

Data Warehousing

Contents

2.PREPARATION OF DATA SOURCES.....	5
3.SOLUTION ARCHITECTURE	10
4.DATA WAREHOUSE DESIGN & DEVELOPMENT.....	11
i. Design	11
ii. Assumptions.....	12
Slowly changing dimensions	13
5.ETL DEVELOPMENT.....	14
i. Data Extraction & Load into Staging tables	14
iii. Data Transformation and Loading.....	20
6.ETL Development – Accumulating fact table.....	26

1. DATA SET SELECTION

Data Set Name: Synthetic Cannabis Dispensary Database

Provided by: kaggle.com

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

About Dataset:

The selected data source is a collection of transactional data.

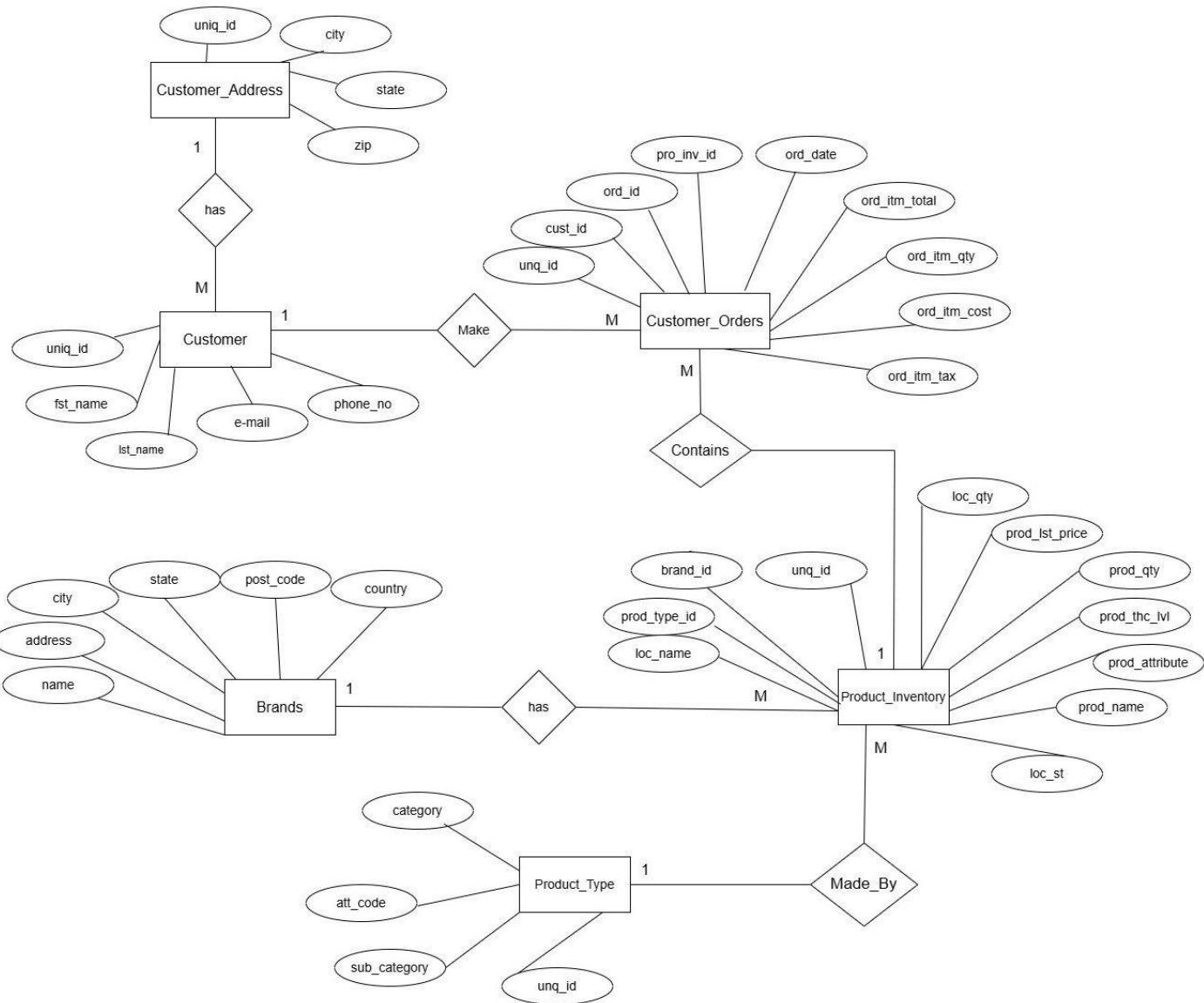
This dataset simulates a comprehensive data set of cannabis dispensary registry and operations database, representing various aspects of the legal cannabis in very popular industry in the United States. The synthetic data reflects realistic structures and relationships between dispensaries, product offerings, 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 and facets 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 all 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.
- Assumptions –
 - One Customer can have 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
- Accumulative table that has been taken as a separate dataset: -
 - Acuumalativedataset.csv

DESCRIPTION OF THE DATA SET

1. 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	Nvarchar(50)	zipcode that customer belongs to

2. 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

4. 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

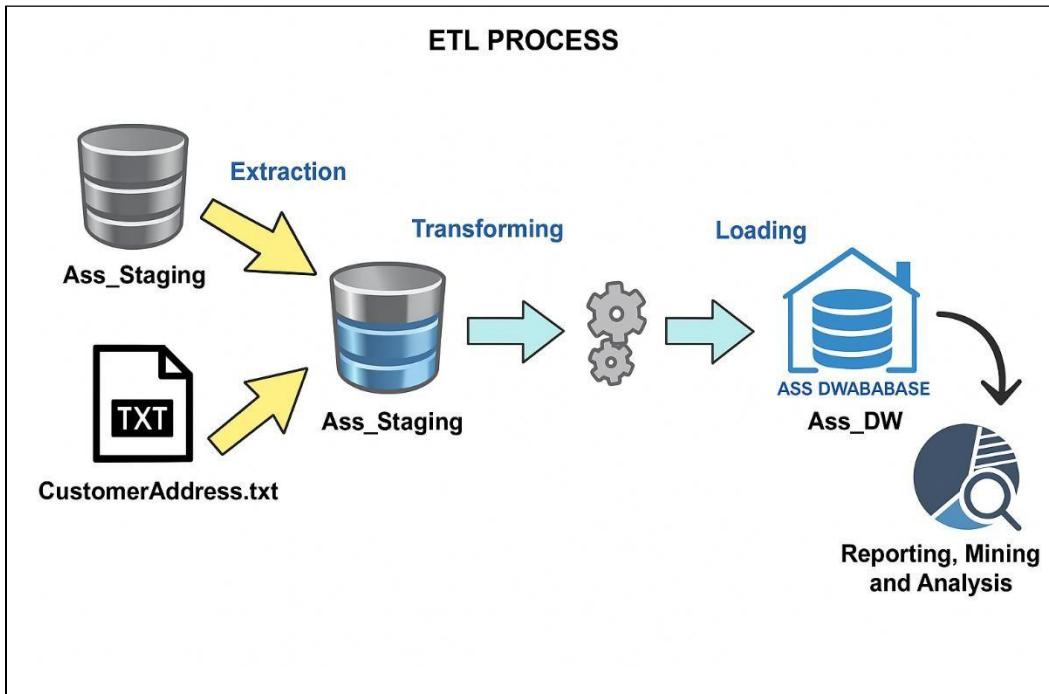
6. Source Type - DWBI_SourceDB

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



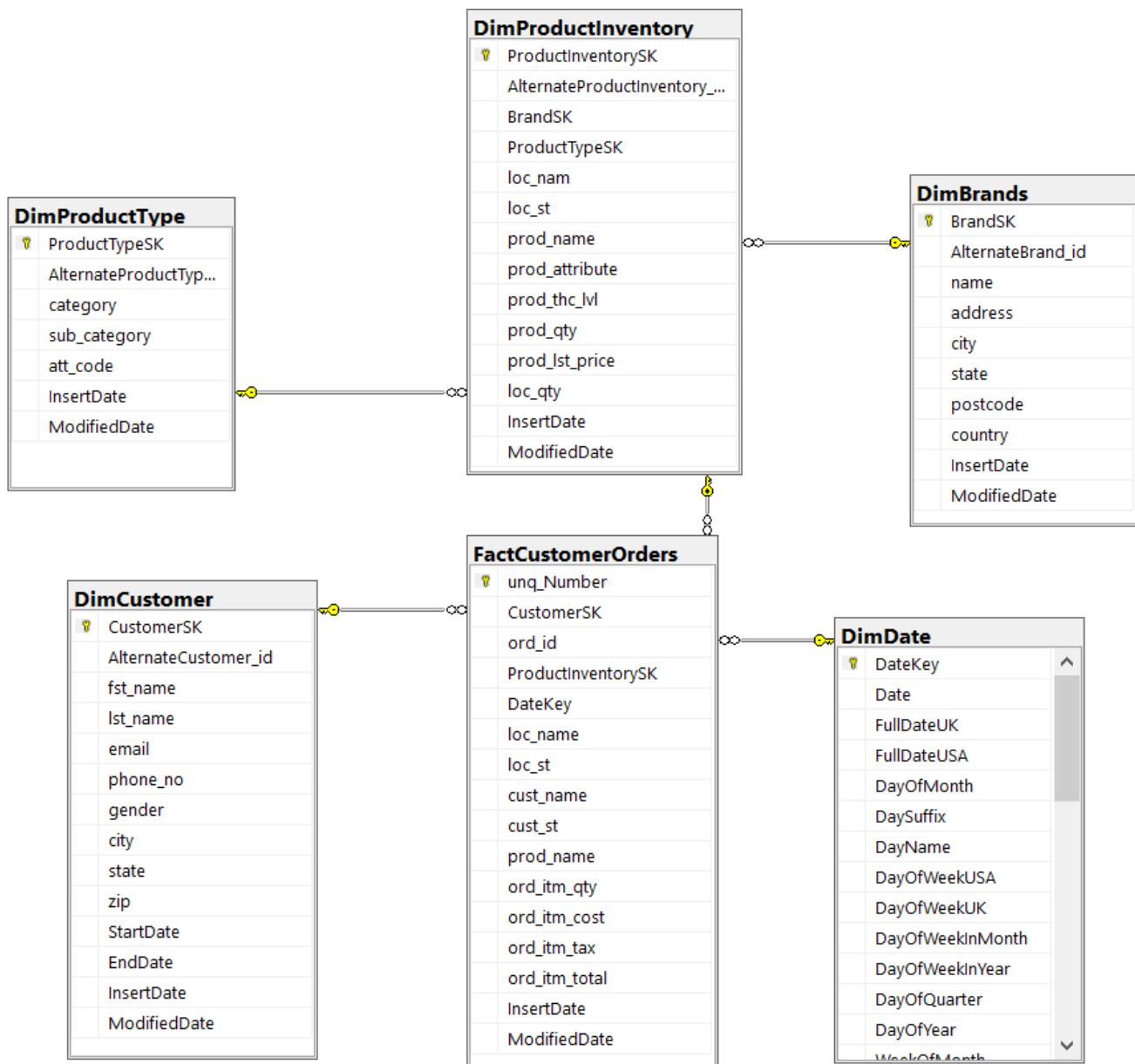
- As this architecture shows above for the ETL processing, first **DWBI_SourceDB** (Source database) and then **CustomerAddress.txt** (Text file) has been used for the data extraction to the staging process. After that staged in **DWBI_Staging** (Staging database) data are transforming and loading in to **DWBI_DW** (Data warehouse) and that data can be used for reporting, visualizing mining and analyzing purposes.

4. DATA WAREHOUSE DESIGN & DEVELOPMENT

i. Design

The DWBI_DW (Data Warehouse) is designed according below show to a snowflake schema as 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_item_total
$$(\text{Ord_itm_qty} * \text{ord_itm_cost}) + \text{ord_itm_tax} = \text{ord_itm_total}$$

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

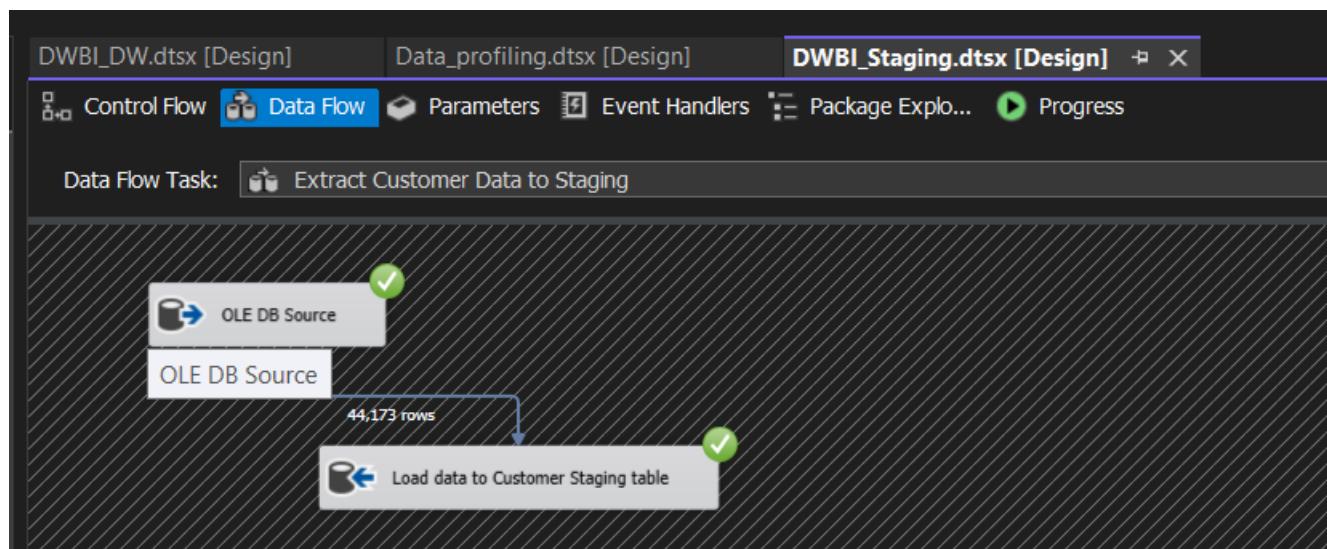
- Customer Details with customer addresses were considered as a slowly changing dimension.

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)

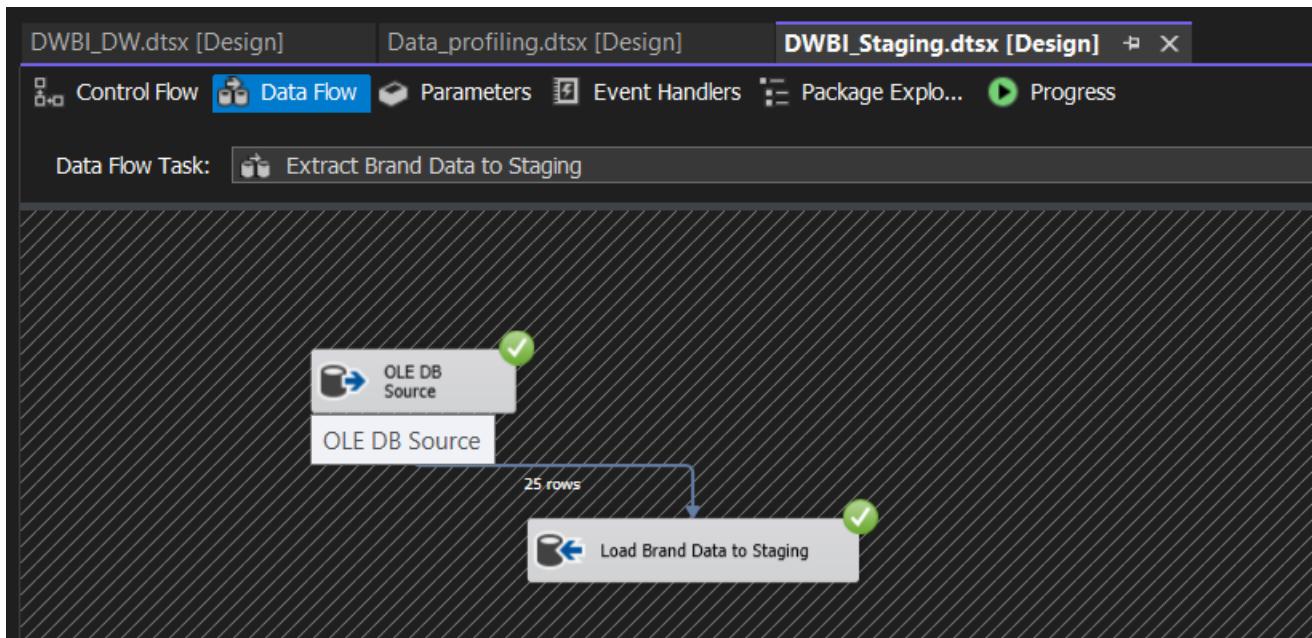
5.ETL DEVELOPMENT

i. Data Extraction & Load into Staging tables

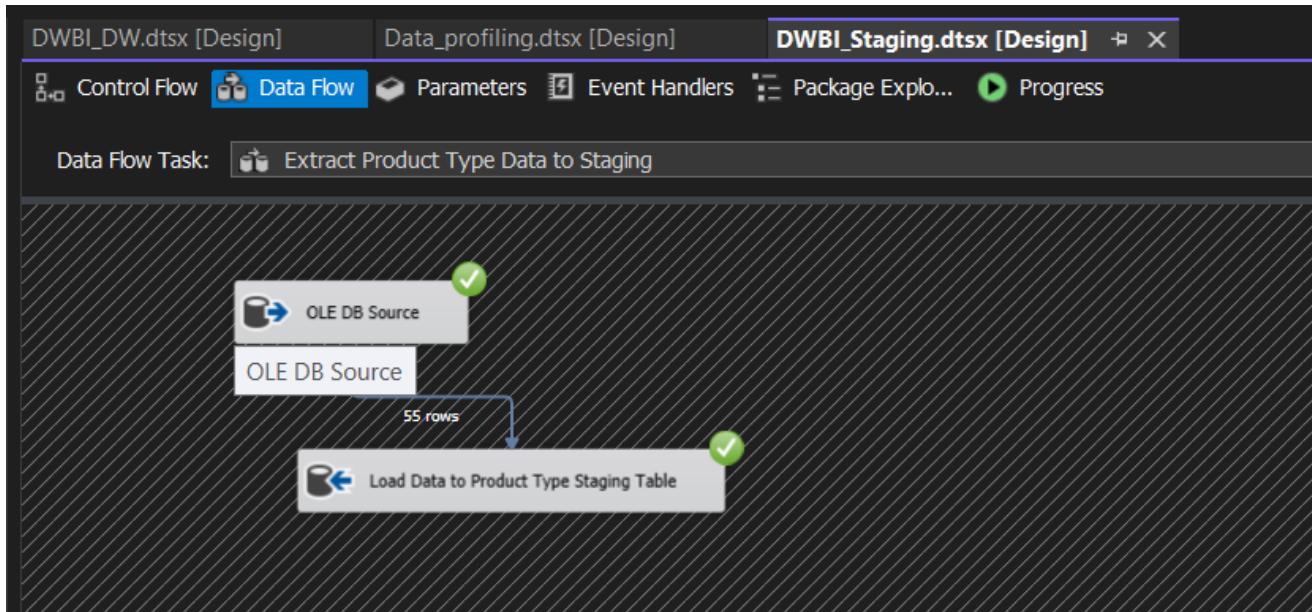
- Data Extraction is done by using the provided data sources mentioned above in Visual Studio 2022 (Data Tool) development environment. The text file and the source database were used here.
- Initially, **OLE DB SOURCE** (for source database) or **FLAT FILE SOURCE** (for flat files txt) is used to extract 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**.



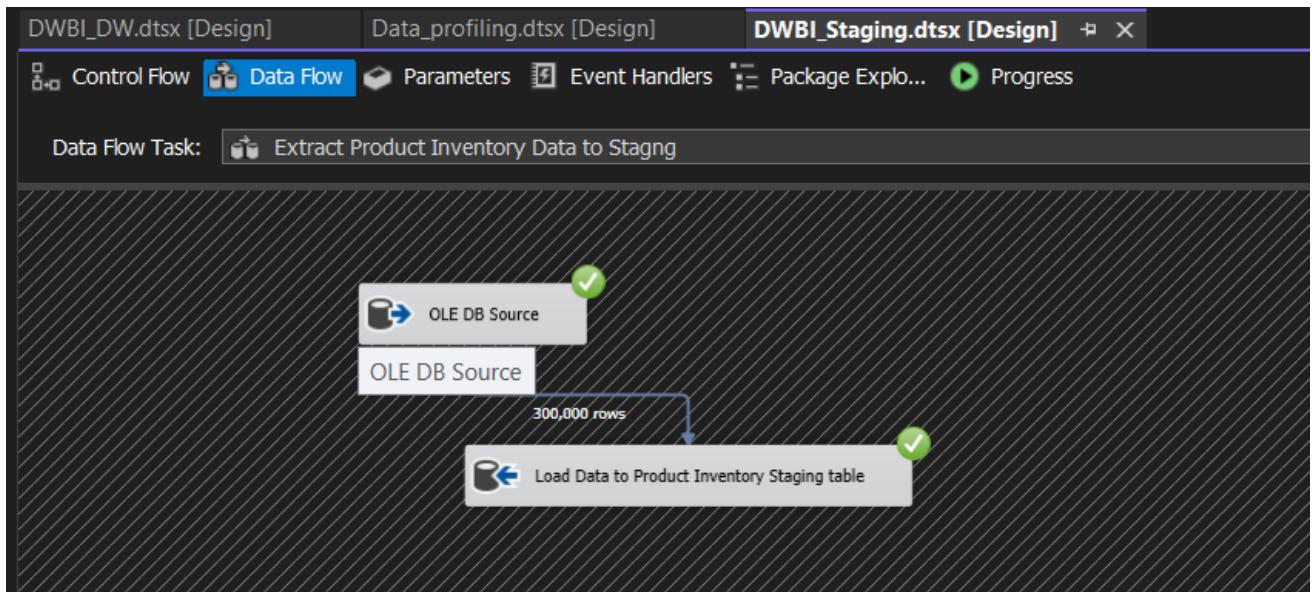
- Customer details data is extracted from Customers table in the source database and inserted to the StgCustomer table.



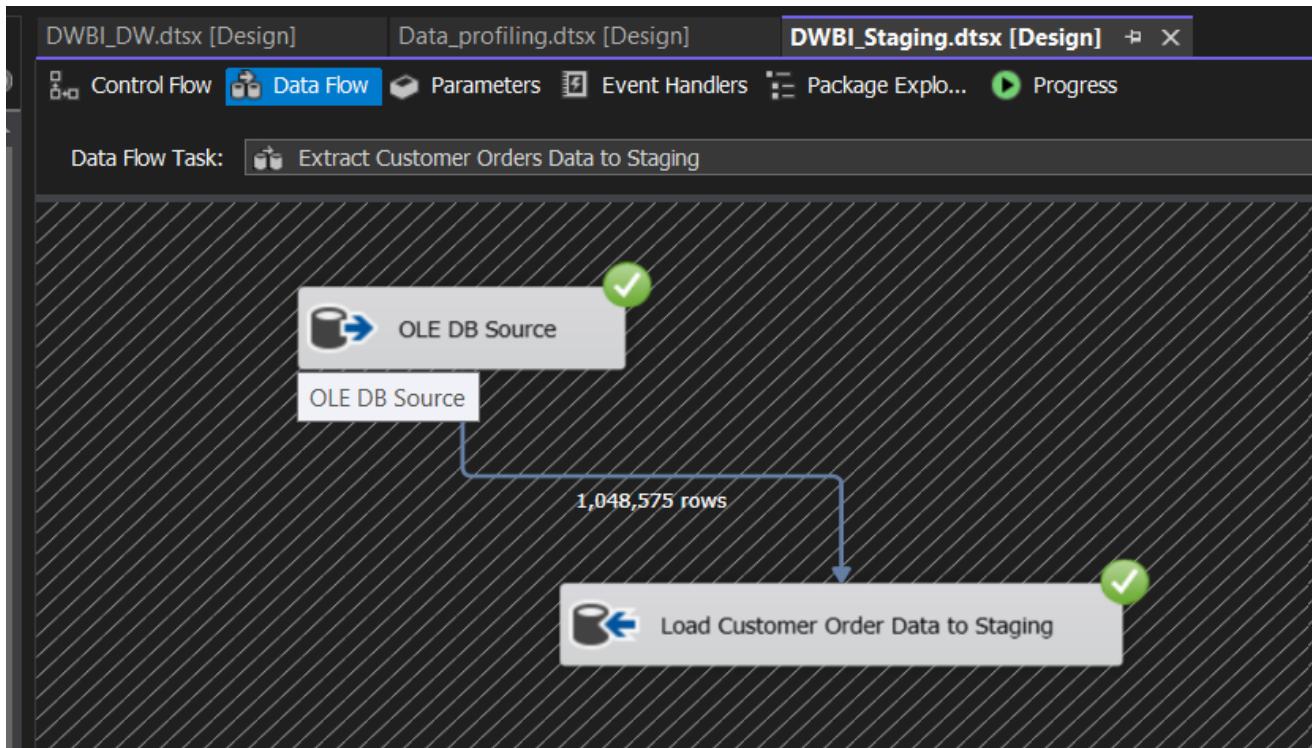
- Brand details data is extracted from Brand table in the source database and inserted to the StgBrand table.



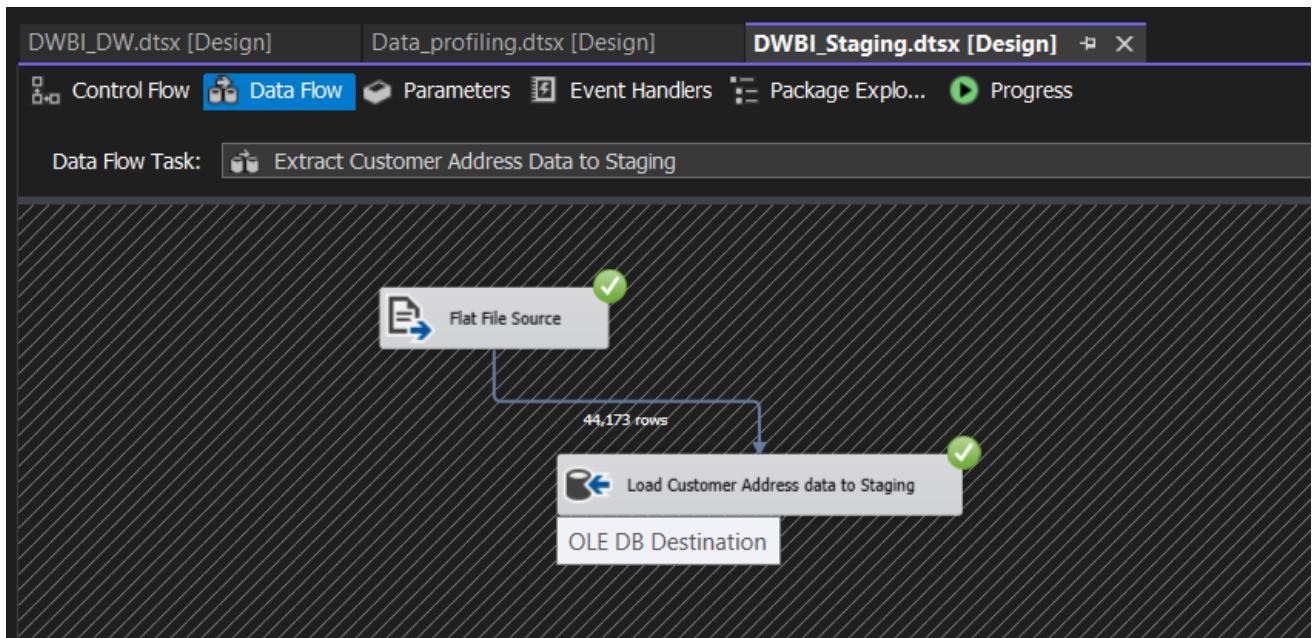
- Product Type data is extracted from ProductType table in the source database and inserted to the StgProductType table.



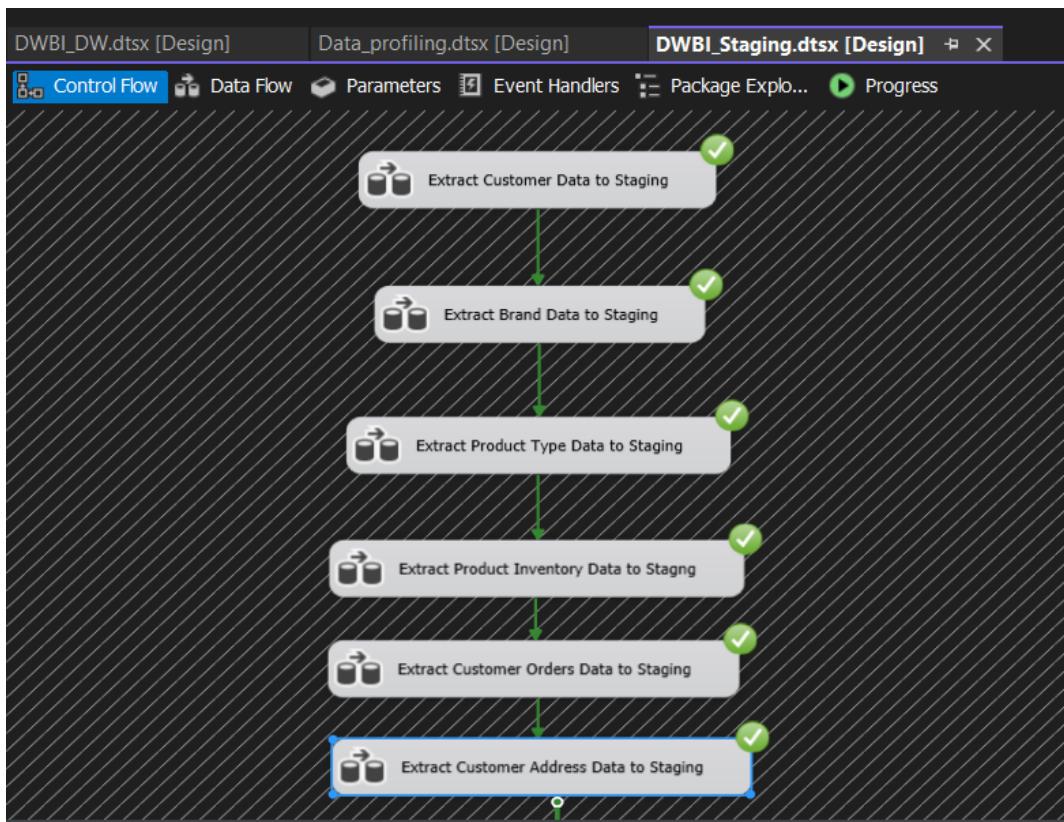
- Product Inventory data is extracted from Product Inventory table in the source database and inserted to the StgProductInventory table.



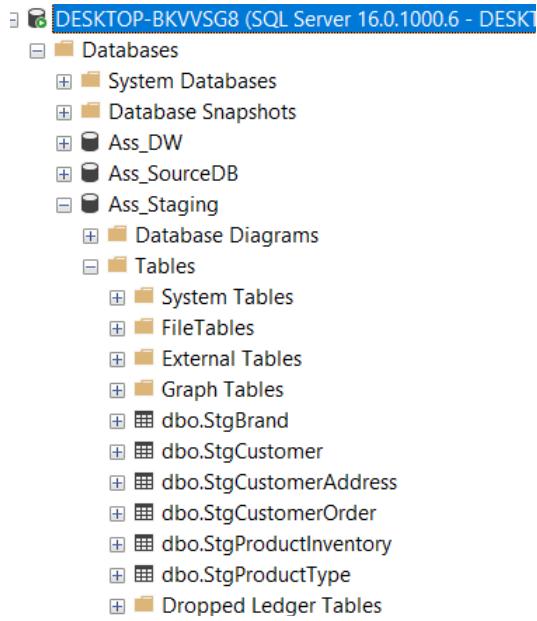
- Customer Orders data is extracted from CustomerOrders table in the source database and inserted to the StgCustomerOrder table.



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



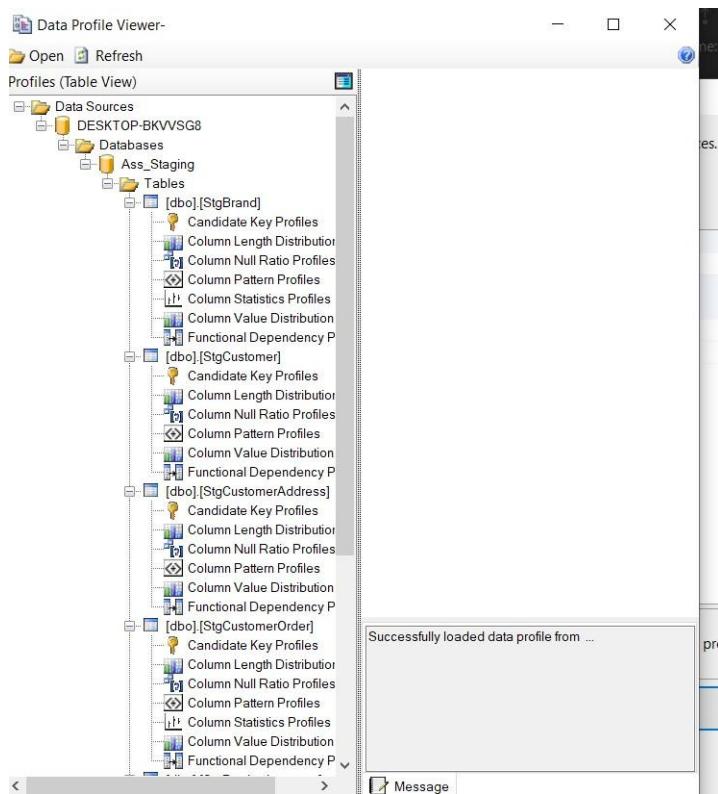
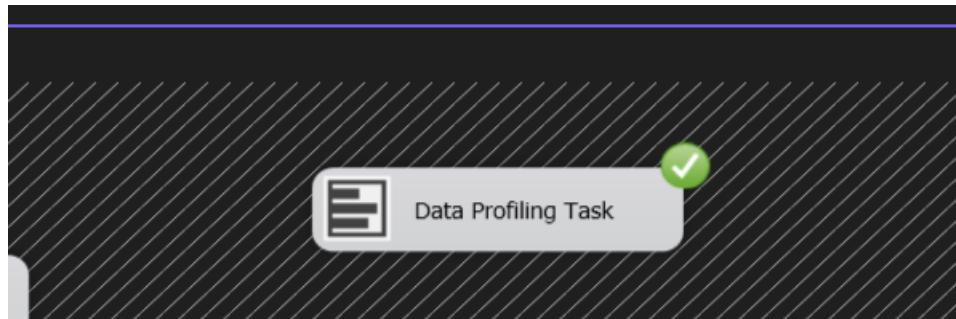
- The Control Flow of 'Extract Data and Load into Staging' is shown as above figure.



- Staging Tables have created and values are inserted.

ii. Data Profiling

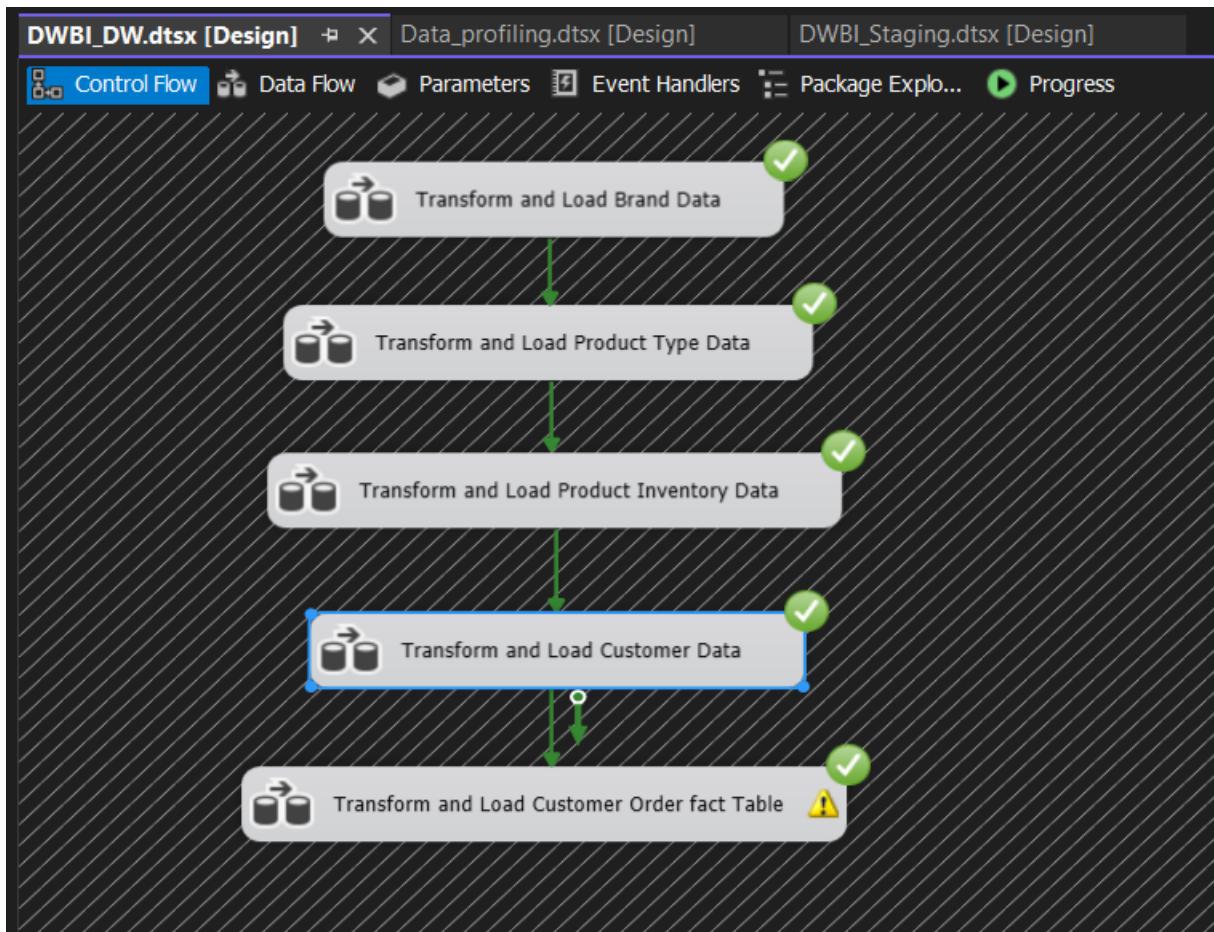
Data Profiling provides the means of analyzing large amount of data using different kind of processes. In this step, null values, repeated values and quality of the data is checked.



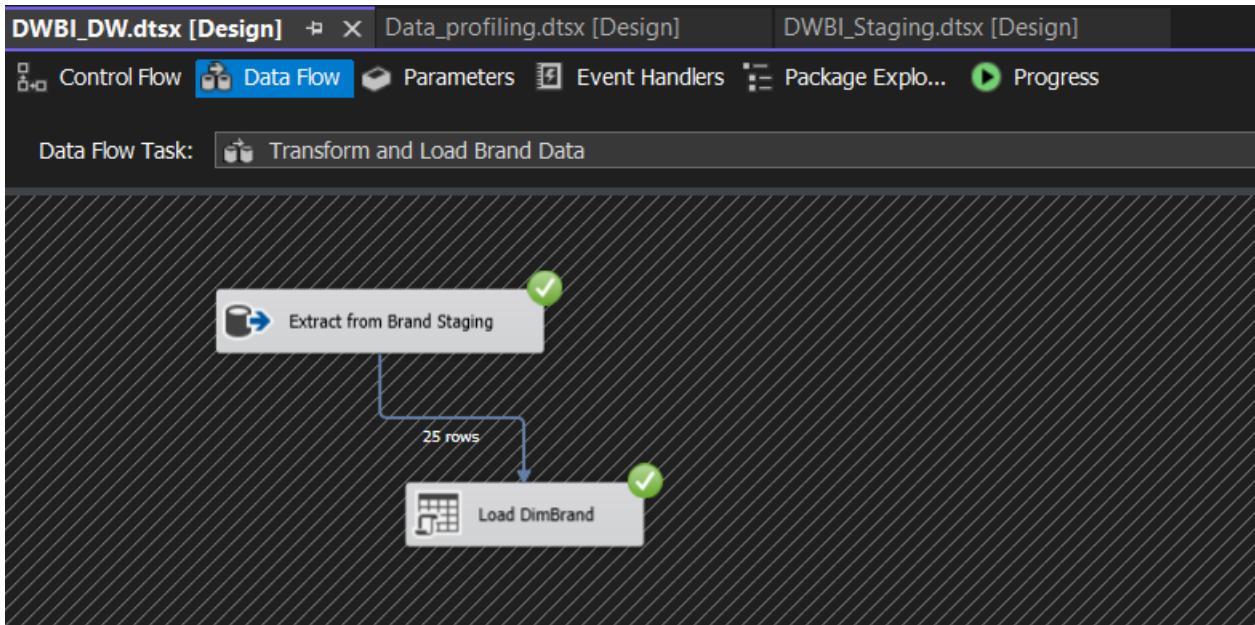
- Every staging table is profiled and saved in a selected location.
- As this shows, after the Staging step doing this task shows the things what the developer must consider about the data which are stored in staging table and the developer is able to identify the issues with staging data by data profiling (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.



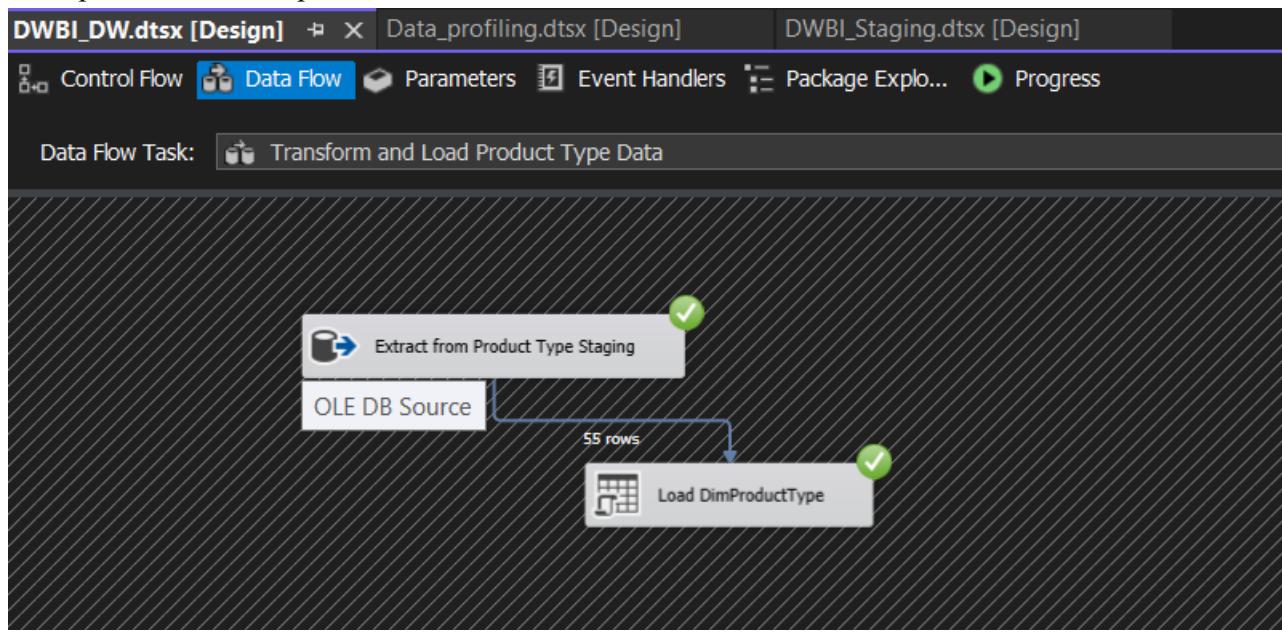
- In this step, the Dimension Tables are created in DWBI_DW are loaded with the data of relevant staging tables.



```
USE [DWBI_DW]
GO
/***** Object: StoredProcedure [dbo].[UpdateDimBrands] Script Date: 5/1/2025 6:04:25 PM *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
ALTER PROCEDURE [dbo].[UpdateDimBrands]
    @unq_id int,
    @name nvarchar(50),
    @address int,
    @city nvarchar(50),
    @state nvarchar(50),
    @postcode int,
    @country nvarchar(50)
AS
BEGIN
    -- If the brand does not exist, INSERT
    IF NOT EXISTS (
        SELECT BrandSK
        FROM dbo.DimBrands
        WHERE AlternateBrand_id = @unq_id
    )
    BEGIN
        INSERT INTO dbo.DimBrands
        (
            AlternateBrand_id,
            name,
            address,
            city,
            state,
            postcode,
            country,
            InsertDate,
            ModifiedDate
        )
        VALUES
        (
            @unq_id,
            @name,
            @address,
            @city,
            @state,
            @postcode,
            @country,
            GETDATE(),
            GETDATE()
        );
    END
    -- If the brand exists, UPDATE
    ELSE
    BEGIN
        UPDATE dbo.DimBrands
        SET
            name = @name,
            address = @address,
            city = @city,
            state = @state,
            postcode = @postcode,
            country = @country,
            ModifiedDate = GETDATE()
        WHERE AlternateBrand_id = @unq_id;
    END
END;
```

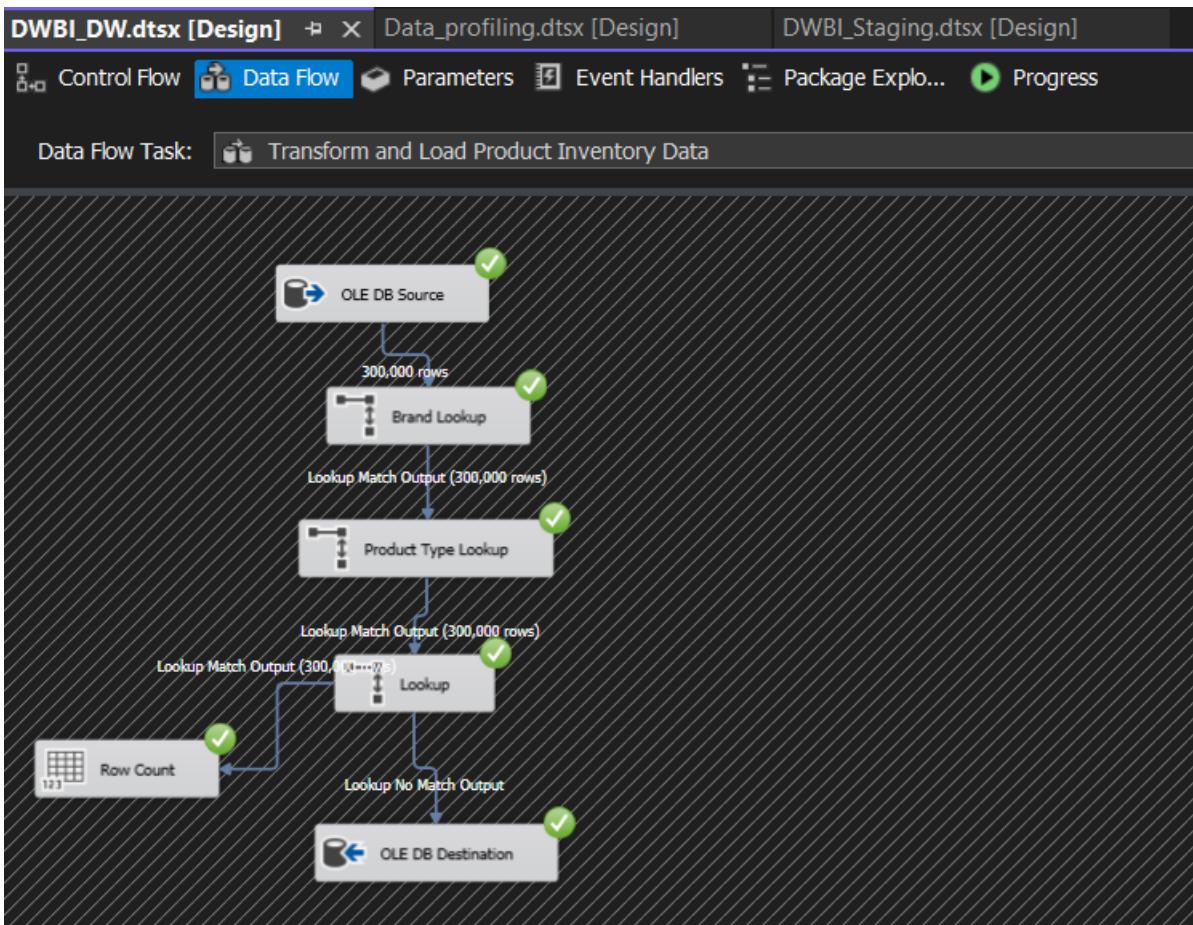
- Brand Details data are loaded to DimBrands

- UpdateDimBrands procedure is used to check whether the data is inserted or not.



```
SQLQuery2.sql - D...-BKVVSG8\HI (117)  X
USE [DWBI_DW]
GO
/******** Object:  StoredProcedure [dbo].[UpdateDimProductType]
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
ALTER PROCEDURE [dbo].[UpdateDimProductType]
@unq_id int,
@category nvarchar(50),
@sub_category nvarchar(50),
@att_code int
AS
BEGIN
if not exists (select ProductTypeSK
from dbo.DimProductType
where AlternateProductType_id = @unq_id)
BEGIN
insert into dbo.DimProductType
(AlternateProductType_id, category, sub_category, att_code,
InsertDate, ModifiedDate)
values
(@unq_id, @category, @sub_category, @att_code,
GETDATE(), GETDATE())
END;
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;
```

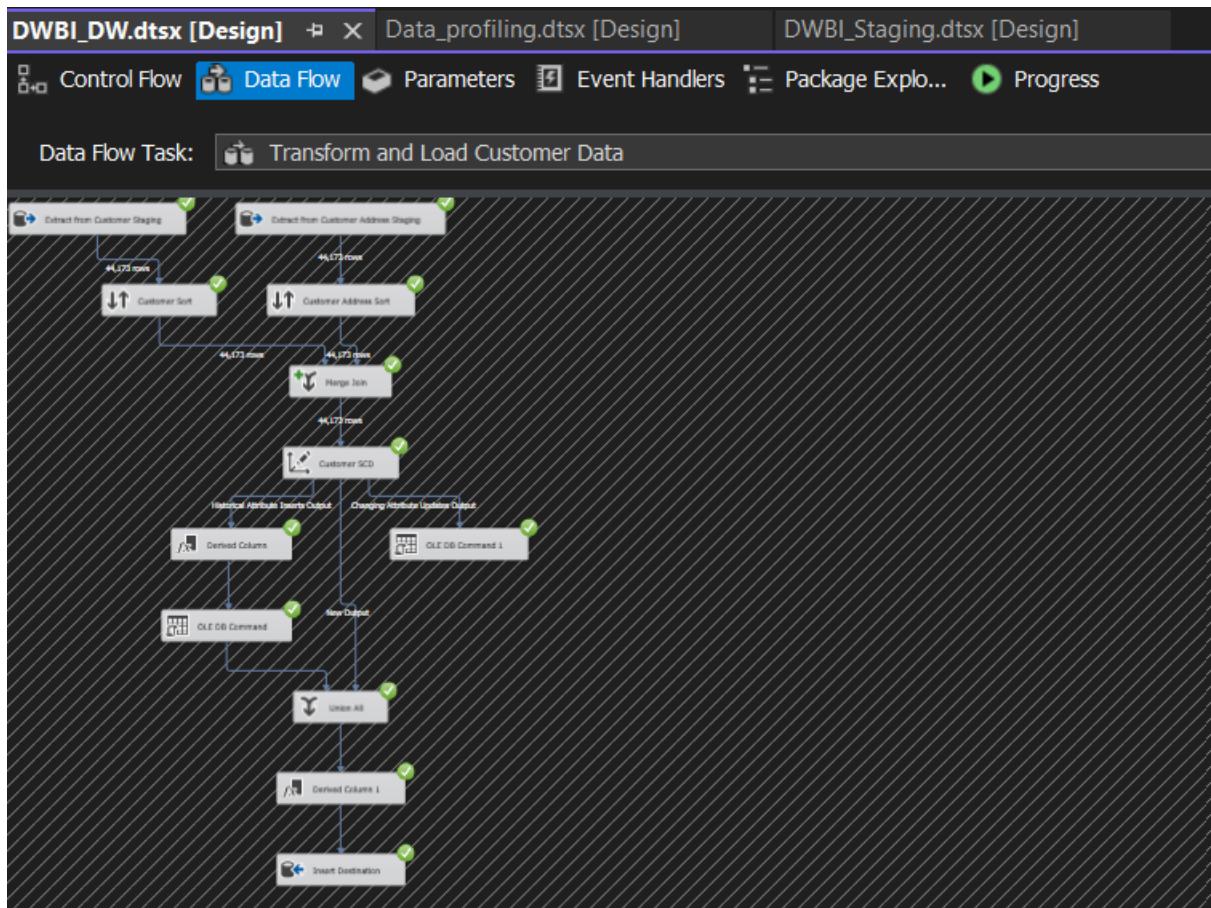
- Product Type data are loaded to DimProductType
- UpdateDimProductType procedure is used to check whether the data is inserted or not.



- 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.

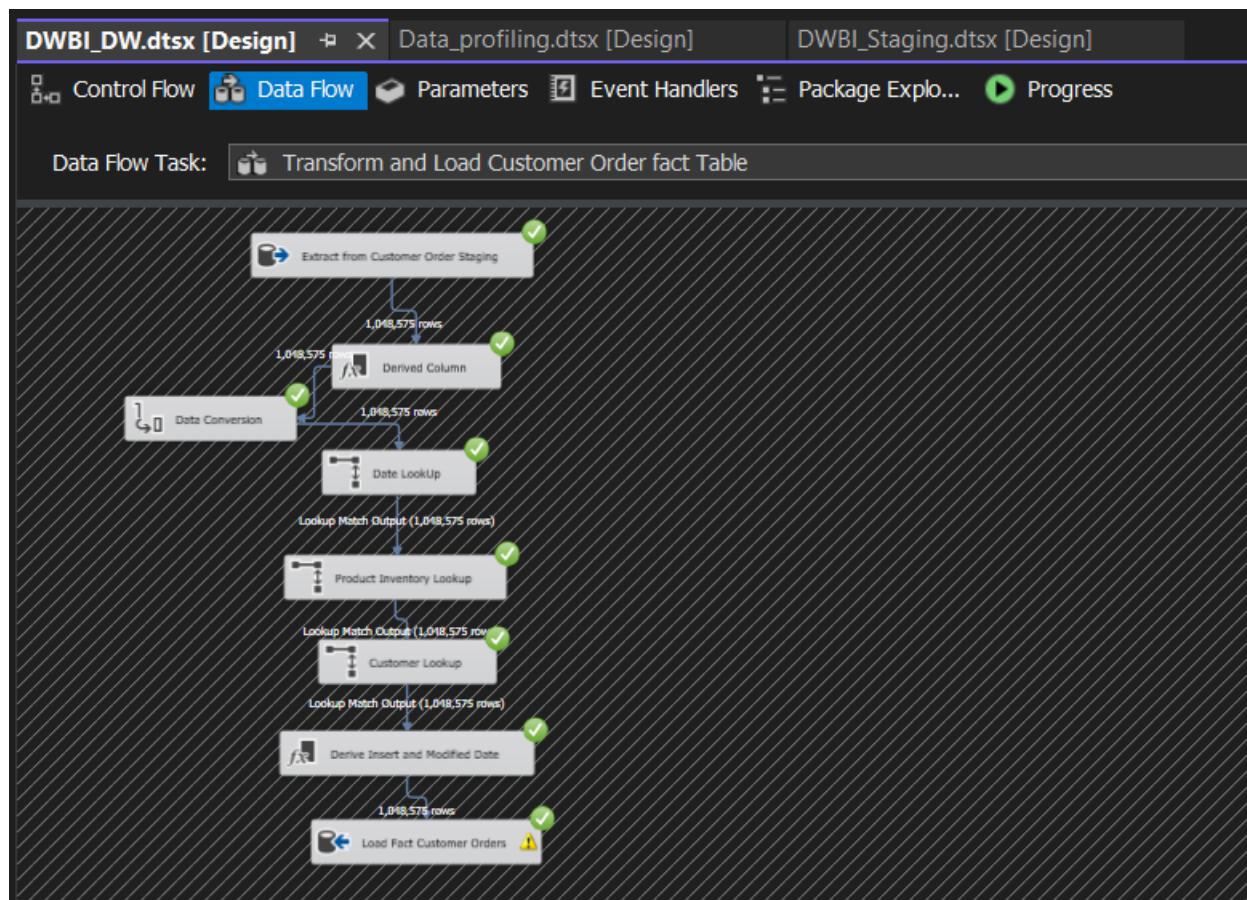


- As mentioned earlier under assumptions, customer details were considered as slowly changing details.
- The below mentioned columns were set as changing attributes:
 - Phone_no
 - E-mail

- 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.

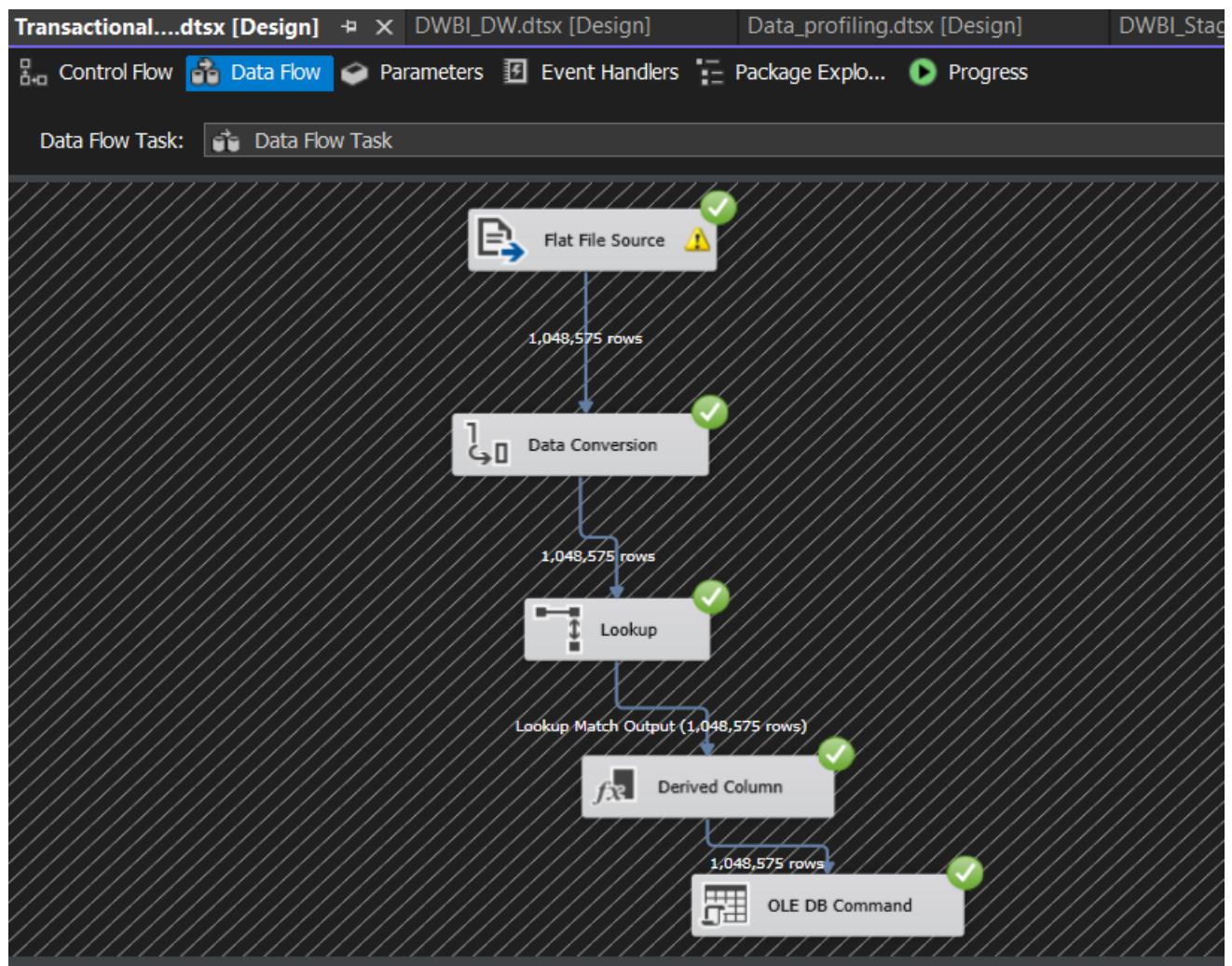
FactCustomerOrder

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 below Sequel Server Management Studio.

82 %

Results Messages

unq_Number	CustomerSK	ord_id	ProductInventorySK	DateKey	loc_name	loc_st	cust_name	cust_st	prod_name	ord_itm_qty	ord_itm_cost	ord_itm_t	
1	1	32136	1705482022	77446	20220604	Karing Kind - Adult Use	CO	Deborah Young	NE	Rutgers Smoke preroll shorties	1	20	1.60€
2	2	39450	1705512022	143316	20220604	Ultra Health - Bernillo	NM	Sara Benjamin	NM	Headdies concentrate cured sugar	4	120	34.75
3	3	27669	1705522022	132214	20220604	GreenHouse - Deerfield	IL	Linda Morton	IL	English Tobacco tincture dropper	1	18	2.16€
4	4	3223	1705532022	118029	20220604	Walla Walla Weedy	WA	Joseph Torres	NJ	Nutz concentrate cured sugar	1	105	24.14
5	5	22531	1705542022	81400	20220604	Lighthshade - Sheridan Recreational	CO	Brooke Ford	MI	Remedy concentrate live budder	1	105	30.45
6	6	41186	1705562022	49566	20220604	The Lakeside Collective	CA	Danielle Williams	CA	Rutgers Smoke concentrate live badder rosin	2	210	65.05
7	7	39884	1705572022	292389	20220604	DESERT ORGANIC SOLUTIONS - Palm Springs 92258	CA	Sarah Johnson	CA	Aladdin Smoke concentrate diamond sauce	1	15	4.65€
8	8	11606	1705582022	50279	20220604	The Red Door 30 CAP	CA	Susan Rogers	NC	Amazing concentrate live diamonds	1	105	23.1€
9	9	24395	1705592022	114167	20220604	Terrain Care Station - 33rd Ave. - Adult Use	CO	Crystal Hill	WI	Smokers Expo flower infused	1	280	28
10	10	22664	1705602022	83946	20220604	MC Caregivers	CO	Daniel Rose	MI	Nutz vape disposable distillate	1	40	11.6€
11	11	42610	1705612022	197776	20220604	Paz Dispensary	OR	Laure Allen	OR	English Tobacco concentrate jam	1	60	13.8€
12	12	37274	1705632022	72213	20220604	Green Tree (Medicinals) of Berthoud	CO	Melissa Green	CO	Cigarette preroll shorties	6	48	4.32€
13	13	11712	1705642022	197629	20220604	Patients Helping Patients	OR	Kimberly Mahoney	NC	Nutz concentrate jam	1	30	6.59€
14	14	34314	1705652022	29859	20220604	NHC: NATURAL HEALTH	CA	Rebecca Stephens	OK	Cigarrilos edible gummy	3	45	13.5
15	15	24725	1705672022	105214	20220604	Quality Choice Alternative Care Center	CO	Stephen King	WI	Old Glory vape disposable live resin	1	20	4.19€
16	16	43374	1705712022	265080	20220604	Sativa Sisters	WA	Barbara Garcia	WA	Head to Toe tincture dropper	1	18	3.59€
17	17	16675	1705722022	145368	20220604	Blum Las Vegas - Desert Inn	NV	Stephen Martinez	TN	Utopia concentrate budder	2	420	92.4€
18	18	39413	1705732022	143595	20220604	Ultra Health - Hobbs	NM	Lindsay Little	NM	Remedy edible mint	4	60	15.6€
19	19	39413	1705732022	143476	20220604	Ultra Health - Hobbs	NM	Lindsay Little	NM	Amazing preroll infused	4	32	4.15€

Query executed successfully.

82 %

Results Messages

ust_name	cust_st	prod_name	ord_itm_qty	ord_itm_cost	ord_itm_tax	ord_itm_total	InsertDate	ModifiedDate	accm_txn_complete_time	txn_process_time_hours
1 Deborah Young	NE	Rutgers Smoke preroll shorties	1	20	1.60000002384186	21.6000003814697	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 22:30:55.847	52
2 Sara Benjamin	NM	Headdies concentrate cured sugar	4	120	34.799999237065	154.800003051758	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 17:17:24.480	47
3 Linda Morton	IL	English Tobacco tincture dropper	1	18	2.16000008583069	20.1599998474121	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 20:34:15.967	50
4 Joseph Torres	NJ	Nutz concentrate cured sugar	1	105	24.1499996185303	129.1499993896484	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 01:14:23.577	31
5 Brooke Ford	MI	Remedy concentrate live budder	1	105	30.4500007629395	135.1499996948242	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 01:22:21.067	31
6 Danielle Williams	CA	Rutgers Smoke concentrate live badder rosin	2	210	65.0999994741211	275.100006103516	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-04 04:17:30.293	58
7 Sarah Johnson	CA	Aladdin Smoke concentrate diamond sauce	1	15	4.65000009536743	19.6499998185303	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-04 22:18:05.907	76
8 Juanan Rogers	NC	Amazing concentrate live diamonds	1	105	23.1000003814697	128.100006103516	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 03:54:33.150	33
9 Crystal Hill	WI	Smokers Expo flower infused	1	280	308	86.4	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-02 22:37:59.837	28
10 Janiel Rose	MI	Nutz vape disposable distillate	1	40	11.6000003814697	51.5999984741211	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 15:54:27.093	45
11 Laurie Allen	OR	English Tobacco concentrate jam	1	60	13.8000001907349	73.8000030517578	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 17:32:14.517	47
12 Melissa Green	CO	Cigarette preroll shorties	6	48	4.32000017166138	52.3199996948242	2025-05-01 18:43:37.063	2025-05-02 19:14:06.053	25	
13 Kimberly Mahoney	NC	Nutz concentrate jam	1	30	6.59999990463257	36.59999984741211	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 23:17:29.703	101
14 Rebecca Stephens	OK	Cigarrilos edible gummy	3	45	13.5	58.5	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-02 02:32:31.417	8
15 Stephen King	WI	Old Glory vape disposable live resin	1	20	4.19999980926514	24.2000007629395	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-02 00:11:57.157	6
16 Barbara Garcia	WA	Head to Toe Tincture dropper	1	18	3.59999990463257	21.6000003814697	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-05 08:14:36.840	86
17 Stephen Martinez	TN	Utopia concentrate budder	2	420	92.4000015258789	512.400024414063	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-03 19:57:27.270	49
18 Lindsay Little	NM	Remedy edible mint	4	60	15.6000003814697	75.5999984741211	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-04 08:57:25.337	62
19 Lindsay Little	NM	Amazing preroll infused	4	32	4.15999984741211	36.15999984741211	2025-05-01 18:43:37.063	2025-05-01 18:43:37.063	2025-05-04 15:45:29.037	69

Query executed successfully.

*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.