**Sri Lanka Institute of Information**

**Technology**



**Data Warehousing & Business Intelligence**

**Assignment 01**

**Submitted by:**

**Gunathilaka U.K.R.S.W(it20188818)**

Contents

[1.Data Set Selection. 3](#_Toc103618829)

[Background 3](#_Toc103618830)

[Content 3](#_Toc103618831)

[ER Diagram 4](#_Toc103618832)

[2. Preparation of Data Sources. 5](#_Toc103618833)

[3. Solution Architecture. 6](#_Toc103618834)

[4. Data warehouse design & development. 8](#_Toc103618835)

[5. ETL Development. 9](#_Toc103618836)

[5.1 ETL –Source To Staging 9](#_Toc103618837)

[5.1.1 Load data User to staging 9](#_Toc103618838)

[5.1.2 Load data Review to staging 10](#_Toc103618839)

[5.1.3 Load data Hotel to staging 10](#_Toc103618840)

[5.1.4 Load data Hotel Category to staging 11](#_Toc103618841)

[5.1.5 Load data location to staging(.txt file) 11](#_Toc103618842)

[5.1.6 Load data accm\_txn\_complete\_time to staging 12](#_Toc103618843)

[6. Staging To DW. 13](#_Toc103618844)

[6.1 ETL System to Datawarehouse 13](#_Toc103618845)

[6.1.1. Transfer and Load DimUser Data from staging 13](#_Toc103618846)

[6.1.2. Transfer and Load DimReview Data from staging 14](#_Toc103618847)

[6.1.3. Transfer and Load DimHotelCategory Data from staging 14](#_Toc103618848)

[6.1.4. Transfer and Load DimLocation Data from staging (Slowly changing diamention) 15](#_Toc103618849)

[6.1.5. Load FactHotel Data from staging 16](#_Toc103618850)

[7. Datawarehouse Updating 17](#_Toc103618851)

[7.1. Datawarehouse updating 17](#_Toc103618852)

[7.1.1 Update factHotel accm\_txn\_complete\_time 18](#_Toc103618853)

[7.1.2 Update factHotel txn\_process\_time\_hours 18](#_Toc103618854)

[7.2 Accumulated Fact Table (FactHotel) 19](#_Toc103618855)

# 1.Data Set Selection.

## Background

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?

## Content

The data set was downloaded from -:

[**https://www.kaggle.com/datasets/datafiniti/hotel-reviews?select=Datafiniti\_Hotel\_Reviews.csv**](https://www.kaggle.com/datasets/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews.csv)

## Diagram Description automatically generatedER Diagram

# Preparation of Data Sources.

In order to data extraction need to prepare the data sources. From my main data source, I have extracted to type of data sources.

1. Text file (.txt)

* Location text file

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

From these dataset, I import above csv files to SQL SERVER MANAGEMENT STUDIO and create the database called **IT20188818\_sourceDB** Database.

**IT20188818\_sourceDB** have following tables -:

* + accm\_txn\_complete\_time table.
  + Hoteltbl Table.
  + HotelCategorytbl Table.
  + Usertbl Table.
  + Reviewtbl Table.

**Text file** –This text file include all the hotel address details including address, Postal code ,city, province and the country.

**ETL**

IT20188818\_SourceDB

**S**

**T**

**A**

**G**

**I**

**N**

**G**

**ETL**

**ETL**

Location.txt

**ETL**

DatawareHouse

(IT20188818\_DW )

accm\_txn\_complete\_time.csv

IT20188818\_Staging

# Solution Architecture.

**3.1 IT20188818\_StagingDB.**

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

**3.2 IT20188818\_DW**

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

**3.3 Architure 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.

**4.1 Relational Diagram – Star Schema.**

Diagram, schematic

Description automatically generated

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

**Address -> PostalCode -> City -> Province ->Country This is the Hierachies (Location table.)**

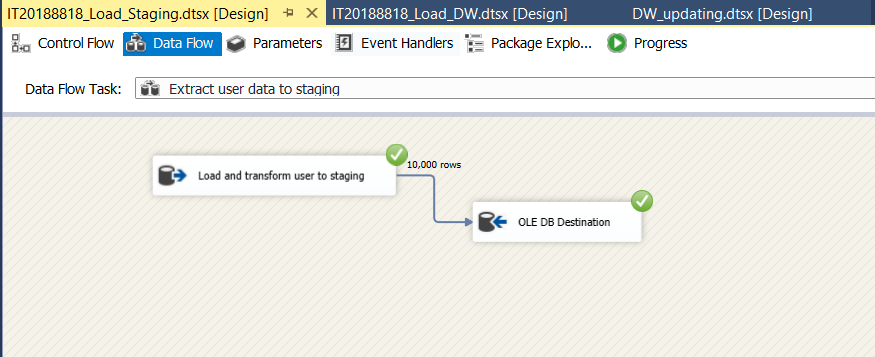
# ETL Development.

## 5.1 ETL –Source To Staging

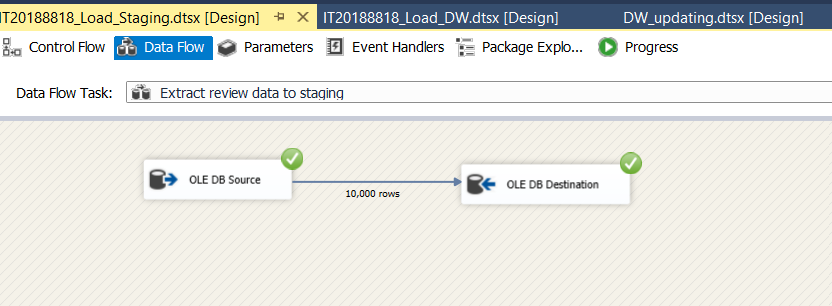
Graphical user interface, text, chat or text message

Description automatically generated

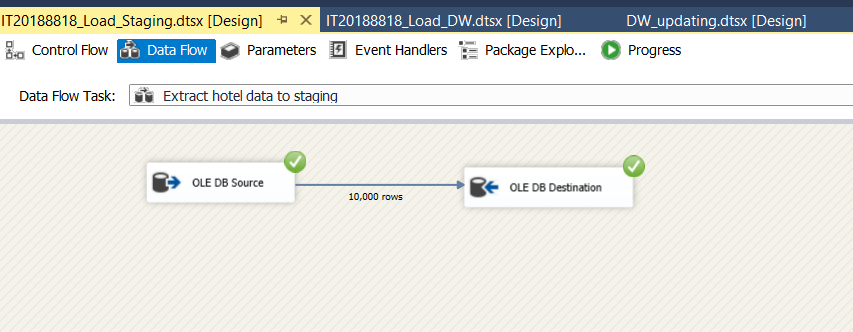
## 5.1.1 Load data User to staging



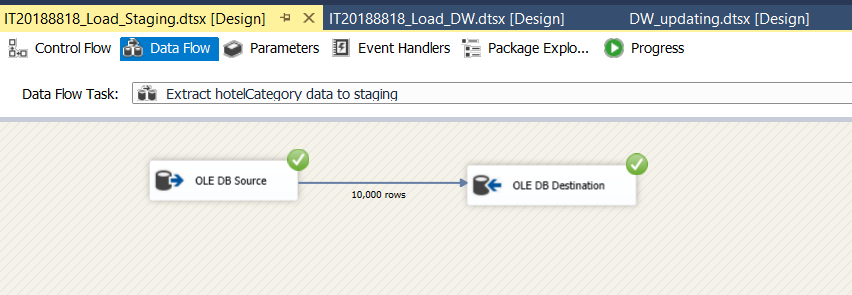
## 5.1.2 Load data Review to staging



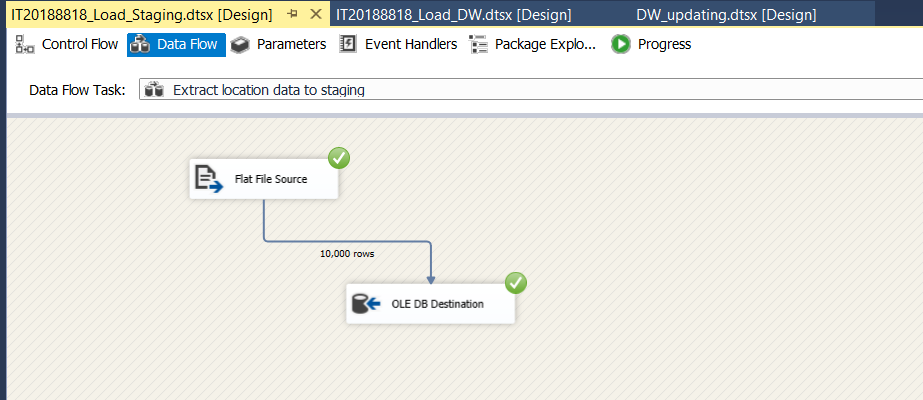
## 5.1.3 Load data Hotel to staging



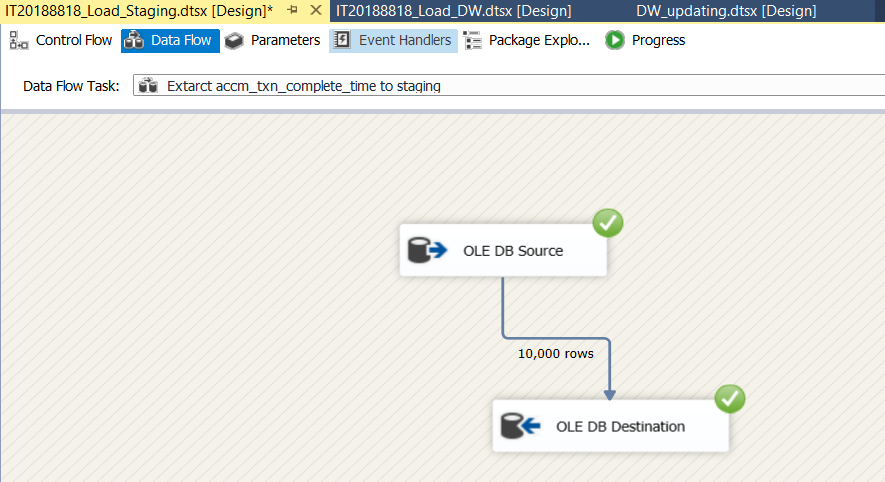
## 5.1.4 Load data Hotel Category to staging



## 5.1.5 Load data location to staging(.txt file)



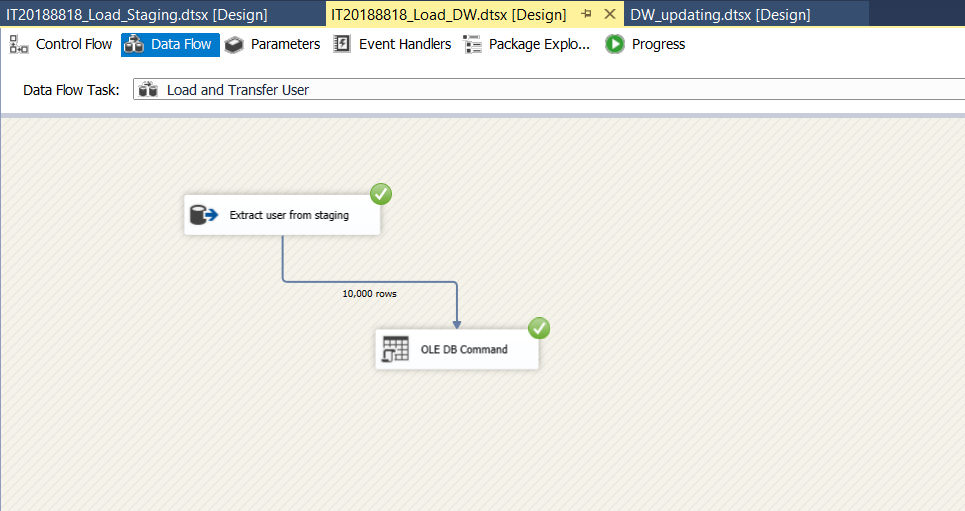
## 5.1.6 Load data accm\_txn\_complete\_time to staging



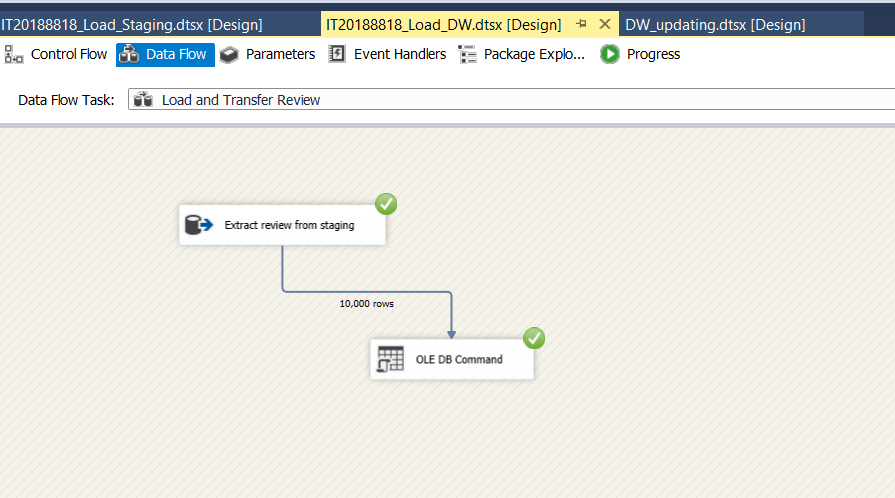
# Staging To DW.

## 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

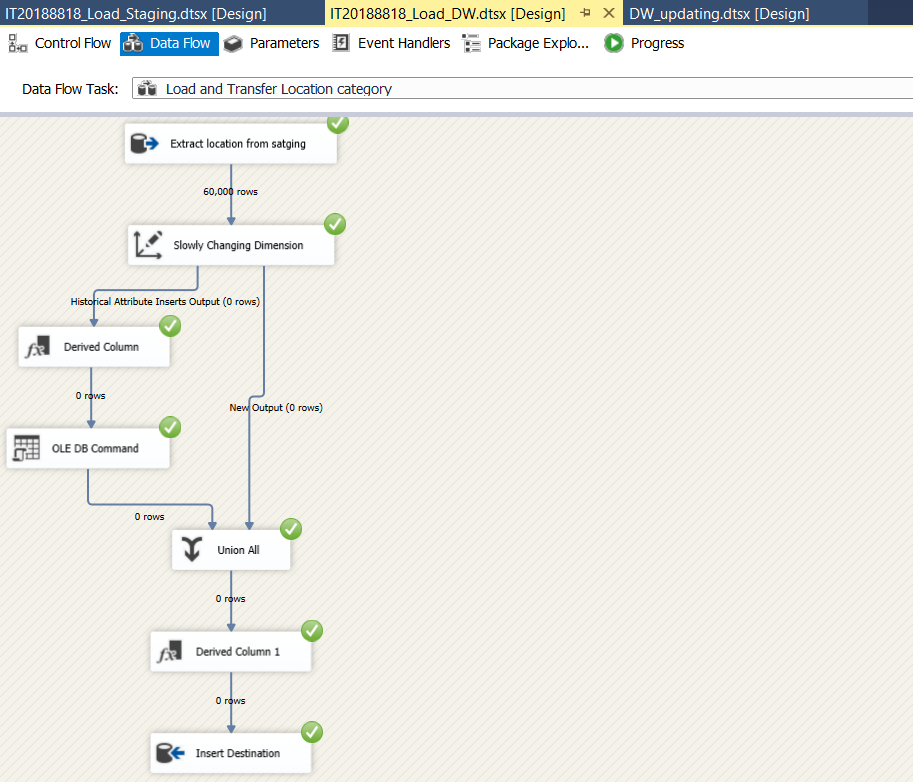


## 6.1.3. Transfer and Load DimHotelCategory Data from staging

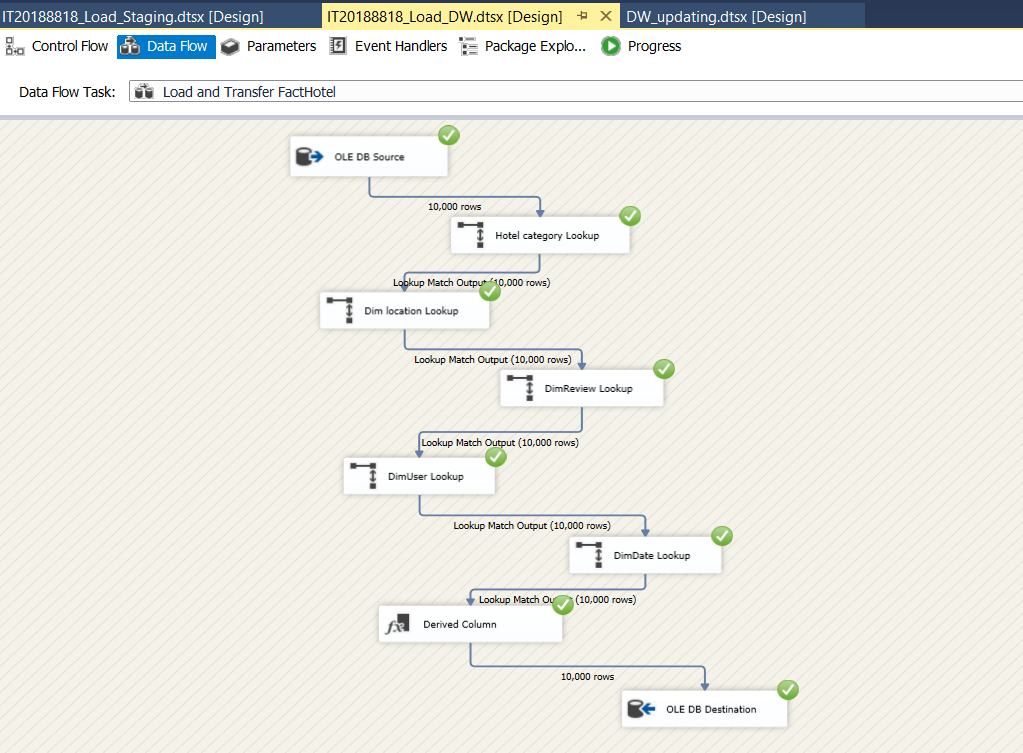
Graphical user interface, text, application, chat or text message

Description automatically generated

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



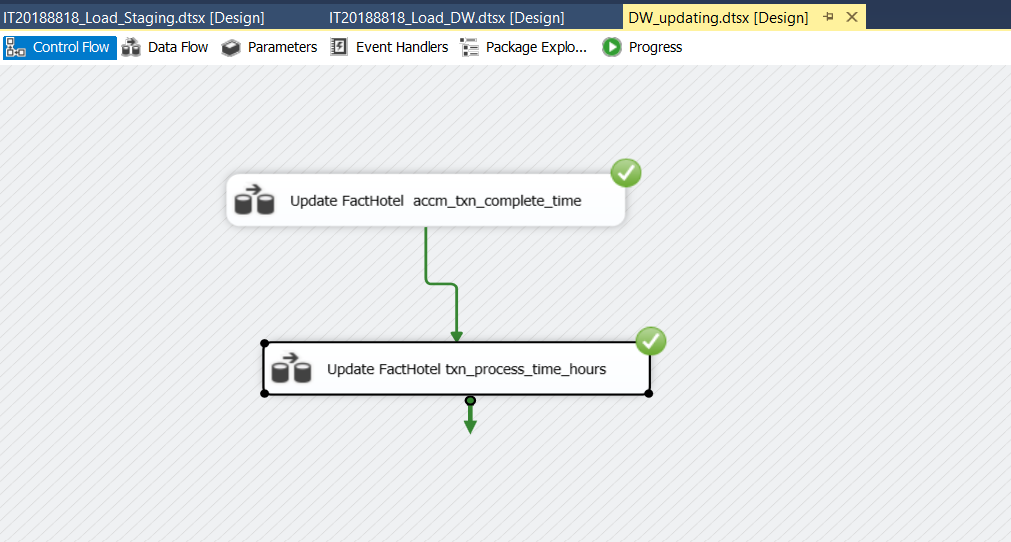
## 6.1.5. Load FactHotel Data from staging



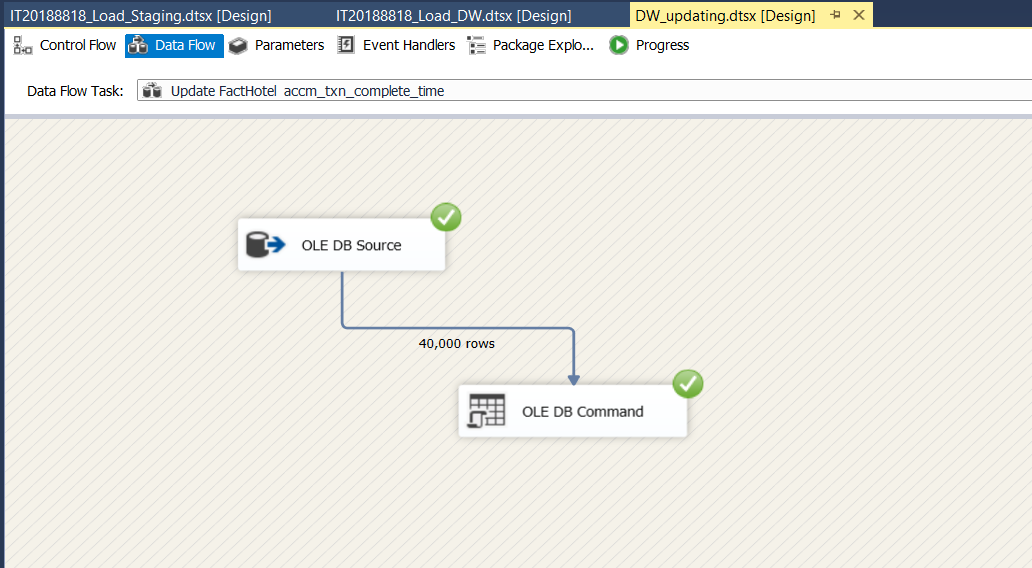
# Datawarehouse Updating

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

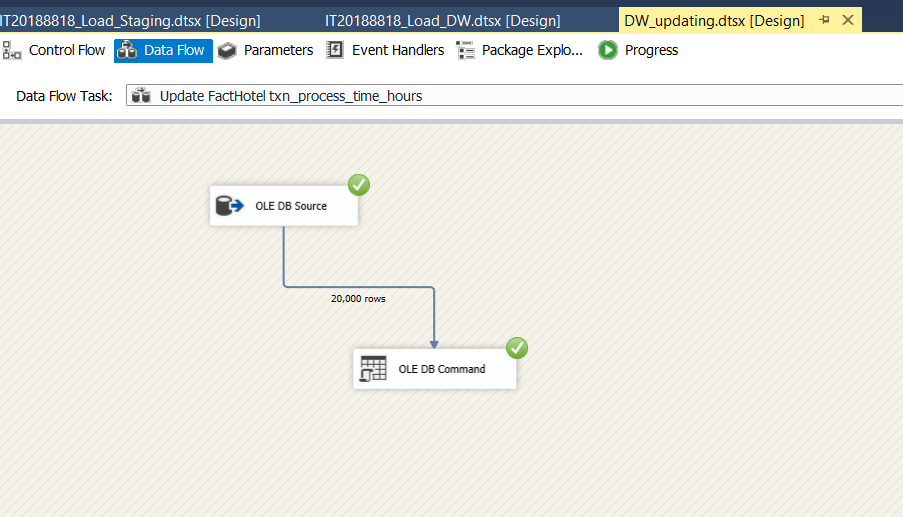
## 7.1. Datawarehouse updating



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



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



## 7.2 Accumulated Fact Table (FactHotel)

