Data Warehousing & Business Intelligence

Customer Credit Card Promotions

Assignment 1

Submitted by:

Jayasuriya K.E – IT17179386

Submitted to:

MR. Jesuthasan Alosius

Data Set Selection

Background

Customers receive coupons under various campaigns and may choose to redeem it. They can redeem the given coupon for any valid product for that coupon as per coupon item mapping within the duration between campaign start date and end date. Next, the customer will redeem the coupon for an item at the retailer store and that will reflect in the transaction table in the column coupon discount.

Content

The dataset was download from this link

- https://www.kaggle.com/bharath901/amexpert-2019/version/1#train.csv

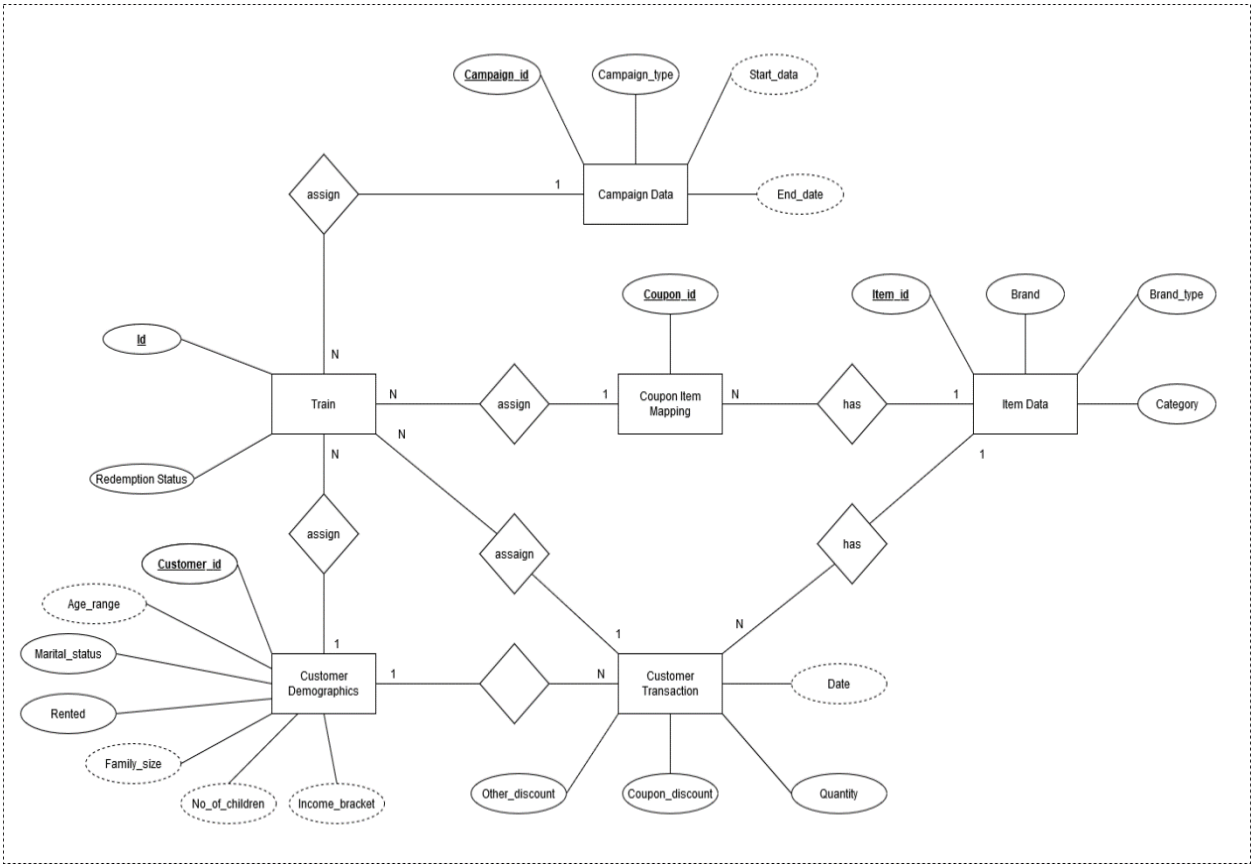
Preparation of Data Sources

For Data Extraction, We have to prepare our data sources. Therefore I have extracted these data sources types.

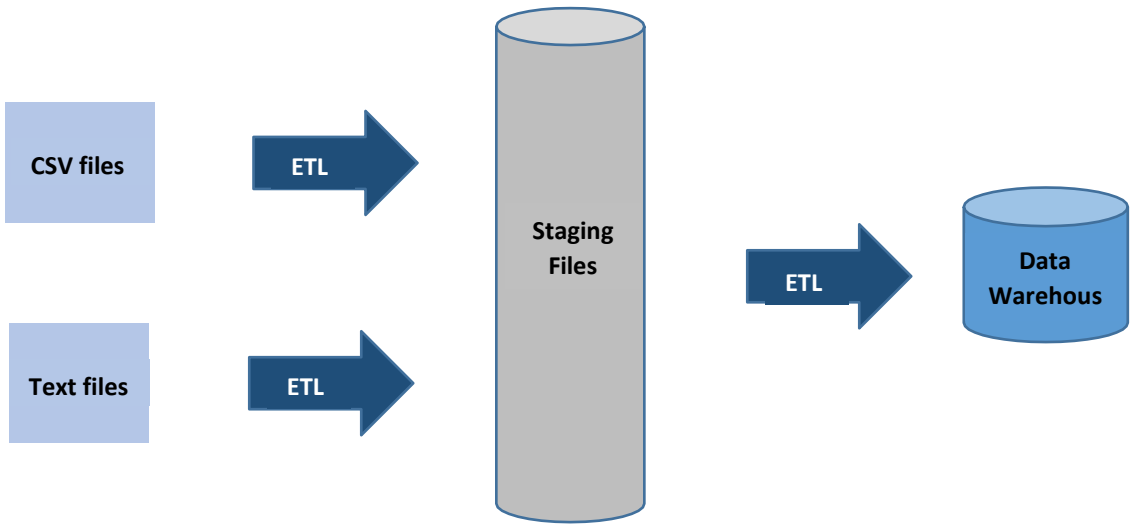
1. Excel files(.csv) – Customer_demographics, Customer_transaction, Item_data, Train_data
2. Text files(.txt) – Campaign_data, Coupan_item_mapping

At the beginning, there are no text files in my data set. There for I have converted few of them into text files for good.

ER Diagram



Solution Architecture



- Item data staging.
- Coupon item mapping staging.
- Customer demographics staging.
- Campaign data to staging.
- Train to staging.
- Customer transaction staging.

Architecture Components

- **Data Sources**

External Sources

- **Extract, Transform and Load**

Extract -: reading data from external sources and staging it.

Transform -: Combine data, which were came from multiple staging sources and De-duplicating it.

Load -: loading data to destination by primary keys or surrogate keys assignment. In addition, foreign keys checks, indexing.

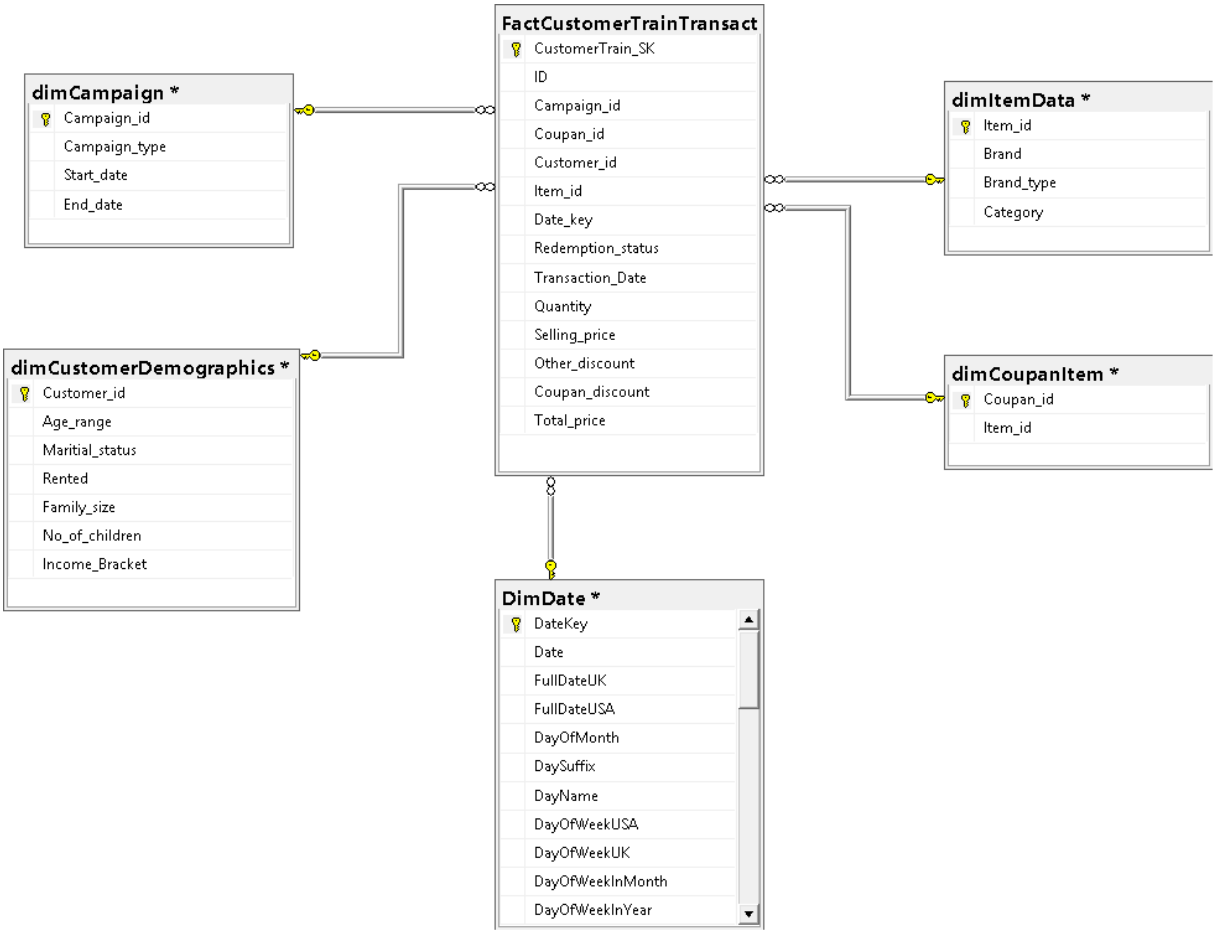
- **Data Warehouse**

Dimensional Modeling -: Facts and Dimensions

Can use many schemas -: In my case, I have used star schema.

Data Warehouse Design and Development

Star Schema



ETL Development

ETL (Extract Transform Load)

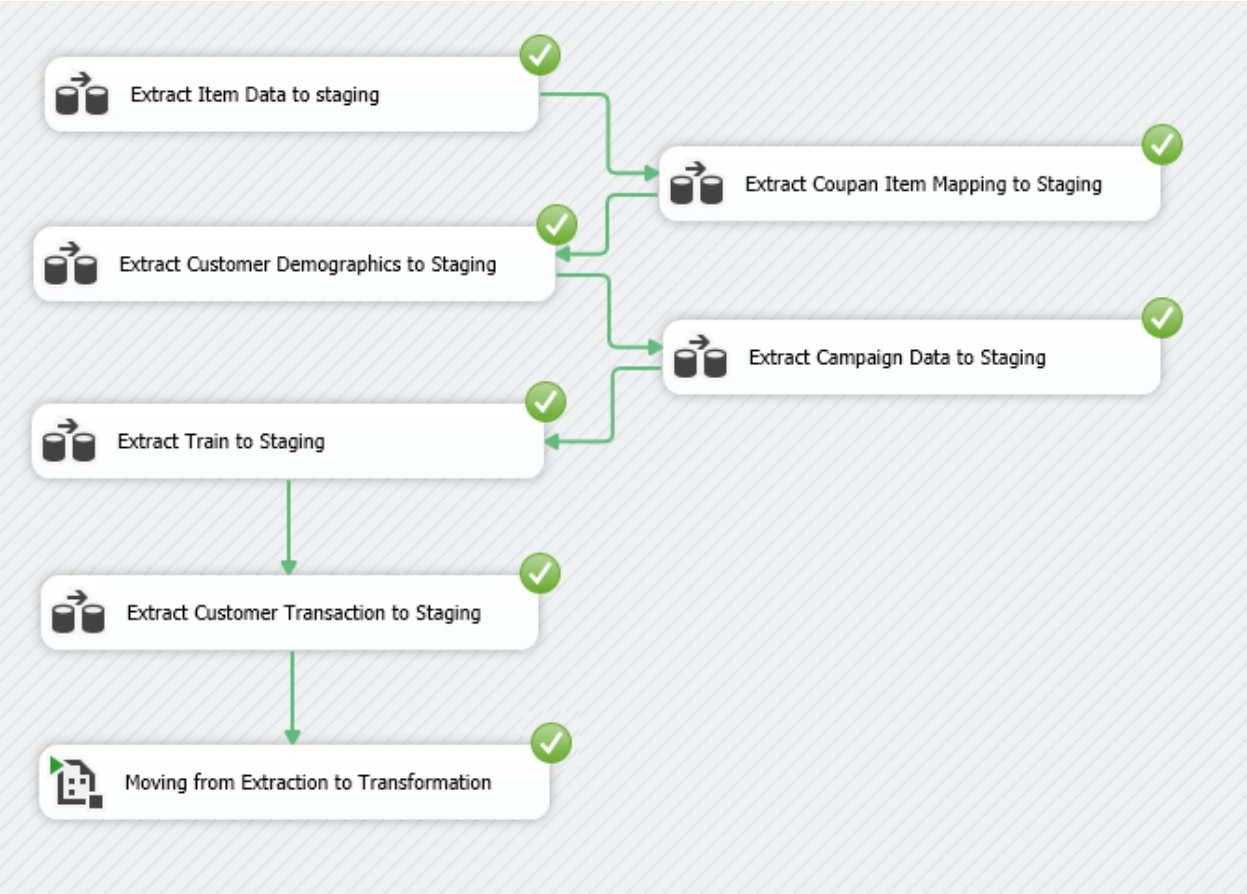
When Data Warehousing we have used ETL process for load data from sources and apply cleaning, merging, sorting etc. After that send that data into our data warehouse which is referring star schema. Now, let us talk about this process by examples.

Staging and extract

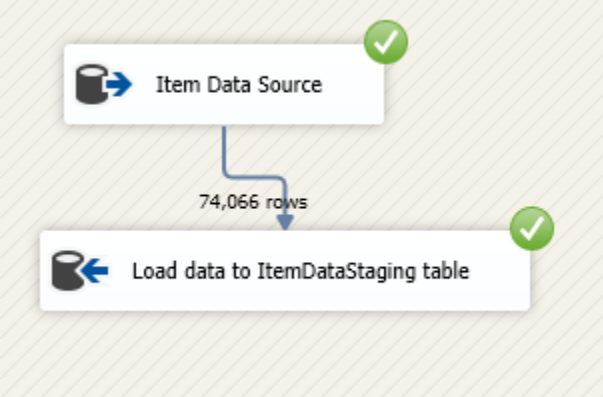
First, we need to create database for our source files (I have created Assignment1_SourceDW database for that). Then we have to implement staging process to send our data to staging tables (I have created Assignment1_Staging database for that). Finally, we can start the process of sending data from our sources to staging.

I have inserted only .csv files to the Assignment1_SourceDW database. I use flat file source to call my text files directly (not from any database) to the Assignment1_Staging database.

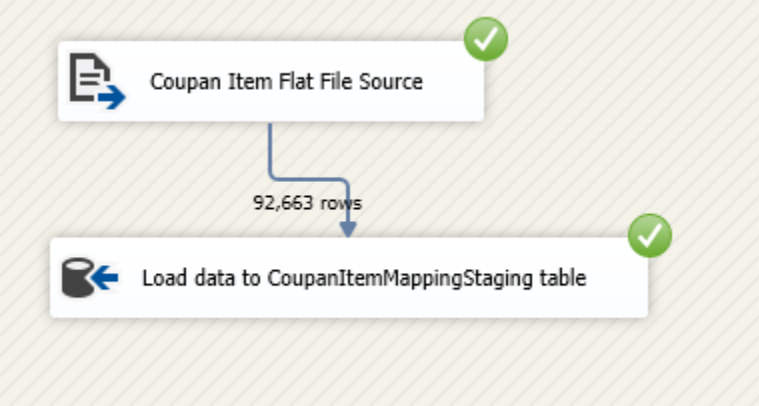
Extract data from sources to staging tables



Data flows of all staging



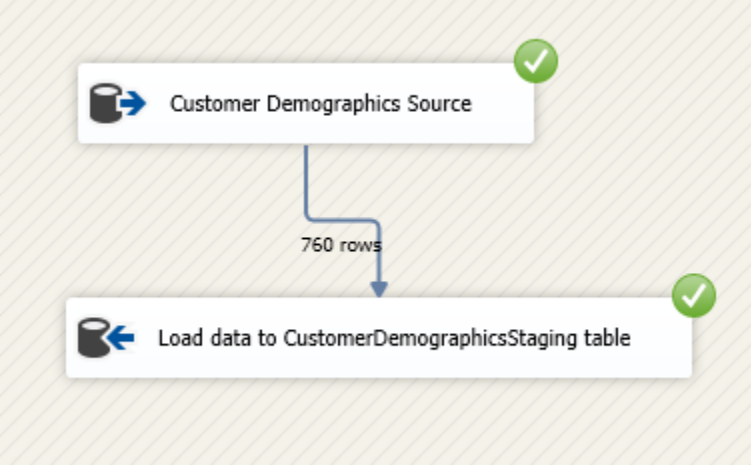
Item data to staging



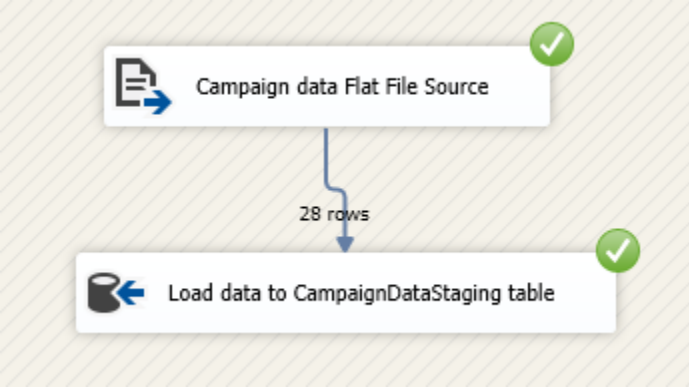
Coupon item mapping to staging

I have used flat file source to read

My text file

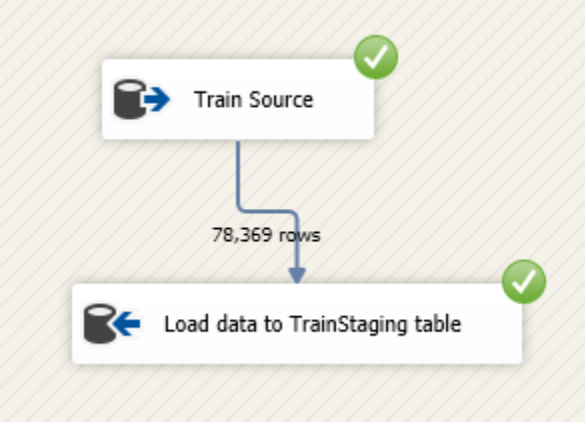


Customer demographics to staging

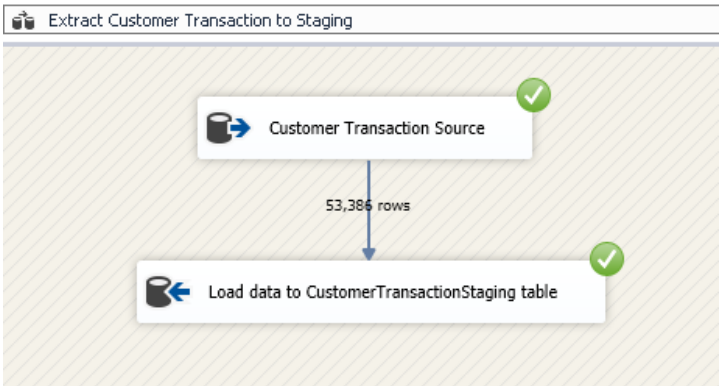


Campaign data to staging

I have used flat file source to read my text file.



Train data to staging

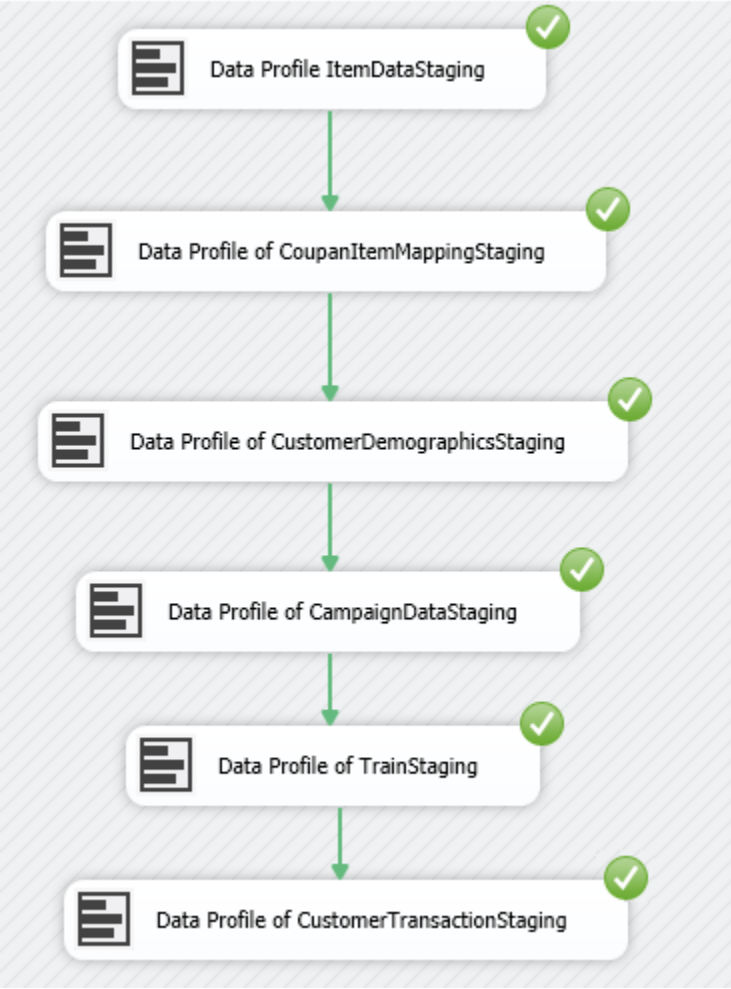


Customer transaction to staging

TRUNCATE TABLE -: I have given one example for this because I need to keep my doc without unwanted extending

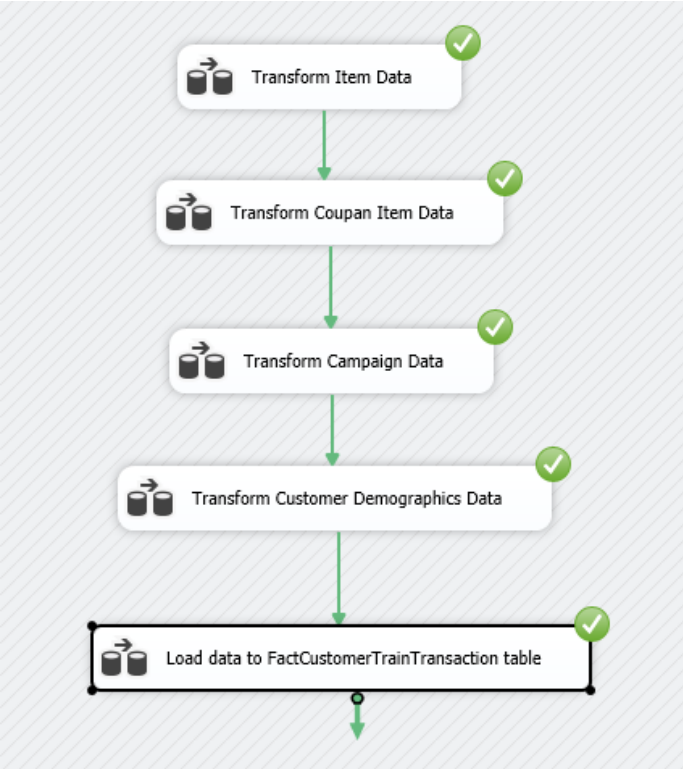
Data Profiling

Data profiles tasks can analysis patterns of our data sources and can determine how clean the data might be, prior to loading it into the data warehouse. By performing a profiling task on incoming data from sources, you will be able to verify your new data meets the quality you expect prior to loading the data into its final location. If the data does not meet the quality, you normally expect data profiling allows you to reject the incoming data.

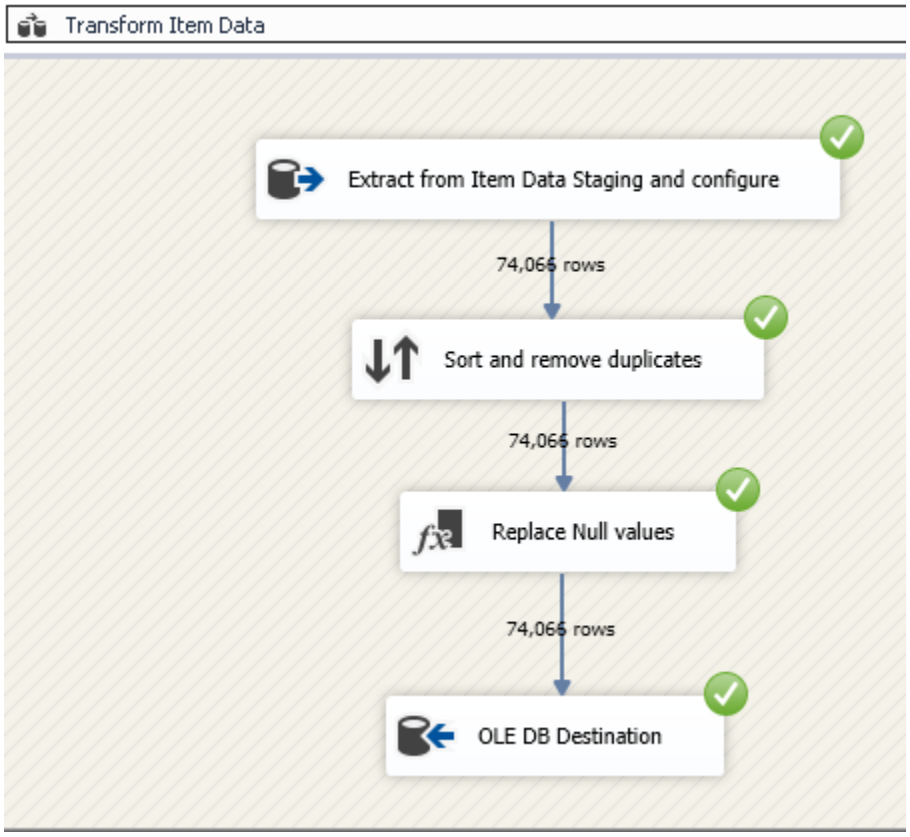


Data Transformation

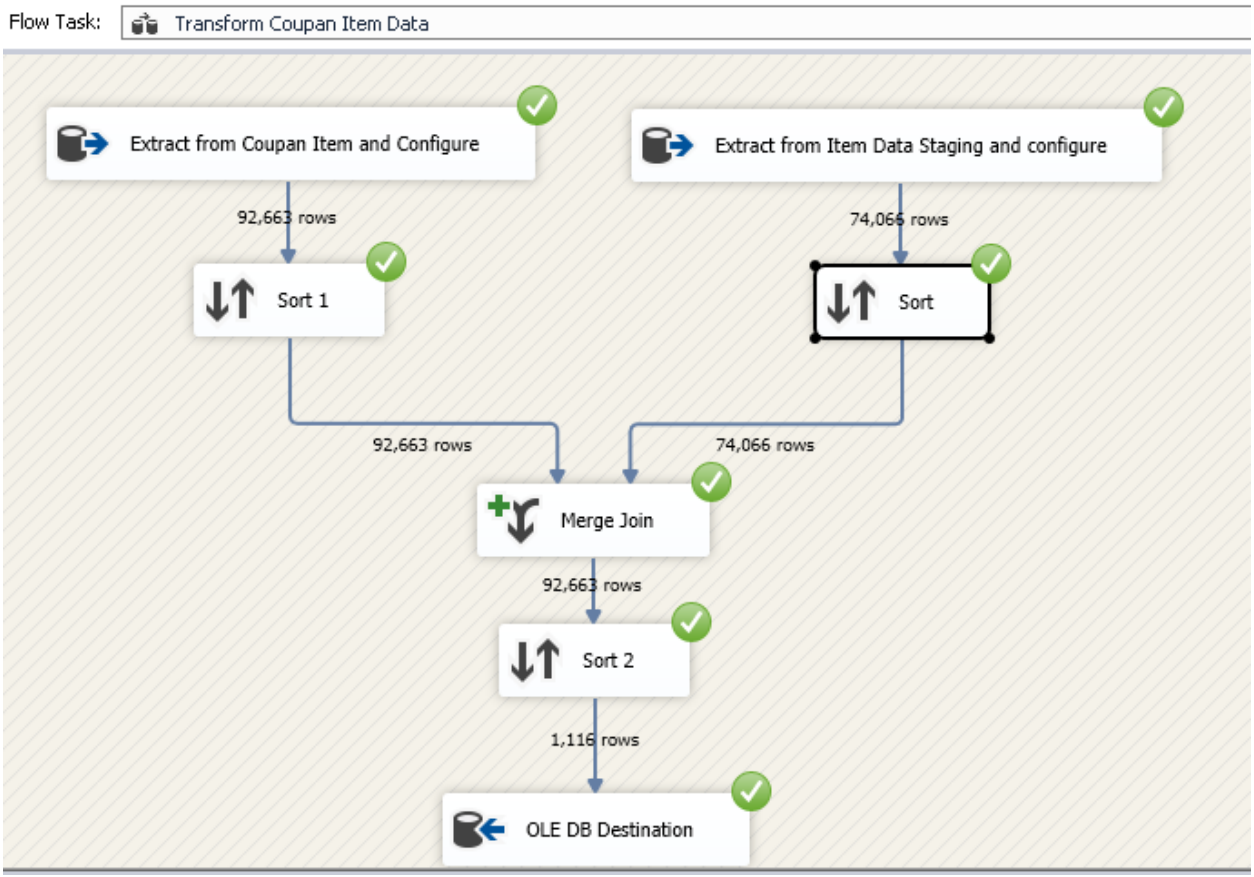
Transformations are the data flow components, which used to perform sorting, merging, union, aggregations, data cleaning and data distribution. By using these components, we are trying to fill our dim tables and fact tables. First, we need to load our staged tables and perform above-mentioned components and transform that data until they are fitting to our dim and fact tables. First, we need to transform and send data to our dim tables and finally we transform and send data to out fact tables.



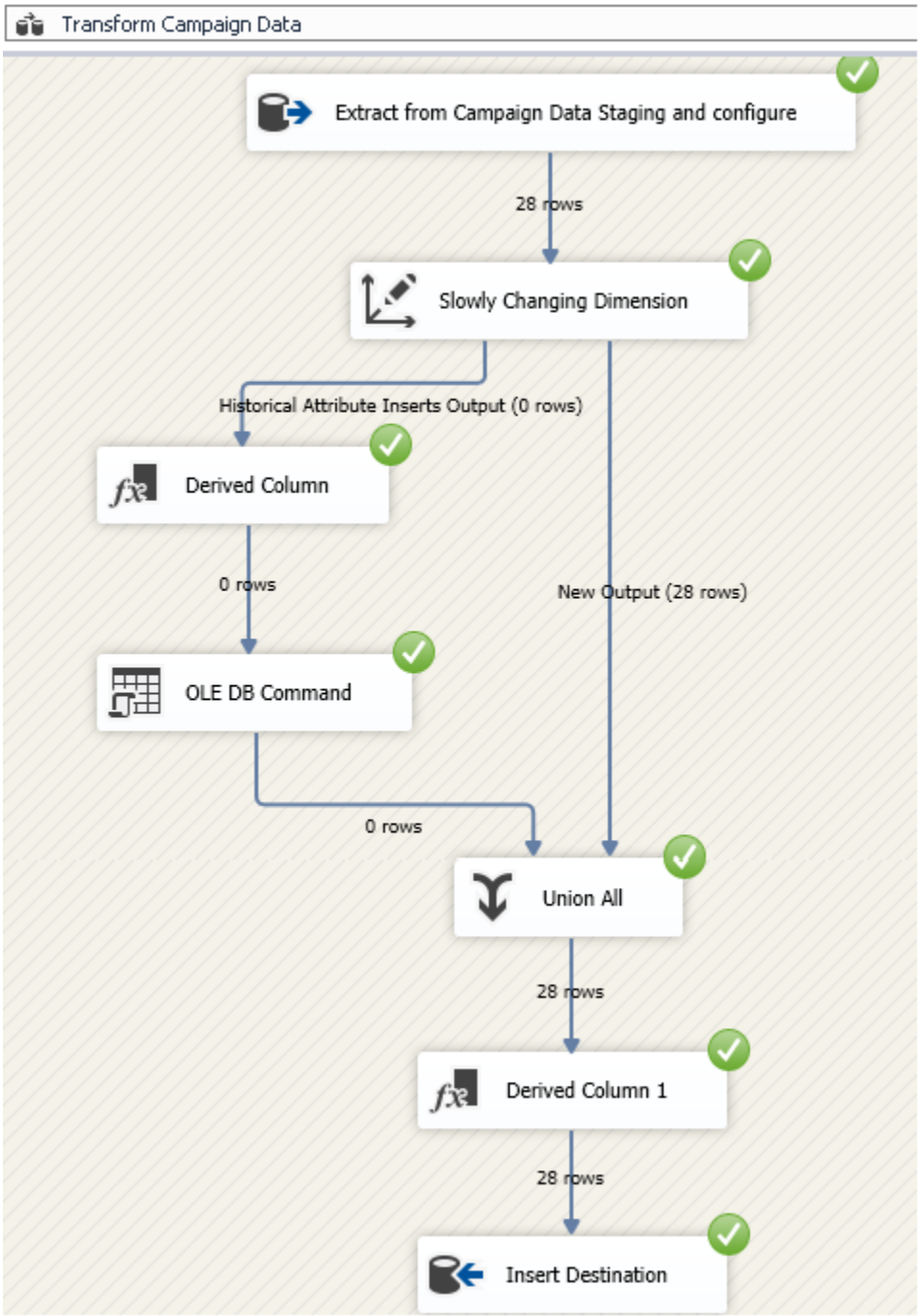
Transform Item data



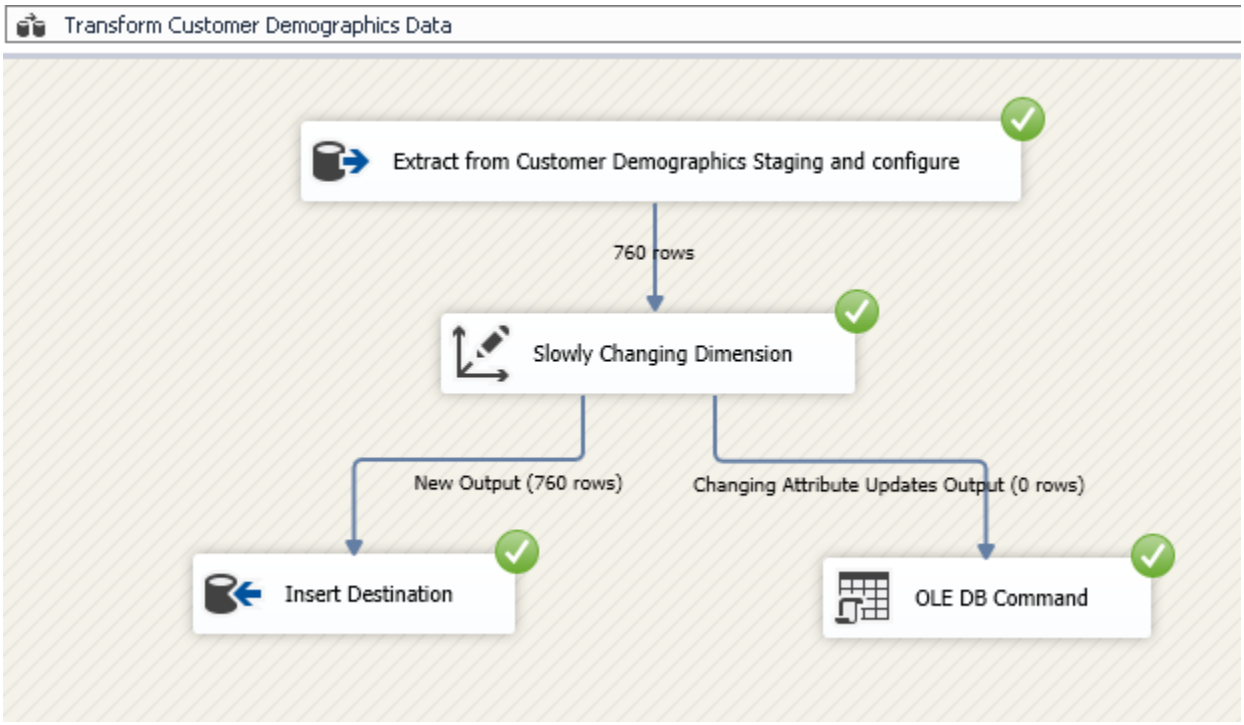
Transform coupon item data



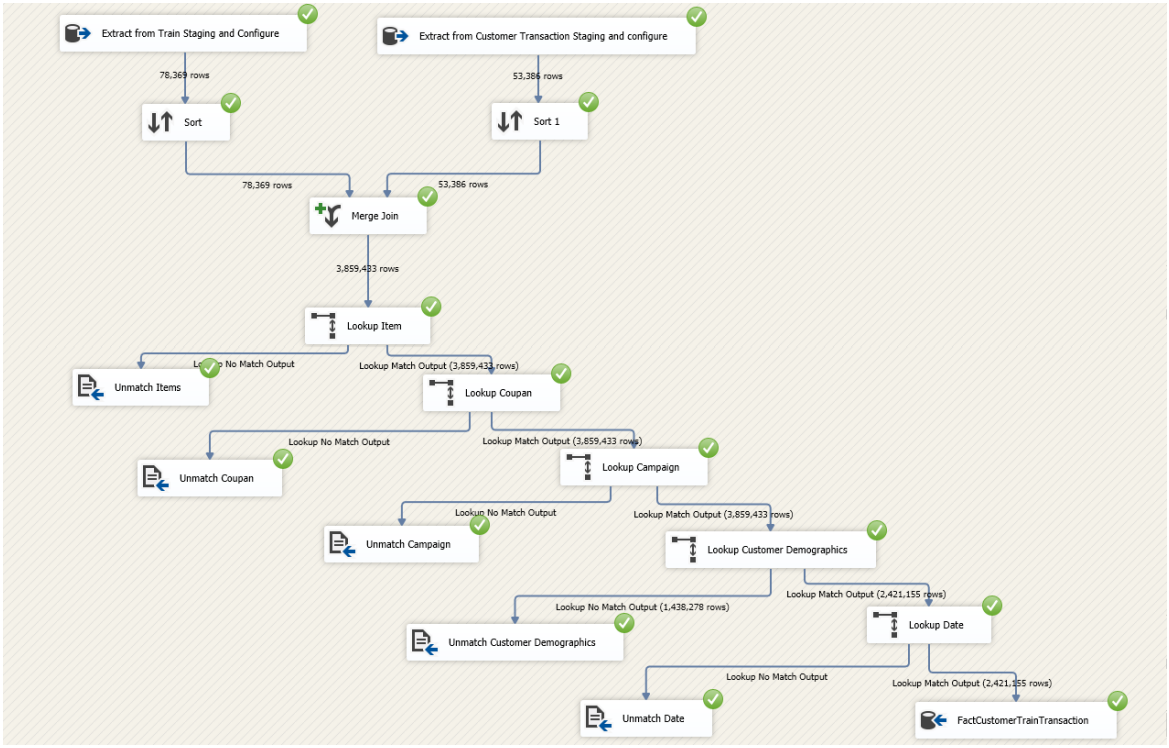
Transform Campaign data



Transform customer demographics data



Load data to FactCustomerTansaction table

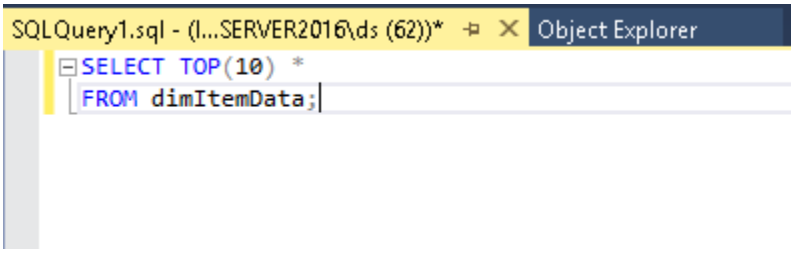


Test Cases

After transformation, I need to assure all staging data send to the dim tables and fact table. There for I have create some simple scripts for whether my dim tables and fact table filled. I write a sql statements to get top 10 rows of my tables.

Test case for dimTrainData table

- **Script**



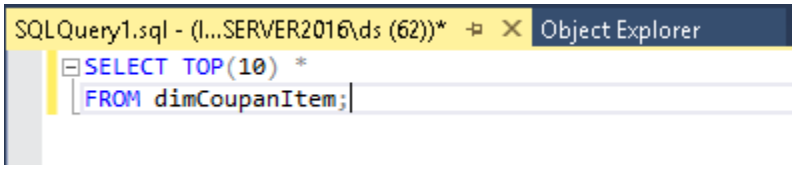
- **Table**

Results | Messages

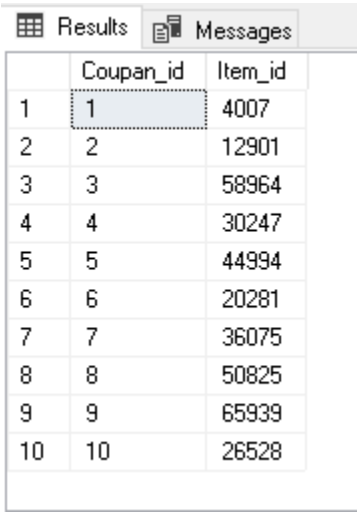
	Item_id	Brand	Brand_type	Category
1	1	1	Established	Grocery
2	2	1	Established	Miscellaneous
3	3	56	Local	Bakery
4	4	56	Local	Grocery
5	5	56	Local	Grocery
6	6	56	Local	Grocery
7	7	56	Local	Pharmaceutical
8	8	56	Local	Bakery
9	9	11	Local	Grocery
10	10	56	Local	Grocery

Test case for dimCouponItem table

- **Script**



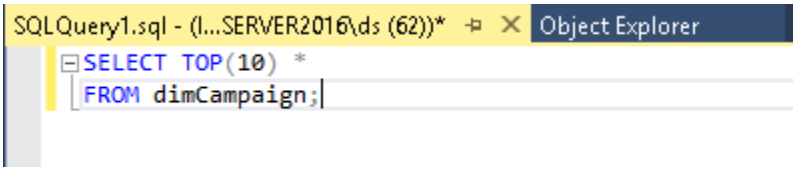
- **Table**



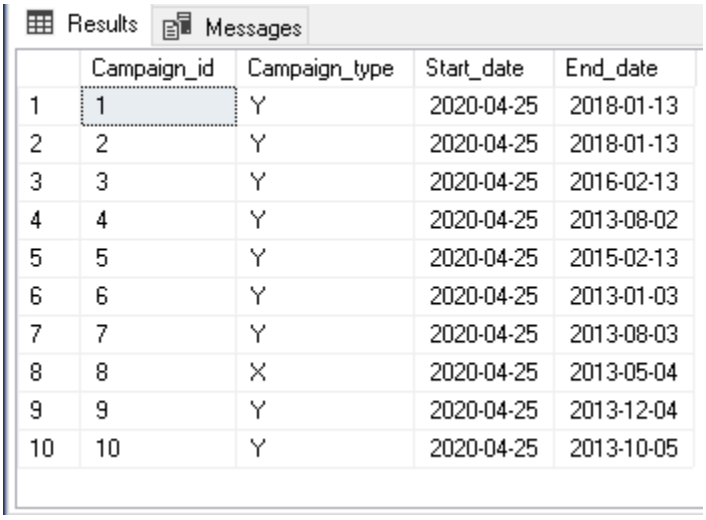
	Coupan_id	Item_id
1	1	4007
2	2	12901
3	3	58964
4	4	30247
5	5	44994
6	6	20281
7	7	36075
8	8	50825
9	9	65939
10	10	26528

Test case for dimCampaign table

- **Script**



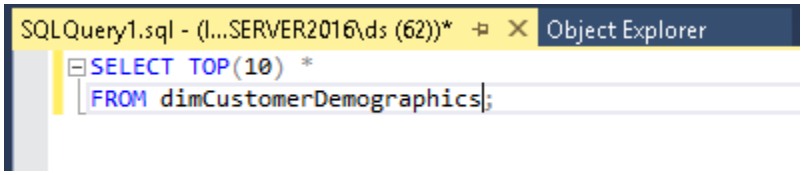
- **Table**



	Campaign_id	Campaign_type	Start_date	End_date
1	1	Y	2020-04-25	2018-01-13
2	2	Y	2020-04-25	2018-01-13
3	3	Y	2020-04-25	2016-02-13
4	4	Y	2020-04-25	2013-08-02
5	5	Y	2020-04-25	2015-02-13
6	6	Y	2020-04-25	2013-01-03
7	7	Y	2020-04-25	2013-08-03
8	8	X	2020-04-25	2013-05-04
9	9	Y	2020-04-25	2013-12-04
10	10	Y	2020-04-25	2013-10-05

Test case for dimCustomerDemographics table

- **Script**



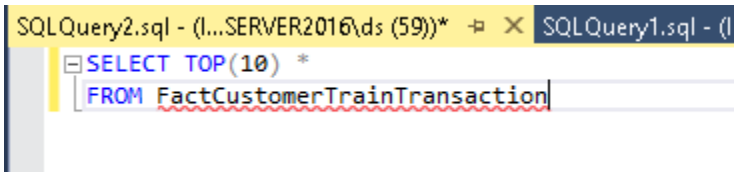
- **Table**

Results | Messages

	Customer_id	Age_range	Maritial_status	Rented	Family_size	No_of_children	Income_Bracket
1	1	70+	Married	0	2		4
2	6	46-55	Married	0	2		5
3	7	26-35		0	3	1	3
4	8	26-35		0	4	2	6
5	10	46-55	Single	0	1		5
6	11	70+	Single	0	2		1
7	12	46-55	Married	0	2		7
8	13	36-45	Single	0	1		2
9	14	26-35	Married	1	2		6
10	15	46-55	Married	0	2		6

Test case for factCustomerTrainTransaction table

- **Script**



- **Train**

Results Messages

	CustomerTrain_SK	ID	Campaign_id	Coupan_id	Customer_id	Item_id	Date_key	Redemption_status	Transaction_Date	Quantity	Selling_price	Other_discount	Coupan_discount	Total_price
1	1	9960	8	700	1	25583	20120221	0	2012-02-21	1	78.00	0.00	0.00	78.00
2	2	9960	8	700	1	53524	20120308	0	2012-03-08	2	178.00	-148.00	0.00	208.00
3	3	9960	8	700	1	28939	20120221	0	2012-02-21	1	66.00	-11.00	0.00	55.00
4	4	9960	8	700	1	56179	20120308	0	2012-03-08	1	83.00	-48.00	0.00	35.00
5	5	9960	8	700	1	28816	20120221	0	2012-02-21	1	142.00	0.00	0.00	142.00
6	6	9960	8	700	1	51608	20120308	0	2012-03-08	1	106.00	-56.00	0.00	50.00
7	7	9960	8	700	1	28801	20120221	0	2012-02-21	1	53.00	0.00	0.00	53.00
8	8	9960	8	700	1	50127	20120308	0	2012-03-08	1	124.00	0.00	0.00	124.00
9	9	9960	8	700	1	27557	20120221	0	2012-02-21	1	38.00	0.00	0.00	38.00
10	10	9960	8	700	1	57921	20120308	0	2012-03-08	1	66.00	-39.00	0.00	27.00

Pop up errors and Solutions

When demonstrate the warehouse I have got some errors in some moment. I have find solutions for all that errors by google searching and by my own knowledge.

1. **Errors without any visible details**

   Sometimes, there is no message displaying on the component when I point my mouse pointer on the error mark. This occurs when you are not correctly done merging, sorting, aggregations, etc. And also it occurs if you changed your source table or staging table.

   Solution -: Correct your merging, sorting, aggregation, etc.

   -: Delete components, which are showing errors and add that component again and configure it again.

2. **Unicode, Non-Unicode Error**

   Sometimes this error occurred on few components because your staged table columns defined in another format specifier and your dim table columns defined as another format specifier.

   Solution -: Click on the component, which have this error.

   -: Then right click on that component and go to advance editor.

   -: Then go to “input and output columns” tab and expand the plus marks and select output columns.

   -: Select a column which was error pointed at and edit it’s data type to suitable which will match with our dim table. Do this for all columns, which have this error. I will add sample data types with their matchings.

SQL Server data type	SSIS data type
INT	DT_I4 (four-byte signed integer)
NVARCHAR(50)	DT_WSTR (Unicode string)
BIT	DT_BOOL (Boolean)
SMALLINT	DT_I2 (two-byte signed integer)
MONEY	DT_CY (currency)
DECIMAL(8,2)	DT_NUMERIC (numeric)
NCHAR(2)	DT_WSTR (Unicode string)
DATETIME	DT_DBTIMESTAMP (database timestamp)
DATE	DT_DBDATE (database date)
UNIQUEIDENTIFIER	DT_GUID (unique identifier)

3. **“SSIS package is being used by another process” Error**

Full error -: Recently, while developing an SSIS package we got the error “The process cannot access the file ‘*.ispac’ because it is being used by another process”.

Reason -: This will occurred when you running your ssis package you have stopped it in middle of the compilation

Solution -: Go to Task Manager –> Details Tab.

-: Locate the process “DtsDebugHost.exe“. Kill this process. There might be multiple instances of this process. Kill all of them.