



Data Warehousing and Business Intelligence

Assignment – 1

IT20476212

Dewmini P.W.K

Table of Contents

1.	Data Set Selection and Scenario.....	3
2.	Preparation of Data Sources	5
3.	Solution Architecture	6
3.1.	Data Sources	7
	Restaurant_SourceDB	7
	Restaurant_Staging_	8
	Restaurant_DataWarehousing.....	8
3.2.	Extract, Transform and Loading	9
4.	Data Warehouse Design & Development.....	11
5.	Test Planning and Design Test Cases.....	13
6.	ETL Development	14
	Extraction.....	14
	Transforming and Loading	16
7.	DataWarehouse Updating.....	21

1. 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:

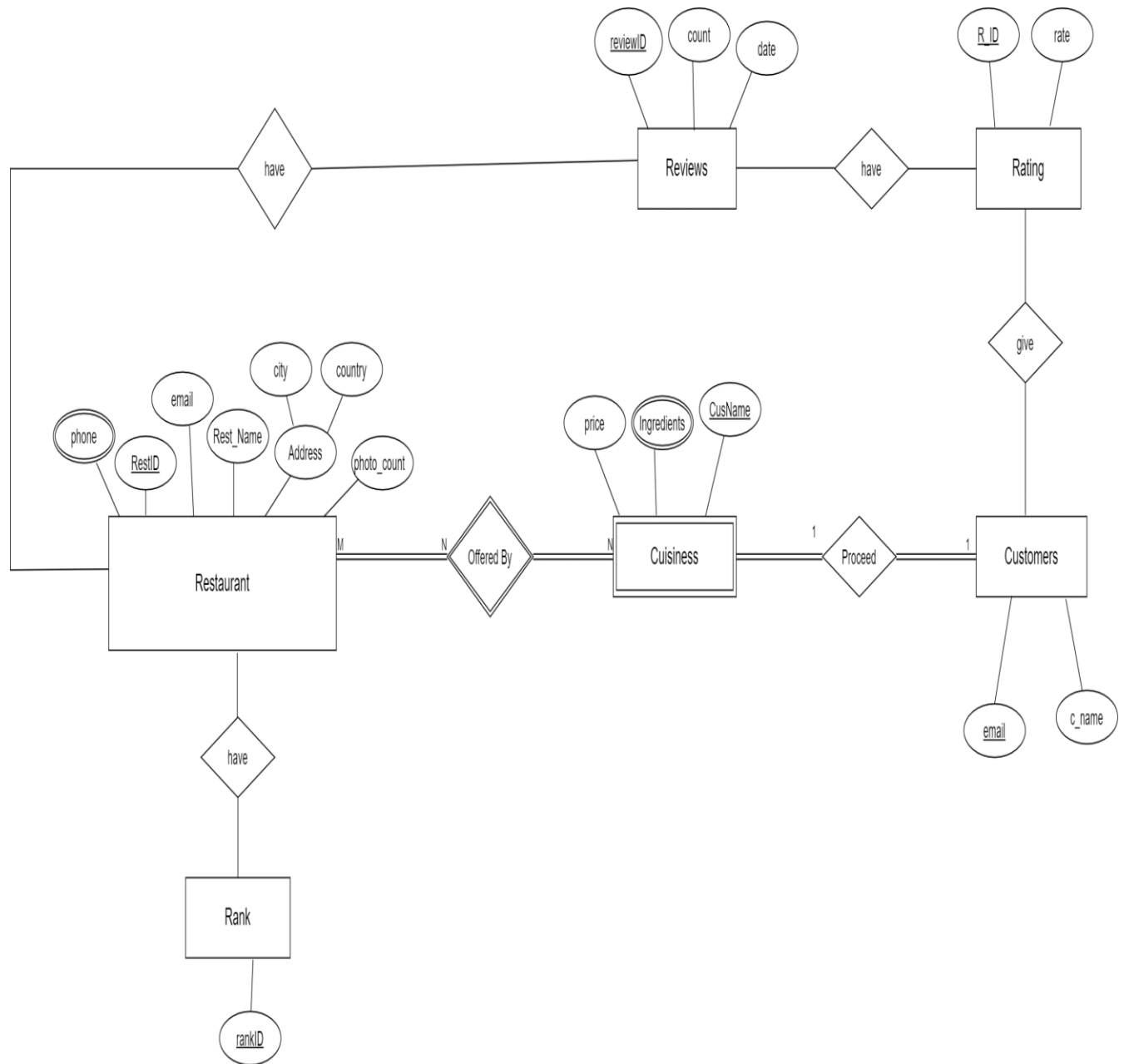
SQL Database

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



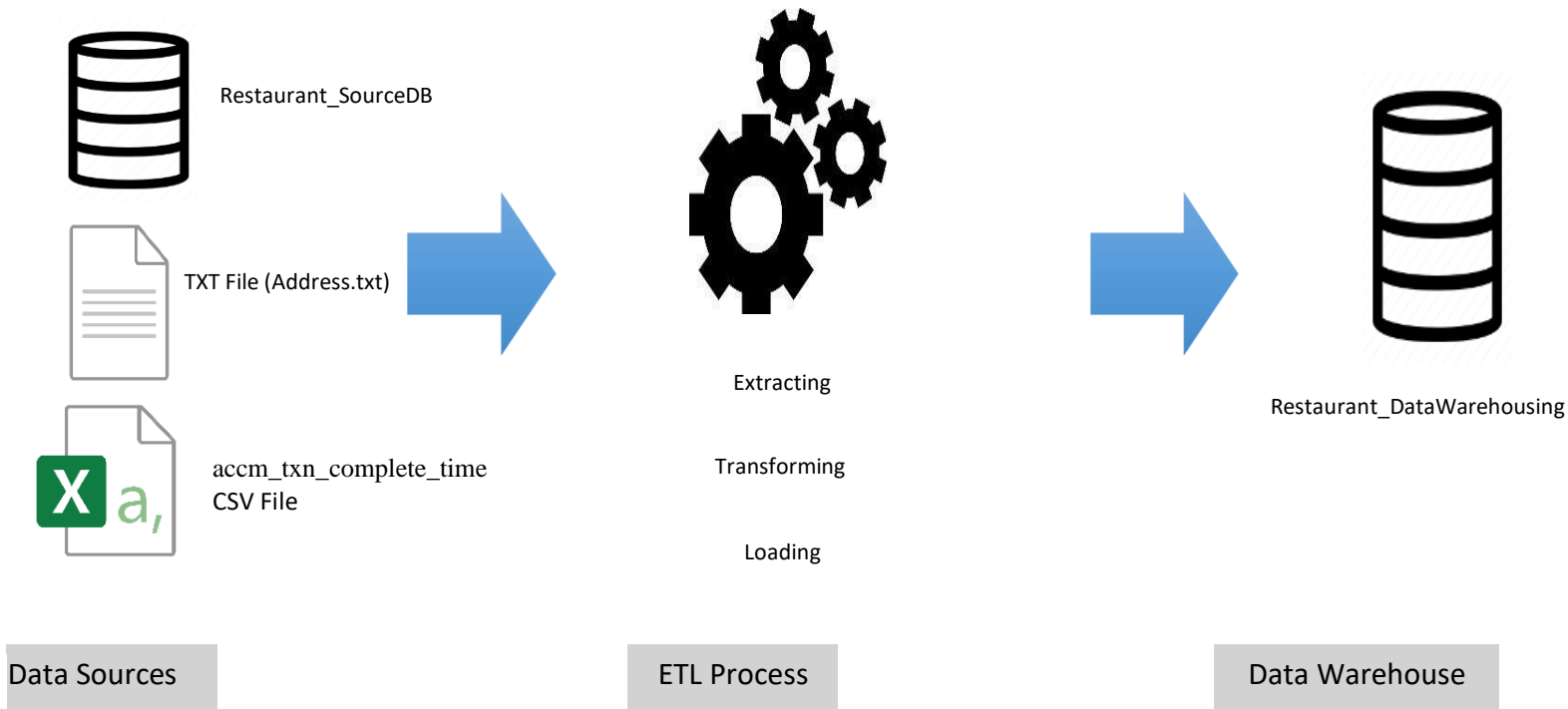
2. Preparation of Data Sources

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

3. Solution Architecture



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.

3.1. Data Sources

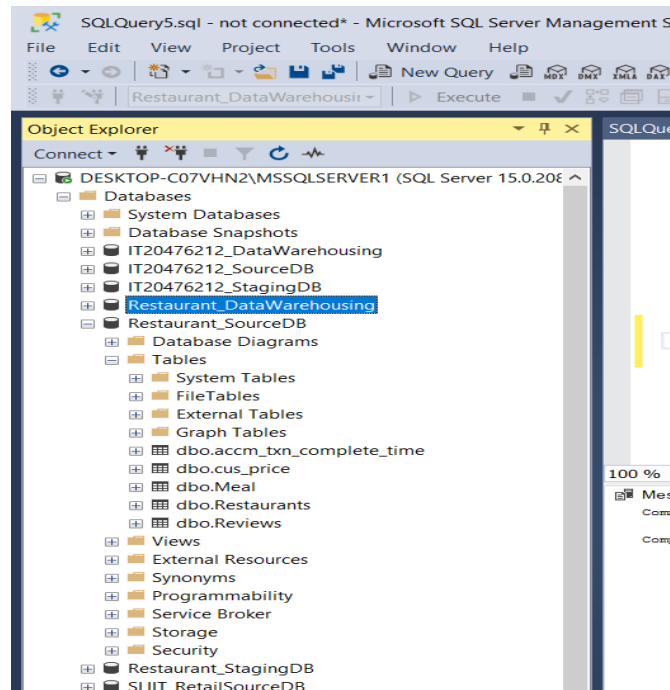
There were five tables in the data set and one as Text (txt) file and the rest of the tables as csv files.

I have created three databases on Microsoft Server Management Studio.

- 1) Restaurant_SourceDB
- 2) Restaurant_StagingDB
- 3) Restaurant_DataWarehousing

Restaurant SourceDB

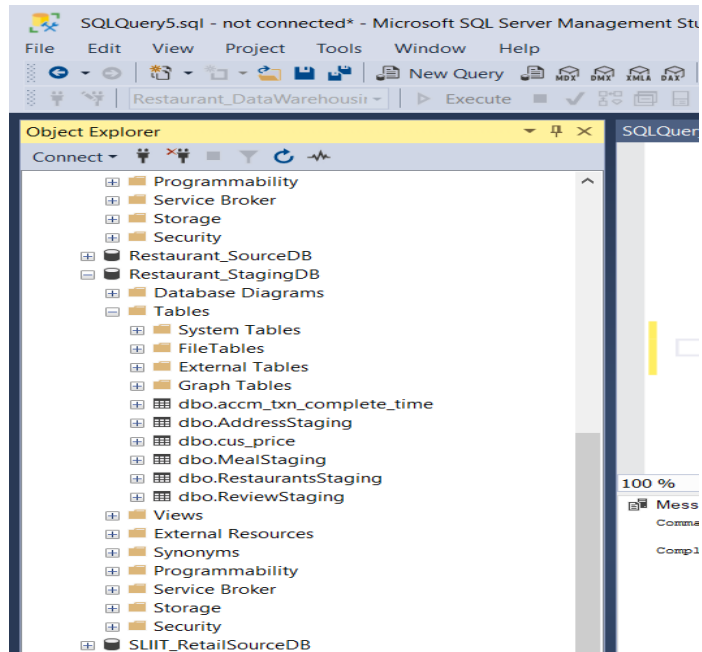
This database is the main source to the Data Warehouse. I imported all the csv files to this database. This database includes the Restaurants, Meal, Reviews and cus_price tables. Address(txt) file was loaded extracted directly through Visual Studio Data Tool.



Restaurant Staging

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

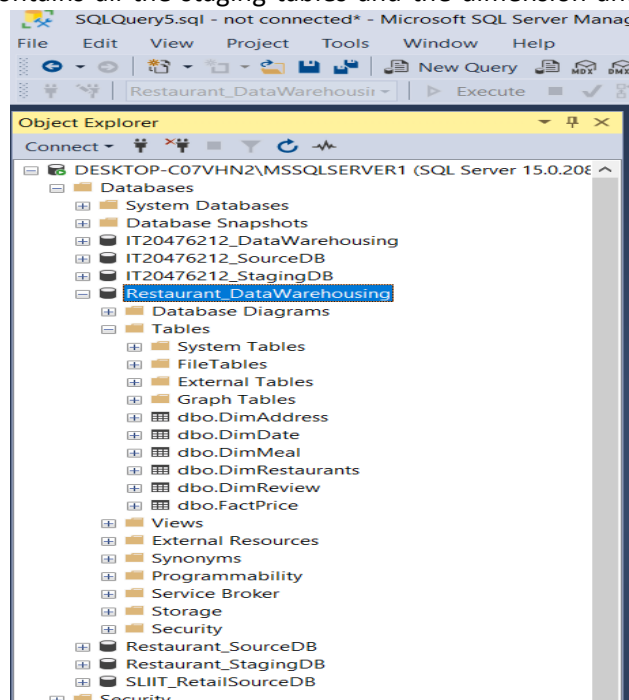
1. Restaurants Staging
2. Meal Staging
3. Reviews Staging
4. cus_price Staging
5. Address Staging



Restaurant DataWarehousing

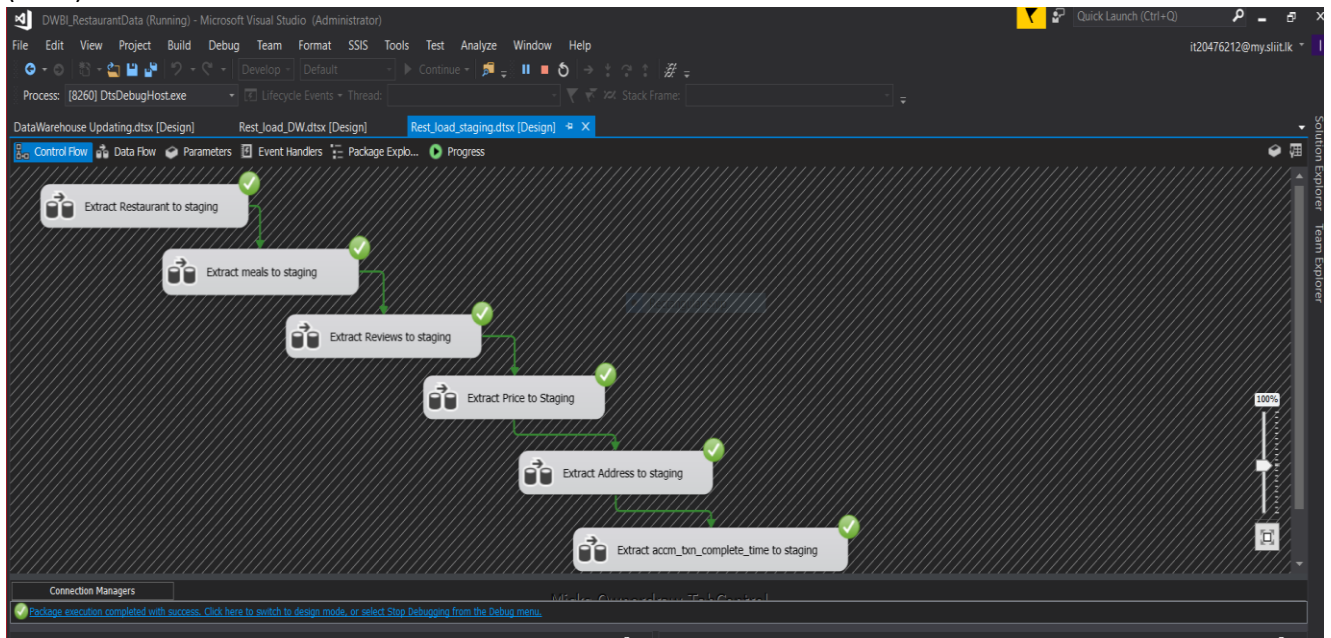
This is the destination of database and this contains all the staging tables and the dimension and fact tables.

1. DimAddress
2. DimDate
3. DimMeal
4. DimRestaurants
5. DimReview
6. FactPrice

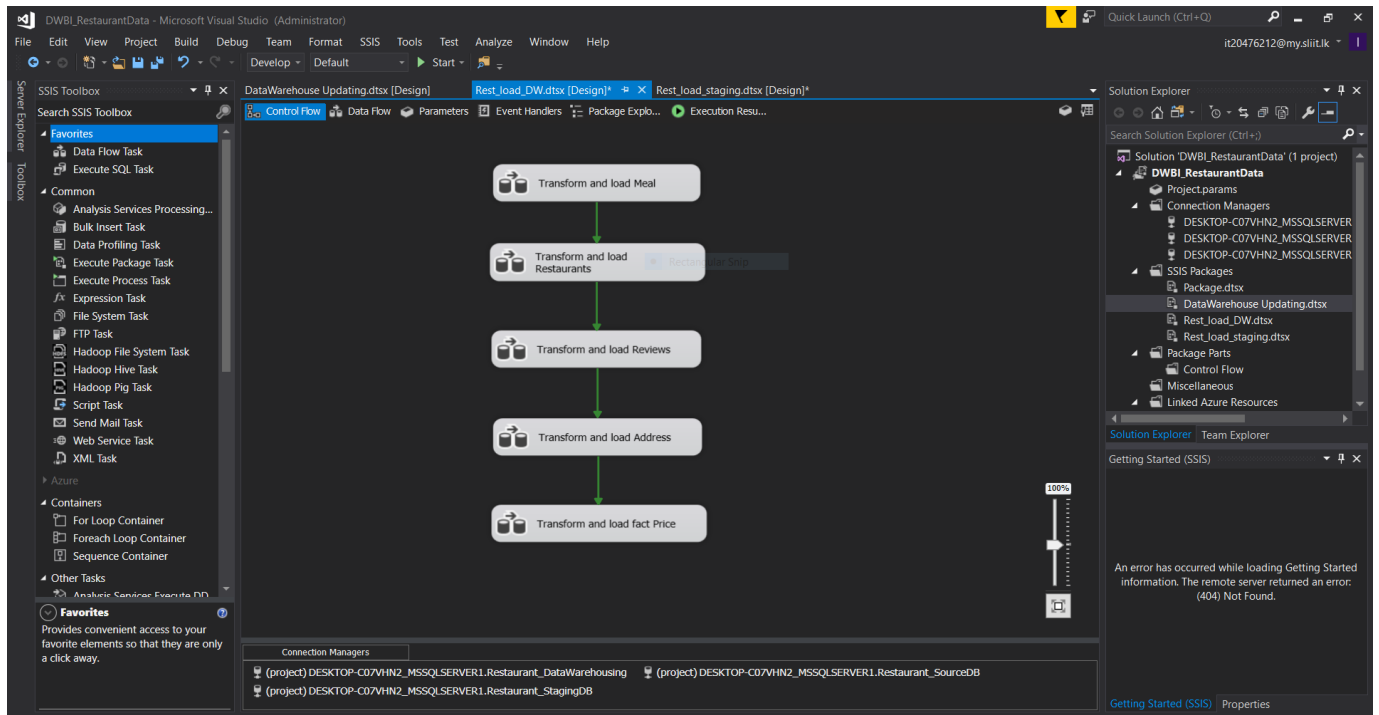


3.2. Extract, Transform and Loading

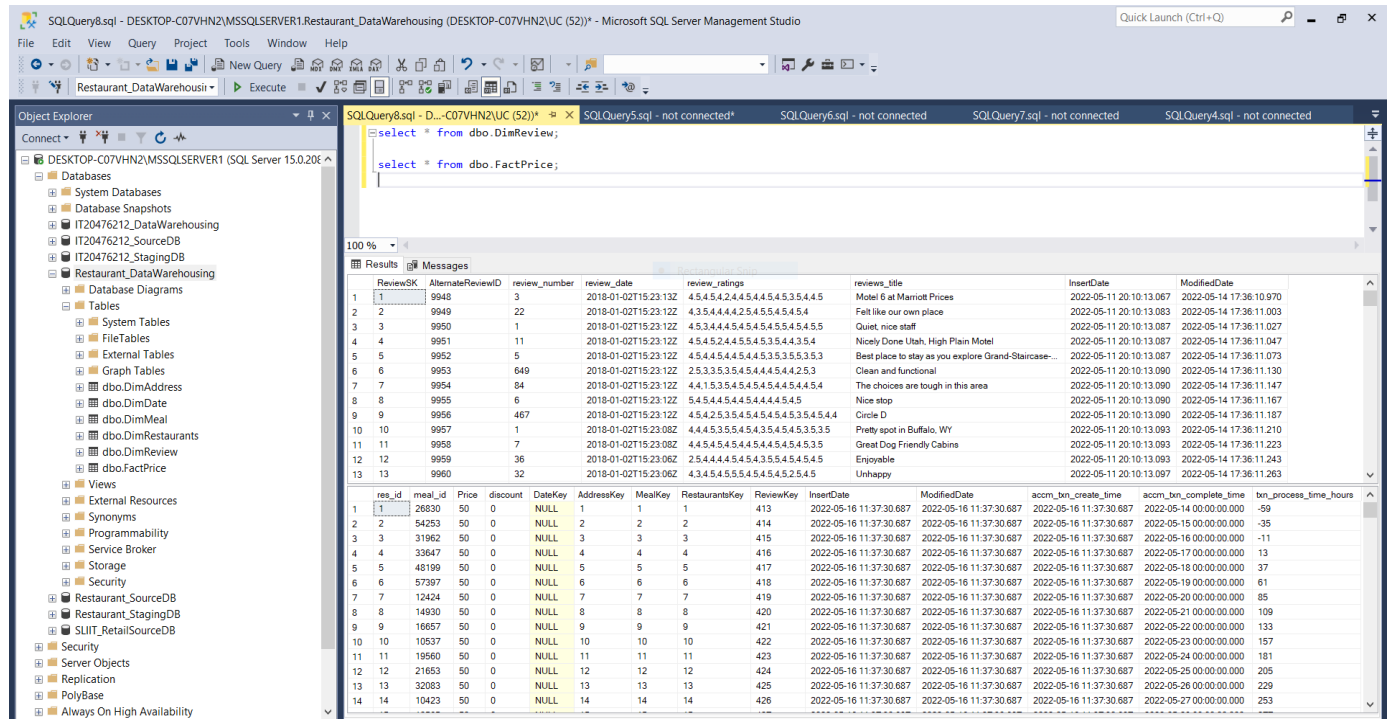
After creating the source database, I have extracted the data to the staging tables through Visual Studio (SSDT).



After extracting the data were transformed.

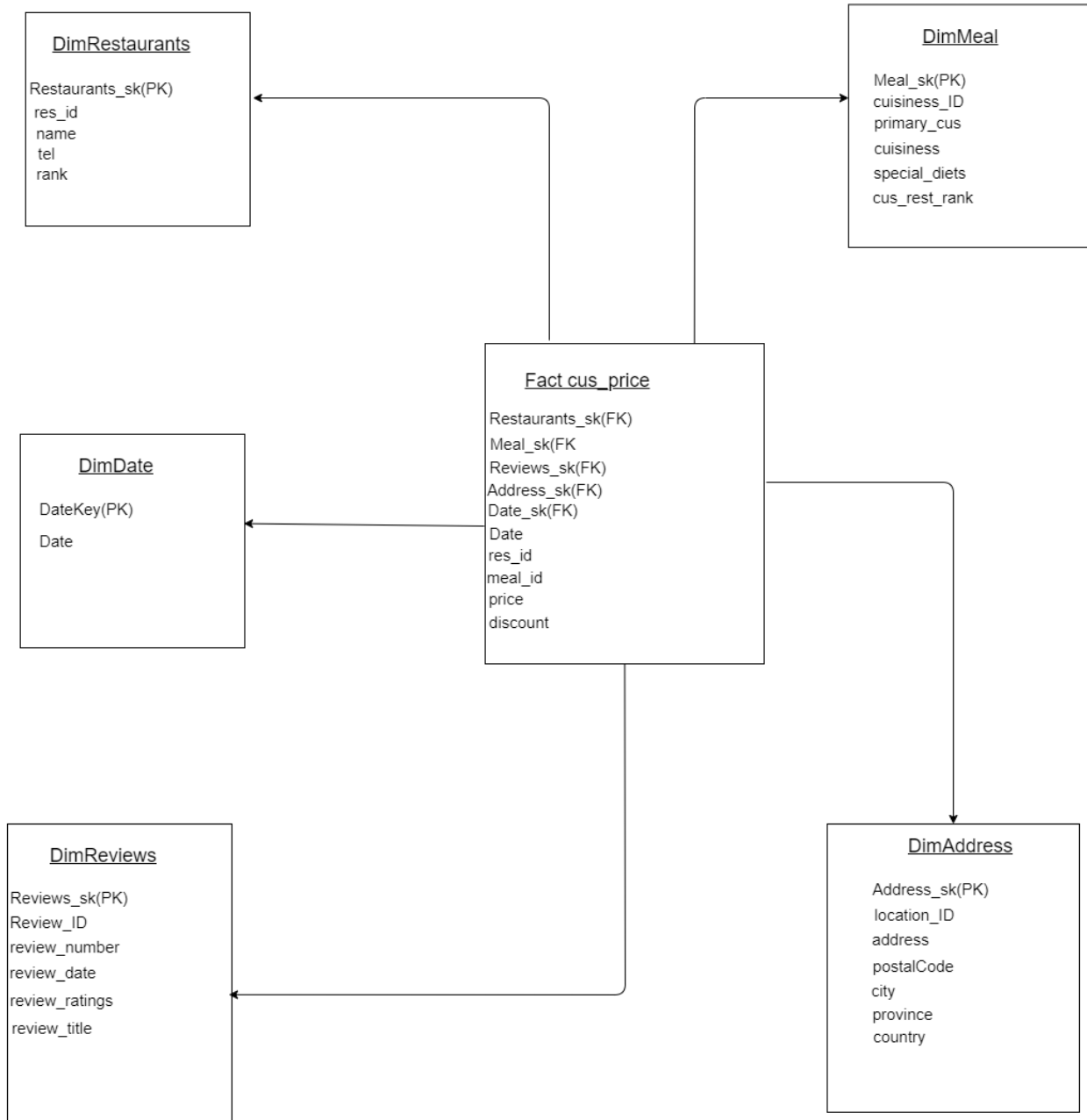


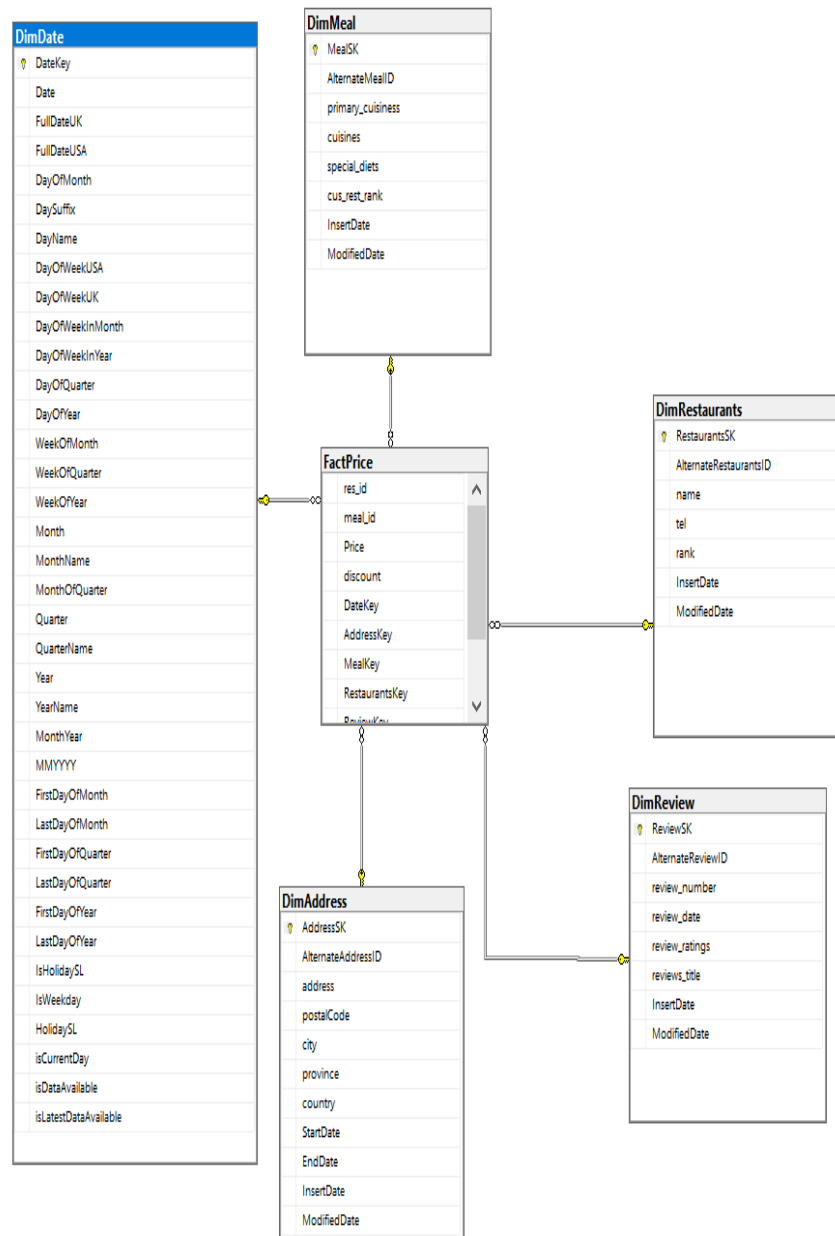
After the extracting and transforming data were loaded to the dimension tables and to fact table.



4. Data Warehouse Design & Development

Relational Diagram (Star Schema)





DimAddress is slowly changing dimension. 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.)

5. 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

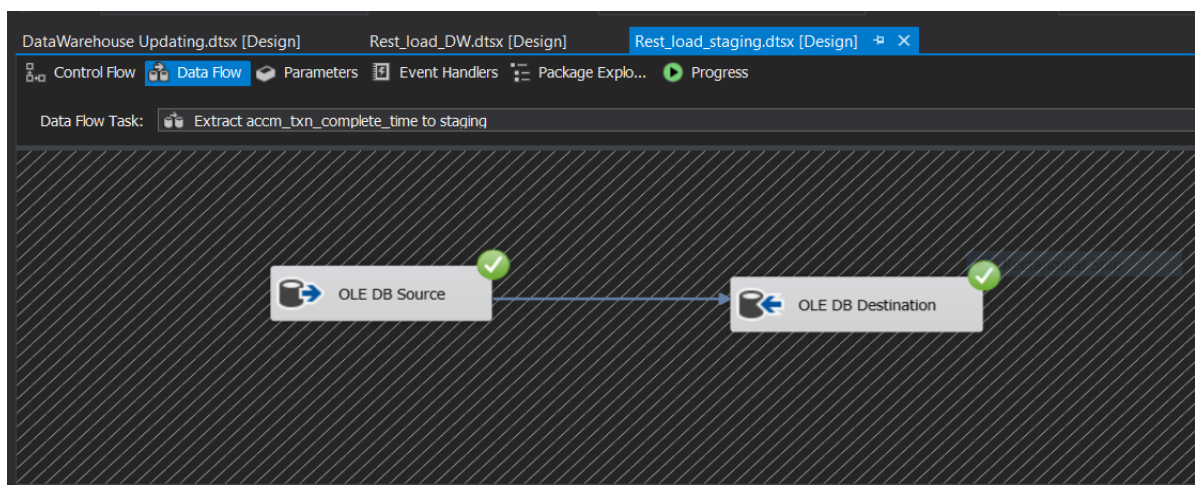
6. ETL Development

Extraction

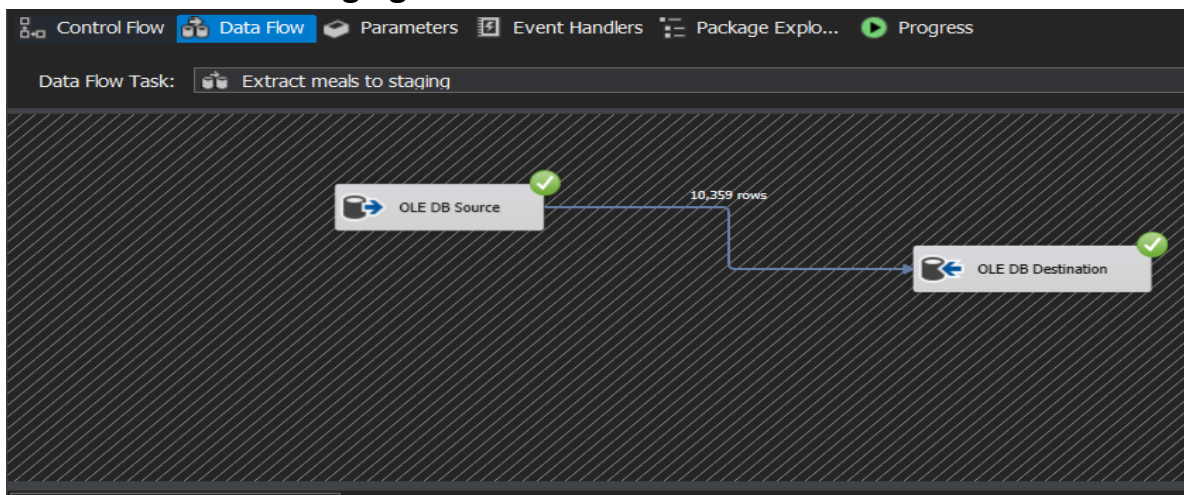
In the extraction all the tables were taken to a staging table in the Restaurant_DataWarehousing. All the csv files were taken through the Restaurant_SourceDB and the Text (txt) files were taken directly to the staging.

Data taken from Restaurant_SourceDB.

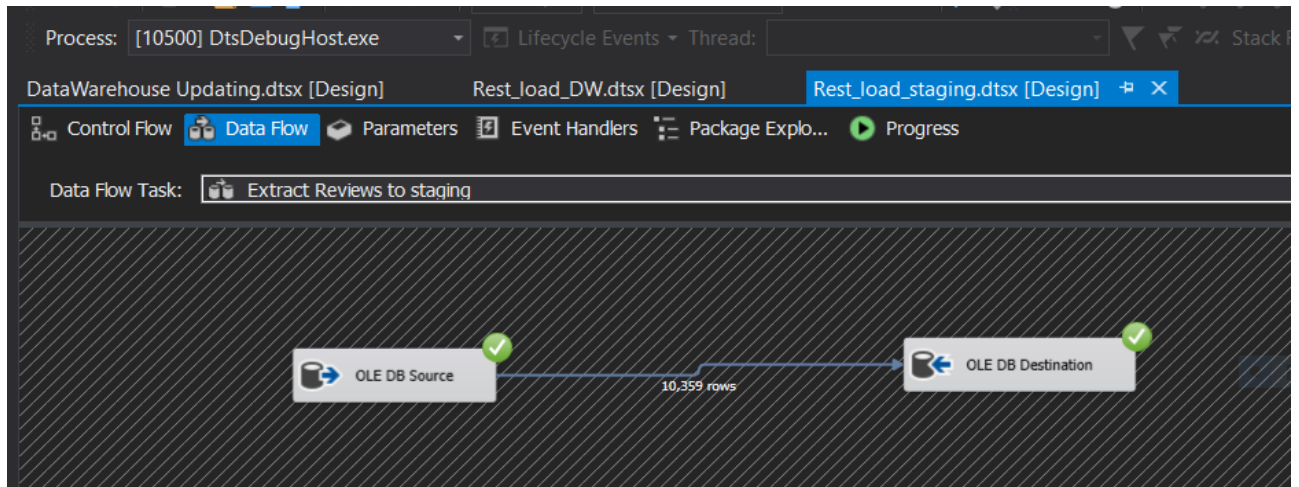
❖ **Load data Restaurants to staging**



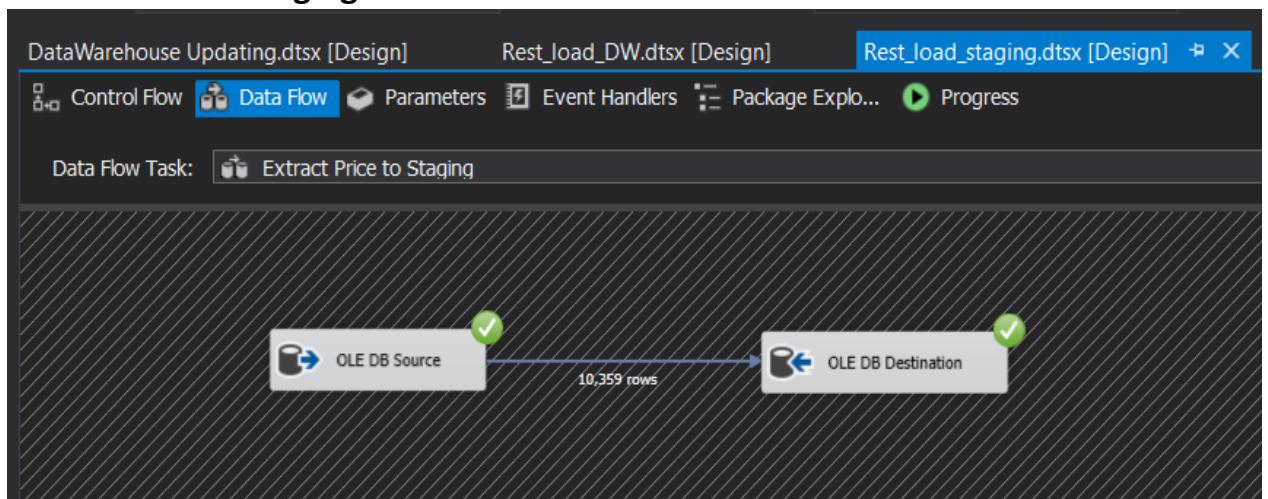
❖ **Load data Meal to staging**



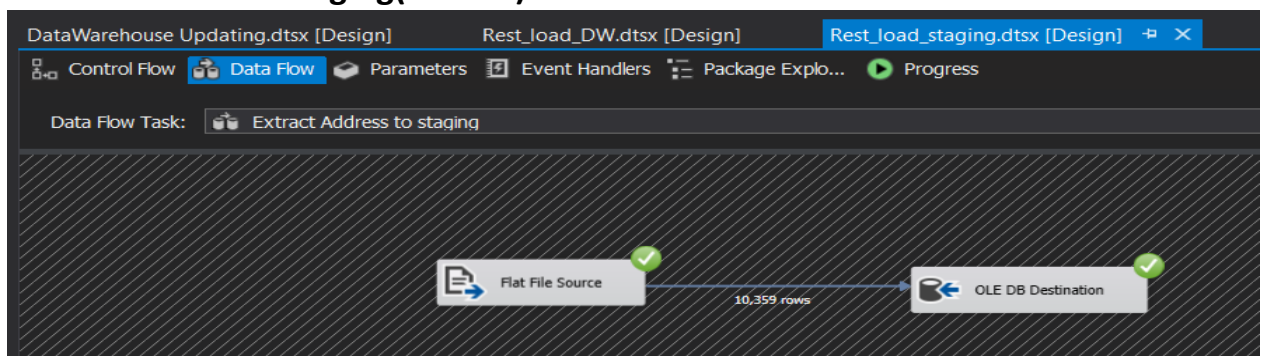
❖ Load data Reviews to staging



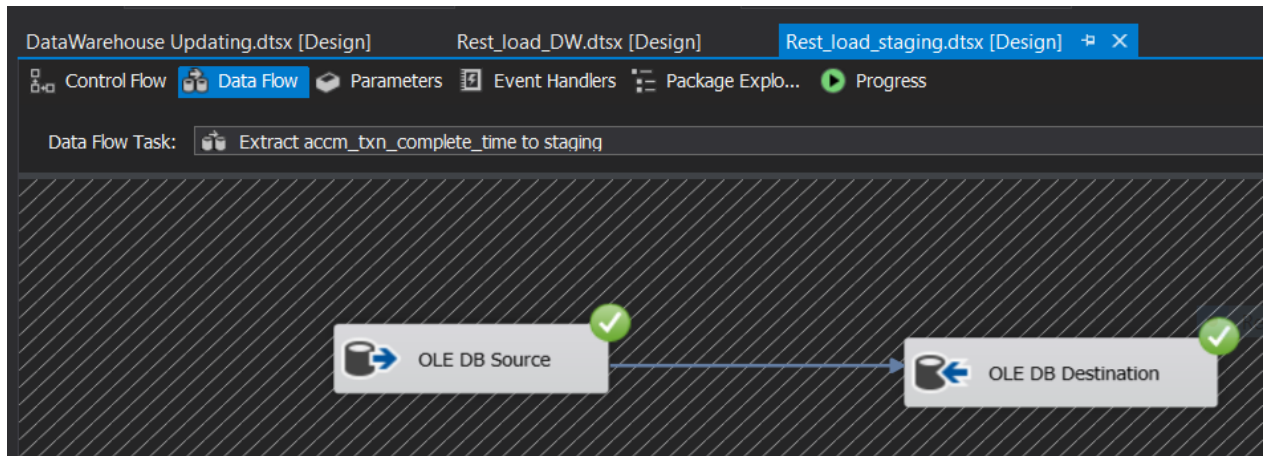
❖ Load data Price to staging



❖ Load data Address to staging(.txt file)



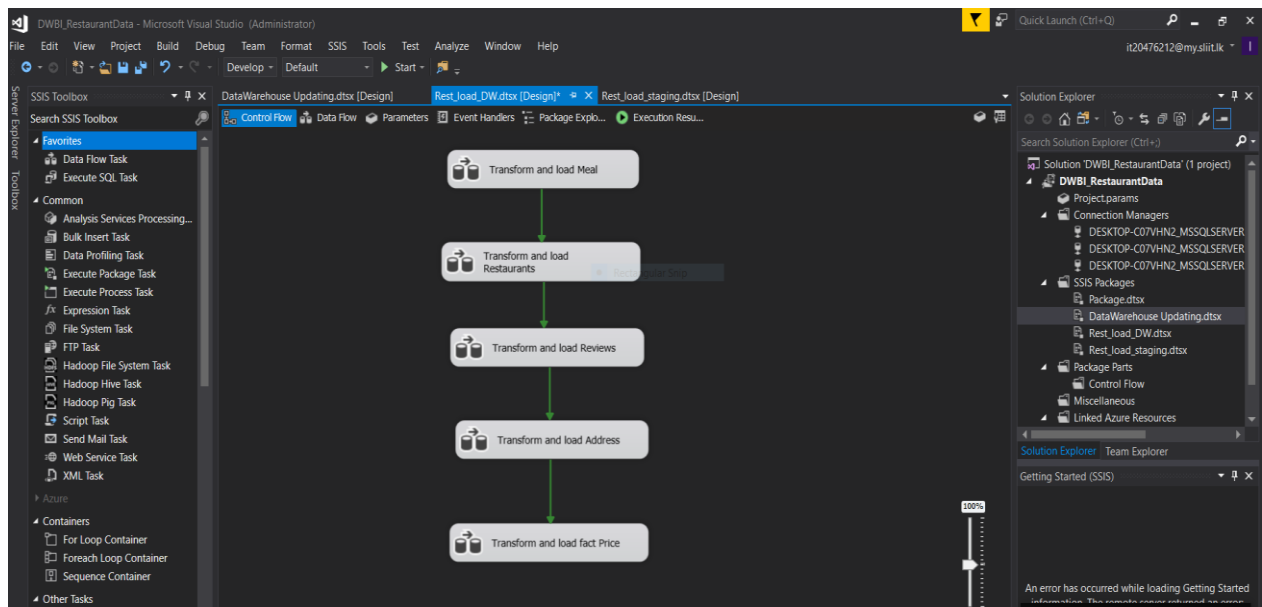
❖ Load data accm_txn_complete_time to staging



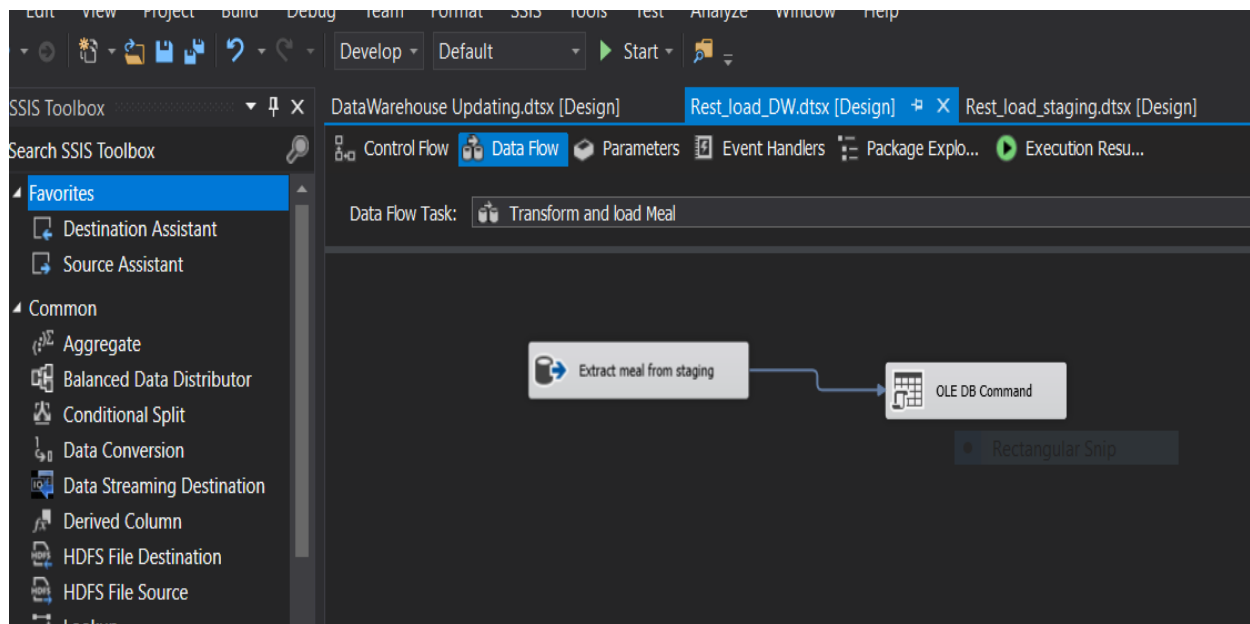
Transforming and Loading

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

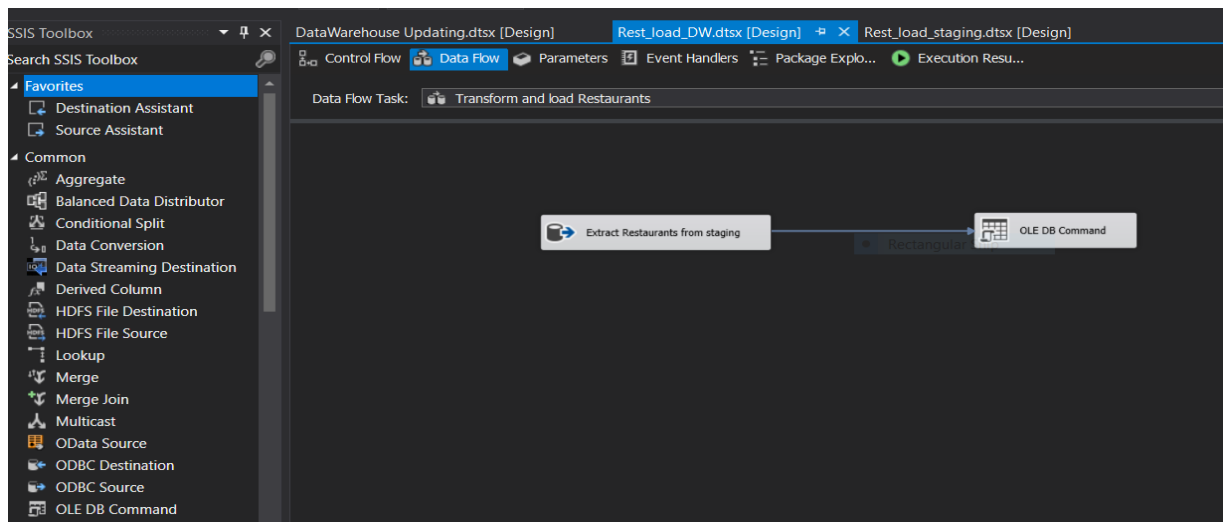
❖ ETL System to Datawarehouse



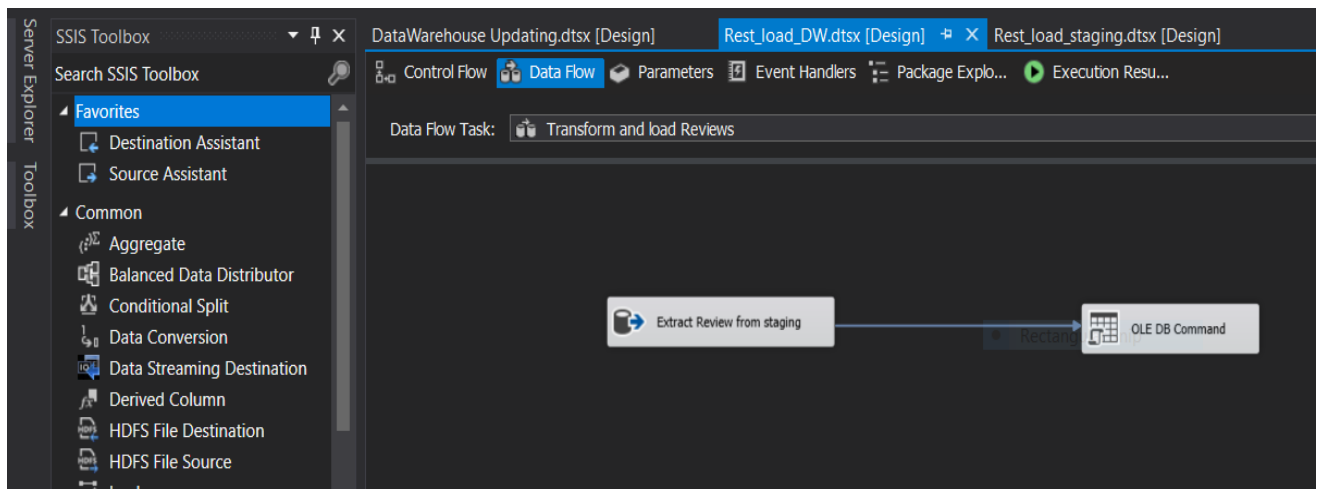
❖ Transform Meal data and load them into DimMeal.



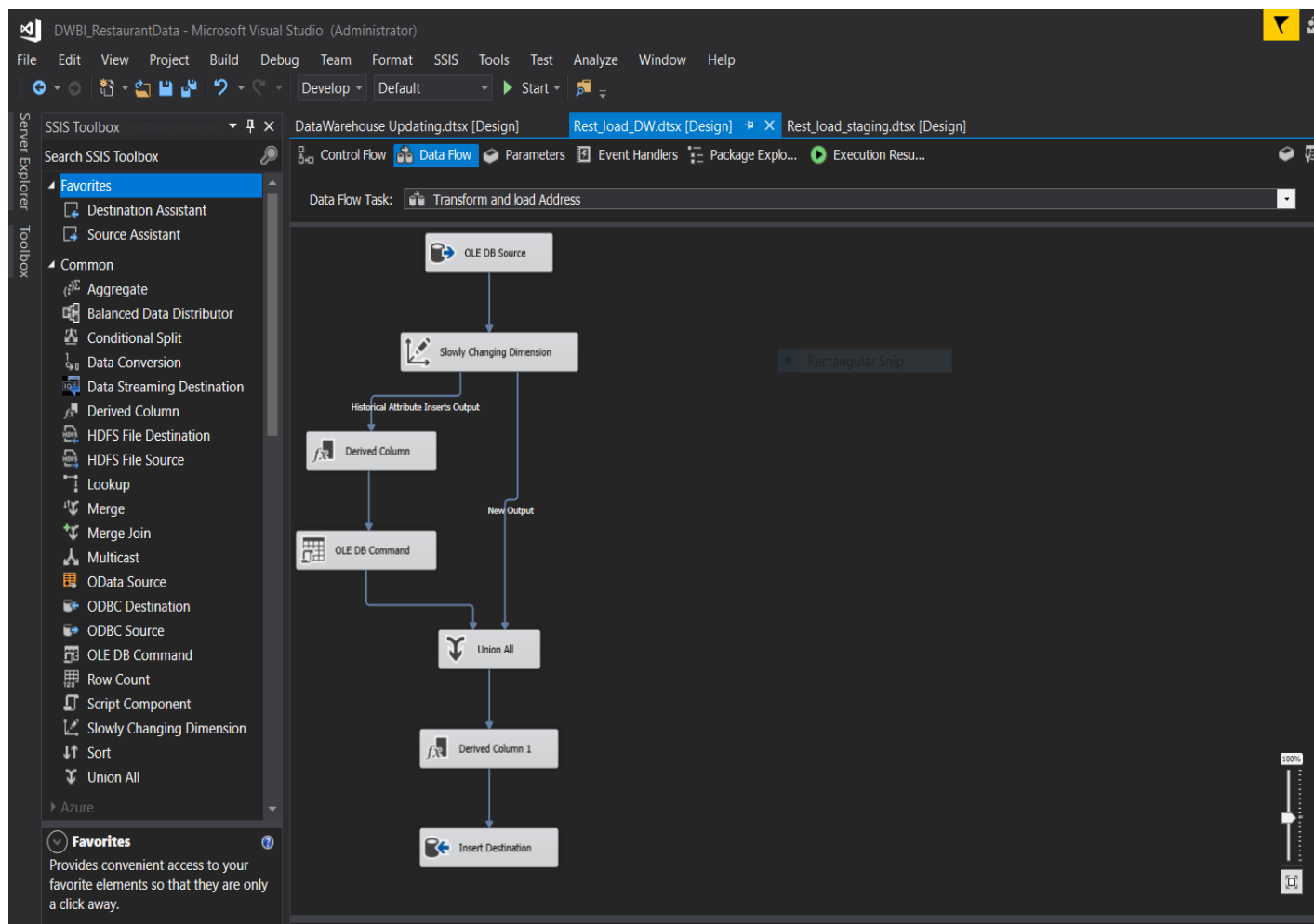
❖ Transform Restaurants data and load them into Restaurants.



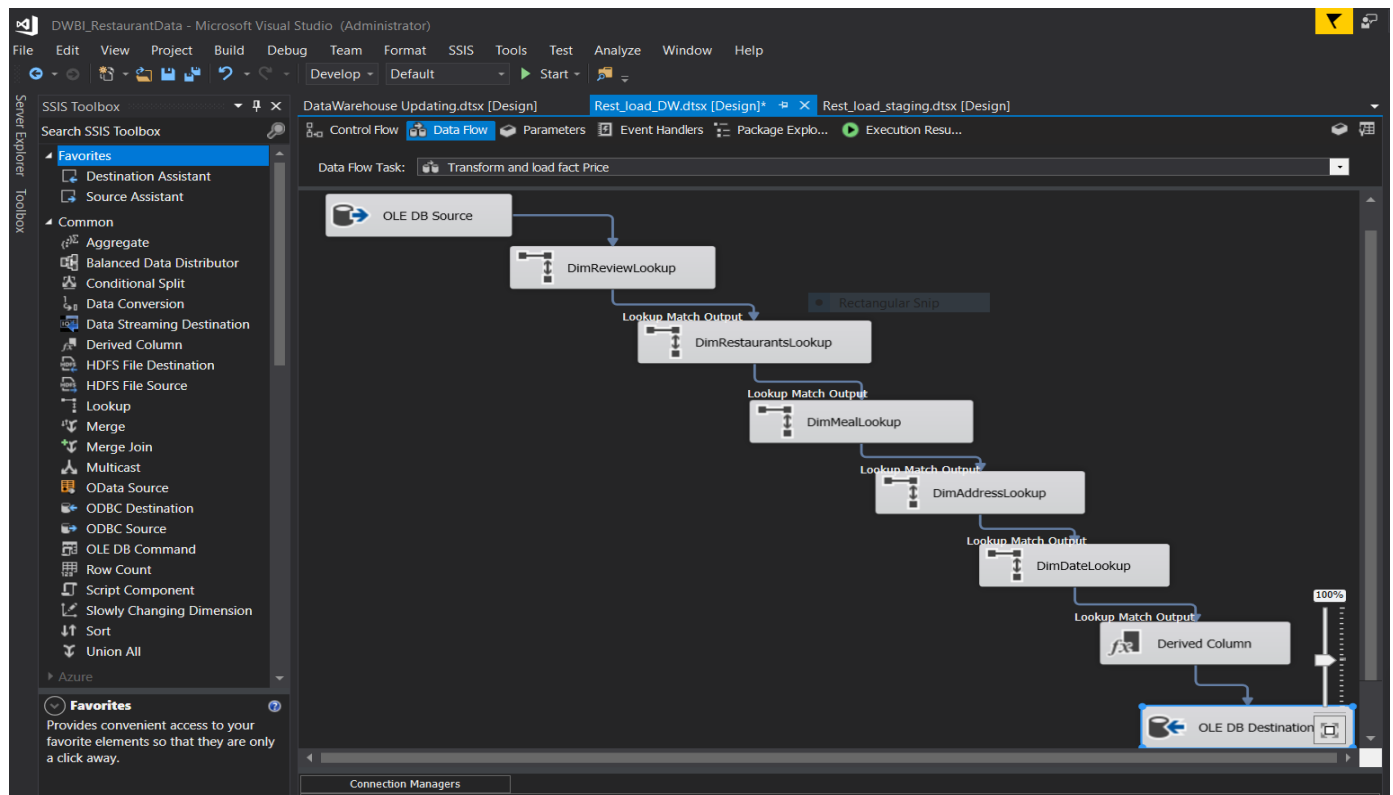
❖ Transform Reviews data and load them into DimReviews



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



❖ Load FactPrice Data from staging



Loaded Data

SQLQuery1.sql - DESKTOP-67F4NUF\SQLEXPRESS.BankingDemo_DW (DESKTOP-67F4NUF\Uvini Wijesinghe (52)) - Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Help

BankingDemo_DW Execute

Object Explorer

Connect

DESKTOP-67F4NUF\SQLEXPRESS (SQL Server 15.0.2000 - DESKTOP-67F4NUF\Uvini Wijesinghe (52))

Databases

System Databases

Database Snapshots

BankingDemo_DW

Database Diagrams

Tables

System Tables

FileTables

Graph Tables

dbo.AccountStaging

dbo.CardStaging

dbo.ClientStaging

dbo.DimCard

dbo.DimClient

dbo.DimDate

dbo.DimDisposition

dbo.DimDistrict

dbo.DimLoan

dbo.DimOrder

dbo.DimTransaction

dbo.DispositionStaging

dbo.DistrictStaging

dbo.FactAccount

dbo.LoanStaging

dbo.OrderStaging

dbo.TransactionsStaging

Views

SQLQuery1.sql - DESKTOP-67F4NUF\Uvini Wijesinghe (52)

```
select * from dbo.DimClient;

select * from dbo.FactAccount;
```

Results

	Client_ID	Sex	DateOfBirth	Day	Month	Year	Age	Social	FirstName	MiddleName	LastName	Phone	Email
1	C00000001	Female	1990-12-13	13	12	1990	29	926-93-2157	Emma	Avaya	Smith	367-171-6840	emma.smith@gmail.com
2	C00000002	Male	1965-02-04	4	2	1965	54	806-94-5725	Noah	Everest	Thompson	212-423-7734	noah.thompson@gmail.com
3	C00000003	Female	1960-10-09	9	10	1960	59	614-70-9100	Olivia	Brooklynne	Johnson	212-425-6932	olivia.johnson@outlook.com
4	C00000004	Male	1976-12-01	1	12	1976	43	580-20-3414	Liam	Irvin	White	951-567-8925	liam.white@gmail.com
5	C00000005	Female	1980-07-03	3	7	1980	39	536-14-5809	Sophia	Danae	Williams	428-265-1568	sophia.williams@gmail.com
6	C00000006	Male	1939-09-22	22	9	1939	80	430-17-5825	Mason	Javen	Lopez	813-629-5038	mason.lopez7@gmail.com
7	C00000007	Male	1949-01-25	25	1	1949	71	305-80-4254	Jacob	Khai	Lee	836-845-8120	jacob.lee7@gmail.com
8	C00000008	Female	1958-02-21	21	2	1958	61	425-96-6358	Ava	Elora	Brown	413-444-9280	ava.brown8@gmail.com

	Account_ID	District_ID	Client_ID	Card_ID	Transaction_ID	Loan_ID	Order_ID	Frequency	ParseDate	Year	Month	Day	ID
1	A00000097	74	C00000116	V00000016	T03533714	L00004986	29562	Monthly Issuance	20160505	2016	5	5	1
2	A00000097	74	C00000116	V00000016	T00030095	L00004986	29562	Monthly Issuance	20160505	2016	5	5	2
3	A00000097	74	C00000116	V00000016	T03533710	L00004986	29562	Monthly Issuance	20160505	2016	5	5	3
4	A00000097	74	C00000116	V00000016	T00030098	L00004986	29562	Monthly Issuance	20160505	2016	5	5	4
5	A00000097	74	C00000116	V00000016	T03533711	L00004986	29562	Monthly Issuance	20160505	2016	5	5	5
6	A00000097	74	C00000116	V00000016	T00030097	L00004986	29562	Monthly Issuance	20160505	2016	5	5	6
7	A00000097	74	C00000116	V00000016	T03533712	L00004986	29562	Monthly Issuance	20160505	2016	5	5	7
8	A00000097	74	C00000116	V00000016	T00030096	L00004986	29562	Monthly Issuance	20160505	2016	5	5	8
9	A00000097	74	C00000116	V00000016	T03533713	L00004986	29562	Monthly Issuance	20160505	2016	5	5	9
10	A00000097	74	C00000116	V00000016	T00030092	L00004986	29562	Monthly Issuance	20160505	2016	5	5	10

Query executed successfully. DESKTOP-67F4NUF\SQLEXPRESS ... DESKTOP-67F4NUF\Uvini ... BankingDemo_DW | 00:00:48 | 1,021,953 rows

Ready Ln 9 Col 1 Ch 1 INS

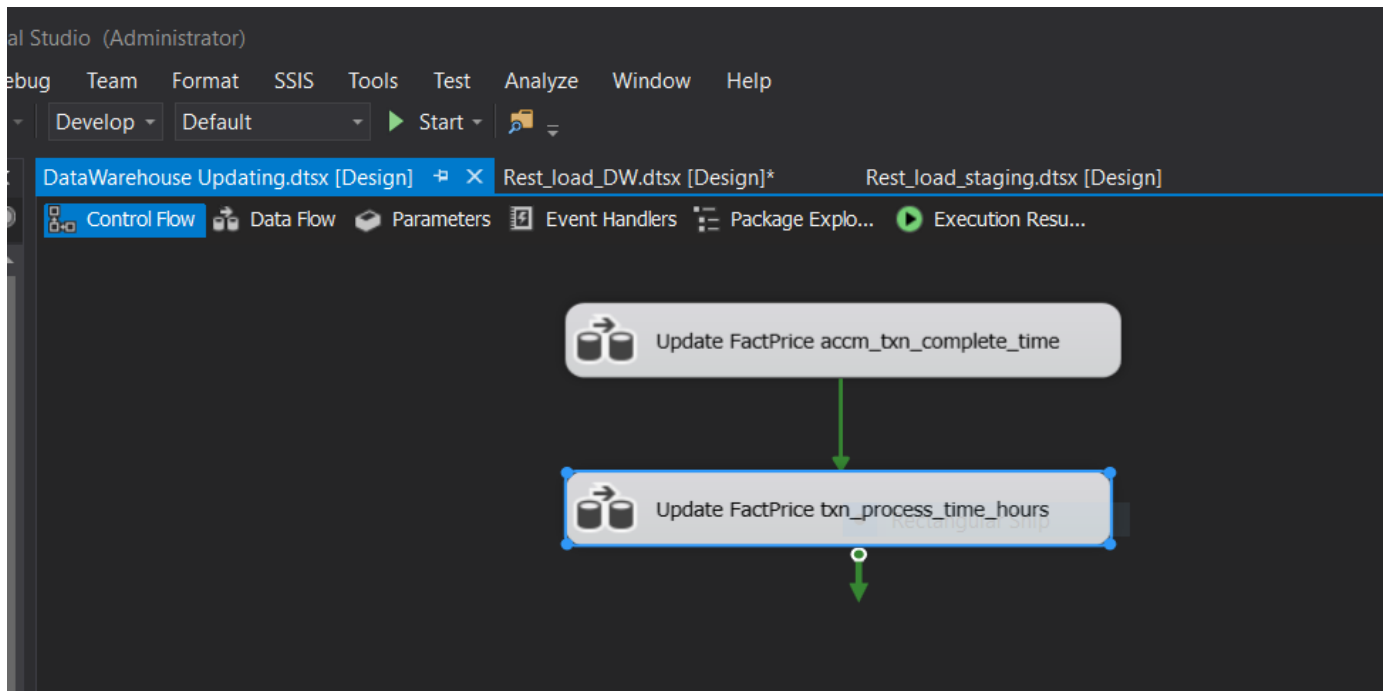
Type here to search

8:37 PM 28-Apr-20

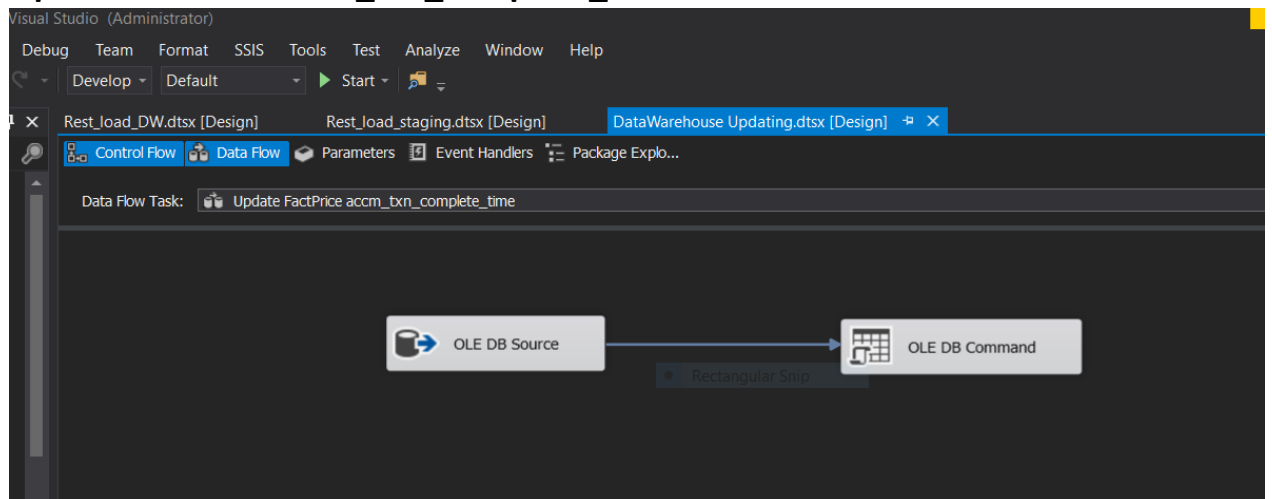
7. DataWarehouse Updating

In order to creating Accumulated fact table I created a new SSIS package and updated accm_txn_complete_time and txn_process_time_hours.

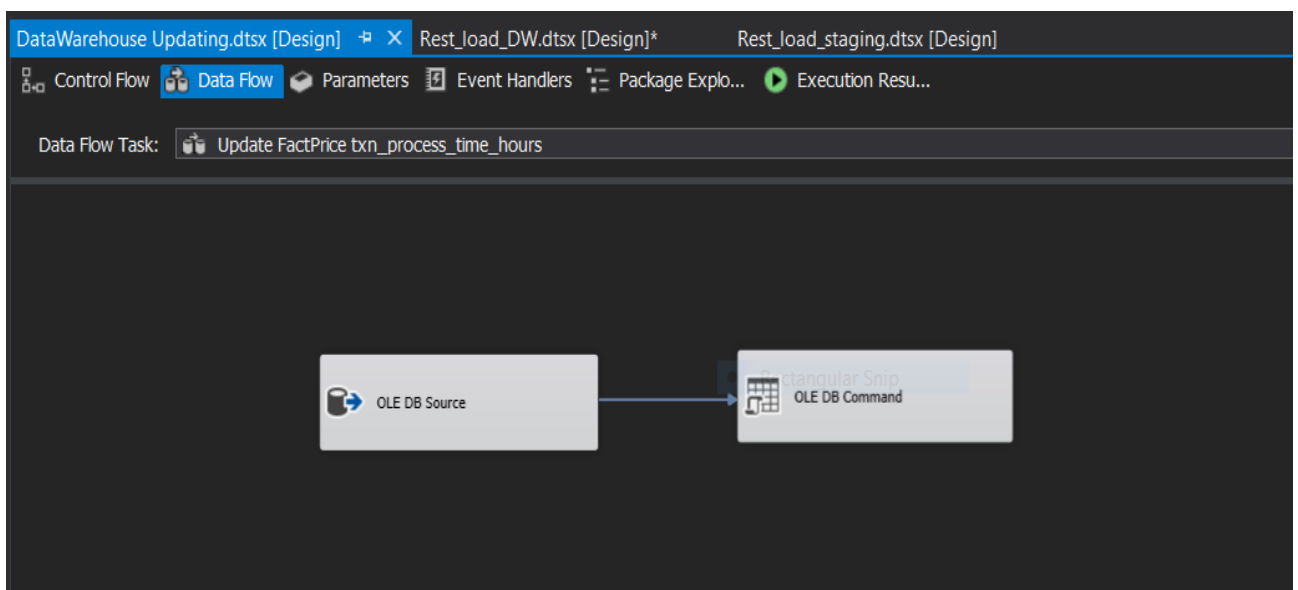
➤ Datawarehouse updating



➤ **Update factPrice accm_txn_complete_time**



➤ **Update factPrice txn_process_time_hours**



7.2 Accumulated Fact Table (FactPrice)

100 %

Results Messages

	res_id	meal_id	Price	discount	DateKey	AddressKey	MealKey	RestaurantsKey	ReviewKey	InsertDate	ModifiedDate
1	1	26830	50	0	NULL	1	1	1	413	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
2	2	54253	50	0	NULL	2	2	2	414	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
3	3	31962	50	0	NULL	3	3	3	415	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
4	4	33647	50	0	NULL	4	4	4	416	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
5	5	48199	50	0	NULL	5	5	5	417	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
6	6	57397	50	0	NULL	6	6	6	418	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
7	7	12424	50	0	NULL	7	7	7	419	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
8	8	14930	50	0	NULL	8	8	8	420	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
9	9	16657	50	0	NULL	9	9	9	421	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
10	10	10537	50	0	NULL	10	10	10	422	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
11	11	19560	50	0	NULL	11	11	11	423	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
12	12	21653	50	0	NULL	12	12	12	424	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
13	13	32083	50	0	NULL	13	13	13	425	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
14	14	10423	50	0	NULL	14	14	14	426	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
15	15	10565	50	0	NULL	15	15	15	427	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
16	16	17035	50	0	NULL	16	16	16	428	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
17	17	27801	50	0	NULL	17	17	17	429	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
18	18	30645	50	0	NULL	18	18	18	430	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
19	19	30832	50	0	NULL	19	19	19	431	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
20	20	34047	56....	0	NULL	20	20	20	432	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
21	21	34139	65....	0	NULL	21	21	21	433	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687
22	22	45502	55.5	0	NULL	22	22	22	434	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687

Query executed successfully. DESKTOP-C07VHN2\MSSQLSERVER... DESKTOP

y	InsertDate	ModifiedDate	accm_txn_create_time	accm_txn_complete_time	txn_process_time_hours
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-14 00:00:00.000	-59
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-15 00:00:00.000	-35
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 00:00:00.000	-11
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-17 00:00:00.000	13
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-18 00:00:00.000	37
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-19 00:00:00.000	61
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-20 00:00:00.000	85
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-21 00:00:00.000	109
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-22 00:00:00.000	133
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-23 00:00:00.000	157
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-24 00:00:00.000	181
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-25 00:00:00.000	205
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-26 00:00:00.000	229
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-27 00:00:00.000	253
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-28 00:00:00.000	277
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-29 00:00:00.000	301
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-30 00:00:00.000	325
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-31 00:00:00.000	349
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-06-01 00:00:00.000	373
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-06-02 00:00:00.000	397
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-06-03 00:00:00.000	421
	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-05-16 11:37:30.687	2022-06-04 00:00:00.000	445