

**Data Warehousing and**

**Business Intelligence**

**Assignment – 1**

**IT20476212**

**Dewmini P.W.K**

Table of Contents

[1. Data Set Selection and Scenario 3](#_Toc103661723)

[2. Preparation of Data Sources 5](#_Toc103661724)

[3. Solution Architecture 6](#_Toc103661725)

[3.1. Data Sources 7](#_Toc103661726)

[**Restaurant\_SourceDB** 7](#_Toc103661727)

[**Restaurant\_Staging\_** 8](#_Toc103661728)

[**Restaurant\_DataWarehousing** 8](#_Toc103661729)

[3.2. Extract, Transform and Loading 9](#_Toc103661730)

[4. Data Warehouse Design & Development 11](#_Toc103661731)

[5. Test Planning and Design Test Cases 13](#_Toc103661732)

[6. ETL Development 14](#_Toc103661733)

[Extraction 14](#_Toc103661734)

[Transforming and Loading 16](#_Toc103661735)

[7. DataWarehouse Updating 21](#_Toc103661736)

# Data Set Selection and Scenario

The selected data source is a collection of restaurants data. The link to the source data set is mentioned below:

<https://www.kaggle.com/datasets/mikhailpustovalov/scraped-data-from-ta>

Modifications were done accordingly to the data set derived from the source. This data set reflects combinations between restaurants details and their ranking on the countries. Restaurant specific details involved in restaurants, Food types customers are keen to order, Reviews and ratings that customers are given to each restaurant are some of the key details included in the data set.

The two main sources are listed below:

SQL Database

One text file(.txt) – Location Data

Also, the below mentioned CSV files were imported to the SQL source database.

* accm\_txn\_complete\_time CSV File
* Restaurants CSV File
* Reviews CSV File
* Meal CSV File
* Cusiness Price CSV File

**ER Diagram**

Diagram

Description automatically generated

# Preparation of Data Sources

The original tables taken from the dataset were in the .csv format. Therefore, 2 tables were converted into excel files (.xlsx) and one table to a text file (.txt).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source Type** | **Table Name** | **Column Names** | **Data Type** | **Description** |
| Microsoft Excel  Comma Separated  Values File (.csv) | Restaurants | res\_id | smallint | This table holds the details of the Restaurants in the countries. |
| name | Varchar(100) |
| tel | Varchar(50) |
| rank | smallint |
| Microsoft Excel  Comma Separated  Values File (.csv) | Meal | cuisiness\_ID | smallint | This table includes the Meal details of each restaurant. |
| primary\_cuisiness | Varchar(50) |
| cuisines | Varchar(100) |
| special\_diets | Varchar(100) |
| cus\_rest\_rank | smallint |
| Microsoft Excel  Comma Separated  Values File (.csv) | Reviews | Review\_ID | smallint | This table contains the details of the reviews made by customers for each restaurant . |
| review\_number | smallint |
| review\_date | Varchar(50) |
| review\_ratings | Varchar(50) |
| reviews.title | Varchar(150) |
| Text Document  (.txt) | Address | location\_ID | smallint | This table holds the information about the address of each restaurant. |
| address | Varchar(50) |
| postalCode | Varchar(50) |
| city | Varchar(50) |
| province | Varchar(50) |
| country | Varchar(50) |
| Microsoft Excel  Comma Separated  Values File (.csv) | cus\_price | Date | datetime | This table stores the meal prices in each restaurant. |
| res\_id | smallint |
| meal\_id | int |
| Price ($) | Float |
| discount | tinyint |
| KSymbol | Varchar(50) |

*Table 1-1*

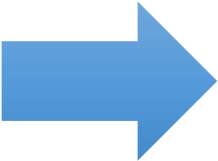
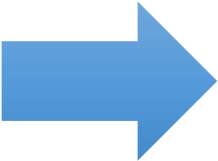
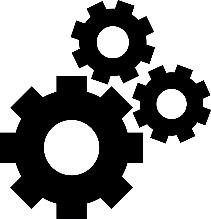
# Solution Architecture

A picture containing dark, night, night sky

Description automatically generated

Restaurant\_DataWarehousing

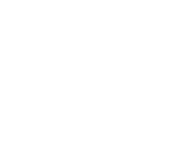
Restaurant\_SourceDB



accm\_txn\_complete\_time CSV File



TXT File (Address.txt)



Extracting

Transforming

Loading

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Sources |  | ETL Process |  | Data Warehouse |

**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.

* As explained First step is staging the source data set.
* Next staged tables are profiled and aggregations are performed when necessary. As the next step data is transformed and loaded.
* After completing the described stages, data is tested and validated and the Datawarehouse is created.
* After the warehouse is created BI results such as OLAP analysis, Reports, Data visualization, Data mining can be obtained as results after further modifications.

## Data Sources

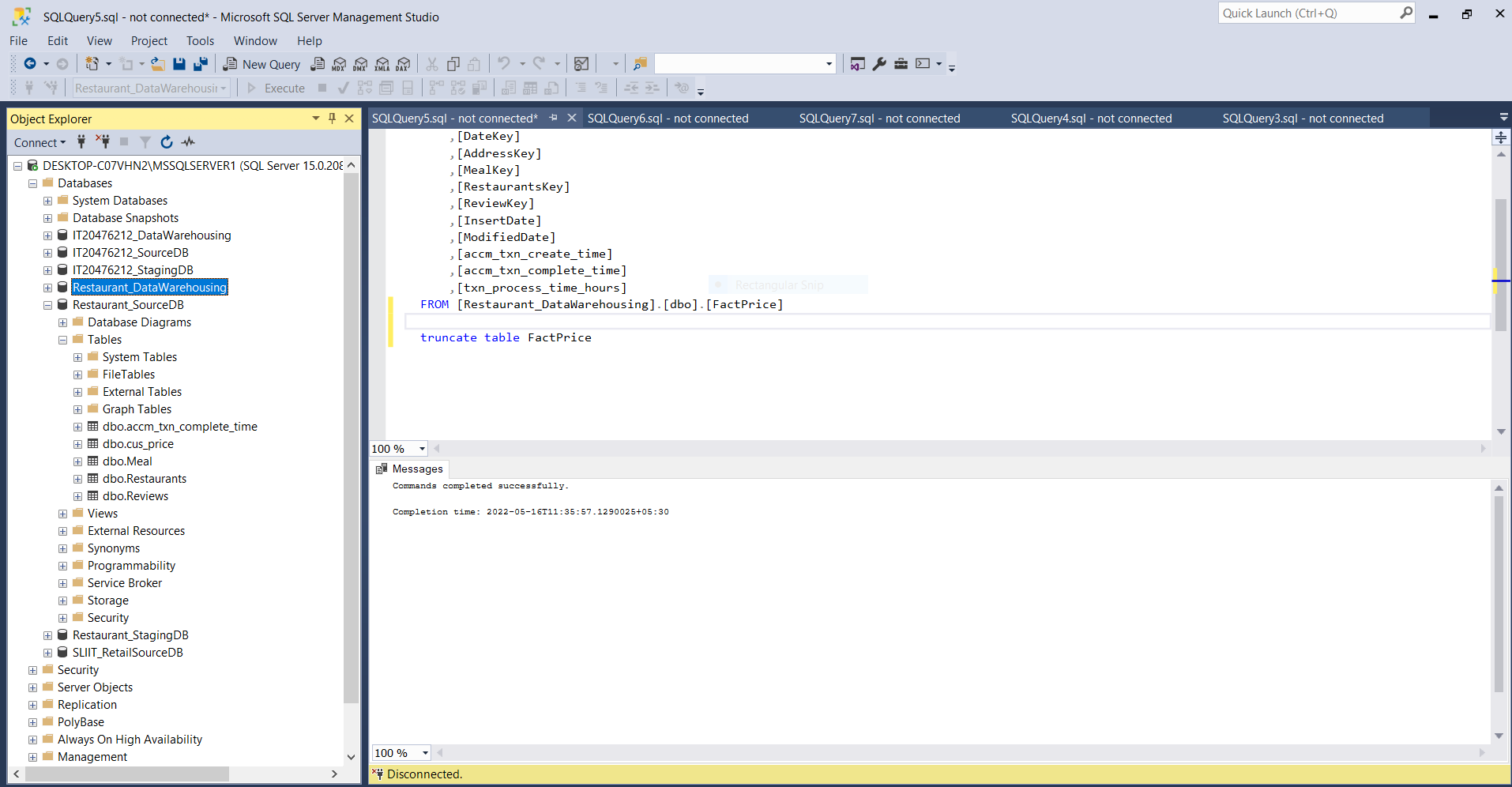
There were five tables in the data set and one as Text (txt) file and the rest of the tables as csv files.

I have created three databases on Microsoft Server Management Studio.

1. Restaurant\_SourceDB
2. Restaurant\_StagingDB
3. Restaurant\_DataWarehousing

### **Restaurant\_SourceDB**

This database is the main source to the Data Warehouse. I imported all the csv files to this database. This database includes the Restaurants, Meal, Reviews and cus\_price tables. Address(txt) file was loaded extracted directly through Visual Studio Data Tool.



### **Restaurant\_Staging\_**

Graphical user interface, text, application

Description automatically generatedThe first step of solution architecture is staging the source data set. After the staging layer the below mentioned staging tables are created:

1. Restaurants Staging

2. Meal Staging

3. Reviews Staging

4. cus\_price Staging

5. Address Staging

### **Restaurant\_DataWarehousing**

Graphical user interface, text, application, email

Description automatically generatedThis is the destination of database and this contains all the staging tables and the dimension and fact tables.

1.DimAddress

2.DimDate

3.DimMeal

4.DimRestaurants

5.DimReview

6.FactPrice

## Extract, Transform and Loading

After creating the source database, I have extracted the data to the staging tables through Visual Studio (SSDT). A screenshot of a video game

Description automatically generated

After extracting the data were transformed.

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

Description automatically generated

After the extracting and transforming data were loaded to the dimension tables and to fact table.

Graphical user interface, application, table

Description automatically generated

# Data Warehouse Design & Development

***Relational Diagram (Star Schema)***

Diagram

Description automatically generated

A screenshot of a computer

Description automatically generated

**DimAddress** 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 (Address table.)**

# Test Planning and Design Test Cases

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test Case No.** | **Scenario** | **Assumptions** | **Schedules** | **Environment** | **Tools** | **Risk and Risk**  **Management** |
| 1 | Check the number of res\_ID in RestaurantsStaging table and DimRestaurants  table | All the transaction data were loaded into the table | Load data to tables and setup a SQL environment  (SSMS) | SQL based environment | SQL Server  Management  Studio  Visual Studio  Data Tool | No risk |
| 2 | Check the data length of a specific in RestaurantsStaging table and DimRestaurants  table | All the restaurants data were loaded into the table | Load data to tables and setup a SQL environment  (SSMS) | SQL based environment | SQL Server  Management  Studio  Visual Studio  Data Tool | No risk |
| 3 | Check the number of duplicate values in  RestaurantsStaging table and DimRestaurant  table | All the restaurant data were loaded into the table | Load data to tables and setup a SQL environment  (SSMS) | SQL based environment | SQL Server  Management  Studio  Visual Studio  Data Tool | No risk |

# ETL Development

## Extraction

In the extraction all the tables were taken to a staging table in the Restaurant\_DataWarehousing. All the csv files were taken through the Restaurant\_SourceDB and the Text (txt) files were taken directly to the staging.

Data taken from Restaurant\_SourceDB.

* **Load data Restaurants to staging**

**Graphical user interface

Description automatically generated**

* **Load data Meal to stagingA screenshot of a computer

  Description automatically generated with medium confidence**
* **Load data Reviews to staging**

Graphical user interface, website

Description automatically generated

* **Load data Price to staging**

Graphical user interface

Description automatically generated

* **Load data Address to staging(.txt file)**

Graphical user interface

Description automatically generated

* **Load data accm\_txn\_complete\_time to staging**

Graphical user interface, website

Description automatically generated

.

## Transforming and Loading

After staging all the tables from the dataset they were transformed and loaded into the dimension tables and to the fact table.

* **ETL System to Datawarehouse**

A screenshot of a computer

Description automatically generated

* Transform Meal data and load them into DimMeal.

A screenshot of a computer

Description automatically generated

* Transform Restaurants data and load them into Restaurants.

A screenshot of a computer

Description automatically generated

* Transform Reviews data and load them into DimReviews

A screenshot of a computer

Description automatically generated.

* Transform Address data and load them into Dim Address **(Slowly changing dimension)**

A screenshot of a computer

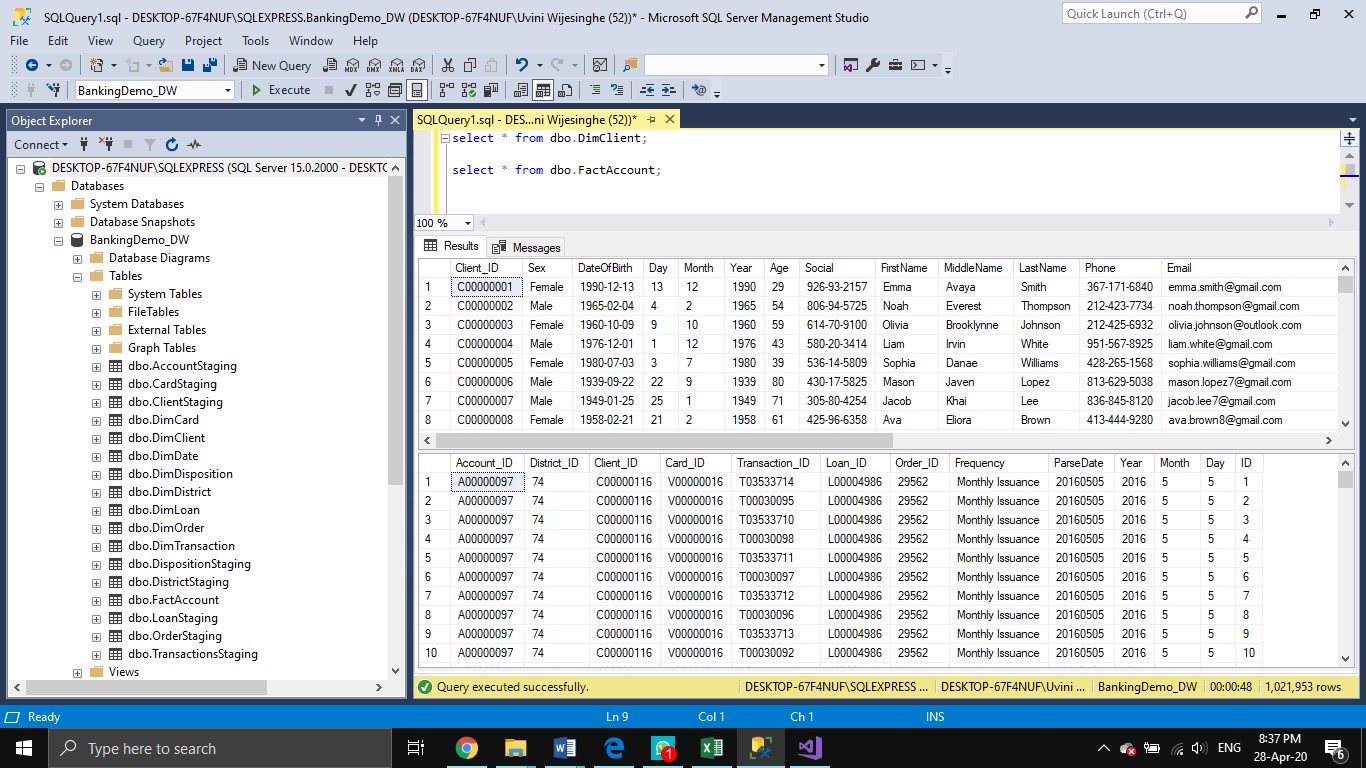
Description automatically generated

* Load FactPrice Data from staging

A screenshot of a computer

Description automatically generated

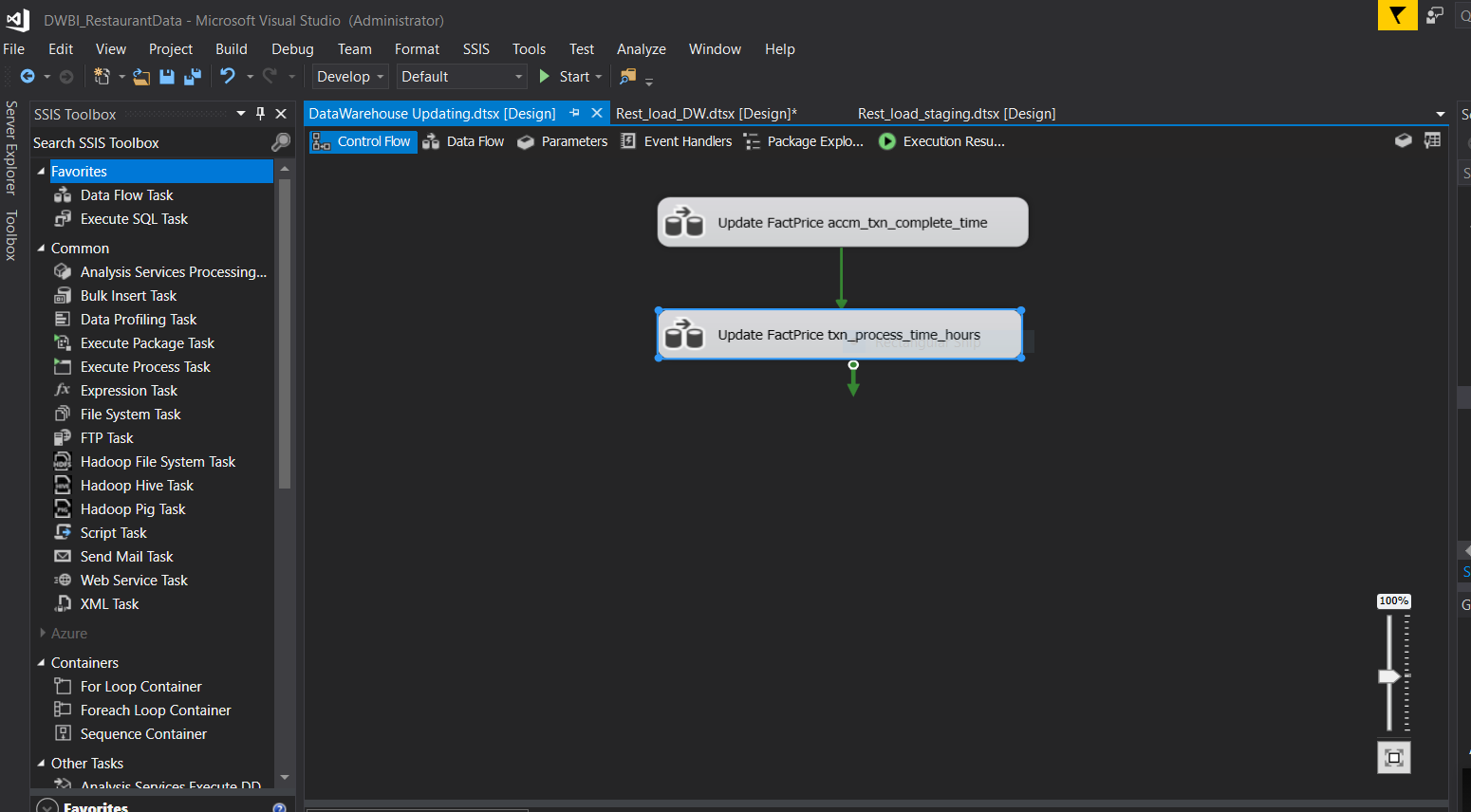
**Loaded Data**



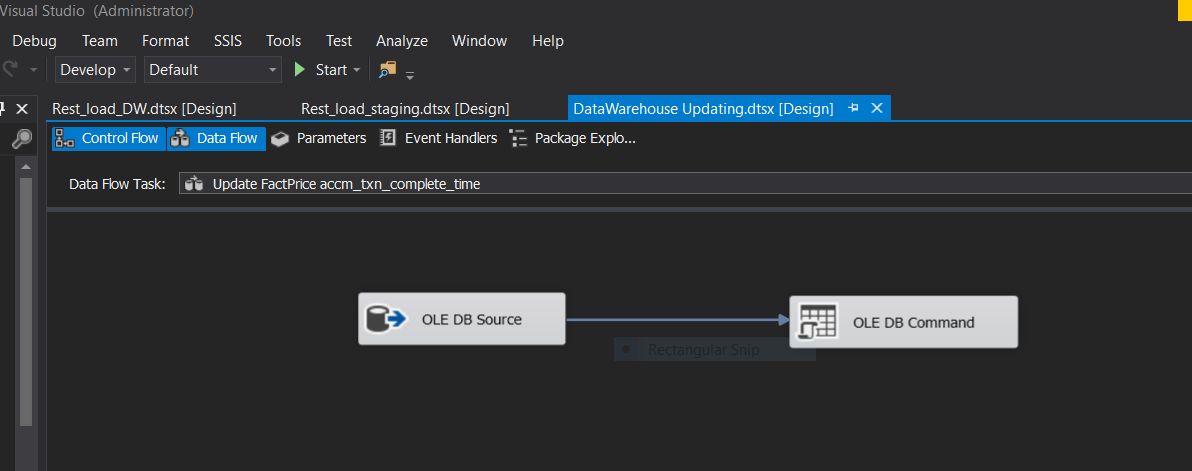
# DataWarehouse Updating

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

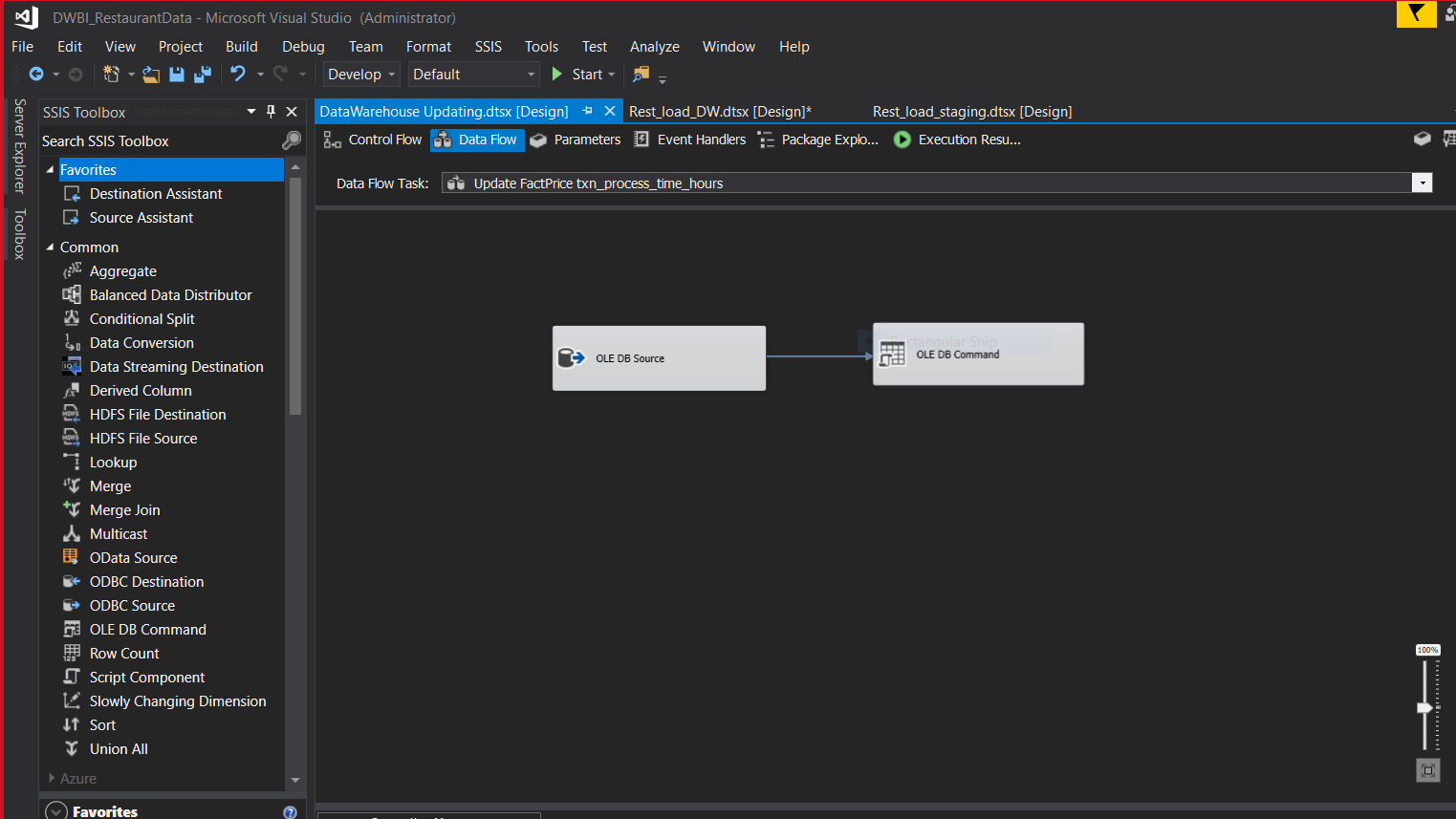
* **Datawarehouse updating**



* **Update factPrice accm\_txn\_complete\_time**



* **Update factPrice txn\_process\_time\_hours**



**7.2 Accumulated Fact Table (FactPrice)**

