Sri Lanka Institute of Information Technology
Data warehousing and Business Intelligence

Assignment 1
2020

Submitted By: IT18115208 – M.C.P Mendis

1. Data set selection

This Dataset is from a Superstore sale. They sell their goods in different geographical locations by online. In this Scenario Customer who orders from the superstore and sellers who provide products to the superstore details were stored.

Inside the Superstore table there were Order details, Customer details, Seller details, Product details columns. Because of that I created separate tables and added more columns from other retails datasets and modified my dataset.

Following ER- diagram will describe the scenario of my selected dataset.

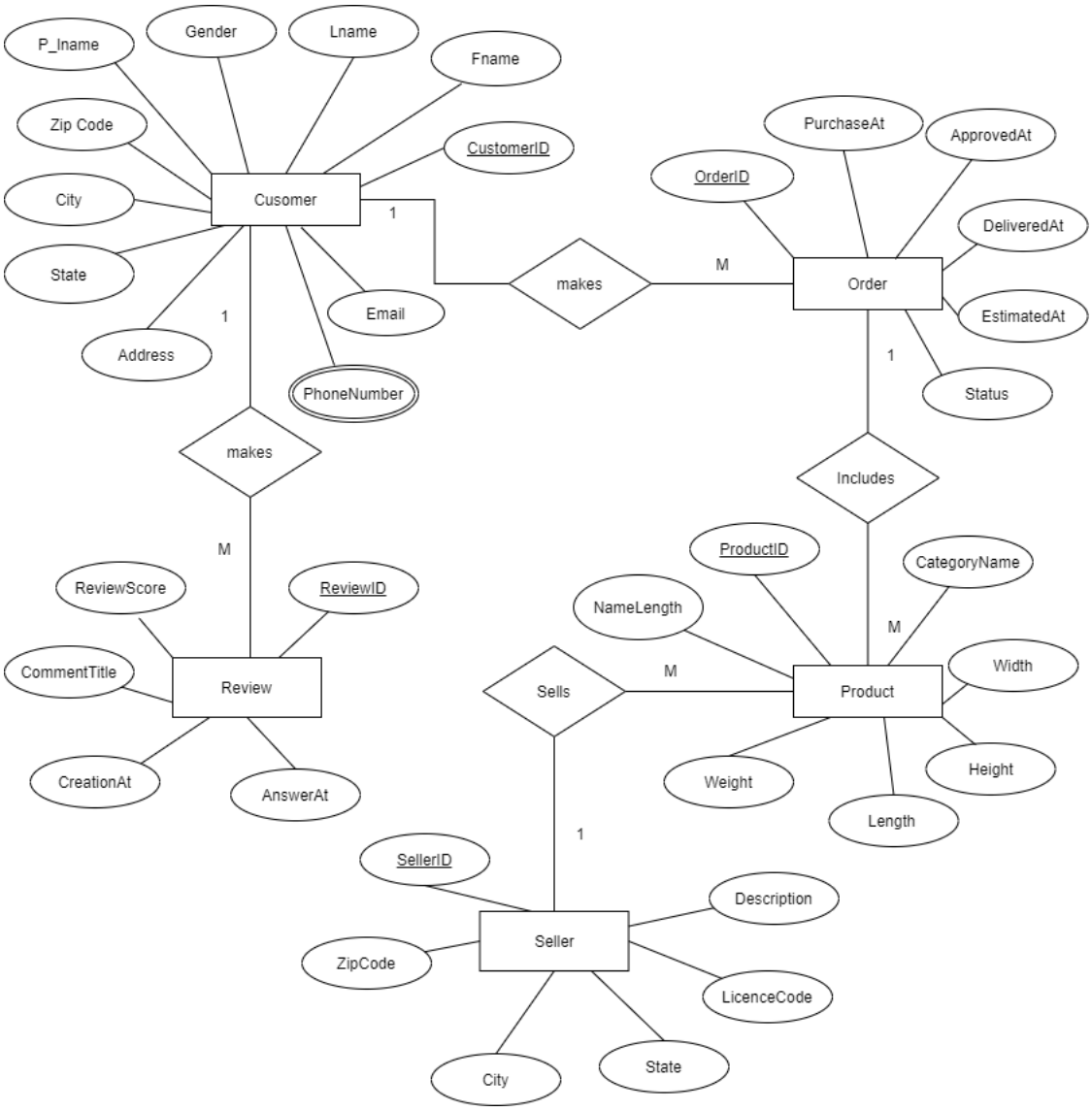


Figure 1.1

2. Preparation of Data Sources

Data Source Type	Table name	Column name	Data type	Description
CSV File	Dbo. CustomerDetails	CustomerID	Nvachar (255)	Customer Unique ID
		CustomerUniqueID	Nvachar (255)	Customer code
		Fname	Nvachar (255)	First name of the customer
		Lname	Nvachar (255)	last name of the customer
		Gender	Nvachar (255)	gender of the customer
		PhoneNumber	Nvachar (255)	Phone number of the customer
		Email	Nvachar (255)	Email of the customer
	Dbo. ProductDetails	ProductID	Nvachar (255)	Product Unique ID
		CategoryName	Nvachar (255)	Product Category name
		ProductNameLength	float	Length of name
		DescriptionLength	float	Length of description
		ProductPhotosQty	float	Product Photos
		ProductWeight	float	Product Weight
		ProductHeight	float	Product Height
		ProductLength	float	Product Length
		ProductWidth	float	Product Width

CSV File	Dbo. OrderDetails	OrderID	Nvachar (255)	Order Unique ID
		OrderStatus	Nvachar (255)	Order status
		PurchaseAt	datetime	Purchased timestamp
		ApprovedAt	datetime	Order approved timestamp
		EstimatedAt	datetime	Order estimates timestamp
		DeliveredCustomerAt	datetime	Order delivered timestamp
		DeliveredCarrierAt	datetime	Order delivered carrier timestamp
		PaymentSequential	float	Payment Sequential
		PaymentValue	float	Payment Value
		PaymentType	Nvachar	Payment Type
		PaymentInstallments	float	Payment Installments
	Dbo. SellerDetails	SellerID	Nvachar (255)	Seller Unique ID
		ZipCode	float	Seller Zip Code
		City	Nvachar (255)	Seller city name
		State	Nvachar (255)	Seller state name
		LicenseCode	float	Seller's License Code
		Description	Nvachar (255)	Seller Title/Description
	Dbo.Review	ReviewID	Nvachar (255)	Review Unique ID
		ReviewScore	float	Score given to review

		ReviewCommentTitle	Nvachar (255)	Review Title

		review_creation_date	datetime	Timestamp for review created date
		review_answer_timestamp	datetime	Timestamp for answered date
Text File	CustomerAddress.tbl.txt (Directly sent to separate staging table and combined then in transformation part)	CustomerID	Nvachar (255)	Customer Unique ID
		ZipCode	Nvachar (255)	Customer's Zip code
		City	Nvachar (255)	Customer's city
		State	Nvachar (255)	Customer's state
		Address	Nvachar (255)	Customer's address

At First, I created separate tables for my source dataset by dividing my main superstore.CSV file. Because it includes all the details of customers, orders etc. Then I imported those csv files into my newly created database (SourceDB).

And Customer Address details saved into text file format. This Text file contains all the customers address informations.

After I imported my CSV files into SorceDB, I created Data Warehouse named "SourceDB_DW" and created my dimension tables and fact table inside data warehouse.

3. Solution Architecture

Following architecture shows the high-level BI solution to the warehouse design.

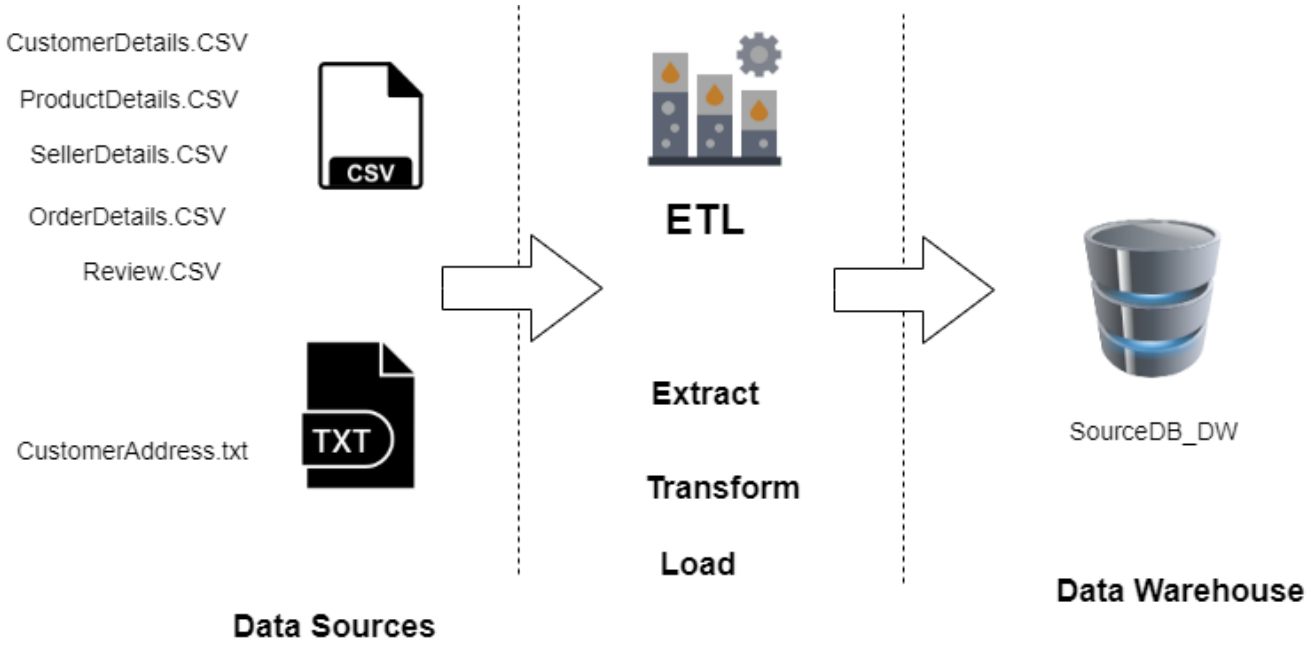


Figure 3.1

1.Data Source

I used two main data sources to the data warehouse (SourceDB_DW).

- CSV Files
  In SourceDB_DW, CSV files are the main data source. I imported 5 CSV files to my SourceDB_DW.

- Text File
  The second data source is a Text file. This text file includes the Customer address information.

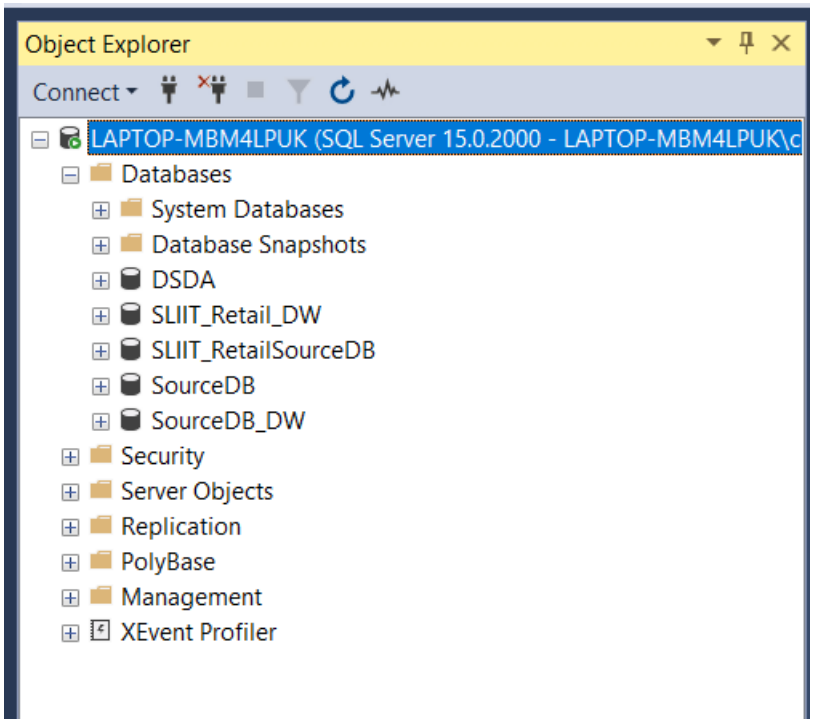


Figure 3.2

2. ETL (Extract, Transform, Load)

1.Extract

In this part I extracted my data source (csv files) in SourceDB. And My text file which includes addresses information directly sent to a separate staging level in SourceDB_DW. All my extracted data store in SourceDB_DW.

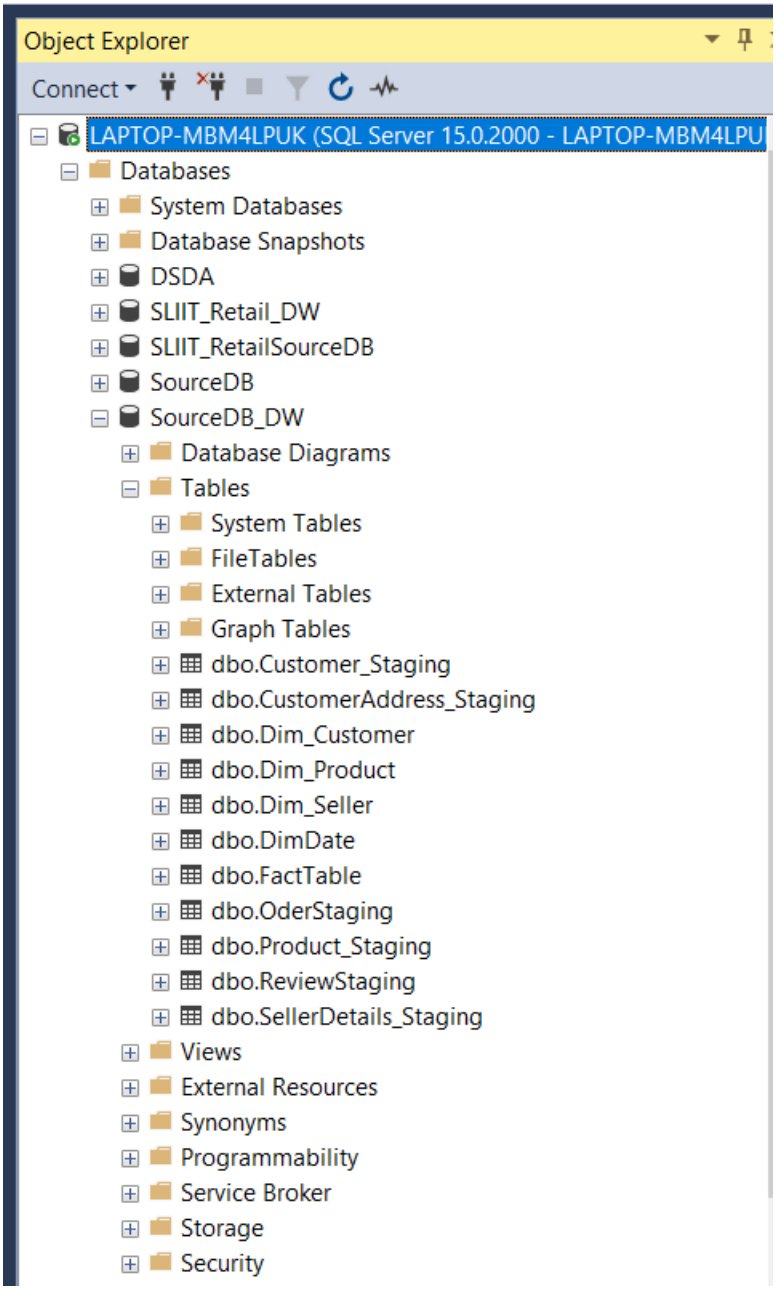


Figure 3.3

2.Transform

Following figure shows the transformation process which includes four steps.

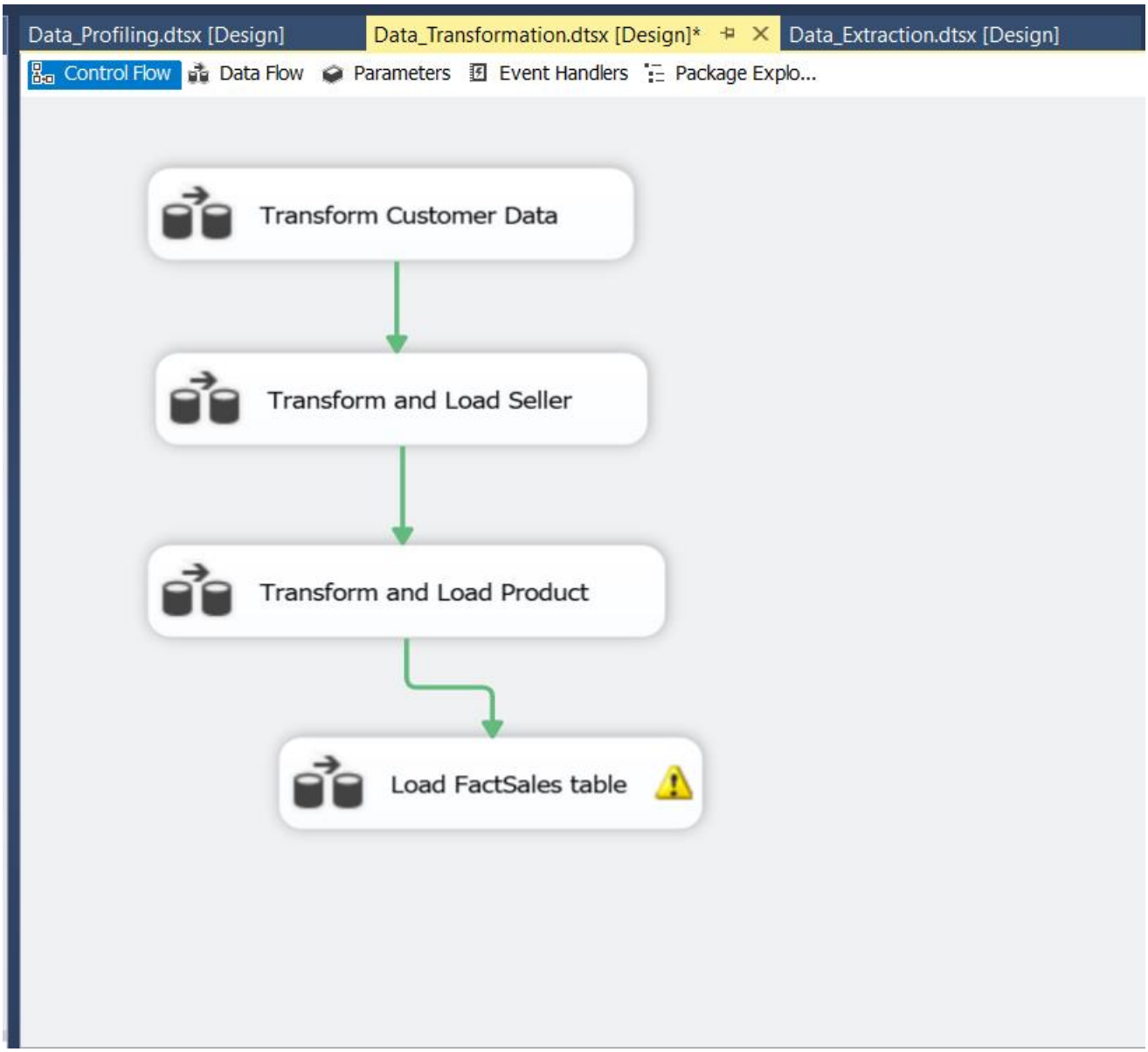
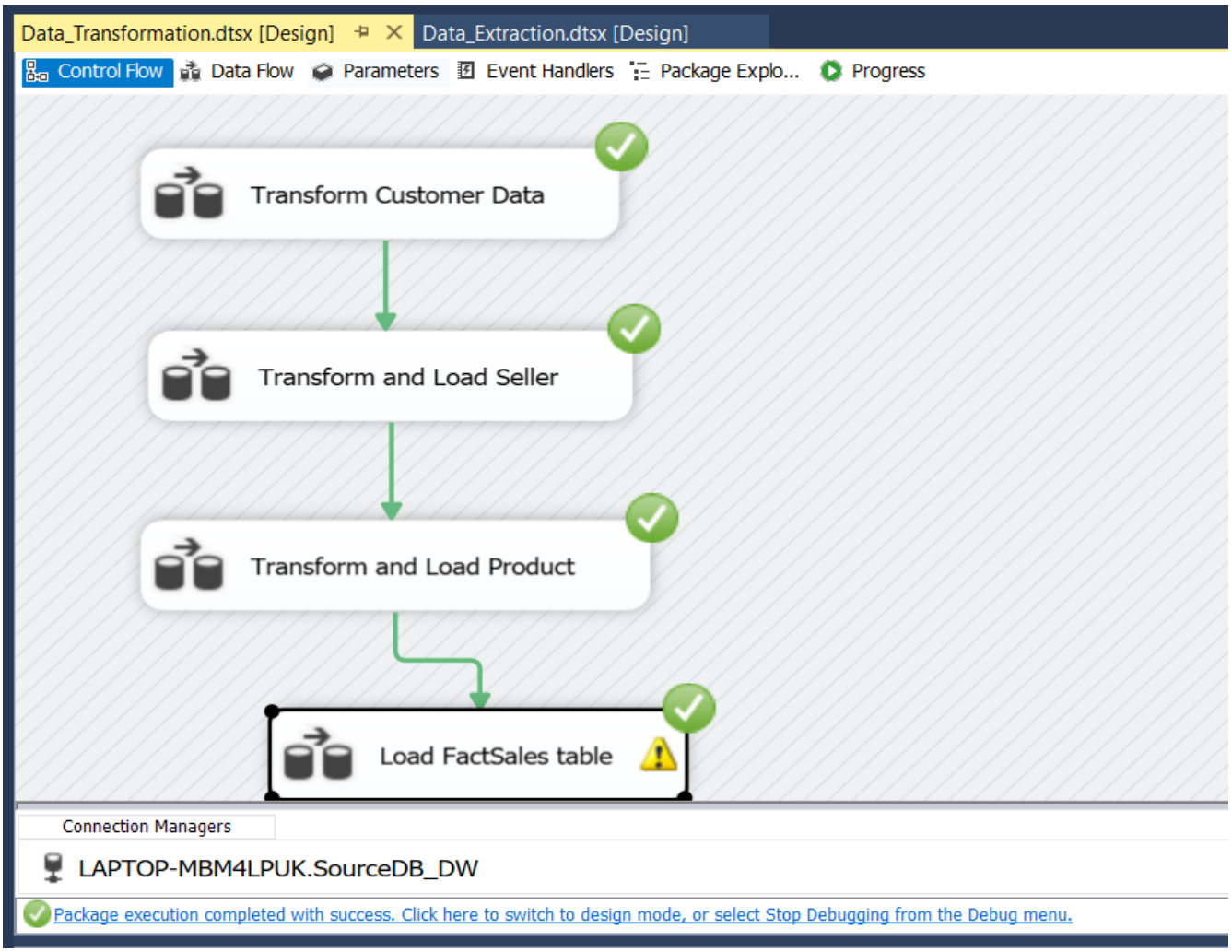


Figure 3.4

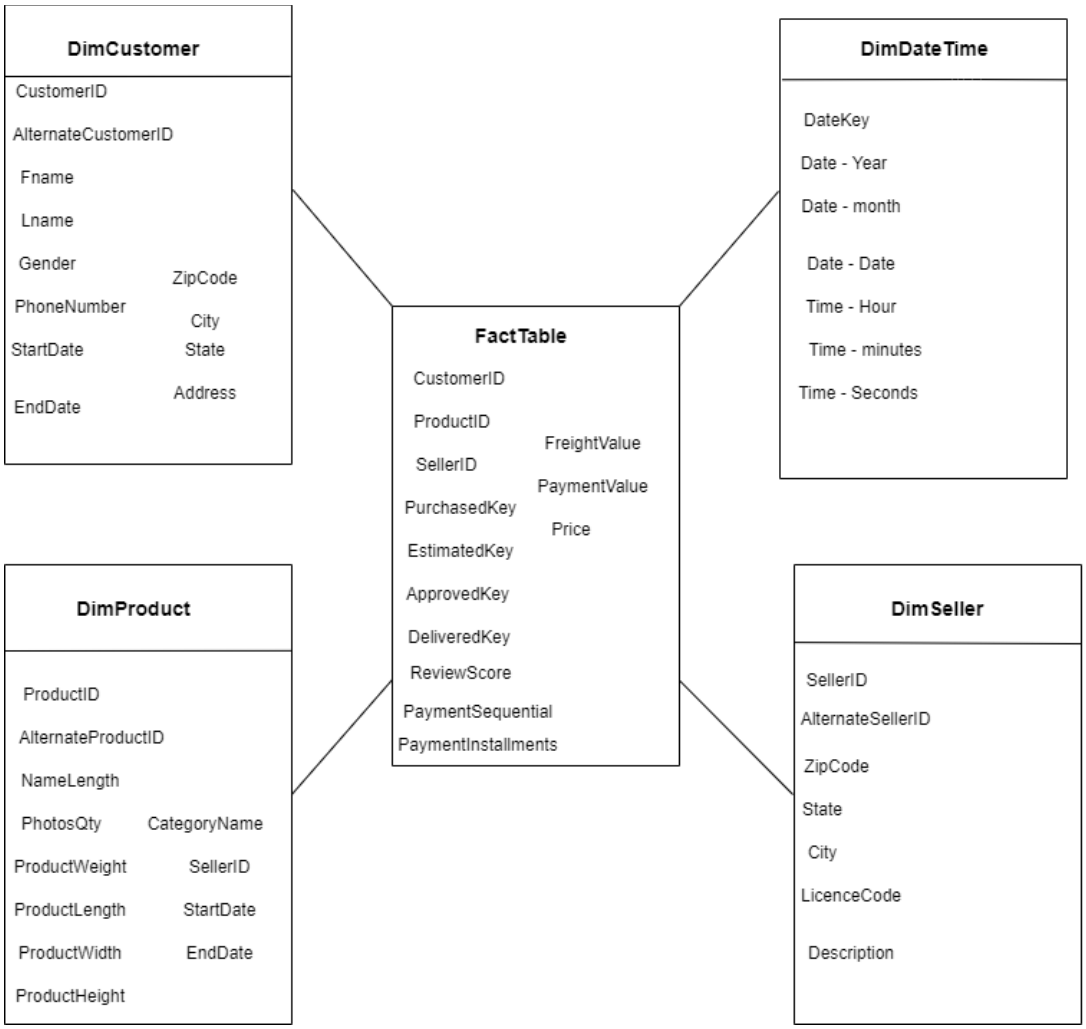
3.Load

In this process, data will load to destination to the dimension tables in the SourceDB_DW.



4. Data warehouse design & development

Following figure will show how the fact table and dimension tables are joined in a logical manner. In the fact table we stored the measures of interest and key values. (Price, Payment Value, Freight Value, and Review Score)



- In this design customer dimension table is a slowly changing dimension table which I have implemented as type two. Because of that I added two columns named StartDate and EndDate to describe more about customer details.

  And it has four historical attributes (City, ZipCode, State, and Address) which can be changed over time.

- I got Product dimension table as a slowly changing dimension which I implemented as type two Because this method adds a new row to the new value and maintains the existing row for historical purposes.

5. Test Planning and Design Test Cases

Scope	1. Duplicate Values checking	Values in column are unique
	2.Record counts validation checking	The number of the records in the dimension table is the same number in source table.
	3.Data type checking	The data type of the dimension table is the same data type in source table.
	4 Index check	Ensure that index created with required columns.
Out of scope	Validations	
Test Environment	Microsoft SQL Server Management Studio 18	
Test Tools	Microsoft SQL Server for Visual Studio 2017 (SSDT)	
Roles and responsibilities	Create test plan Create test cases Execute	
Schedule	4/28/2020 – 4/29/2020	

Test cases

Test Scenario ID	01	Test Case ID	01
Test Case Description	Data Completeness	Test Priority	High
Pre-Requisite		Post-Requisite	NA

Test Execution Steps:

S.No	Action	Inputs	Expected Output	Actual Output	Test Browser	Test Result	Test Comments
1	Check Duplicate values	Figure 7.1	Values in columns are unique	Customer ID – 2 Customer_id = 00012a2ce6f8dcda20d059ce98491703	-	Succe ssful	No duplicate customer IDs
2	Record counts validatio n checking	Figure 7.2 Figure 7.3	source table records = Dimension table records	No of columns in dimension – 96479 No of columns in source - 96478	-	Failed	Dimension table has extra row.
3	Data type checking	Figure 7.4	Same data types	Figure 7.4	-	Succe ssful	Data types same.
4	Index and check	Figure 7.5	Index created with requirement columns.	Figure 7.5	-	Succe ssful	Successful

Sample test data

```
SQLQuery1.sql - L...M4LPUK\chame (56))*

select *
from Dim_Seller

select *
from FactTable

-----Source Table -----
select *
from SourceDB.dbo.CustomerDetails

select *
from SourceDB.dbo.Product

select *
from SourceDB.dbo.Review

select *
from SourceDB.dbo.SellerDetails

select *
from SourceDB.dbo.OrderUpdate
```

100 %

Results | Messages

	CustomerID	AlternateCustomerID	CustomerUniqueID	fname	lname	Gender	PhoneNumber	Email
13	13	00062b33cb9f6fe976afdcff967ea74d	dbdab35c90de88d44f96fd05b8688cea	Jennifer Braxton	Ross	F	647-555-0146	destiny28@adventu
14	14	00066ccbe787a588c52bd5ff404590e3	514f8ca0f04813ed414874f8f422e6c3	Delfina Latchford	Cavallari	NULL	695-555-0161	matthew1@adventu
15	15	00072d033fe2e59061ae5c3aff1a2be5	6a5ecf25eae9db640b117f6a67c67790	Sanjit Chand	Kumar	F	1 (11) 500 555-0112	alisha32@adventur
16	16	0009a69b72033b2d0ec8c69fc70ef768	39ee665787cdce6191c4b41431bdc4c2	Eric Murdock	Shan	F	1 (11) 500 555-0172	april8@adventure-w
17	17	000bf8121c3412d3057d32371c5d3395	a753e2043d1bab6426cfc45022073647	Frank Merwin	Liu	M	493-555-0141	louis42@adventure-
18	18	000e943451fc2788ca6ac98a682f2f49	dc869c4d42ab0d63664fdcea4e7b7440	Ionia McGrath	Berry	NULL	471-555-0181	john11@adventure-
19	19	000f17e290c26b28549908a04cfe36c1	e98e98f29f69ce5beba405c65acec38c	Ionia McGrath	Jai	F	921-555-0165	nicole46@adventur
20	20	000fd45d6fedae68fc6676036610f879	0bf1278577a2acf8532a031d98cd517c	John Grady	Chandra	M	990-555-0126	levi1@adventure-wo

	ProductID	AlternateProductID	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_heig
13	13	508d48ea59be64138f0015d2cb9c75e7	NULL	1212	1	5400	18	10
14	14	df473738565b52f77b4e22b328b41576	NULL	369	1	400	16	2
15	15	302a19dacdbb5ed2f74f9dee8126ef79	NULL	882	5	300	17	7
16	16	f41b27c06a91e6f554c113a7e702ee7a	NULL	1275	2	2750	60	10
17	17	47969dd948e918289f809be899ddfb4c	NULL	501	1	258	19	12
18	18	835fdb74fa8c0da45cd0879b1307fcd0	NULL	563	1	1100	35	25
19	19	8097e6d8de77768d9f72295263e440fa	NULL	601	1	224	16	13
20	20	5ec665da4518623a3a2d7731d56bfd68	NULL	543	2	1600	31	14

	SellerID	AlternateSellerID	seller_zip_code_prefix	seller_city	seller_state	LicenceCode	LicenceDescription
1	1	5670f4db5b62c43d542e1b2d56b0cf7c	3694	sao paulo	SP	1010	Limited Business License
2	2	7142540dd4c91e2237acb7e911c4eba2	16301	penapolis	SP	1010	Limited Business License
3	3	4a3ca9315b744ce9f8e9374361493884	14940	ibitinga	SP	1625	Raffles
4	4	40ec8ab6cdafbcc4f544da38c67da39a	85603	francisc...	PR	1010	Limited Business License
5	5	8ae520247981aa06bc94abddf5f46d34	88370	navega...	SC	4404	Regulated Business Lic...
6	6	cd68562d3f44870c08922d380acae552	14050	ribeirao ...	SP	1010	Limited Business License
7	7	cd68562d3f44870c08922d380acae552	14050	ribeirao ...	SP	4409	Itinerant Merchant
8	8	8b321bb669392f5163d04c59e235e066	1212	sao paulo	SP	4406	Peddler License

	CustomerID	ProductID	SellerID	order_purchase_DateKey	order_approved_DateKey	order_delivered_carrier_DateKey	order_delivered_customer_DateKey	order_estimated_d
16	110	167	2654	20180418	20180418	20180420	20180510	20180518
17	112	1226	63	20180604	20180604	20180605	20180606	20180628
18	122	839	3017	20170808	20170809	20170810	20170818	20170905
19	128	2535	228	20180203	20180206	20180207	20180207	20180221
20	131	35	783	20180329	20180329	20180404	20180508	20180425
21	139	832	1507	20180601	20180602	20180704	20180709	20180725

Query executed successfully. | LAPTOP-MBM4LPUK (15.0 RTM) | LAPTOP-MBM4LPUK\chame ... | SourceDB_DW | 00:00:01 | 132,509 rows

6. ETL Development

a) Extraction

First, I have extracted all my data from the tables which were in the SourceDB to separate staging tables as shown below.

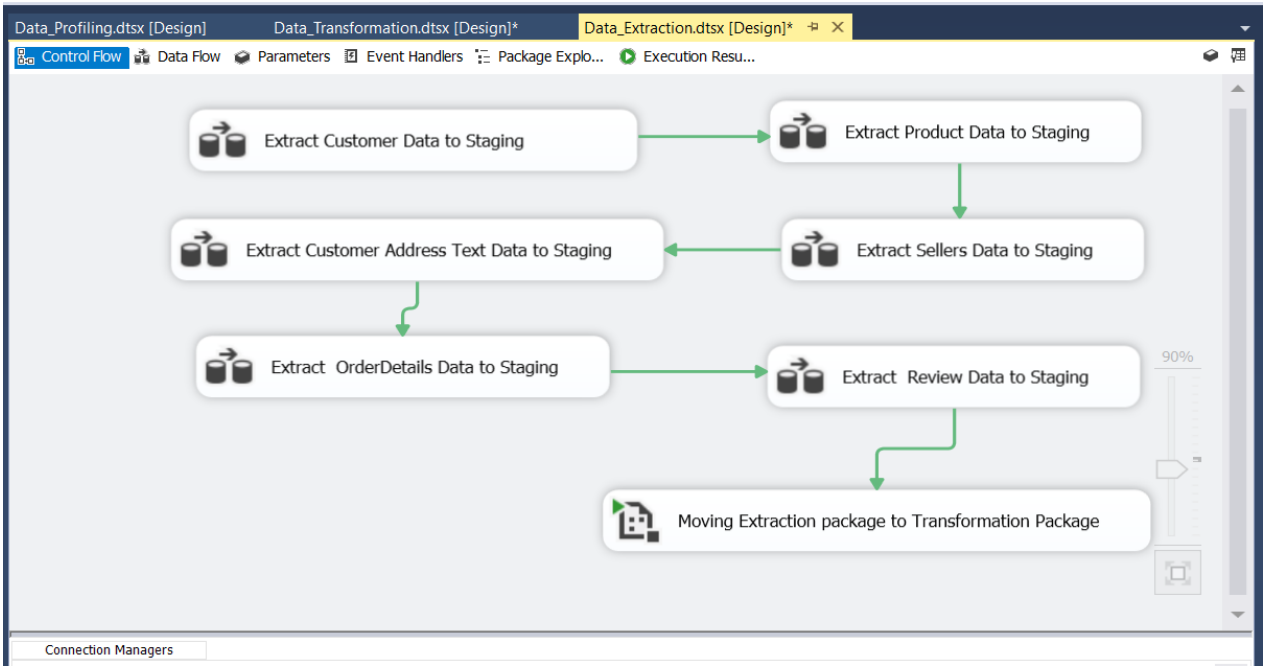


Figure 6.1

Following figure shows inside the Data flow of “Extract Customer Data to Staging”. I used OLE DB Source and select the table which I want to extract. In OLE DB Destination I created the staging table in SourseDB_DW data warehouse.

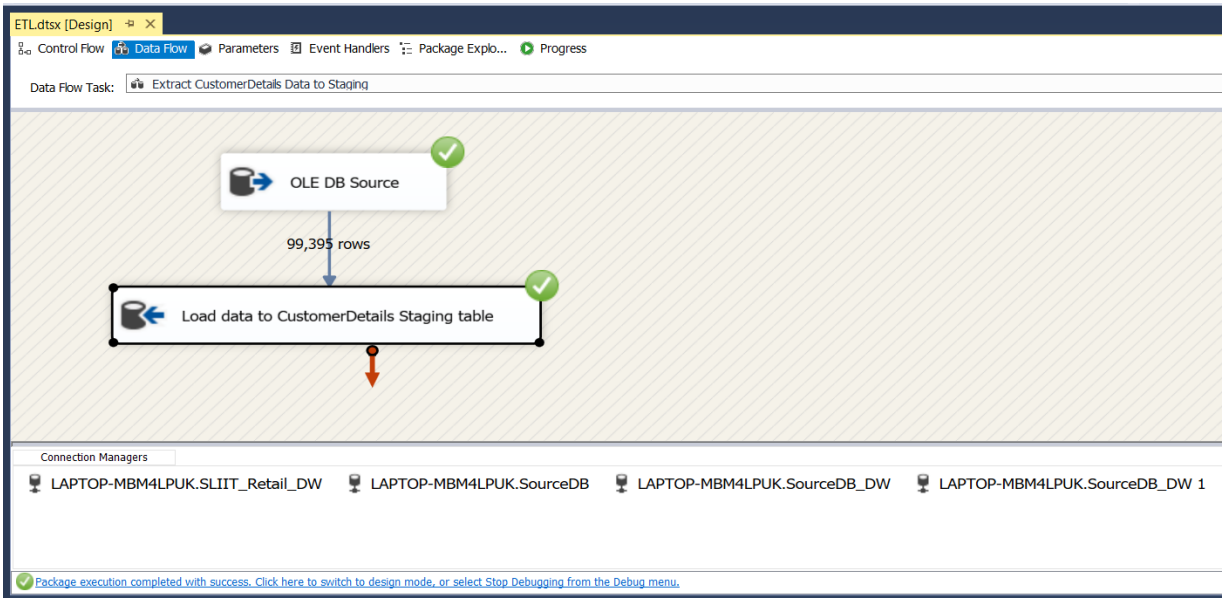


Figure 6.2

This figure shows inside “CustomerAddress Text Data to staging” dataflow. Here I used Flat File Source to extract my text file data and in OLE DB Destination I created the staging table in data warehouse.

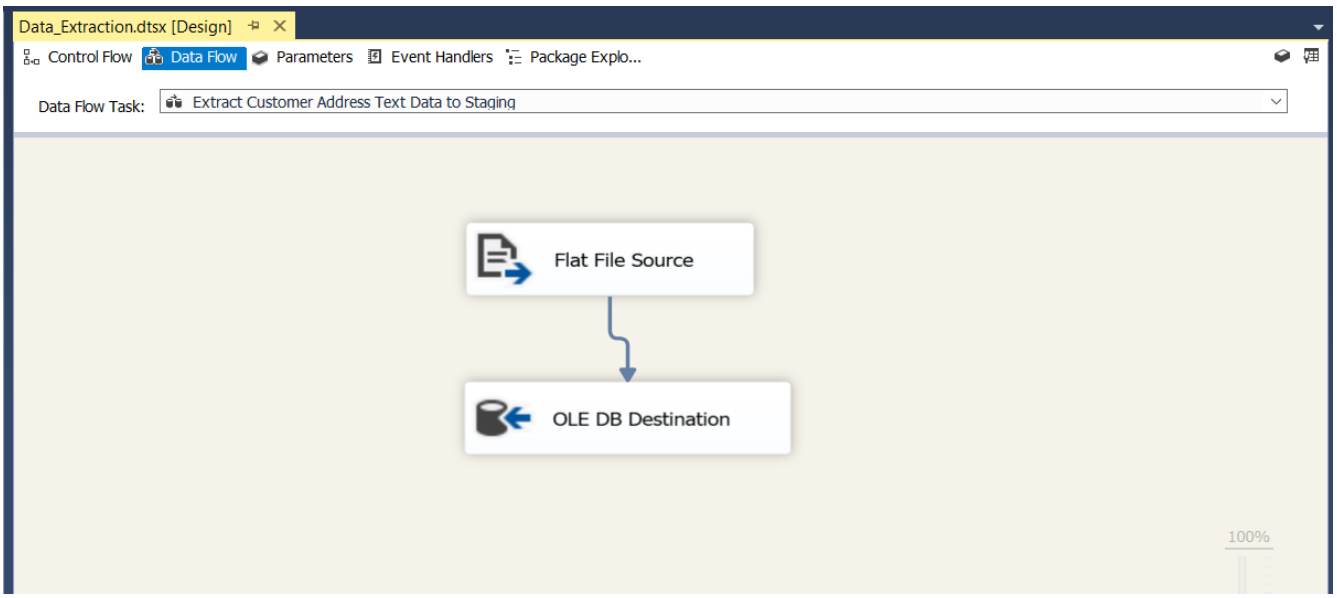


Figure 6.3

After creating the staging table in data warehouse, I used an Execute SQL Task Component to truncate each staging table which I created before.

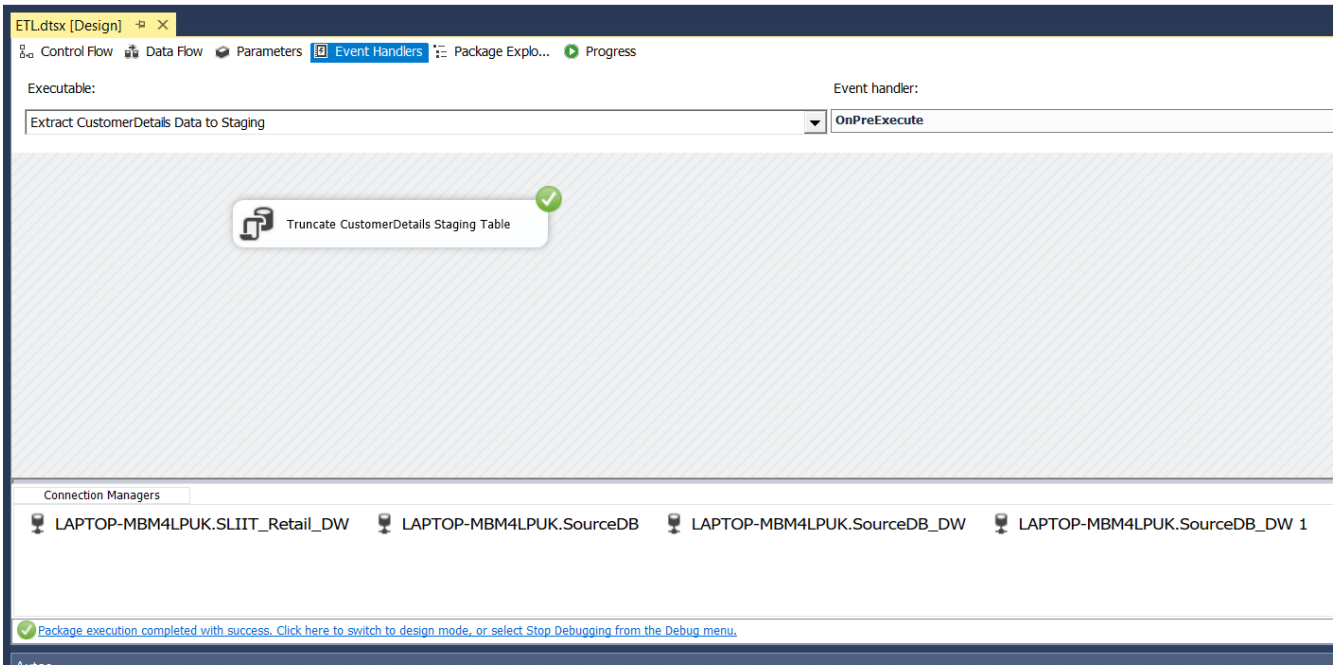


Figure 6.4

Above steps were followed to each table in SourceDB and extract those tables into a separate staging table.

b) Transform and Load

1. Transform load Customer Data

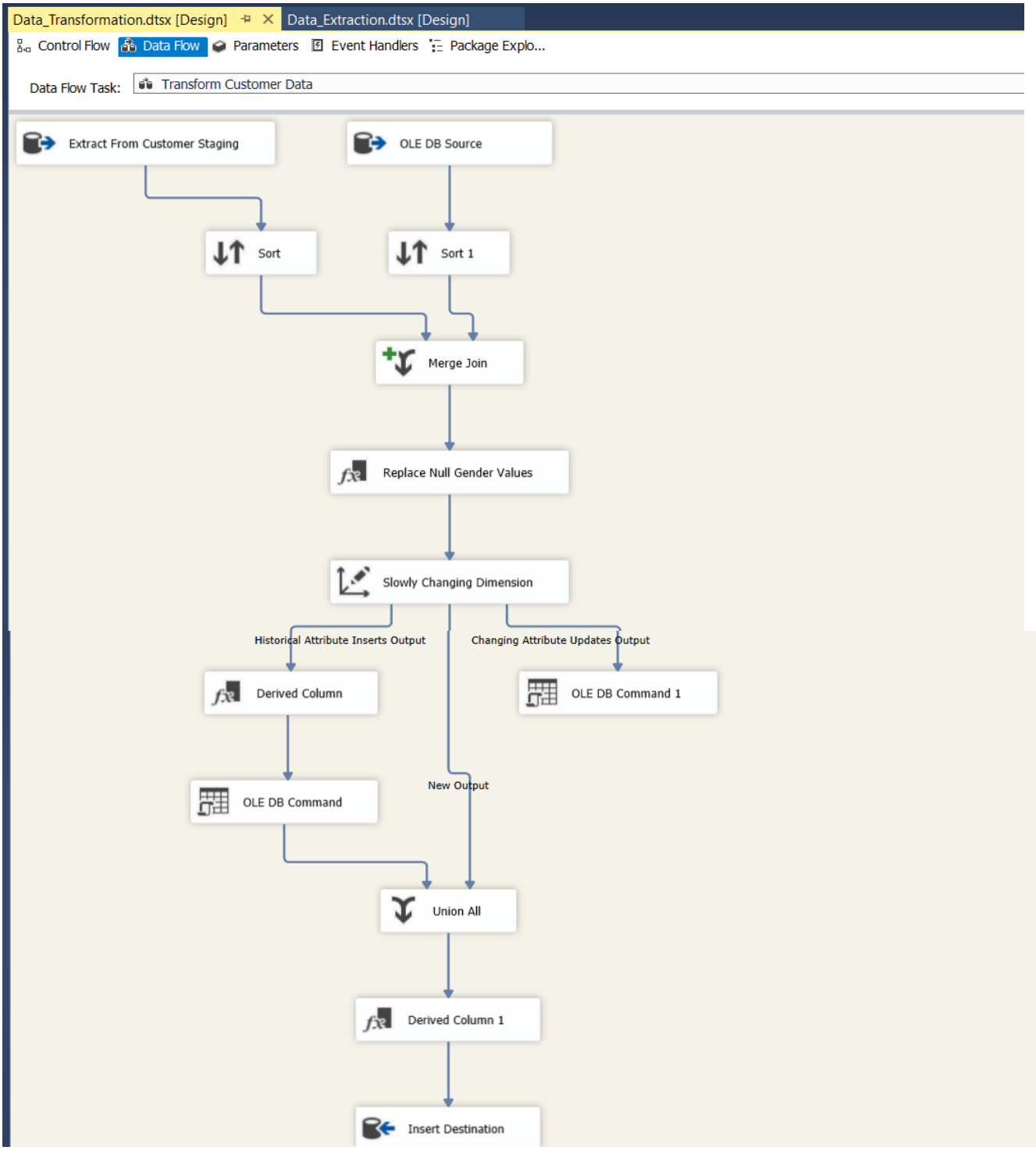


Figure 6.5

In this Transform and load Process I have to get CustomerStaging and CustomerAddressStaging tables to load the DimCustomer table.So I take two sort components using CustomerID and merge them(CustomerStaging table and CustomerAddressStaging table) and by using derived column component, I replace the null values in Gender column.

Then I added a slowly changing dimension component to update the DimCustomer table records.

So as shown in the figure 3.4 I have transformed and loaded data in to the DimSeller and DimProduct dimension tables.

2. Transform and Load Seller

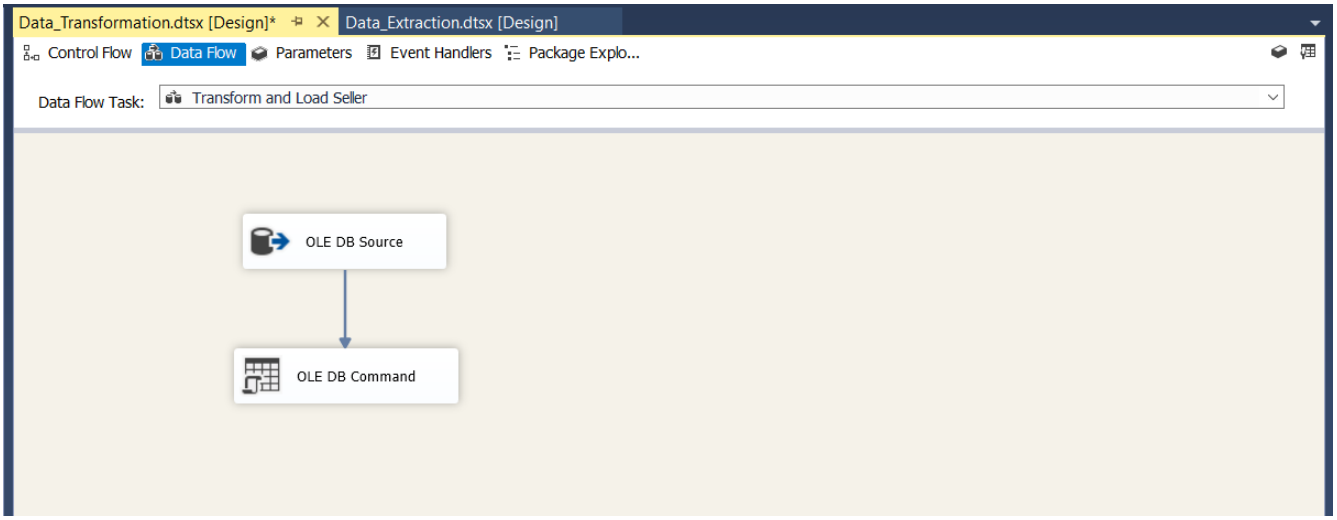


Figure 6.6

SQL query used to update the Seller details,

```
Create Procedure dbo.UpdateSellerDetails
@SellerID nvarchar(255),
@seller_zip_code_prefix int,
@seller_city nvarchar(50),
@seller_state nvarchar(5),
@LicenceCode int,
@LicenceDescription nvarchar(50)

as
Begin
if not exists (select SellerID
from dbo.Dim_Seller
where AlternateSellerID = @SellerID
and seller_zip_code_prefix= @seller_zip_code_prefix
and seller_city  = @seller_city
and seller_state = @seller_state
and LicenceCode = @LicenceCode
and LicenceDescription = @LicenceDescription  )
begin
insert into dbo.Dim_Seller
(AlternateSellerID ,seller_zip_code_prefix, seller_city,seller_state,LicenceCode,LicenceDescription)
values
(@SellerID, @seller_zip_code_prefix,
@seller_city,@seller_state,@LicenceCode,@LicenceDescription)
end;
End;
```

3. Transform and Load ProductData

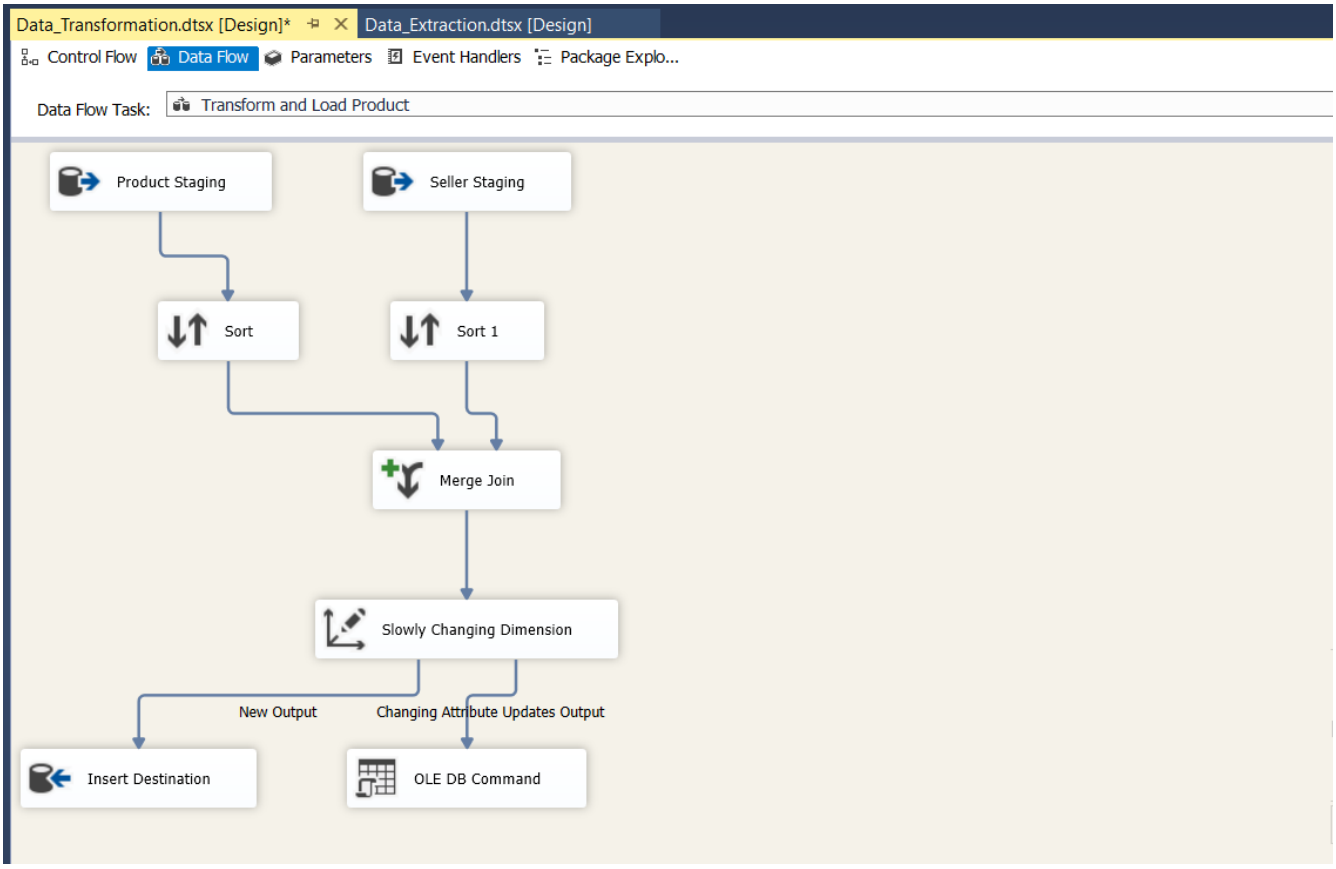


Figure 6.7

4. Load the Fact table

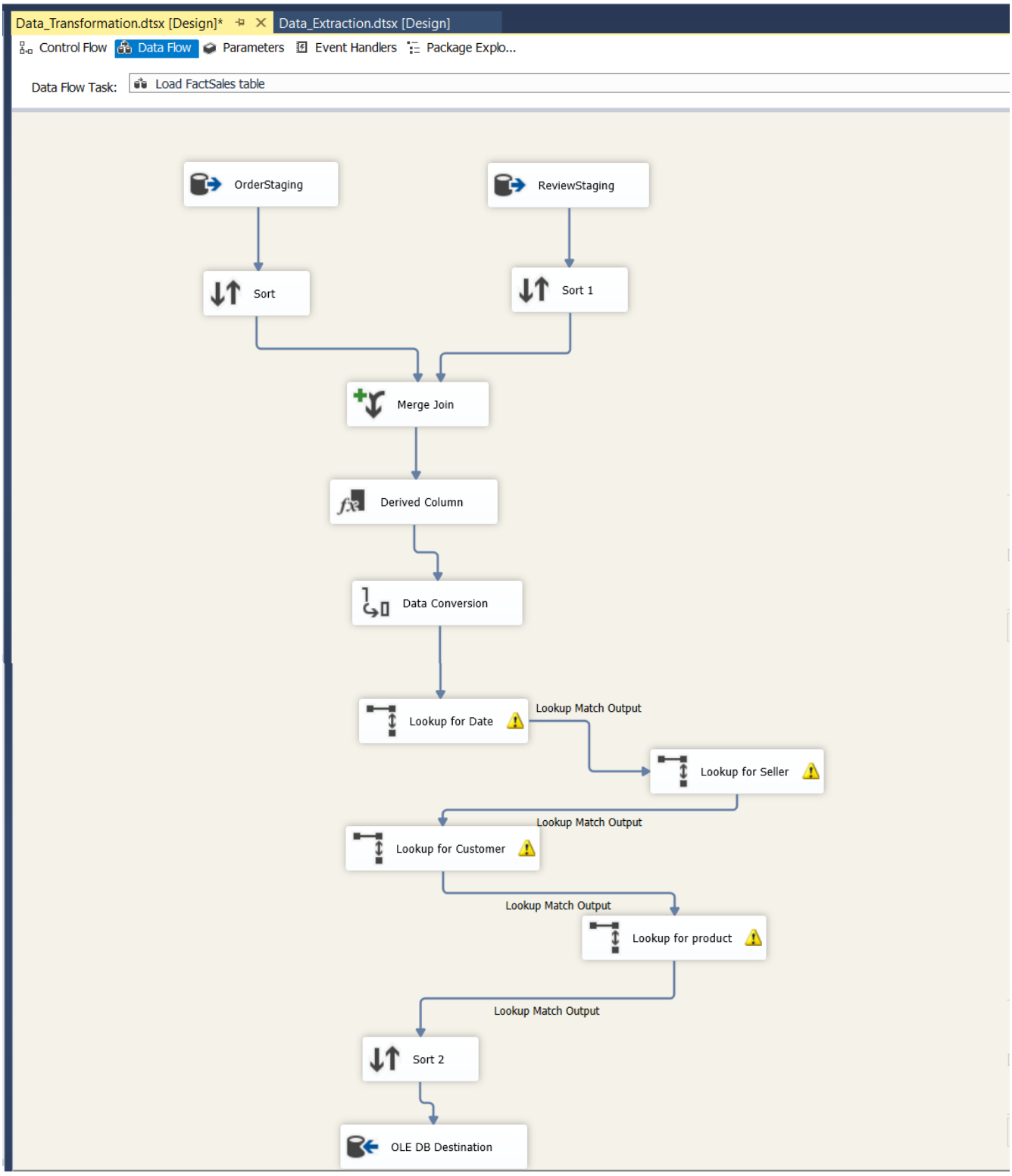
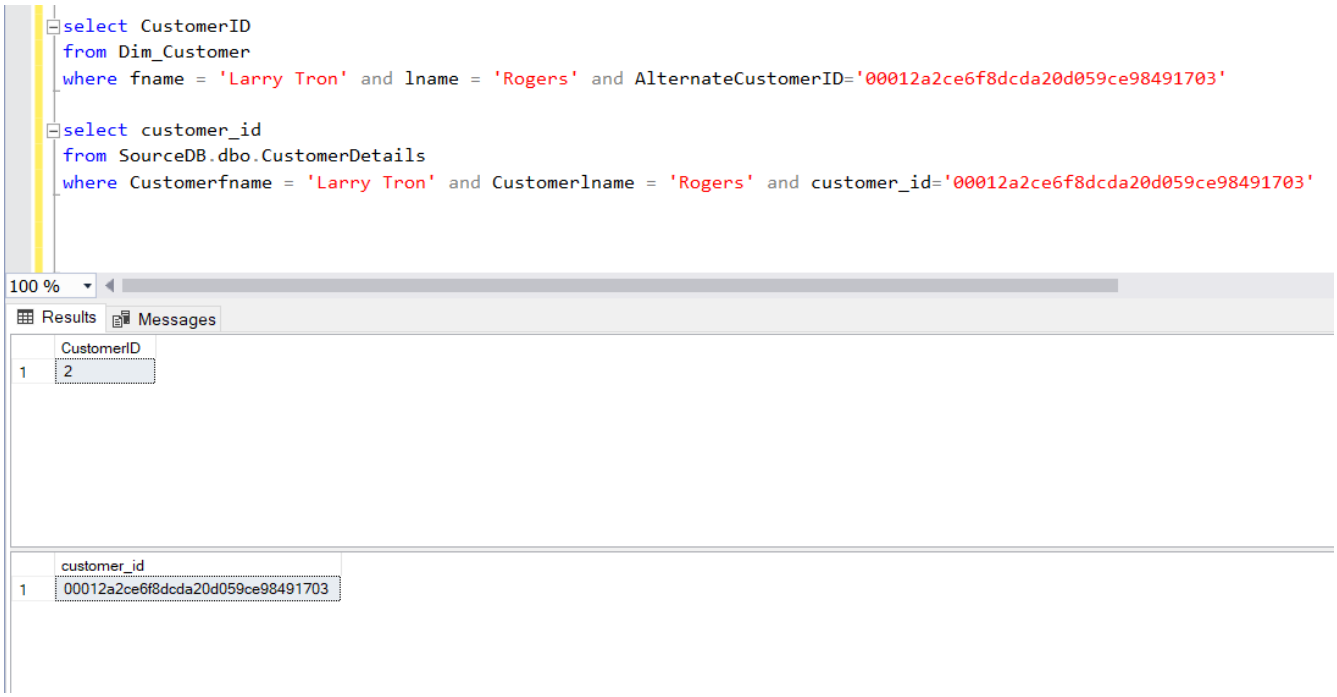


Figure 6.8

7. Execution of Test Cases and create a Test Summary Report

1.Check Duplicate values in Source file and dimension table



Note – In DimProduct table there is no duplicate values.

Figure 7.1

2. Record counts validation checking

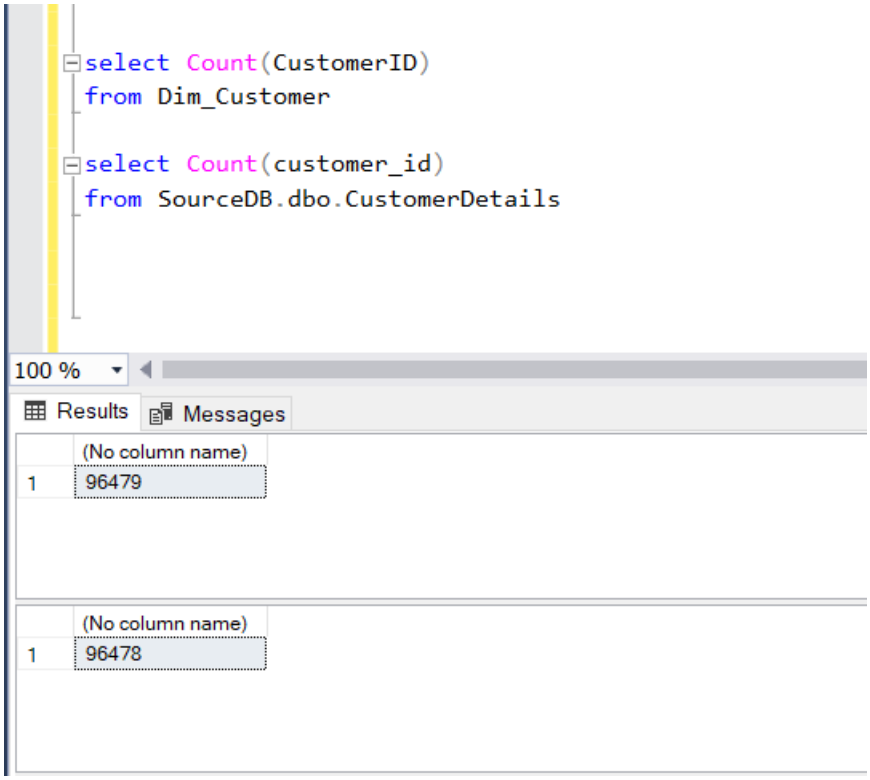


Figure 7.2

Product table after removing duplicate values (Dim_Product table has no duplicate values)

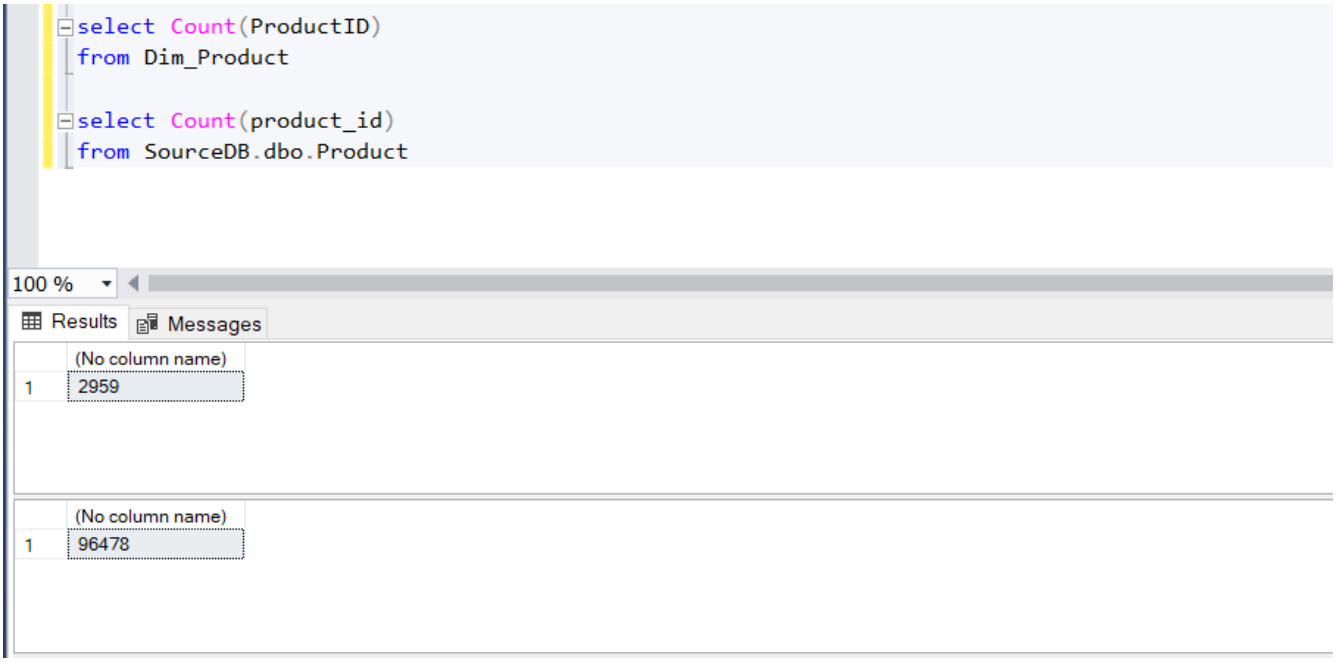


Figure 7.3

3. Data type checking

```
SELECT *
FROM INFORMATION_SCHEMA.COLUMNS
WHERE table_name = 'CustomerDetails'
```

100 %

Results | Messages

	TABLE_CATALOG	TABLE_SCHEMA	TABLE_NAME	COLUMN_NAME	ORDINAL_POSITION	COLUMN_DEFAULT	IS_NULLABLE	DATA_TYPE	CHARACTER_MAXIMUM_LENG
1	SourceDB	dbo	CustomerDetails	customer_id	1	NULL	YES	nvarchar	255
2	SourceDB	dbo	CustomerDetails	customer_unique_id	2	NULL	YES	nvarchar	255
3	SourceDB	dbo	CustomerDetails	Customerfname	3	NULL	YES	nvarchar	255
4	SourceDB	dbo	CustomerDetails	Customerlname	4	NULL	YES	nvarchar	255
5	SourceDB	dbo	CustomerDetails	Gender	5	NULL	YES	nvarchar	255
6	SourceDB	dbo	CustomerDetails	PhonrNumber	6	NULL	YES	nvarchar	255
7	SourceDB	dbo	CustomerDetails	Email	7	NULL	YES	nvarchar	255

```
SELECT *
FROM INFORMATION_SCHEMA.COLUMNS
WHERE table_name = 'Dim_Customer'
```

100 %

Results | Messages

	TABLE_CATALOG	TABLE_SCHEMA	TABLE_NAME	COLUMN_NAME	ORDINAL_POSITION	COLUMN_DEFAULT	IS_NULLABLE	DATA_TYPE	CHARACTER_MAXIMUM
1	SourceDB_DW	dbo	Dim_Customer	CustomerID	1	NULL	NO	int	NULL
2	SourceDB_DW	dbo	Dim_Customer	AlternateCustomerID	2	NULL	YES	nvarchar	255
3	SourceDB_DW	dbo	Dim_Customer	CustomerUniqueID	3	NULL	YES	nvarchar	255
4	SourceDB_DW	dbo	Dim_Customer	fname	4	NULL	YES	nvarchar	255
5	SourceDB_DW	dbo	Dim_Customer	lname	5	NULL	YES	nvarchar	255
6	SourceDB_DW	dbo	Dim_Customer	Gender	6	NULL	YES	nvarchar	255
7	SourceDB_DW	dbo	Dim_Customer	PhoneNumber	7	NULL	YES	nvarchar	255
8	SourceDB_DW	dbo	Dim_Customer	Email	8	NULL	YES	nvarchar	255
9	SourceDB_DW	dbo	Dim_Customer	customer_zip_code_prefix	9	NULL	YES	nvarchar	25
10	SourceDB_DW	dbo	Dim_Customer	customer_city	10	NULL	YES	nvarchar	25
11	SourceDB_DW	dbo	Dim_Customer	customer_state	11	NULL	YES	nvarchar	5
12	SourceDB_DW	dbo	Dim_Customer	Address	12	NULL	YES	nvarchar	100
13	SourceDB_DW	dbo	Dim_Customer	StartDate	13	NULL	YES	datetime	NULL
14	SourceDB_DW	dbo	Dim_Customer	EndDate	14	NULL	YES	datetime	NULL

Figure 7.4

4. Index check

```
select * from sys.indexes
where object_id = (select object_id from sys.objects where name = 'Dim_Product')

select * from sys.indexes
where object_id = (select object_id from sys.objects where name = 'Dim_Customer')

select * from sys.indexes
where object_id = (select object_id from sys.objects where name = 'Dim_Seller')
```

100 %

Results | Messages

	object_id	name	index_id	type	type_desc	is_unique	data_space_id	ignore_dup_key	is_primary_key	is_unique_constraint	fill_factor	is_padded	is_disable
1	1669580986	PK_Dim_Product	1	1	CLUSTERED	1	1	0	1	0	0	0	0

	object_id	name	index_id	type	type_desc	is_unique	data_space_id	ignore_dup_key	is_primary_key	is_unique_constraint	fill_factor	is_padded	is_disab
1	1637580872	PK_Dim_Customer	1	1	CLUSTERED	1	1	0	1	0	0	0	0

	object_id	name	index_id	type	type_desc	is_unique	data_space_id	ignore_dup_key	is_primary_key	is_unique_constraint	fill_factor	is_padded	is_disabled
1	949578421	PK_Dim_Seller	1	1	CLUSTERED	1	1	0	1	0	0	0	0

Query executed successfully. | LAPTOP-MBM4LPUK (15.0 RTM) | LAPTOP-MBM4LPUK\chame ... | SourceDB_DW | 00:00:00 | 3 rows

Figure 7.5

Test Summary Report

Test case ID	Test case Title	Expected Output	Actual Output	Status
01	Check Duplicate values in Source file and dimension table	No duplicates in Dimension	No duplicates in Dimension	Successful
02	Record counts validation checking			
02	Data type checking	Source and Dimension have the same data type	Source and Dimension have the same data type	Successful
04	Index check	Index created with required columns.	Index created with required columns.	Successful

End