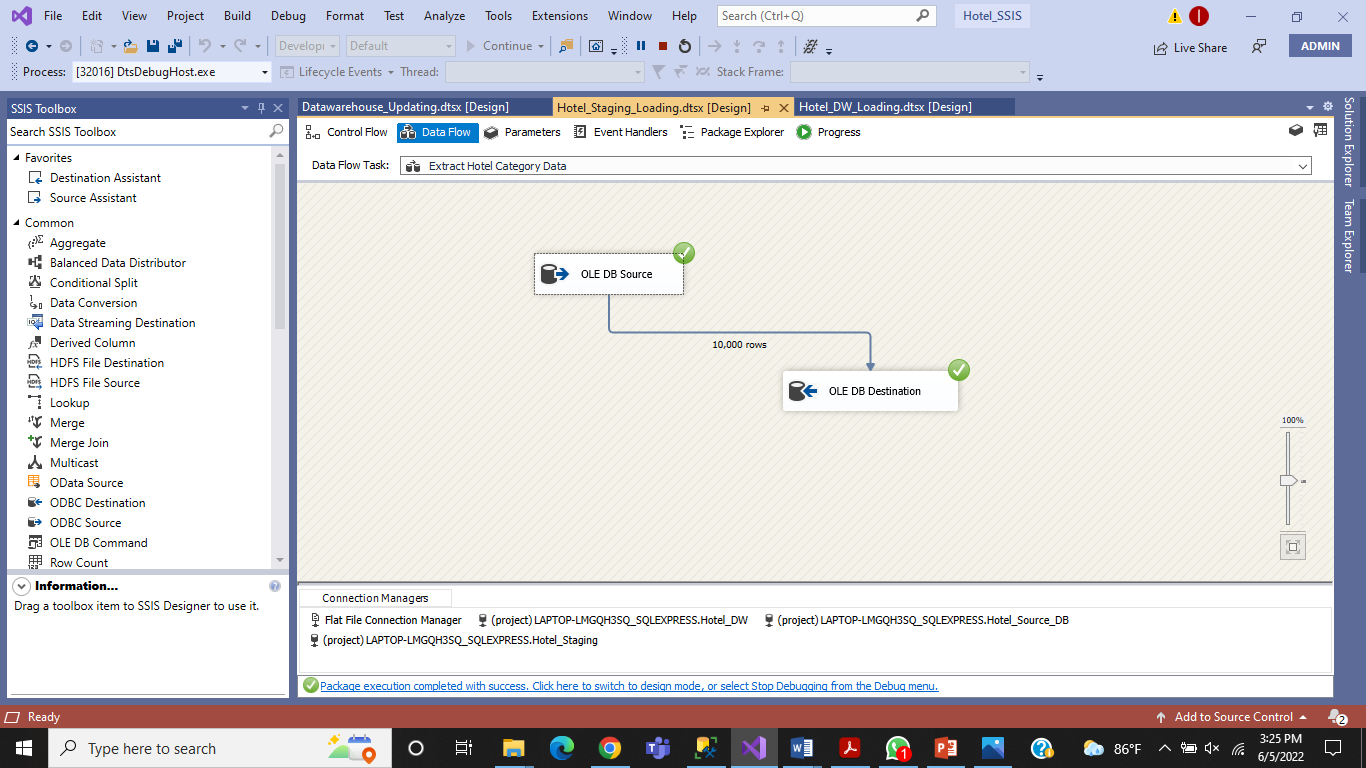


5.1.3 Load data Location to staging (.txt file)

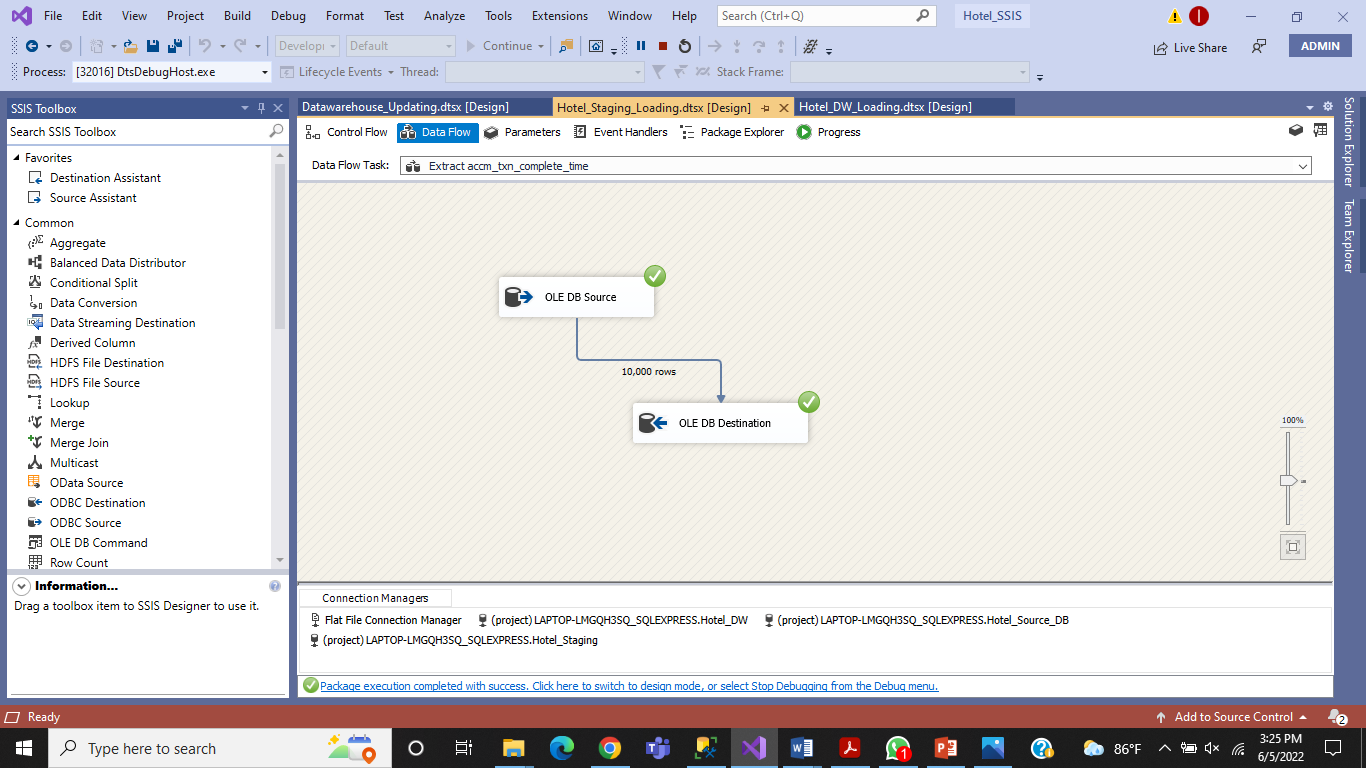
### 5.1.4 Load data Hotel to staging



### 5.1.5 Load data Hotel Category to staging

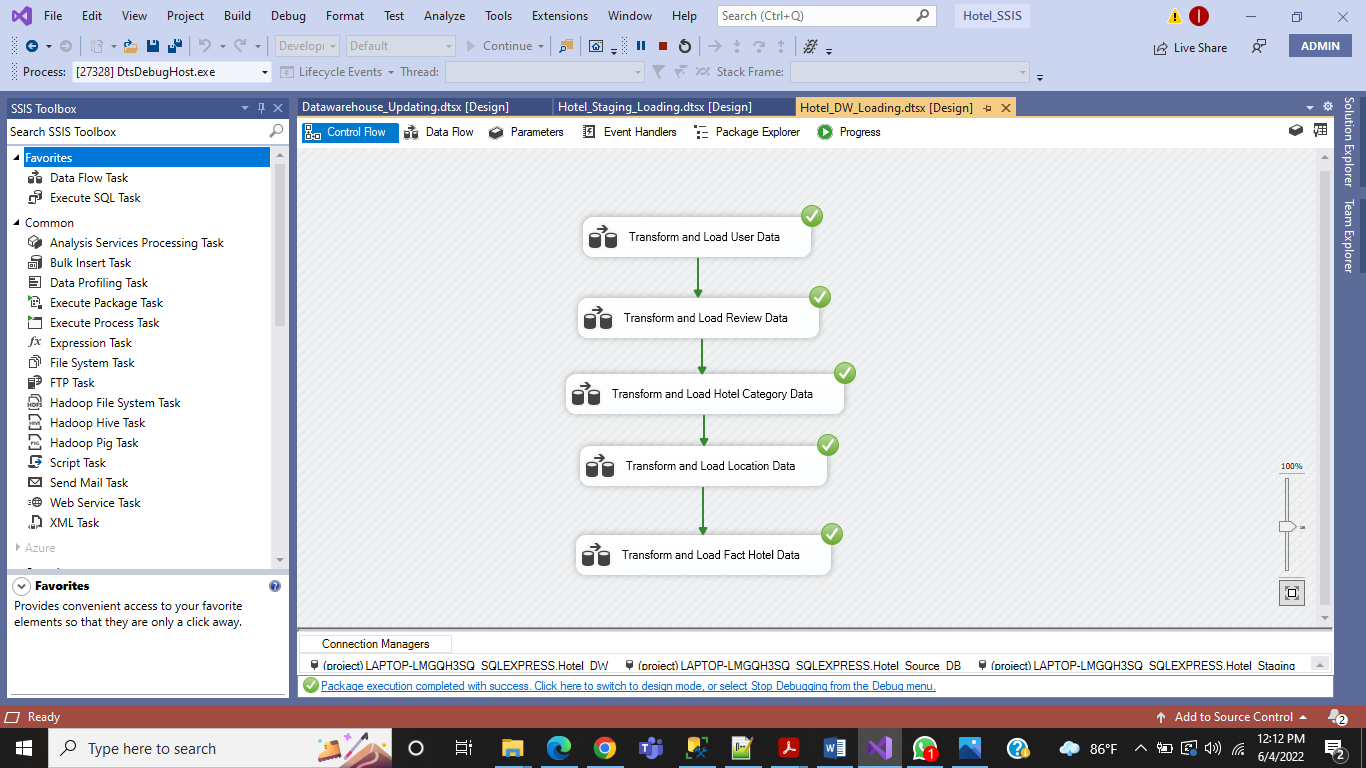


### 5.1.6 Load data accm\_txn\_complete\_time to staging

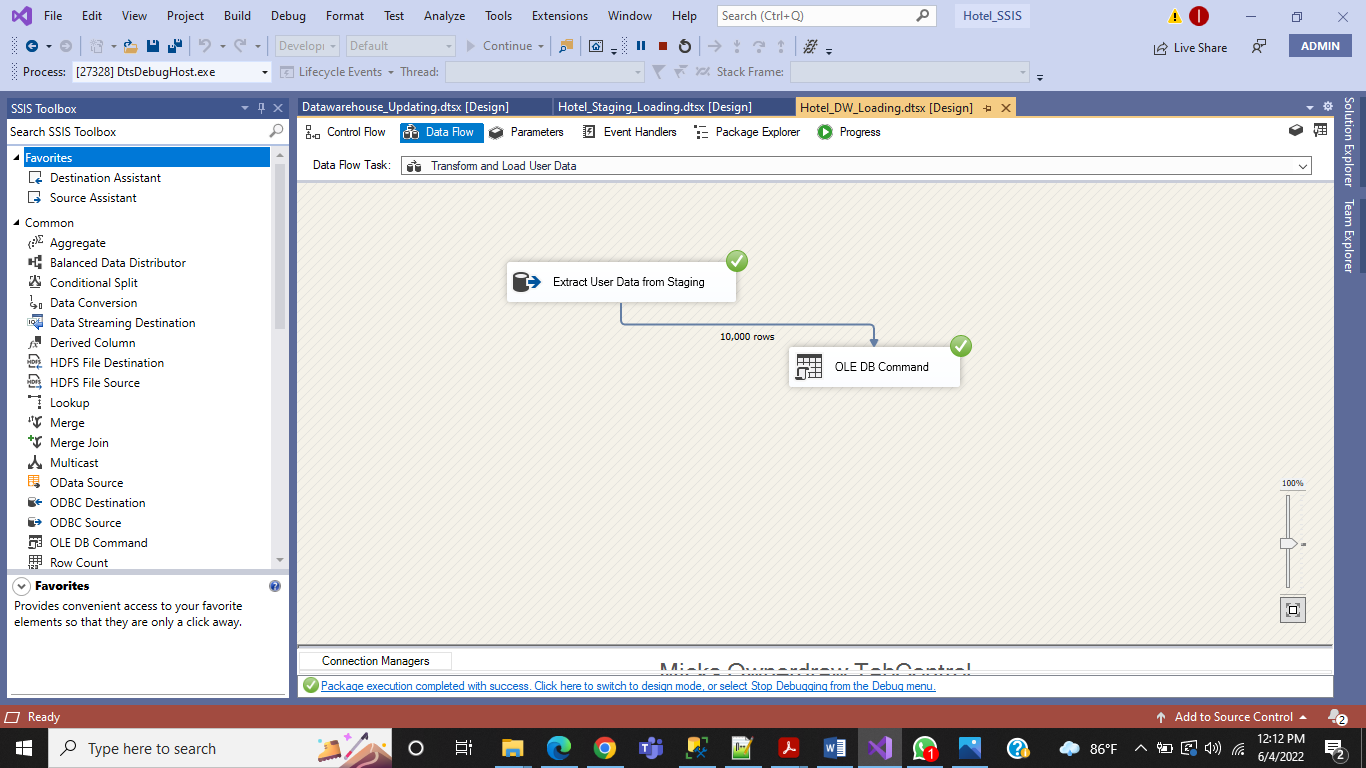


# Staging to DW

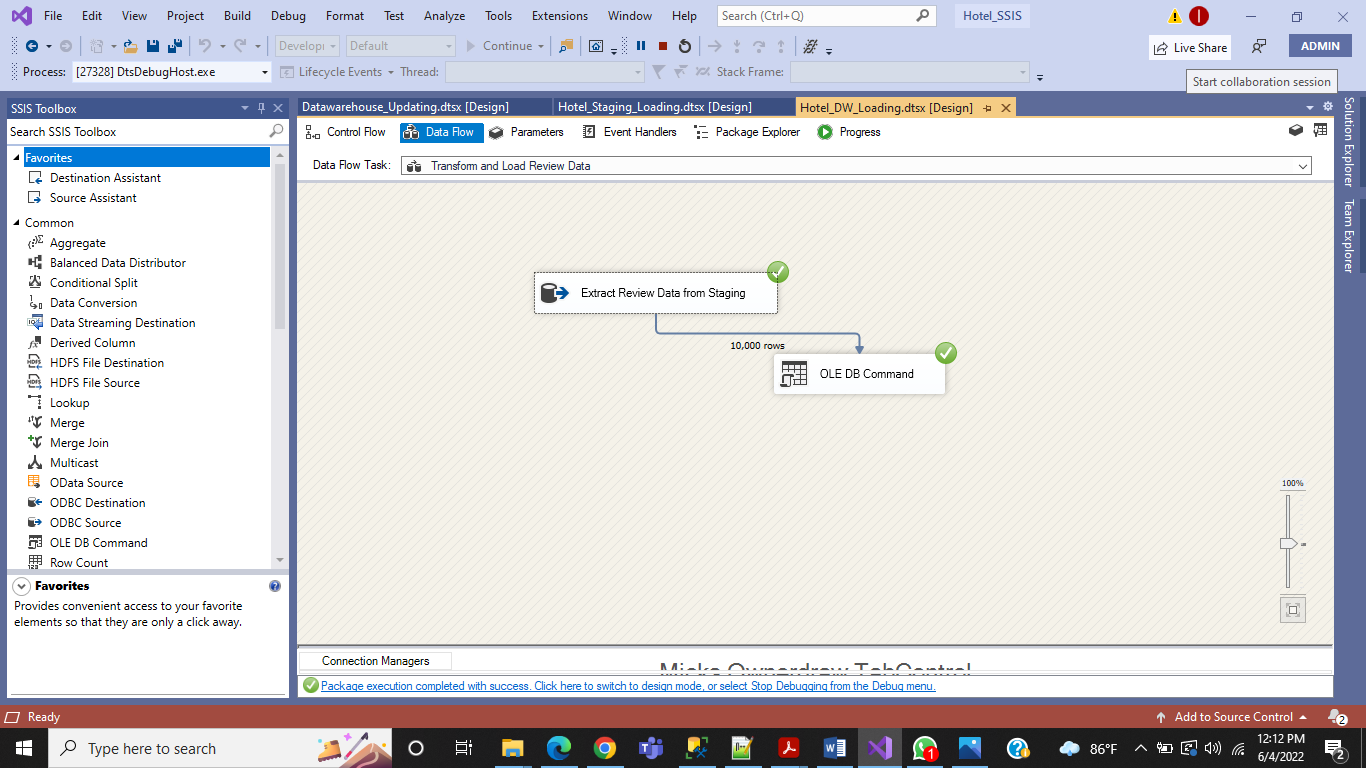
## 6.1 ETL System to Datawarehouse



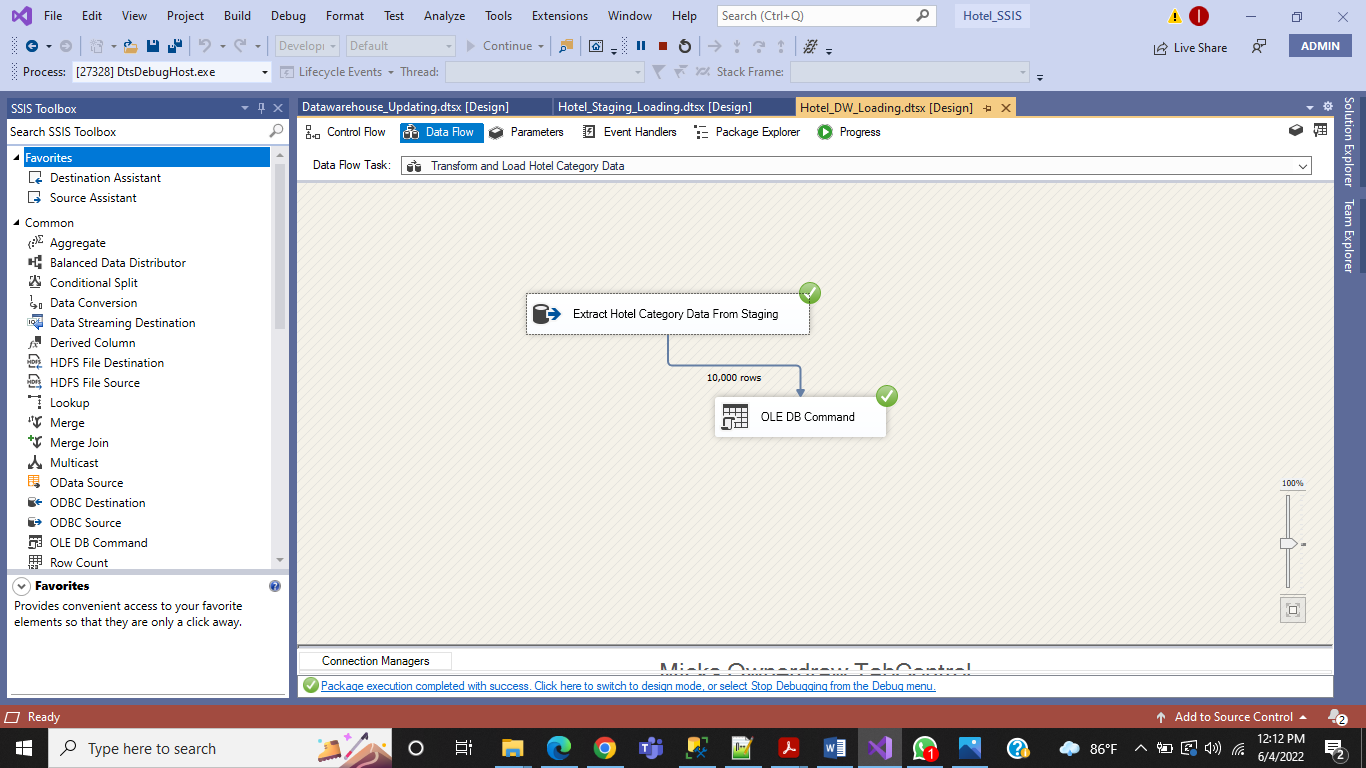
### 6.1.1 Transfer and Load DimUser Data from staging



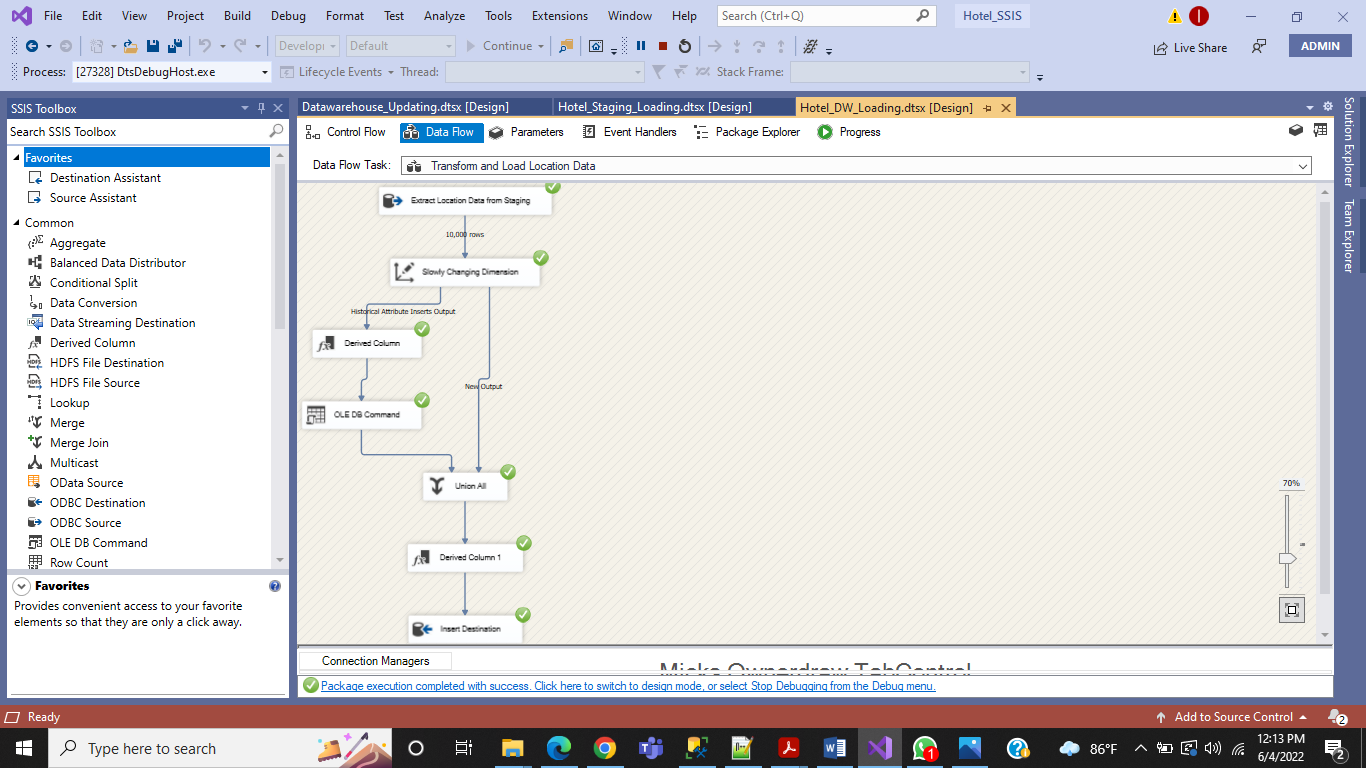
### 6.1.2 Transfer and Load DimReview Data from staging



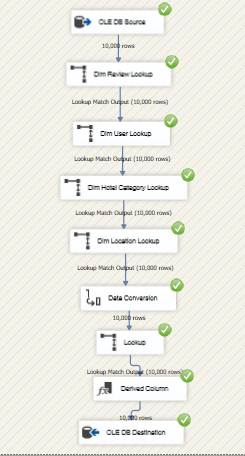
### 6.1.3 Transfer and Load DimHotelCategory Data from staging



### 6.1.4 Transfer and Load DimLocation Data from staging (Slowly changing dimension)



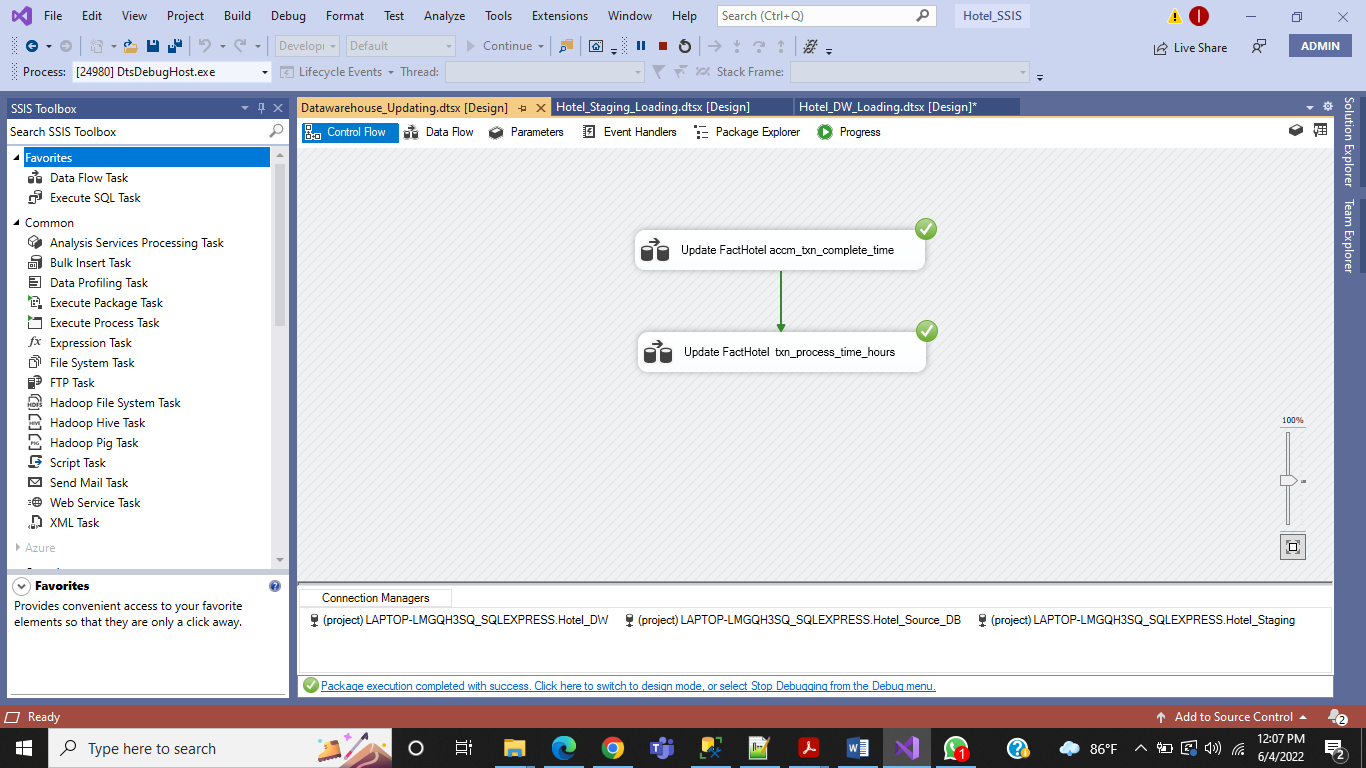
### 6.1.5 Load FactHotel Data from staging

****

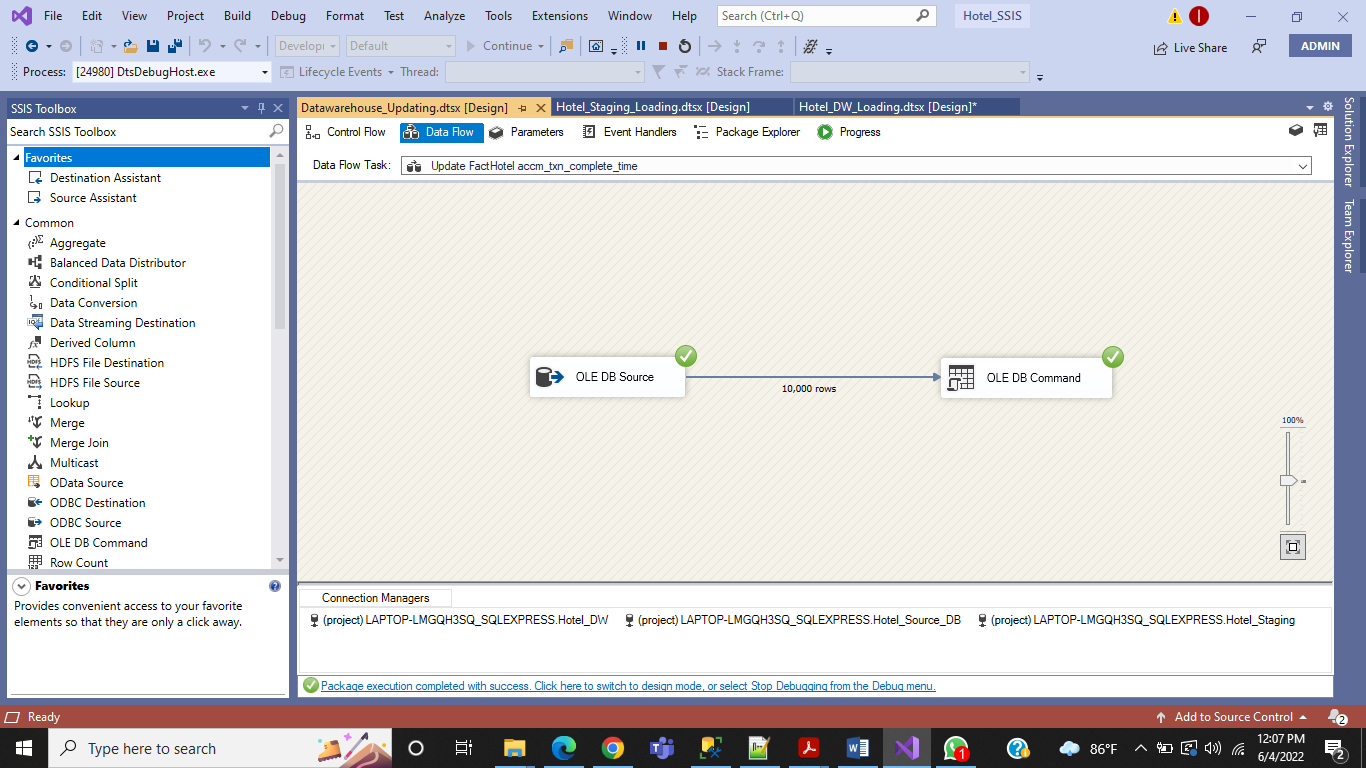
# Datawarehouse Updating

In order to creating Accumulated fact table I created a new SSIS package and updated accm\_txn\_complete\_time and txn\_process\_time\_hours.

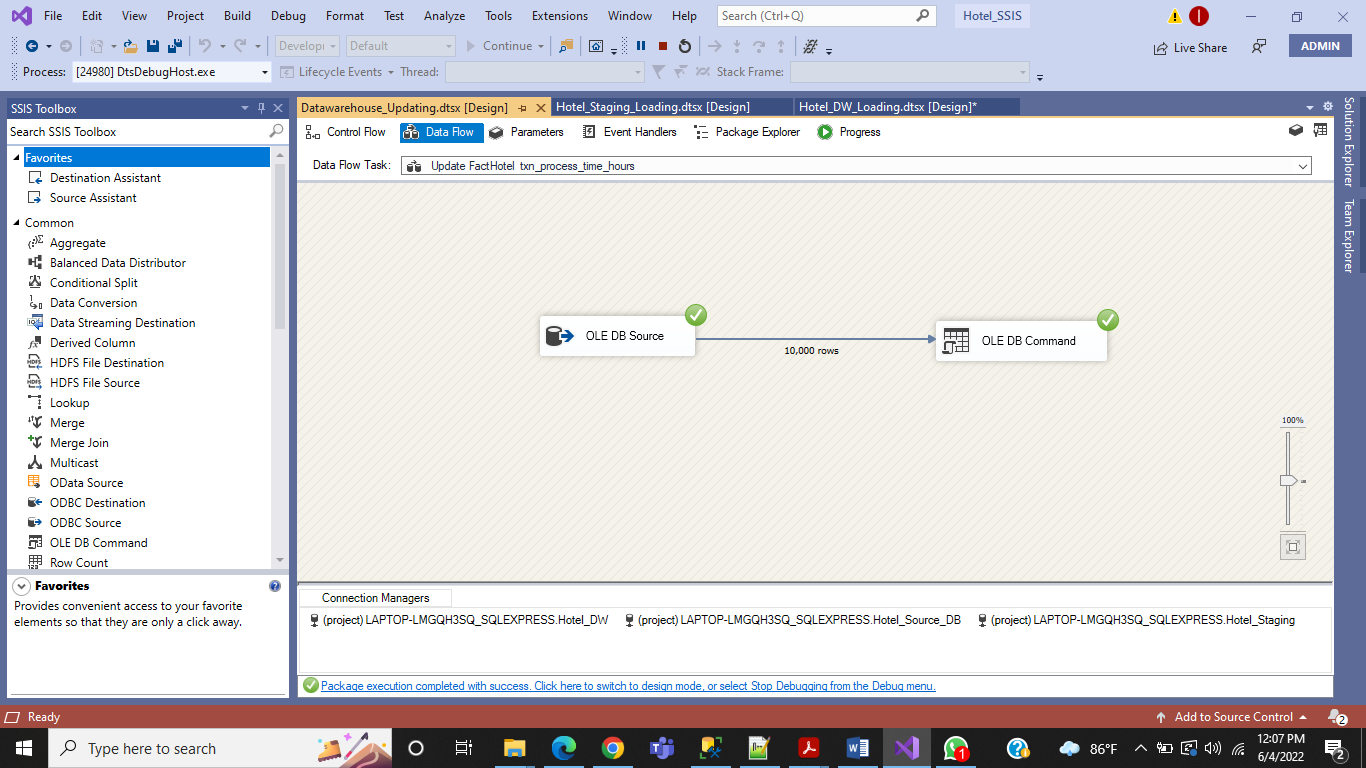
## 7.1 Datawarehouse updating

****

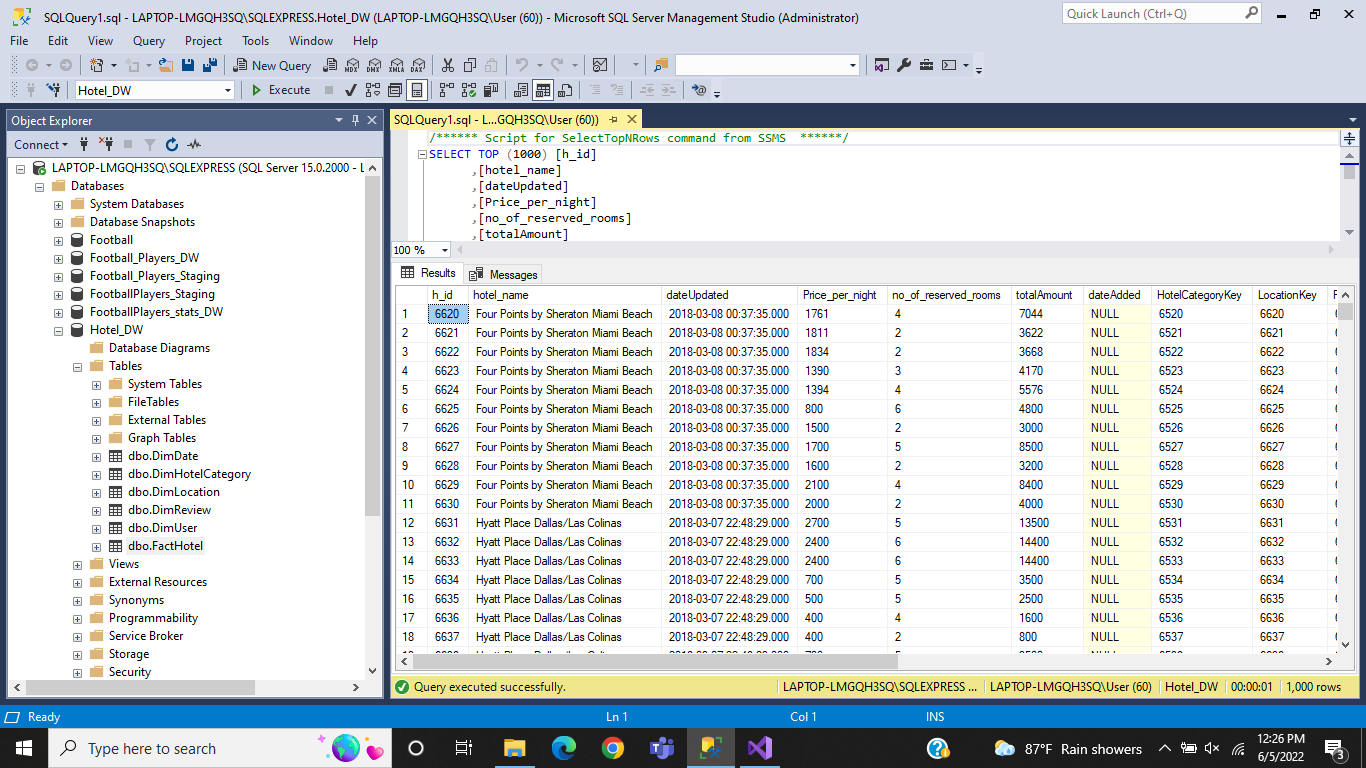
### 7.1.1 Update FactHotel accm\_txn\_complete\_time

****

### 7.1.2 Update FactHotel txn\_process\_time\_hours

****

## 7.2 Accumulated Fact Table (FactHotel)

****

