## Sri Lanka Institute of Information Technology



# Data Warehouse and Business Intelligence - Assignment 1

Name: Deemantha P.H.H.C

Reg-No: IT22560162

Batch: Y3.S2.WE.DS.02.01

## **Contents**

1. Data Set Selection	3
2. Preparation of Data Source	5
3. Solution Architecture	11
4. Data Warehouse Design and Development	12
i. Design	12
ii. Assumptions	13
iii. Slowly changing dimensions	14
5. ETL Development	15
i. Data Extraction & Load into Staging tables	15
ii. Data Profiling 19 <b>Error! Bookmark</b>	x not defined.
iii. Data Transformation and Loading Error! Bookmark	not defined.
6. ETL Development – Accumulating fact tables	27

#### 1. DATA SET SELECTION

Data Set Name: Synthetic Cannabis Dispensary Database

Provided by: kaggle.com

Source link: <a href="https://www.kaggle.com/datasets/adampq/synthetic-cannabis-dispensary-database?select=stateReg.csv">https://www.kaggle.com/datasets/adampq/synthetic-cannabis-dispensary-database?select=stateReg.csv</a>

#### **About Dataset:**

The selected data source is a collection of transactional data.

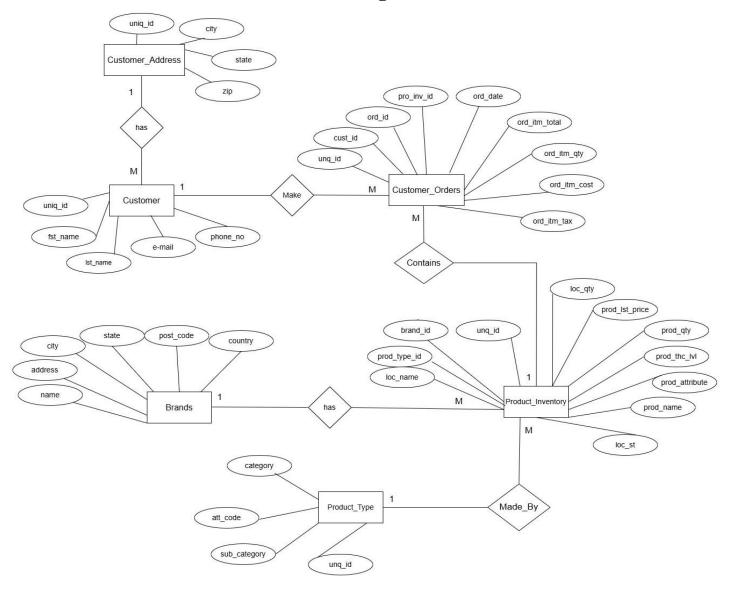
This dataset simulates a comprehensive cannabis dispensary registry and operations database, representing various aspects of the legal cannabis industry in the United States. The synthetic data reflects realistic structures and relationships between dispensaries, product offerings, business licenses, and regulatory frameworks across different states. Customers can access detailed information about dispensaries, including their product inventory, business operations, and state-level regulations.

As the cannabis industry continues to grow and legalize across more regions, businesses and researchers alike are investigating ways to improve transparency, regulatory compliance, and customer satisfaction. This dataset offers valuable insights for analyzing dispensary operations, tracking compliance with local laws, and understanding consumer behavior.

Dataset contains eight csv files but I selected five csv files with information about Customers, products, brands, product inventory and customer Orders. Modifications were done accordingly to the data set derived from the source. This data set contains transactions of the customer orders which are provided by the customers and the other features around it.

- Customers.csv: Contains the details of the customers.
- Brand.csv: Include information about brands of the product.
- ProductType.csv: Details of product category and sub-category.
- ProductInventory.csv: Contains details of the products, brands available, location, prices and the quantities available.
- CustomerOrders.csv: Purchases made by customers and relevant product details.

### **ER-Diagram**



- This diagram shows the relationships between entities in this dataset.
- O Assumptions -
  - One Customer can has only one address.
  - Customers can place many orders.
  - There can be many customers in the same address

## 2. PREPARATION OF DATA SOURCES

Final State of Preparation of the source data formats before Transforming data =>

- Text file that has been taken as a separate source type: -
  - CustomerAddress.txt
- ♣ Ass\_SourceDB (Source Database) Tables: -
  - dbo.Customers
  - dbo.Brands
  - dbo.ProductType
  - dbo.ProductInventory
  - dbo.CustomerOrders

## Description of the data set

Source Type - CustomerAddress.txt
 Table Name - CustomerAddress
 Include

Column	Data Type	Description
uniq_id	Nvarchar(50)	customerID
city	Nvarchar(50)	City that customer belongs to
state	Nvarchar(50)	State that customer belongs to
zipcode		zipcode that customer belongs to

Source Type - DWBI\_SourceDB
 Table Name - Customers
 Include

Column	Data Type	Description
uniq_id	Nvarchar(50)	Customer Uniq ID
fst_name	Nvarchar(50)	First name of the customer
lst_name	Nvarchar(50)	Last name of the customer
email	Nvarchar(50)	E-mail of the customer
phone_no	Nvarchar(50)	Phone number of the customer
gender	Nvarchar(50)	Gender of the customer

3. Source Type - DWBI\_SourceDB

Table Name - Brand

Include -

Column	Data Type	Description
unq_id	Int	Unique ID for Brands
name	Nvarchar(50)	Name of the store
address	Int	Address number where the store is located
city	Nvarchar(50)	City where the store is located
state	Nvarchar(50)	Store located state
postcode	Int	Postcode of the store location
country	Nvarchar(50)	Country of the store

Source Type - DWBI\_SourceDB
 Table Name -Product Type
 Include

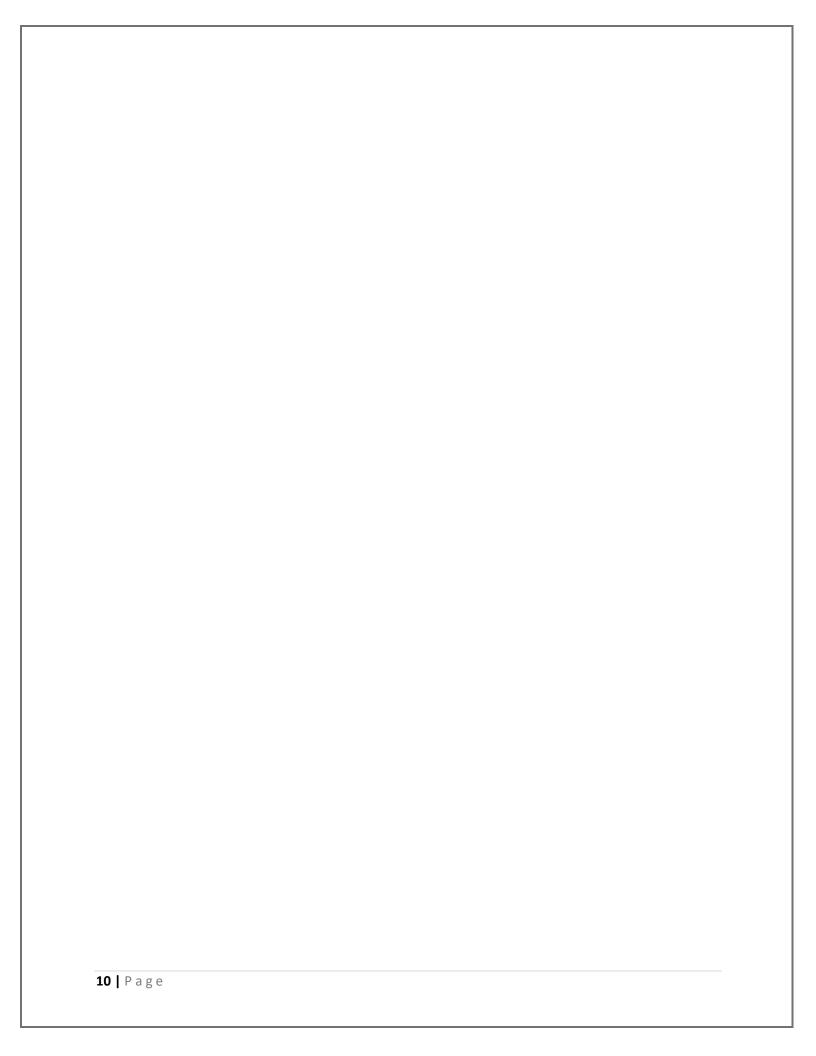
Column Name	Data Type	Description
unq_id	Int	Unique ID for Product Type
category	Nvarchar(50)	Category of the product
sub_category	Nvarchar(50)	Sub category of the product
att_code	Int	Attribute code of the product

## 5. Source Type - DWBI\_SourceDB Table Name - Product Inventory Include -

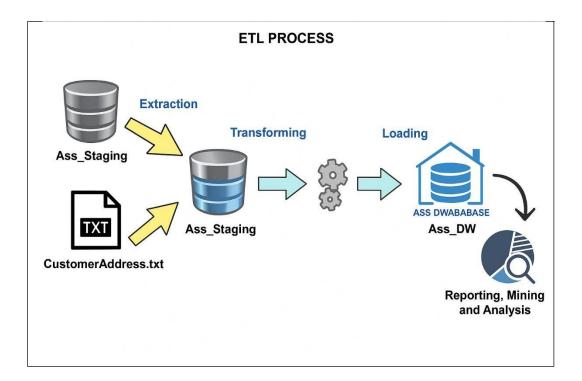
Column Name	Data Type	Description
unq_id	Int	Unique ID for Product Inventory
brand_id	Int	ID for Brand (FK)
prod_type_id	Int	ID for product type (FK)
loc_name	Nvarchar(100)	Location name of the product distributed
loc_st	Nvarchar(50)	Location state
prod_name	Nvarchar(50)	Distributed product name
prod_attribute	Nvarchar(50)	Distributed product attribute
prod_thc_lvl	Nvarchar(50)	Product thc level
prod_qty	Nvarchar(50)	Product quantity
prod_lst_price	money	Last price of the product
loc_qty	Int	Location quantity

## 6. Source Type - DWBI\_SourceDB Table Name - CustomerOrders Include -

Column Name	Data Type	Description
unq_id	Int	Unique ID for Customer Orders
cust_id	Int	Uniq ID from customer table as a Foreign Key
ord_id	bigint	Unique ID for Orders
prod_inv_id	Int	Unique ID from Product Inventory table as a Foreign Key
ord_datetime	Datetime	Date and time of the Order purchase
loc_name	Nvarchar(100)	Location name of the store
loc_st	Nvarchar(50)	Location state of the store
cust_name	Nvarchar(50)	Customer name who made the order
cust_st	Nvarchar(50)	Location state of the customer
prod_name	Nvarchar(50)	Name of the product
ord_itm_qty	Int	Quantity of the ordered item
ord_itm_cost	float	Cost of the ordered item
ord_itm_tax	float	Tax for the ordered item
ord_itm_total	float	Total amount payed by the customer for his/her order



#### 3. SOLUTION ARCHITECTURE

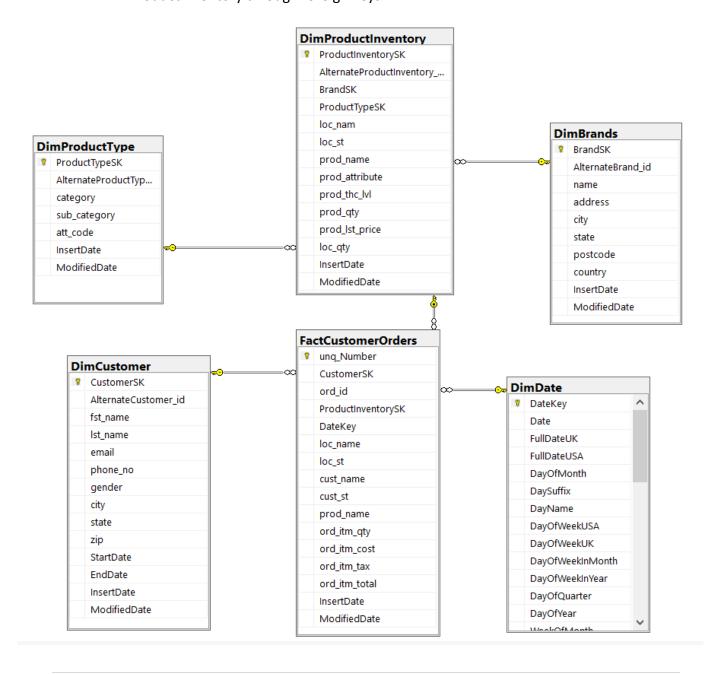


As this architecture shows for the ETL processing, first DWBI\_SourceDB (Source database) and then CustomerAddress.txt (Text file) has been used for the data extraction to the staging process. After that staged in DWBI\_Staging (Staging database) data are transforming and loading to DWBI\_DW (Data warehouse) and that data can be used for reporting, mining and analyzing purposes.

#### 4. DATA WAREHOUSE DESIGN S DEVELOPMENT

#### i. Design

The DWBI\_DW (Data Warehouse) is designed according to a snowflake schema as shown in the following figure with one fact table (dbo.FactCustomerOrders) and five dimension tables including the Date dimension. DimProductType and DimBrands are connected with DimProductInventory through foreign keys.



#### Hierarchies

- DimCustomer is consisted with the hierarchy of address which includes City, State, ZipCode.
- DimBrands is consisted with the hierarchy of address which includes Address, City, State, PostCode, Country
- DimDate is consisted with the hierarchy of dates which includes DayofMonth, Month, Quarter, Year.

#### Calculation

Order Item Total is calculated in dbo.FactCustomerOrder as ord\_itm\_total
 (Ord itm qty \* ord itm cost) + ord itm tax = ord itm total

#### ii. Assumptions

- dbo.DimDate is added to the Data Warehouse for better performance.
- dbo.FactCustomerOrders is used in creating the fact table.

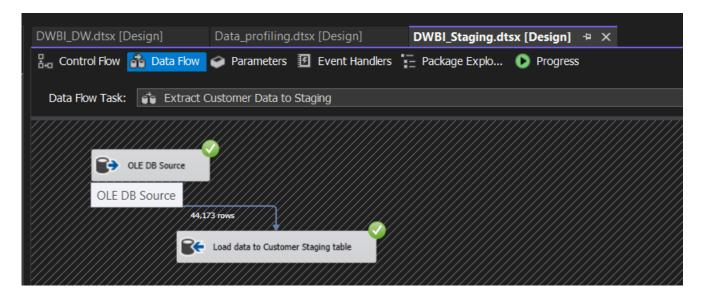
## Slowly changing dimensions

• Customer Details with customer addresses were considered as a slowly changing dimension.

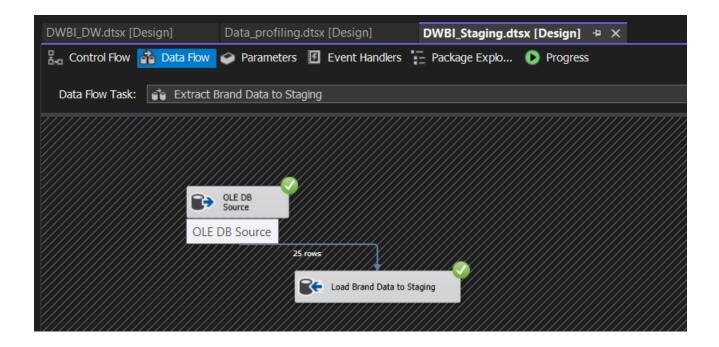
Dimension Table	Attributes
dbo.DimCustomer	Uniq_id (Business Key)
	Fst_name (Fixed attribute)
	Lst_name (Fixed attribute)
	Email (Changing attribute)
	Phone_no (Changing attribute)
	Gender (Fixed attribute)
	City (Historical attribute)
	State (Historical attribute)
	Zip (Historical attribute)

#### **5. ETL DEVELOPMENT**

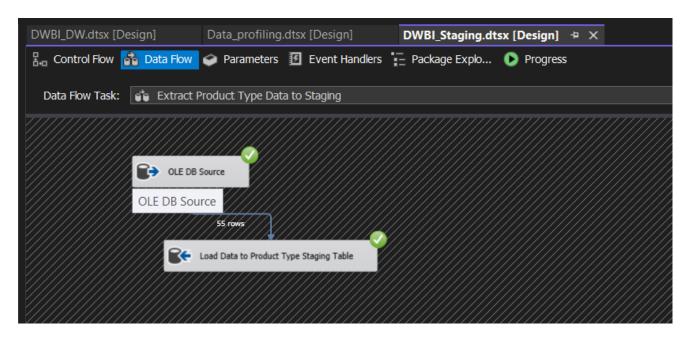
- i. Data Extraction & Load into Staging tables
  - Data Extraction is done by using the provided data sources mentioned above in Visual Studio 2022 (Data Tool) development environment. The text file and the source database were used here.
  - Initially, **OLE DB SOURCE** (for source database) or **FLAT FILE SOURCE** (for flat files txt) is used to extract data for the Staging criteria. In this step developer can select the columns what would be included in the Staging from available data columns. As the next step of Staging, **OLE DB DESTINATION** has applied here to storing data in the Staging tables of **DWBI\_Staging.**



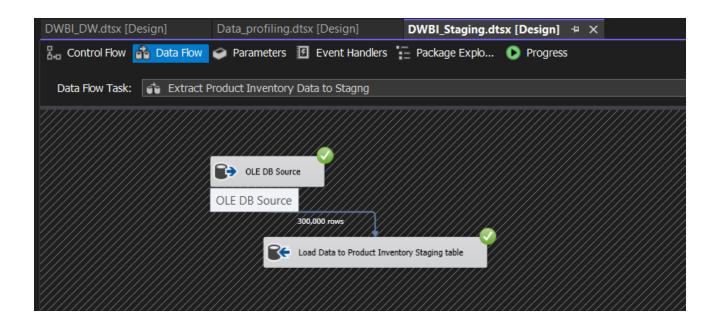
Customer details data is extracted from Customers table in the source database and inserted to the StgCustomer table.



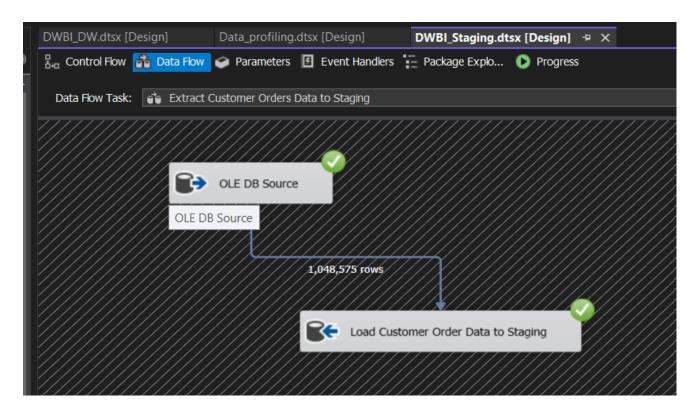
Brand details data is extracted from Brand table in the source database and inserted to the StgBrand table.



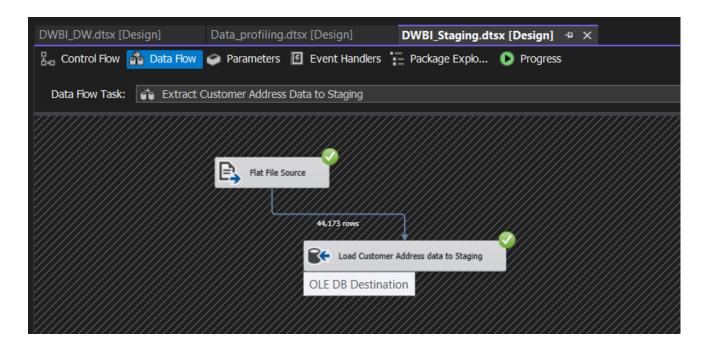
Product Type data is extracted from ProductType table in the source database and inserted to the StgProductType table.



Product Inventory data is extracted from Product Inventory table in the source database and inserted to the StgProductInventory table.



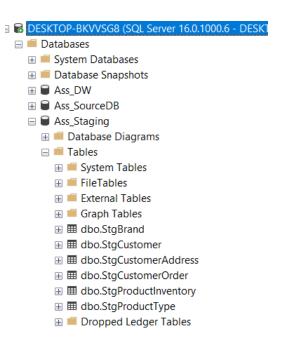
Customer Orders data is extracted from CustomerOrders table in the source database and inserted to the StgCustomerOrder table.



Customer address data is extracted from CustomerAddress.txt (text file) in and inserted to the StgCustomerAddress table.



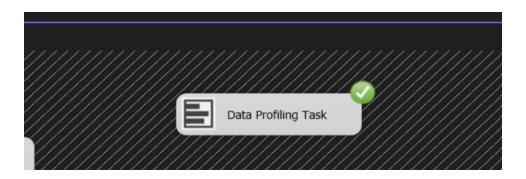
The Control Flow of 'Extract Data and Load into Staging' is shown as above figure.

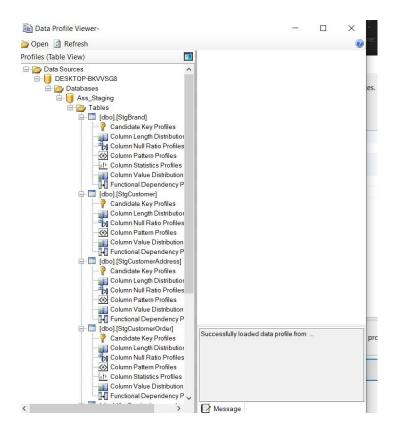


• Staging Tables have created and values are inserted.

#### ii. Data Profiling

Data Profiling provides the means of analyzing large amount of data using different kind of processes. In this step, null values, repeated values and quality of the data is checked.

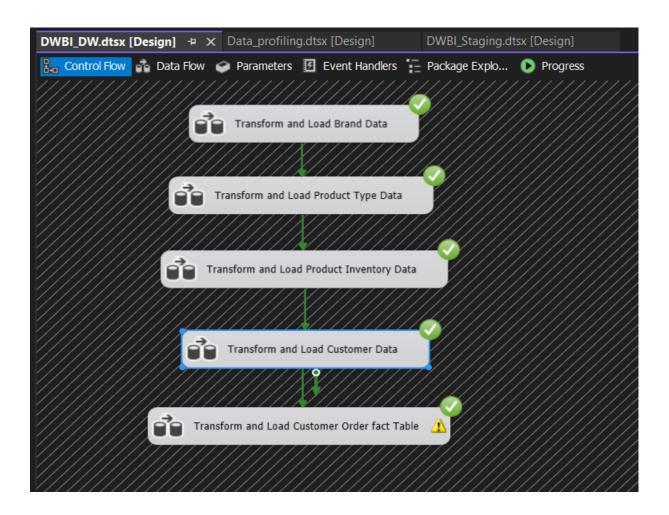




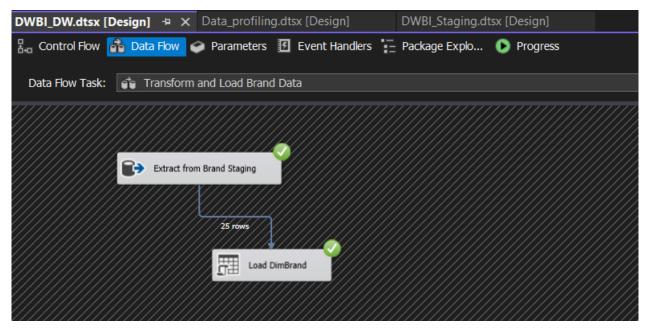
- Every staging table is profiled and saved in a selected location.
- As this shows, after the Staging step doing this task shows the things what the
  developer must consider about the data which are stored in staging table and the
  developer is able to identify the issues with staging data by data profiling (such as
  null values).
- Complete part of Data Profiling relevant to the Staging is shown in this figure.

### iii. Data Transformation and Loading

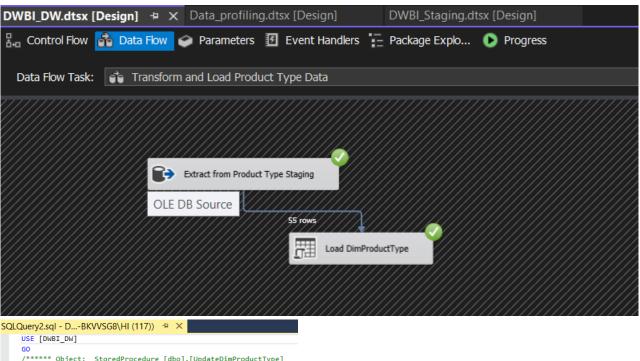
• Data Transformation is developed according to the dimensional modeling designed above.



• In this step, the Dimension Tables are created in DWBI\_DW are loaded with the data of relevant staging tables.

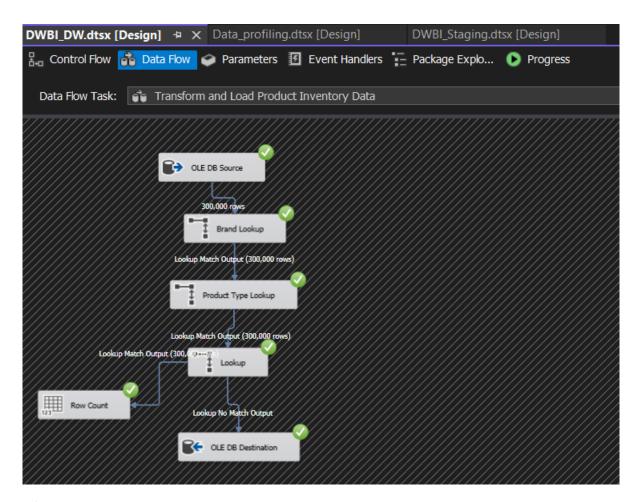


- O Brand Details data are loaded to DimBrands
- O UpdateDimBrands procedure is used to check whether the data is inserted or not.



```
/****** Object: StoredProcedure [dbo].[UpdateDimProductType]
 SET ANSI NULLS ON
 SET QUOTED_IDENTIFIER ON
ALTER PROCEDURE [dbo].[UpdateDimProductType]
 @unq_id int,
 @category nvarchar(50),
 @sub_category nvarchar(50),
 @att_code int
BEGIN
if not exists (select ProductTypeSK
 from dbo.DimProductType
 where AlternateProductType_id = @unq_id)
BEGIN
insert into dbo.DimProductType
 ({\tt AlternateProductType\_id},\ {\tt category},\ {\tt sub\_category},\ {\tt att\_code},\\
 InsertDate, ModifiedDate)
 values
 (@unq\_id, \ @category, \ @sub\_category, \ @att\_code,\\
   ETDATE(), GETDATE())
if exists (select ProductTypeSK
 from dbo.DimProductType
 where AlternateProductType_id = @unq_id)
BEGIN
update dbo.DimProductType
 set category = @category,
sub_category = @sub_category,
att_code = @att_code,
 ModifiedDate = GETDATE()
 where AlternateProductType_id = @unq_id
 END;
 END;
```

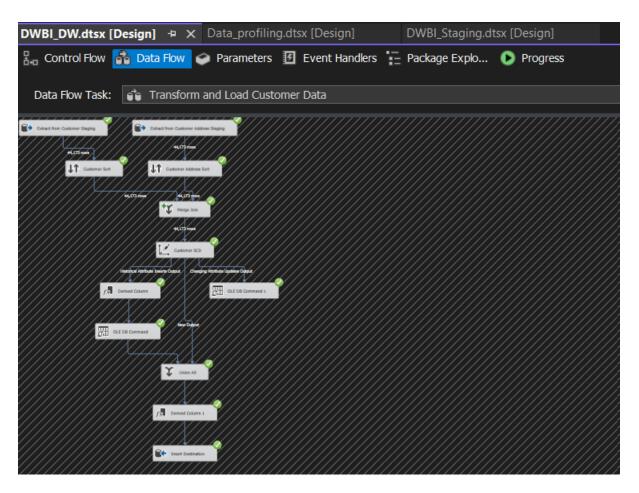
- O Product Type data are loaded to DimProductType
- UpdateDimProductType procedure is used to check whether the data is inserted or not.



O Product Inventory are loaded to DimProductInventory

#### **Loading Slowly Changing Dimension**

- DimCustomer is the slowly changing dimension in this dimensional modeling.
- In order to load data to Dimension table, the slowly changing dimensions (historical) have two specific columns as StartDate & EndDate to ensure that the data is valid at the moment.
- Slowly changing dimension wizard let the developer to select the dimension table, business keys of the dimension and what would be the slowly changing attributes.

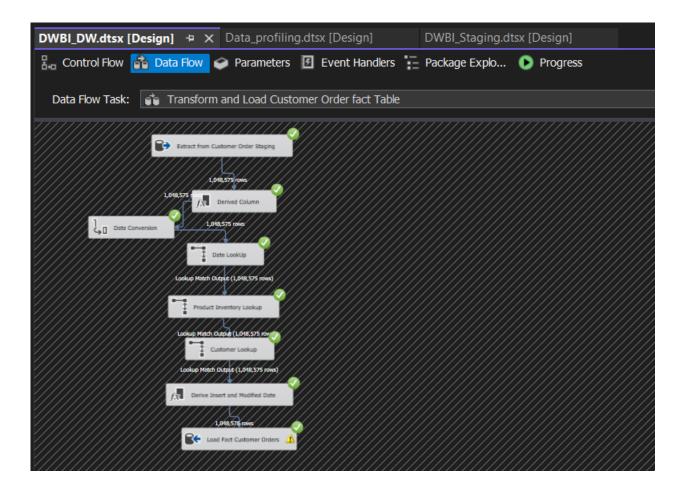


- As mentioned earlier under assumptions, customer details were considered as slowly changing details.
- **O** The below mentioned columns were set as changing attributes:
  - 1. Phone\_no
  - 2. E-mail
- The below mentioned columns were set as historical attributes:
  - 1. City

- 2. State
- 3. Zip Code
- After extracting data from the StgCustomer table, it was sorted according to the uniq id and as it was identified as a slowly changing dimension, it was connected as shown above and loaded data to the Customer dimension table.

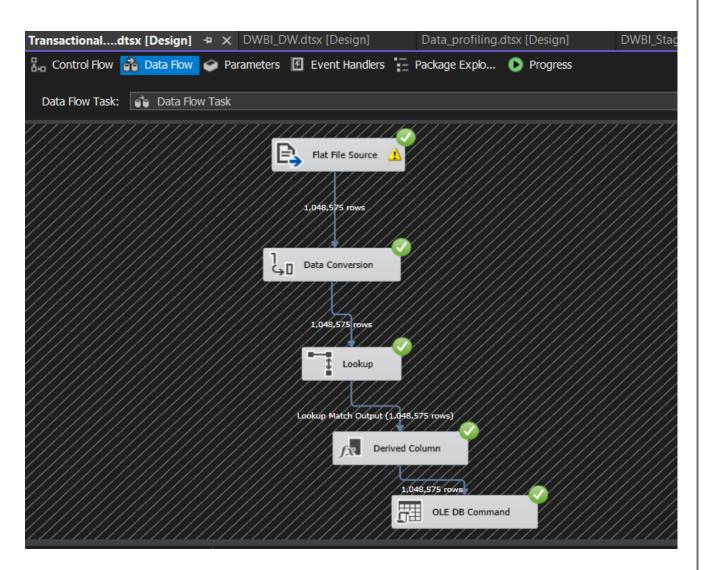
#### **FactCustomerOrder**

After loading data in to dimension tables, fact table was loaded with customer order data.

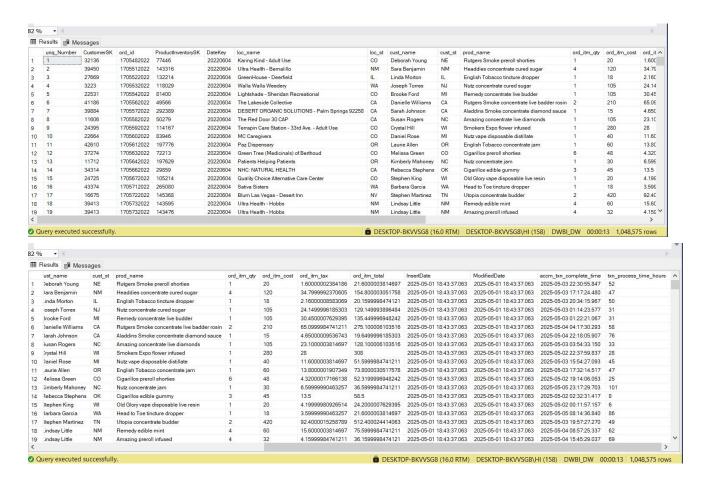


## 6. ETL Development – Accumulating fact table

- The final step of Transformation & Loading is load data to the accumulative fact table.
   According to the dimensional model, StgCustomerOrder table is used to insert values into FactCustomerOrders table.
- InsertDate was set to be equal to the current system date when loading data into the fact table.
- A separate dataset was generated including uniq\_number(FactCustomerOrder key) and completed\_time.
- A separate SSIS package was created, which reads data from the csv file and update the complete\_time in FactCustomerOrder.



• Can view the accumulative fact table as the final step in Sequel Server Management Studio.



<sup>\*</sup>In here there are two figures for the accumulative fact table because the table was too long to capture from a single screenshot.

• Fact details were added to the FactCustomerOrders table.