

Data Warehousing & Business Intelligence (IT)

3rd Year, 1st Semester

# Assignment 1

Submitted to

Sri Lanka Institute of Information Technology

IT20035358

K.N.N.D.S Jayarathna.

Weekday Batch

**STEP 01: Data set selection**

I have selected the predicting Coupon Redemption data set as the data set. It consists of seven csv files. Among them I chose five csv files to implement my solution.

According to the assignment guidelines the data set was initiated with sufficient data. In my data set which I have selected is a transactional data type and it is used to predict whether customer will Redeem coupons such as discounts coupons or other types of coupons released in various marketing campaigns.

Customer, Campaign, Items and coupon redemption status and customer transaction are all detailed in the data set.

In predicting coupon redemption data set, there are also enough records and attributes to build complete data warehouse and perform ETL process. I can receive two sources of data from the data set that I have selected (A database and a text file) as my source data. And the data set is sufficient to build data warehouse and I was able to do ETL process with this data set.

As well as I can correctly identify hierarchies, dimensions, and aggregates in the data set. I also identified that with data set I will be able to generate appropriate reports.

* Data set link: -

<https://www.kaggle.com/datasets/vasudeva009/predicting-coupon-redemption>

* **Data Description**
* Train.CSV

|  |  |
| --- | --- |
| Variable | Definition |
| id | Unique id for coupon customer impression |
| campaign\_id | Unique id for a discount campaign |
| coupon\_id | Unique id for a discount coupon |
| customer\_id | Unique id for a customer |
| Redemption\_status | (target) (0 - Coupon not redeemed, 1 - Coupon redeemed) |

* Campaign\_data.csv

|  |  |
| --- | --- |
| Variable | Definition |
| campaign\_id | Unique id for a discount campaign |
| campaign\_type | Anonymised Campaign Type (X/Y |
| start\_date | Campaign Start Date |
| end\_date | Campaign End Date |

* customer\_demographics.csv

|  |  |
| --- | --- |
| Variable | Definition |
| customer\_id | Unique id for a customer |
| age\_range | Age range of customer family in years |
| marital\_status | Married/Single |
| rented | 0 - not rented accommodation, 1 - rented accommodation |
| family\_size | Number of family members |
| noOfChildren | Number of children in the family |
| income\_bracket | Label Encoded Income Bracket (Higher income corresponds to higher number |

* customertransaction.csv

|  |  |
| --- | --- |
| Variable | Definition |
| date | Date of Transaction |
| customer\_id | Unique id for a customer |
| item\_id | Unique id for item |
| quantity | quantity of item bought |
| selling\_price | Sales value of the transaction |
| other\_discount | Discount from other sources such as manufacturer coupon/loyalty card |
| coupon\_discount | Discount availed from retailer coupon |

* Item.csv

|  |  |
| --- | --- |
| Variable | Definition |
| item\_id | Unique id for item |
| brand | Unique id for item brand |
| brand\_type | Brand Type (local/Established) |
| Category | Item Category |

* **ER Diagram**

**Diagram

Description automatically generated**

**STEP 02: Preparation of Data Sources**

There were seven csv files available in the data set. They are,

* + Item data csv
  + Customer Demographics csv
  + Customer Transaction Data csv
  + Campaign Data csv
  + Train Data csv
  + Coupon Item Mapping csv
  + Test csv
* Among these csv files Item Mapping and Test csv files were not taken to create tables.
* And I decided to convert Campaign data csv into text file format.
* Then I imported all other four csv files into my CouponRedemption\_SourceDB.
* After I imported my csv files into sourceDB, I created Data warehouse named CouponRedemption\_DW and created my dimension tables and fact tables inside the data warehouse.
* Here I converted train csv file data into CouponRedemption table.

Graphical user interface, table, Excel

Description automatically generatedTable

Description automatically generated **campaign \_data.csv customer\_demographics.csv**

Table

Description automatically generated with low confidence

**Customer\_transaction\_data\_new.csv**

A picture containing table

Description automatically generatedTable

Description automatically generated

**item\_data.csv train.csv**

A picture containing text

Description automatically generatedTable

Description automatically generated**campaign\_data.csv campaign\_data.txt**

Graphical user interface, text, application

Description automatically generatedI have loaded customer\_demographics.csv, customer\_transaction\_dat\_new.csv, item\_data.csv and train.csv files to the DB called CouponRedemption\_SourceDB.

**CouponRedemption\_SourceDB**

**The campaign\_data.csv details were taken as s text file, and I used It as another sourceType.**

**Table

Description automatically generated**

**A picture containing text

Description automatically generated**

**Campaign\_data.csv Campaign\_data.txt**

**STEP 03: Solution Architecture**

**Diagram

Description automatically generated**

* + The data warehouse is the core of the BI system. A data warehouse is a purpose of data analysis and reporting. This purpose changes the design of this database as well.
  + This architecture shows the high-level BI solution to the warehouse.

**STEP 04: Data warehouse Design and Development**

****

**Assumptions**

* I used a schema type of **Snowflake schema** for the Data warehouse development. At there,
* There is one Fact Table, it is **FactCustomerTransaction** table.

There are five dimensions table.

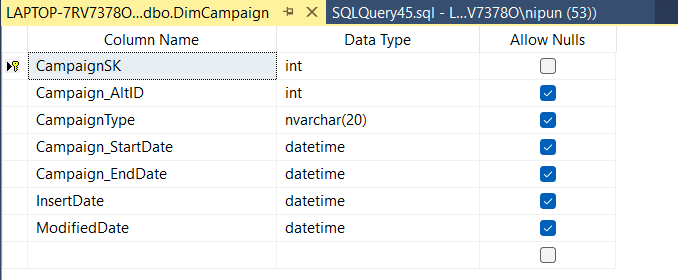
* 1. DimCustomers table
  2. DimItem table
  3. DimDate table
  4. DimCouponRedemption table
  5. DimCampaign table
* I have taken **DimCustomers** details as slowly changing dimension, **Income bracket** can change time to time, and we need to keep track of their historical data.
* I have implemented data warehouse tables in the SQL server and the tables that I have created shown below

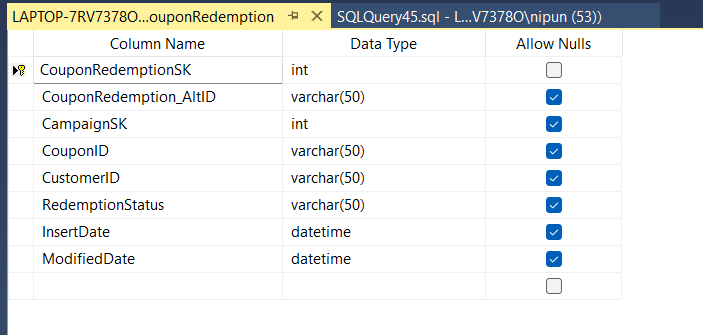
Diagram

Description automatically generated

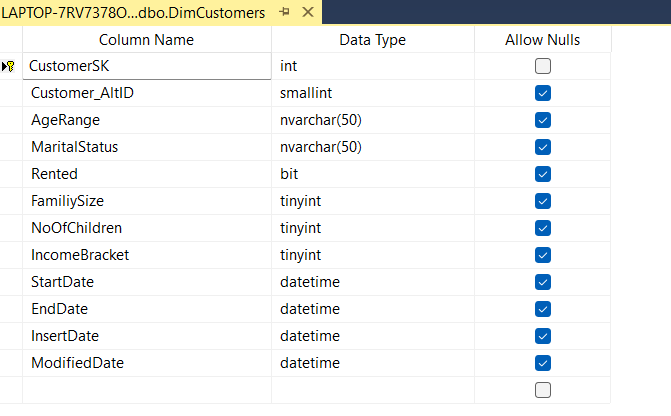
**Data Warehouse Data types**

* Before creating the FactTransaction fact table and other dimensios, start by creating Date dimension. For that I used the code in the “DateMaster.sql” file.





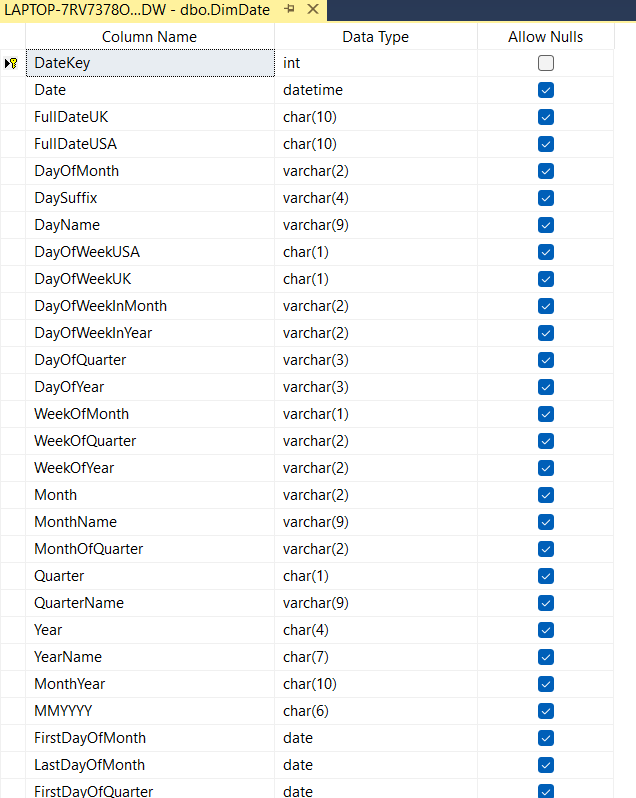
**Campaign Dimension Table**  **Coupon Redemption Dimension Table**



**Customer Dimension Table** **Item Dimension Table**

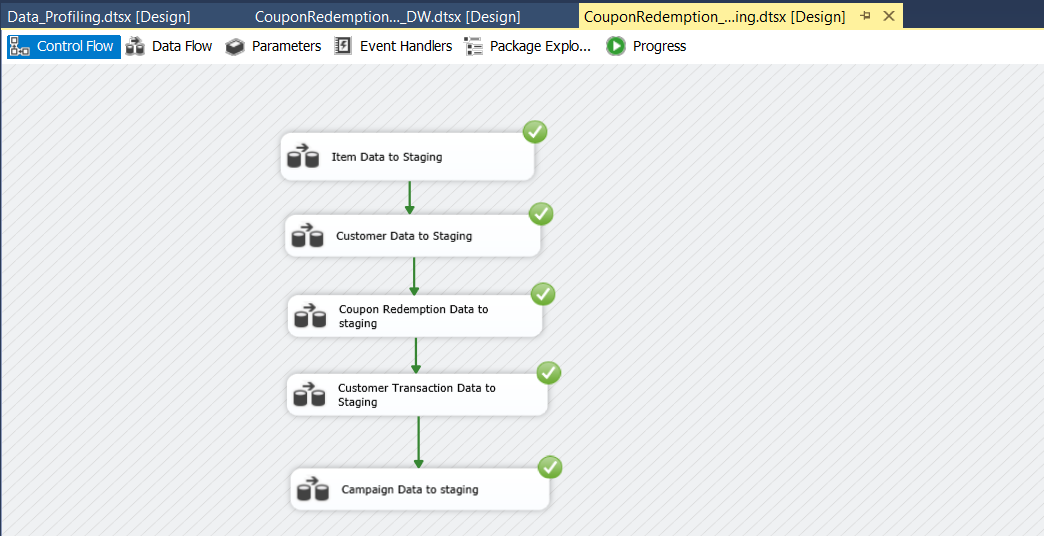
**Table

Description automatically generatedFact Customer Transaction Table**



**Date Dimension Table**

**STEP 05: ETL Development**

****

* First using the SQL Server integration service Software, I have extracted all the data from the Tables which were in the CouponRedmption\_SourceDB and Extract Campaign\_Data.txt to Separate DB called CouponRedemption\_Staging. And also, **here I converted train.csv data into** **CouponRedemptionStaging table**.

Text

Description automatically generated

**CouponRedemption\_Staging DB**

1. Graphical user interface, text

   Description automatically generated with medium confidence **Extract Coupon Redemption Data to Staging**

* Used OLE DB Source as dbo.train data table in CouponRedemption\_SourceDB.
* OLE DB Destination for create new table CouponRedemptionStaging table in the CouponRedemption \_Staging database.

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

Description automatically generated

* Used to execute SQL Task SSIS tool truncate table for SQL command as truncate table CouponRedemptionStaging in CouponRedemption\_Staging database.

1. Graphical user interface, text, chat or text message

   Description automatically generated**Extract Item Data to Staging**

* Used OLE DB Source as dbo.item\_data table in CouponRedmption\_SourceDB.
* OLE DB Destination for create new table ItemStaging table in the CouponRedmption \_Staging database.

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

Description automatically generated

* Used to execute SQL Task SSIS tool truncate table for SQL command as truncate table ItemStaging in CouponRedemption\_Staging database.

1. **Extract Customer Data to Staging**

Graphical user interface, text, application

Description automatically generated

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

Description automatically generated

* Used OLE DB Source as dbo.Customer\_demographocs table in CouponRedmption\_SourceDB. OLE DB Destination for create new table CutomerStaging in the CouponRedmption \_Staging database. Used to execute SQL Task SSIS tool truncate table for SQL command as truncate table CutomerStaging in CouponRedemptio\_Staging database.

1. **Extract Campaign Data to Staging**

Graphical user interface, text

Description automatically generated

* Used OLE DB Source as **Campaign\_data.txt Flat** **File Source.**
* Graphical user interface, text, chat or text message

  Description automatically generatedOLE DB Destination for create new table CampaignStaging in the CouponRedmption\_Staging database.
* Used to execute SQL Task SSIS tool truncate table for SQL command as truncate table CampaignStaging in CouponRedemption\_Staging database.

Graphical user interface, text

Description automatically generated**5. Extract Customer transaction Data to Staging**

* Used OLE DB Source as dbo.customer\_transaction\_data\_new table in CouponRedmption\_SourceDB.
* Graphical user interface, text, application, chat or text message

  Description automatically generatedOLE DB Destination for create new table CustomerTransactionStaging in the CouponRedmption \_Staging database.

Used to execute SQL Task SSIS tool truncate table for SQL command as truncate table CustomerTransactionStaging in CouponRedemption\_Staging database.

**Data Profiling**

* I used the staging table data to analyze how the data looks like to determine what type of transformations I need to perform on the data.

Graphical user interface, text, application

Description automatically generated

A picture containing diagram

Description automatically generated

* Data Profiling was done to all staging tables to get a better insight on the data like checking for unique values or if null values were present. CustomerStaging table data profiling shown below.

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedTable

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

**Data Transformation**

1. **Transform and Load Item Data Details**

I created Item data transformation by below mentioned steps.

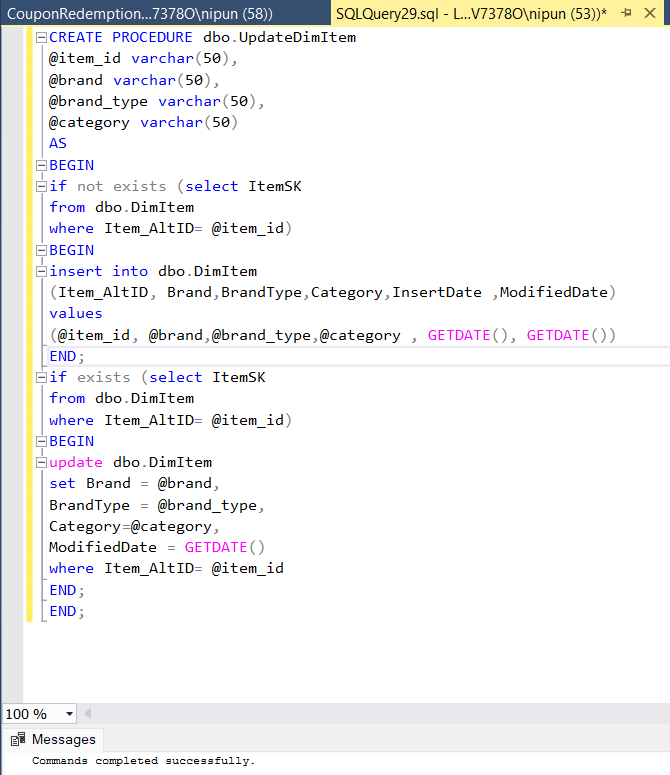
* Created new package called CouponRedemption\_DW.dtsx.
* Then dragged and dropped a data flow task, renamed it as transform and load

Item Data.

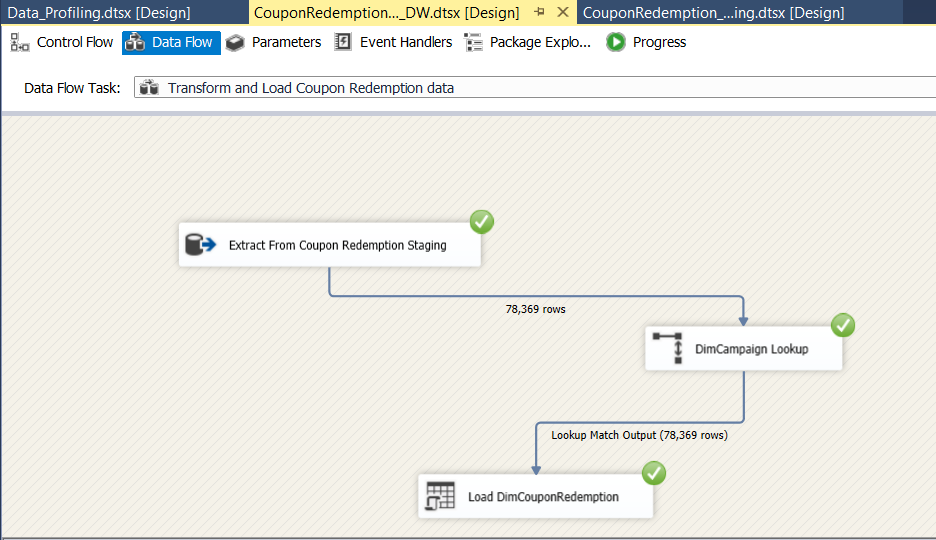
* Then dragged and dropped OLE DB source renamed as Extract from Item Staging and configure it to access the Item staging table.
* After that I dragged and OLE DB Command and connect to the OLE DB source.

Graphical user interface, application

Description automatically generated

* First, I have created a procedure called UpdateDimItem and executed in the CouponRedemption\_DW database.
* **UpdateDimItem procedure**, it is used to insert data from ItemStaging to DimItem without data duplication.
* Then I did same process to DimCouponRedemption table and DimCampaign table as well.

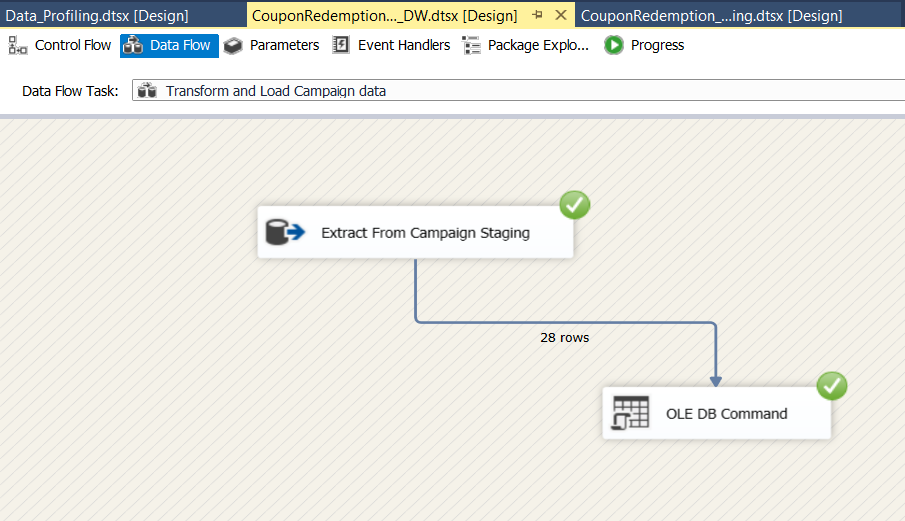
1. **Transform and Load Coupon Redemption Data Details**

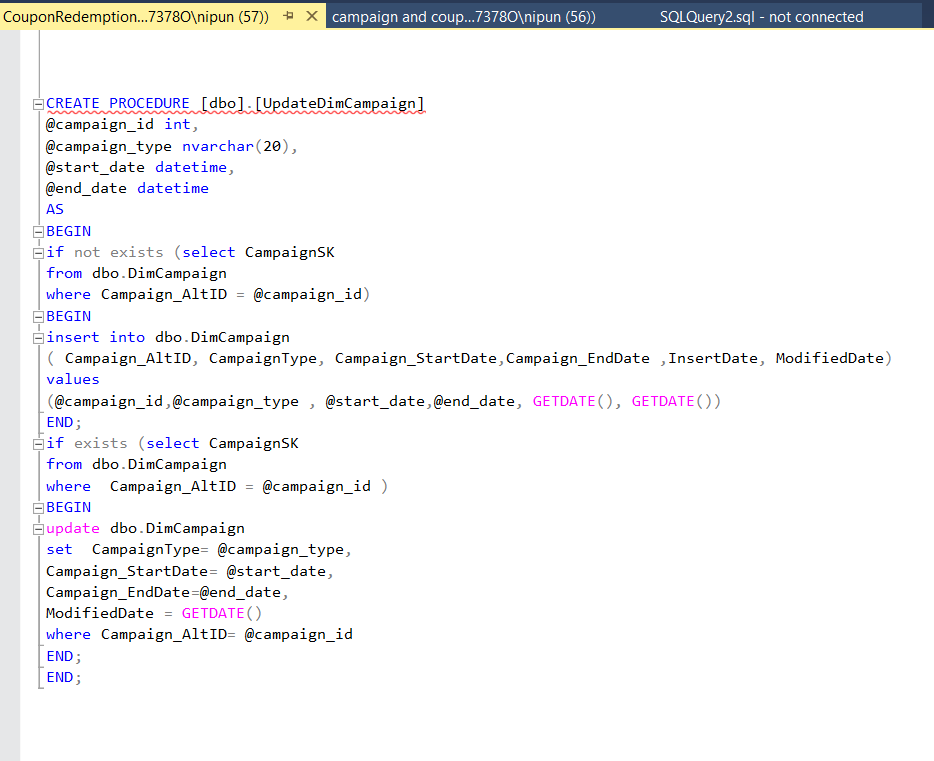


**UpdateDimCouponRedemption Procedure**

Graphical user interface, text, application

Description automatically generated

1. **Transform and Load Campaign Data Details**

 **UpdateDimCampaign Procedure**

**Transform and Load Customer Data Details (Slowly Changing Dimension)**

Graphical user interface

Description automatically generated with medium confidence

* The slowly Changing Dimension was handled when creating the DimCustomers table. For the slowly changing dimension I have used a slowly changing dimension, derived columns, merge, union, and sort in order to do the ETL process of loading data to the warehouse from the staging.

**Transform and Load Customer Transaction Data Details**

Graphical user interface, application

Description automatically generated

Table

Description automatically generated

**CouponRedemption\_DW**

**STEP 06: ETL Development -Accumulating Fact Table**

First, I extended my fact table (FactCustomerTransaction Table) with following 03 coloumns.

* + accm\_txn\_create\_time
  + accm\_txn\_complete\_time
  + txn\_process\_time\_hours

Graphical user interface

Description automatically generated with low confidence

* Then I prepared a dataset which contains fact table natural key(tID) and accm\_txn\_complete time and saved it csv file format.

Table

Description automatically generateds

* Then I created a separate ETL SSIS package named as **Complete\_time.dtsx.** Which reads data from this file and update the corresponding accm\_txn\_complete time in my DW Fact table.

Graphical user interface, text, application

Description automatically generated

Diagram

Description automatically generated

s

Graphical user interface, application

Description automatically generatedDiagram

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, table

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Graphical user interface

Description automatically generated **Final Control CouponRedemption\_DW Control Flow.**

**Finally, successfully loaded data to the FactCustomerTransaction table with updated txn\_process\_time.**

Graphical user interface, application, table

Description automatically generated