Sri Lanka Institute of Information

Technology



# Data Warehousing & Business Intelligence

Assignment 01 IT22918192

Rathnayaka S J

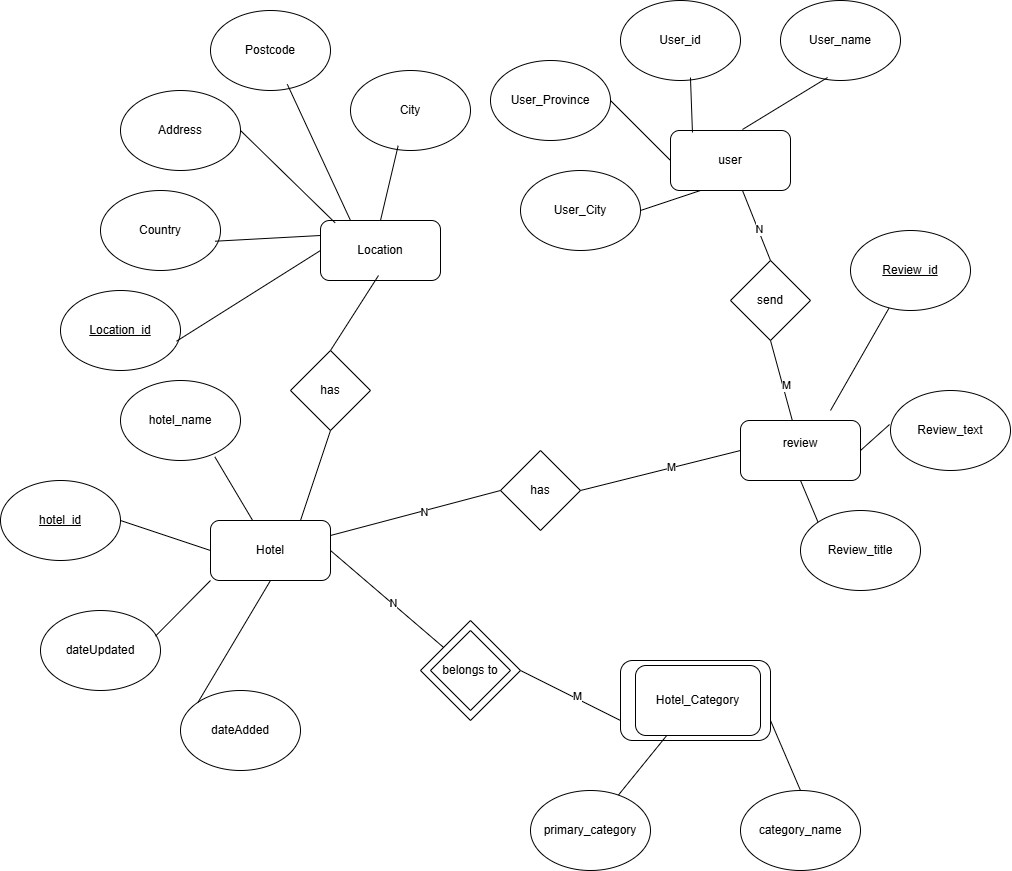
Year 03 semester 2- WE.DS.01.02

## 01. Data set selection

### Dataset- Hotel reviews

Data source- [Hotel Reviews](https://www.kaggle.com/datasets/datafiniti/hotel-reviews)

The Datafiniti Hotel Reviews dataset on Kaggle comprises information on 1,000 hotels, including their locations, names, ratings, review data, titles, usernames, and more. This dataset is suitable for various data analysis and machine learning tasks, such as sentiment analysis and natural language processing. It can be utilized to compare hotel reviews across different states and to experiment with sentiment scoring techniques. The dataset is provided in CSV format, facilitating easy integration into data warehousing and business intelligence projects. Its comprehensive structure makes it a valuable resource for exploring customer feedback and enhancing hotel services



#### 02. Preparation of data

All the data sources were initially provided in CSV format by the website. During the data preparation phase, modifications were made to the original files, such as adding new columns and splitting some information into separate tables. The files were then either converted into text files or imported directly into a source database.

Final Format of Source Data Before Transformation:

1. CSV Files (.csv)
   * Hotel data CSV file
   * Hotel Category CSV file
   * User data CSV file
   * Review data CSV file
   * Accommodation transaction complete time CSV file

These CSV files were imported into SQL Server Management Studio (SSMS) and organized within a newly created database named HotelDB.

1. Text File (.txt)
   * Location data stored in a text file

## 03. Solution Architecture

Hotel\_Staging.

* stgcomplete\_time
* stgHotel
* stgHotalCategory
* stgReview
* StgUser
* stgLocation

Hotel\_DW

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

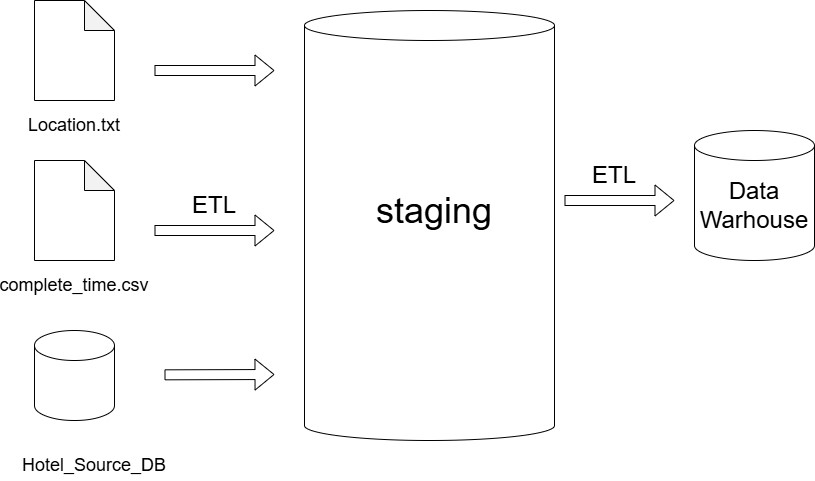
Architecture Components.

* Data Sources.
* Transactional processing systems (operational databases)
* Third-party data providers and external datasets

* Extract, Transform and Load.
* Extraction phase - Retrieving information from source database
* Conversion phase - Merging datasets, eliminating duplicates
* Quality control - Standardizing and cleansing information

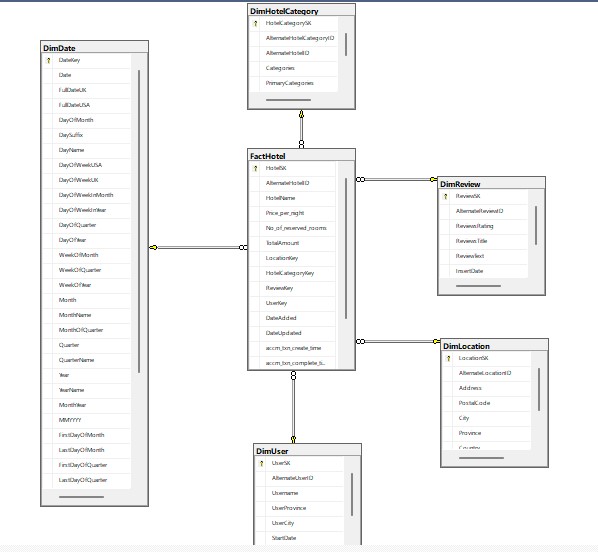
### • Data Warehouse

* EDW and Data Mart.
* Dimensional Modeling- Facts and Dimensions.
* Schema approach - Utilizing star-based structure for relationships



## 04. Data Warehouse Design & Development

### Relational Diagram – Star Schema

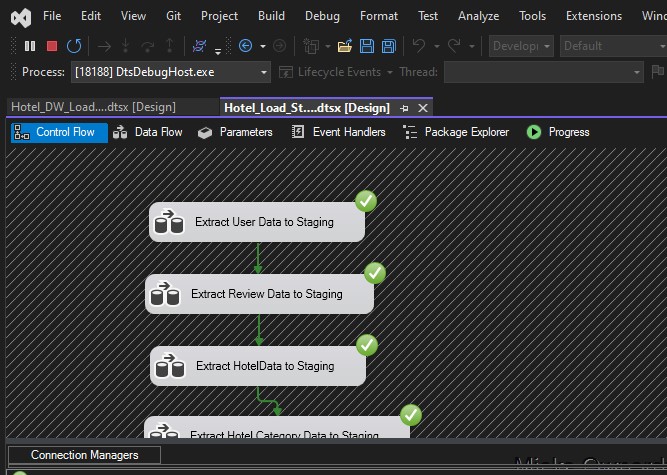


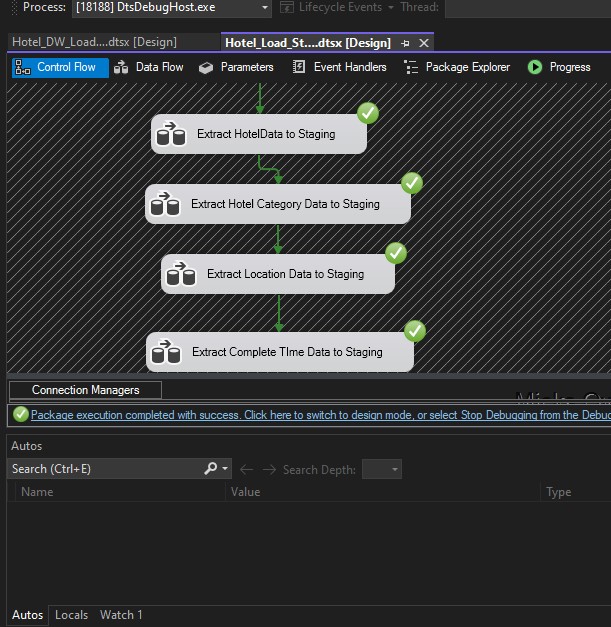
* This schema consists with 5 dimensional tables and 1 fact table.
* I considered DimLocation as a slowly changing dimensions because addresses can be changed in future.

## 05.ETL Development

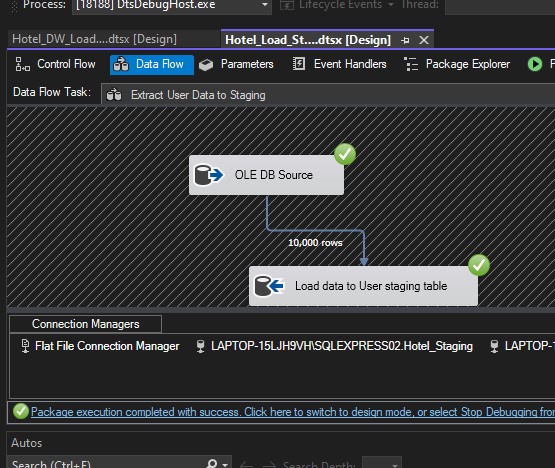
5.1 Staging tables

• Load the tables from data sources to staging

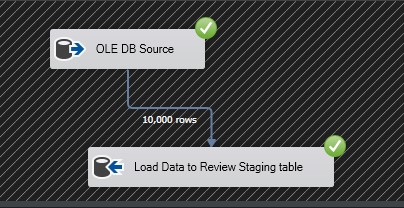




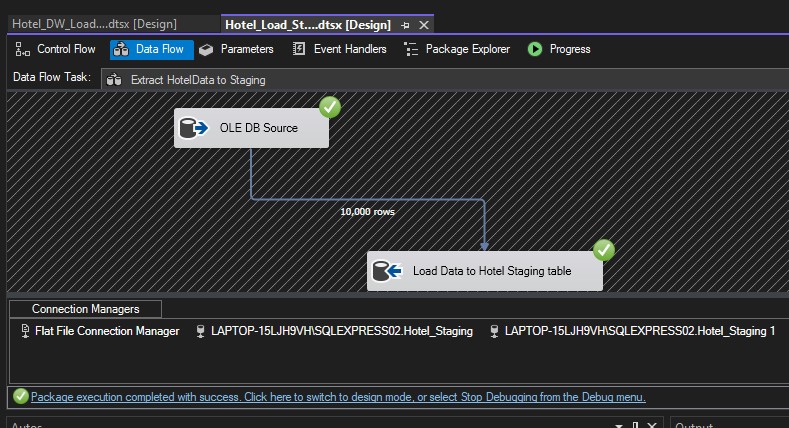
### 5.1.1 Load data User to staging



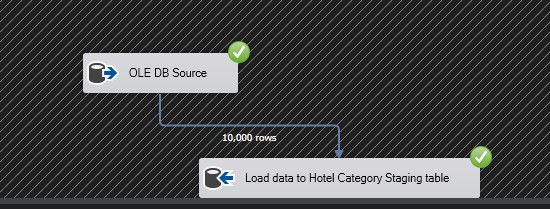
### 5.1.2 Load data Review to staging



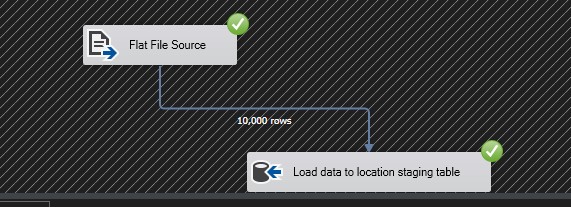
### 5.1.3 Load data Hotel to staging



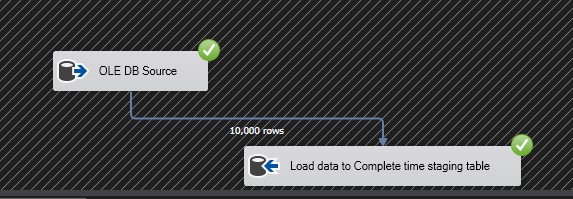
### 5.1.4 Load data Hotel Category to staging



### 5.1.5 Load data Location to staging



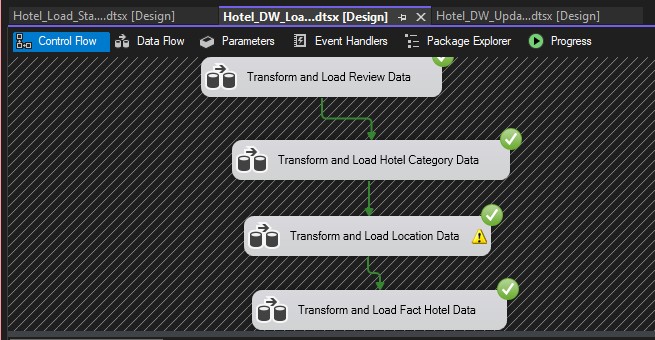
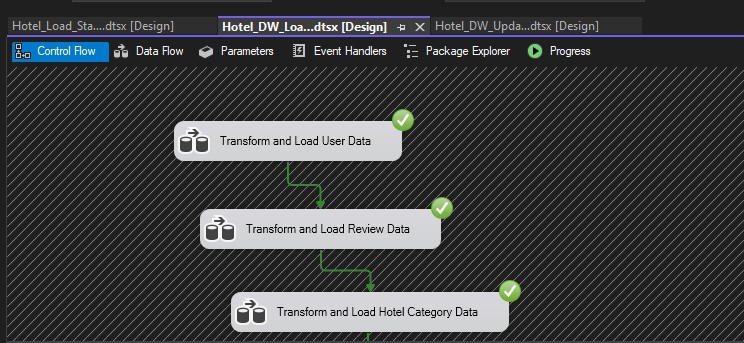
### 5.1.6 Load data complete\_time to staging



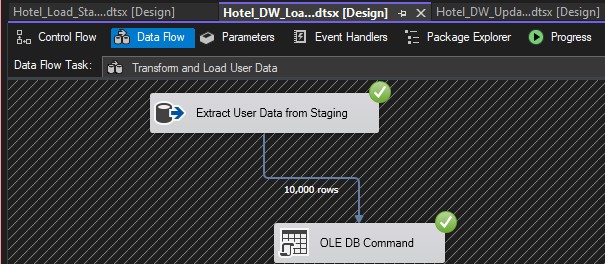
## 06.Staging to DW

• Loading the staging tables to datawarehouse.

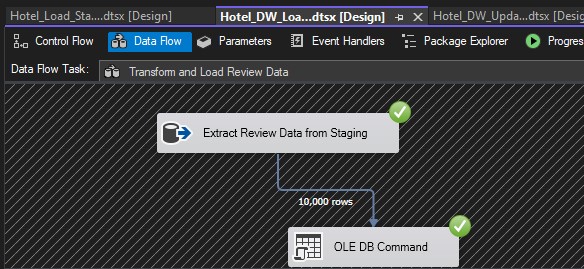
### 6.1 ETL System to Datawarehouse



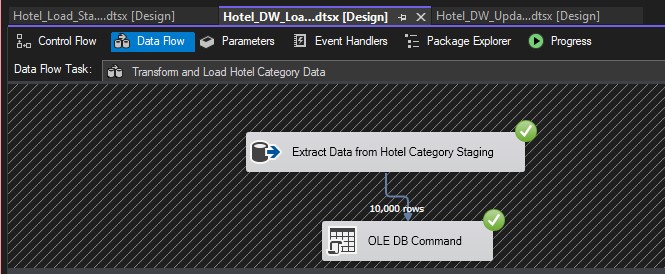
#### 6.1.1 Transfer and Load DimUser Data from staging



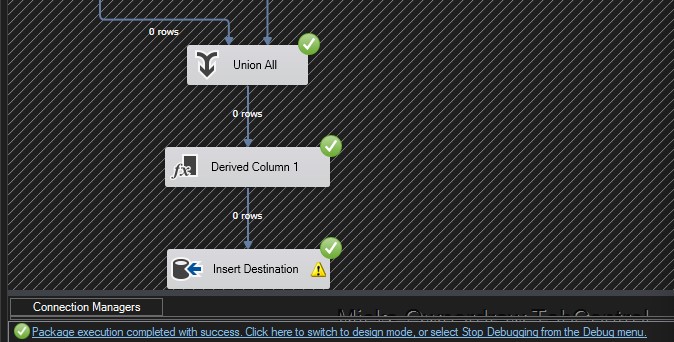
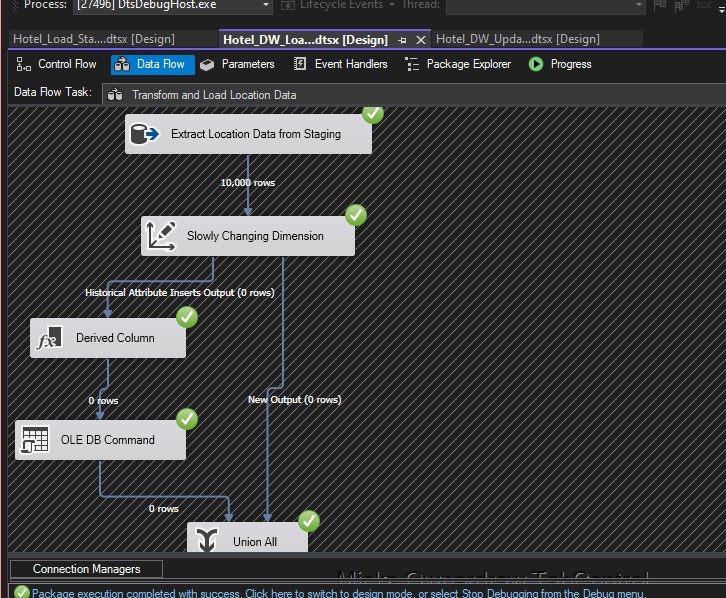
#### 6.1.2 Transfer and Load DimReview Data from staging



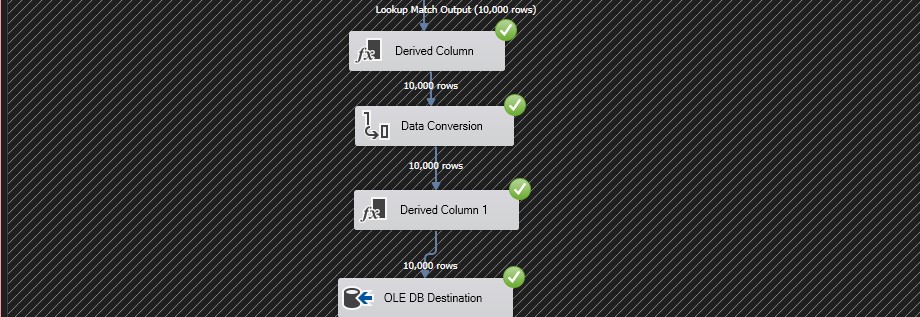
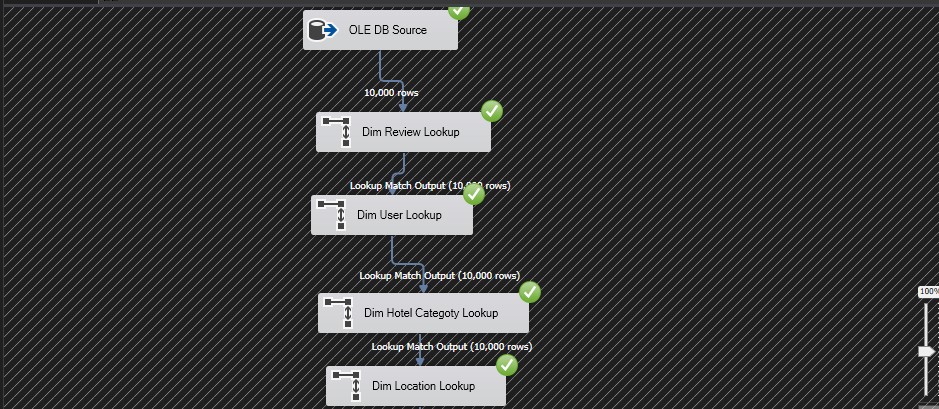
#### 6.1.3 Transfer and Load DimHotelCategory Data from staging



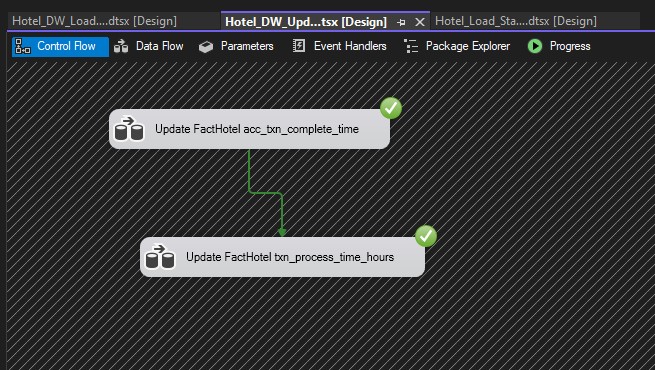
#### 6.1.4 Transfer and Load DimLocation Data from staging



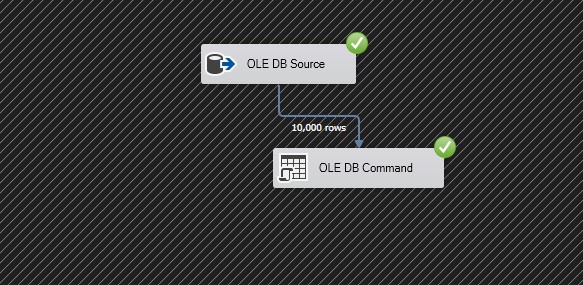
#### 6.1.5 Load FactHotel Data from staging



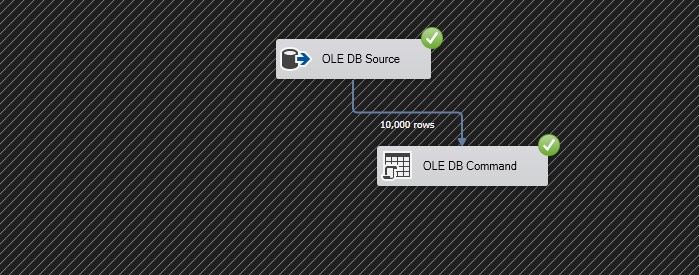
## 07. Datawarehouse Updating



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



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



### 08. Accumulated Fact Table

