Sri Lanka Institute of Information Technology

Data Warehousing & Business Intelligence

Assignment 1

IT Number: IT20142964

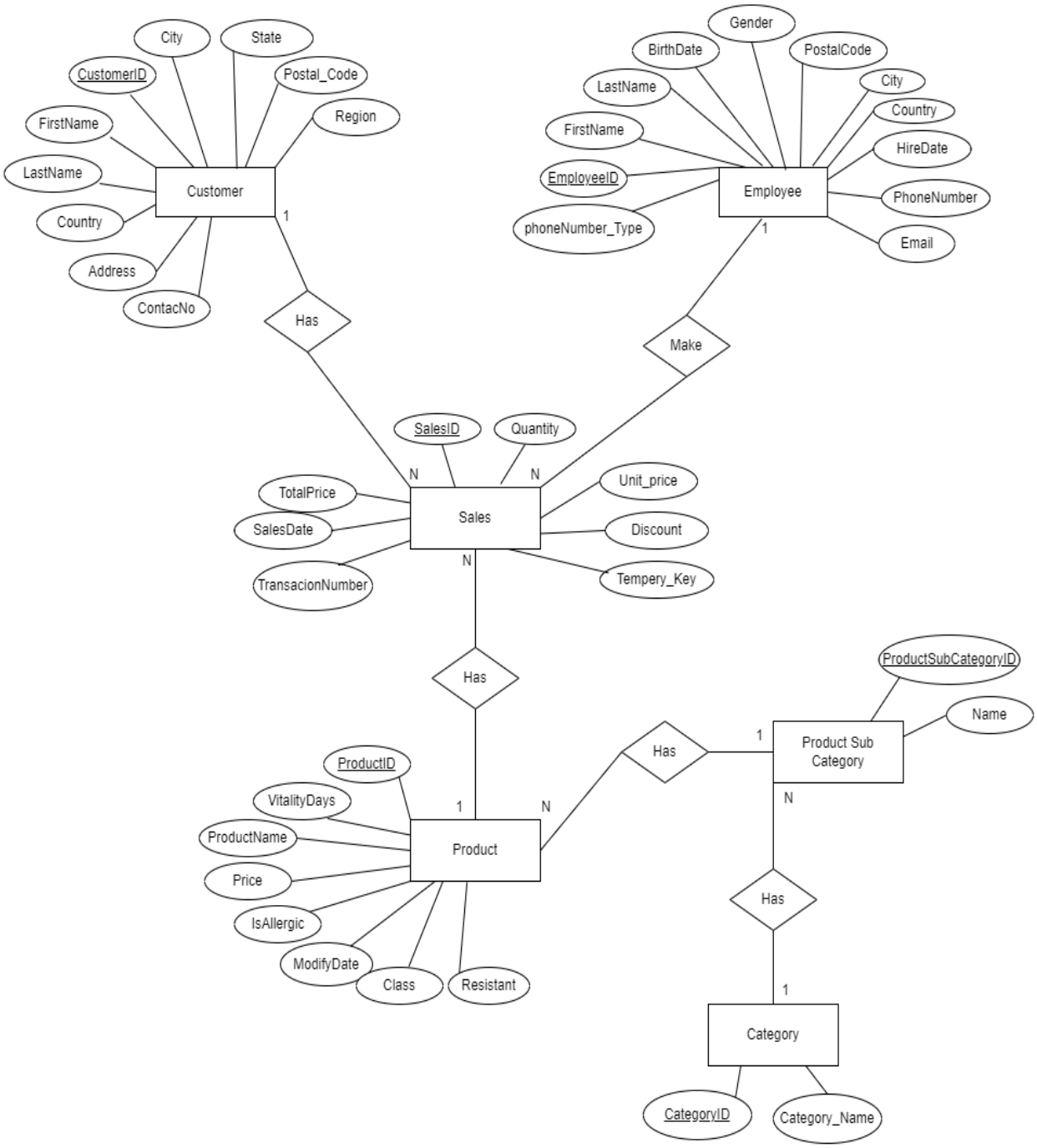
Submitted by: Dulakshi Hansini R.M

Step 01 – Dataset Selection

This data set is about Superstore grocery sales data which is selected from Kaggle. This is a collection of daily sales in superstore grocery store. I separate columns in the original data set and put them into different source tables to get more dimensions and hierarchies. Because of that it enriches the ETL process.

Customized source has 6 tables. It includes Employee's details, Customer's details, all the products details, product sub categories' details and categories' details in the store.

ER Diagram



Step 02 – Preparation of Data Sources

There are 2 main sources in my data set.

1. Grocery superstore sales_sourceDB

    - **Customers :** Details about customer's primary details. (CSV file)
    - **Sales :** All the details about the sales in the store. (CSV file)
    - **Products :** All the details about products(CSV file)
    - **Categories :** All the details about the category details. (CSV file)
    - **Products Sub Categories :** All the product sub category details. (CSV file)

2. Employee.txt
    - **Employee :** Details about Employee's details. (TEXT file)

        This text file include all the customer address details including street,city,country,Postcode and the country

Step 03 – Solution Architecture

In this solution, there are 3 main components up to the ETL process in the process of Datawarehouse design.

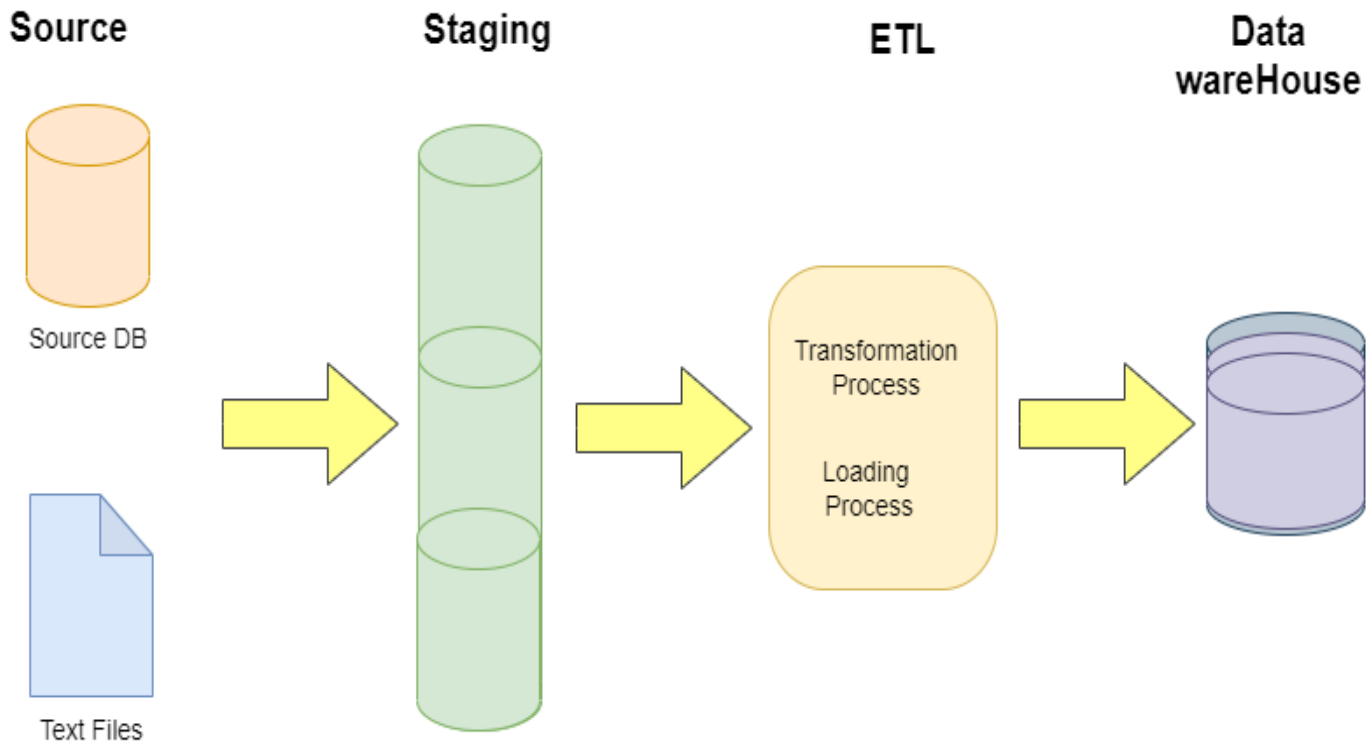
- ❖ Source Files
- ❖ Staging DB
- ❖ Data warehouse

Source Files : There are 6 source tables in two formats (CSV & text) as the data sources.

Staging DB : Extracted the sources files information and change the data types of the attributes that are in source file to relevant data type and loaded that information into the staging database.

Datawarehouse : After the staging, the information in staging database will become the sources to the transformation process. In this process data will transformed & loaded into the tables in Datawarehouse database.

Architectural diagram



- GrocerySuperstoreSale_StagingDB
  - Database Diagrams
  - Tables
    - System Tables
    - FileTables
    - External Tables
    - Graph Tables
    - dbo.Category_staging
    - dbo.CompleteTime_staging
    - dbo.Customer_staging
    - dbo.Employee_staging
    - dbo.Product_Staging
    - dbo.ProductSubCategories_staging
    - dbo.Sales_staging

- GrocerySuperstoreSale_DW
  - Database Diagrams
  - Tables
    - System Tables
    - FileTables
    - External Tables
    - Graph Tables
    - dbo.DimCategory
    - dbo.DimCustomer
    - dbo.DimDate
    - dbo.DimEmployee
    - dbo.DimProduct
    - dbo.DimProductSubCategory
    - dbo.FactSales

Architecture Components.

- Data Sources.

  Operational System (Sales).

  External Sources.

- Extract, Transform and Load.

  Extract – reading data from source systems.

  Transform – Combine data from multiple sources, De-duplicating.

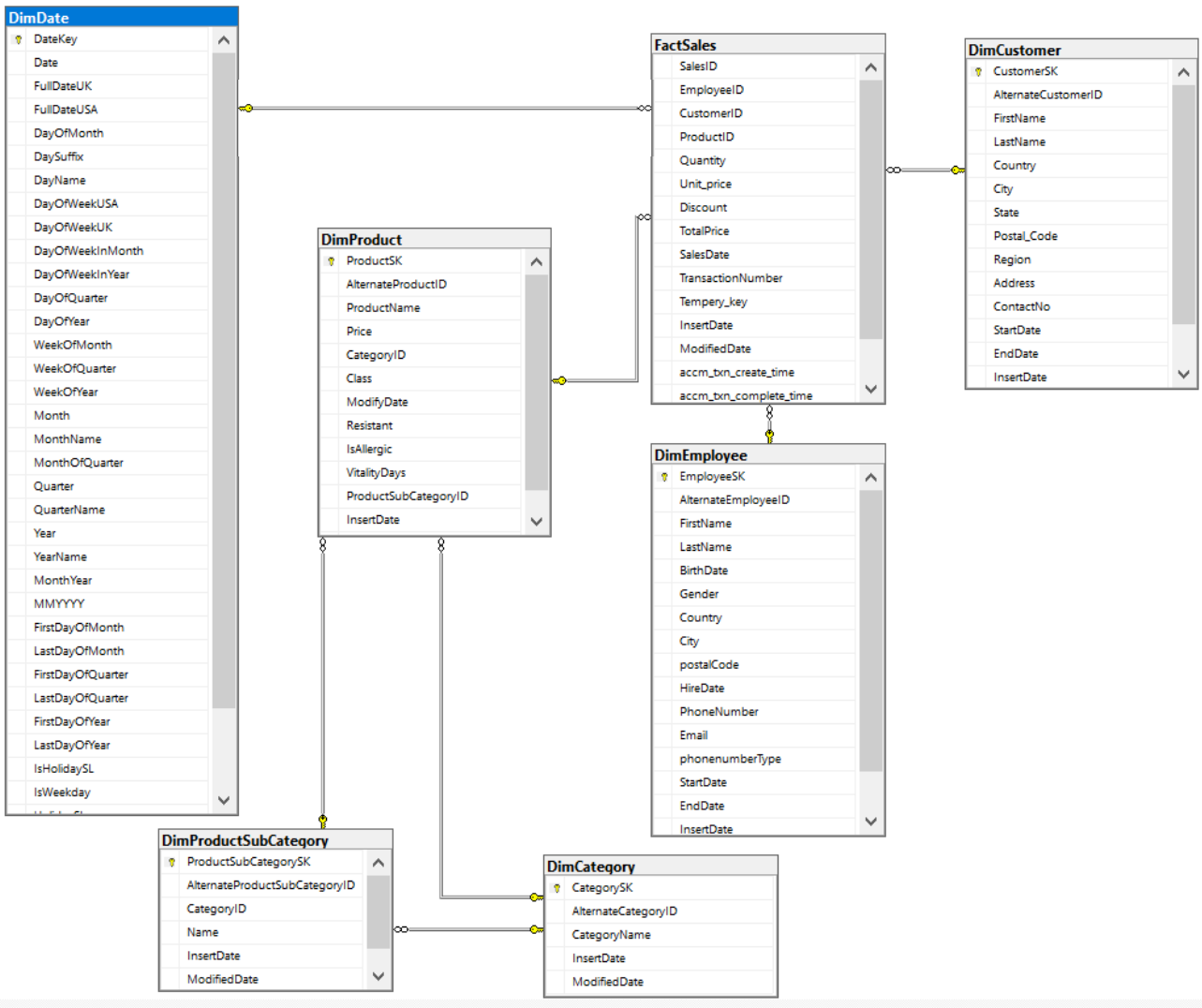
- Data Warehouse EDW and Data Mart.

  Dimensional Modeling- Facts and Dimensions.

Step 04 – Datawarehouse Design & Development

- ❖ My DW design is represented as a Snowflake.
- ❖ Customer and Employee dimensions are slowly changing dimensions.

Schema Relational Diagram(snowflakes schema)



DimCustomer and DimEmployee is slowly changing dimention.ContactNo and PhoneNumber may be changed in future respectively.Therefore, I get it as slowly changing attribute for that.

Product->ProductSubCategory->Category =This is the Hierarchies.

Step 05 – ETL Development

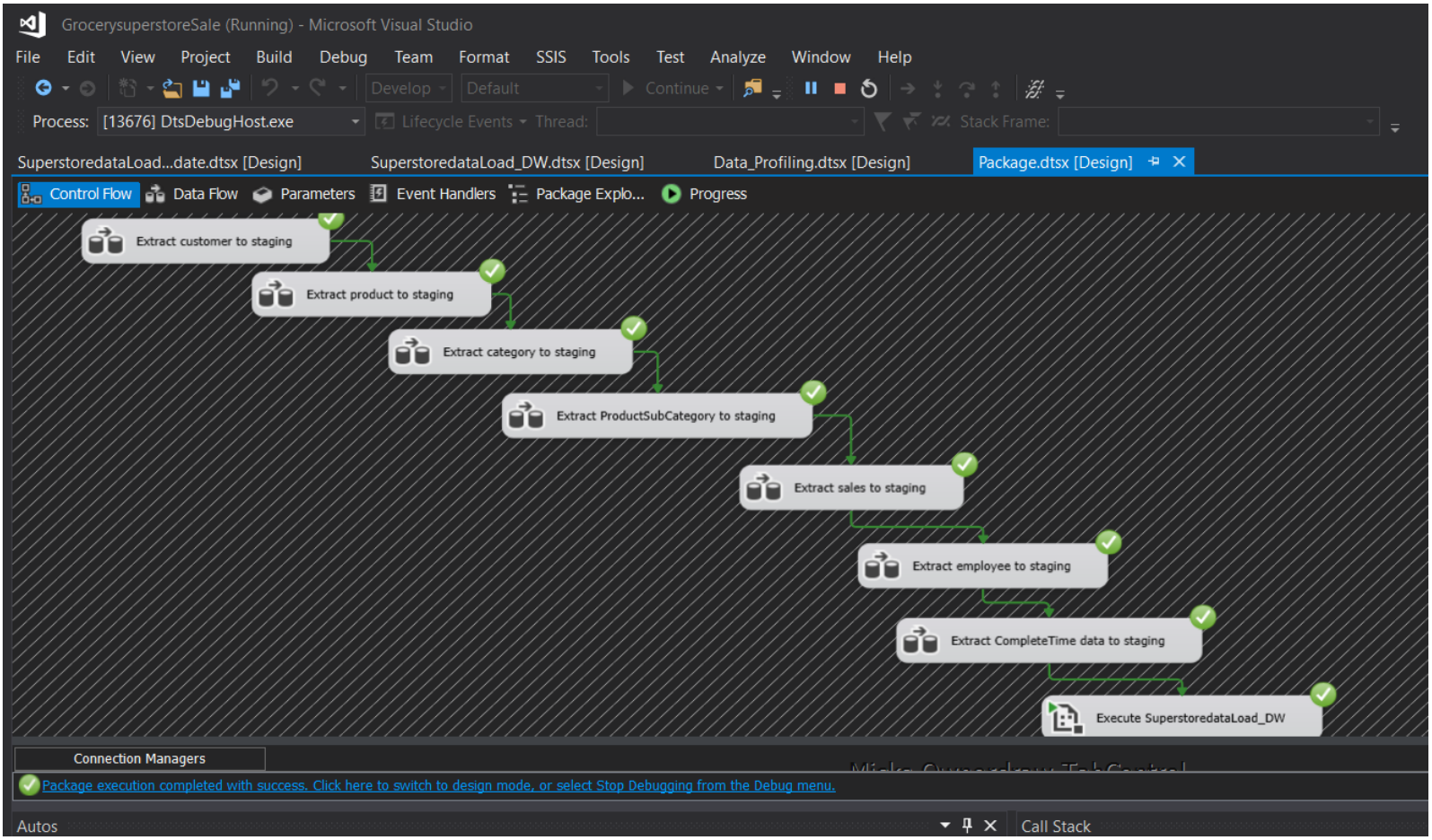
Step 01 : Restored the Grocery superstore sales_SourceDB in Microsoft SQL Server.

Step 02 : Created the Grocery superstore sales _staging DB in the Microsoft SQL Server.

Step 03 : Created the SSIS solution, called SSIS_DWBI_Assignment01.

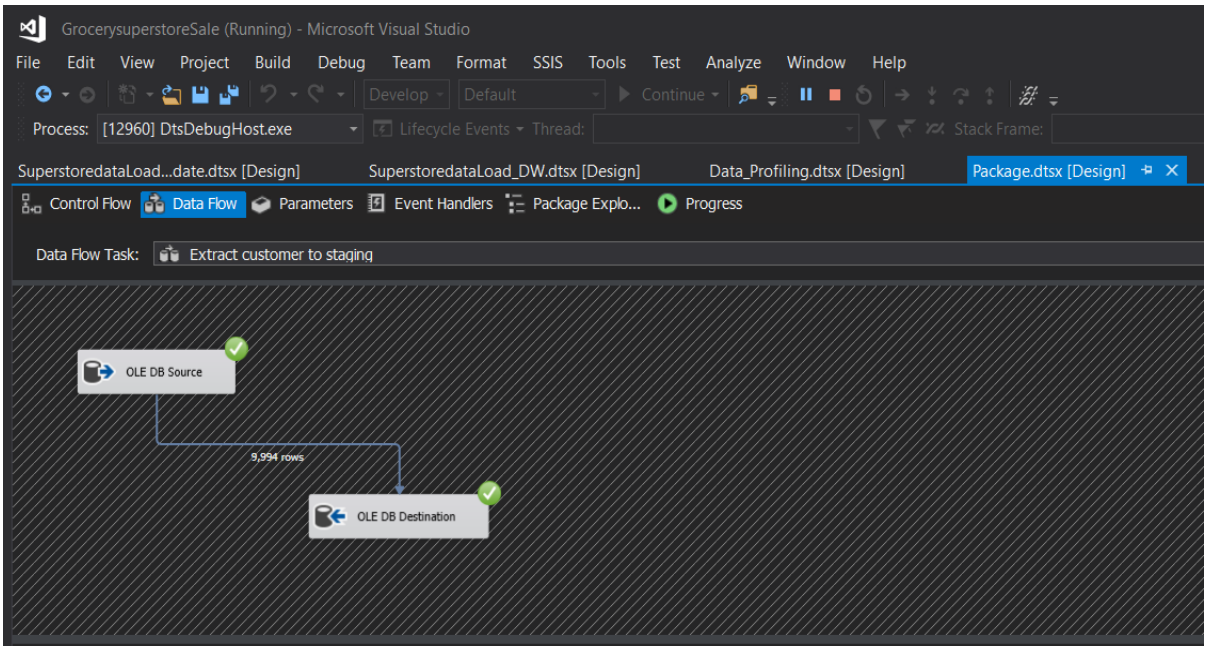
- Created the package called Grocery superstore sales _ETL and extracted data from Grocery superstore sales _SourceDB tables to Grocery superstore sales _staging_DB tables.
- Used OLE DB Sources and OLE DB Destinations to extract data.
- Used Flat File Source and an OLE DB Destination for extract data in Employee.txt

ETL – Source to Development

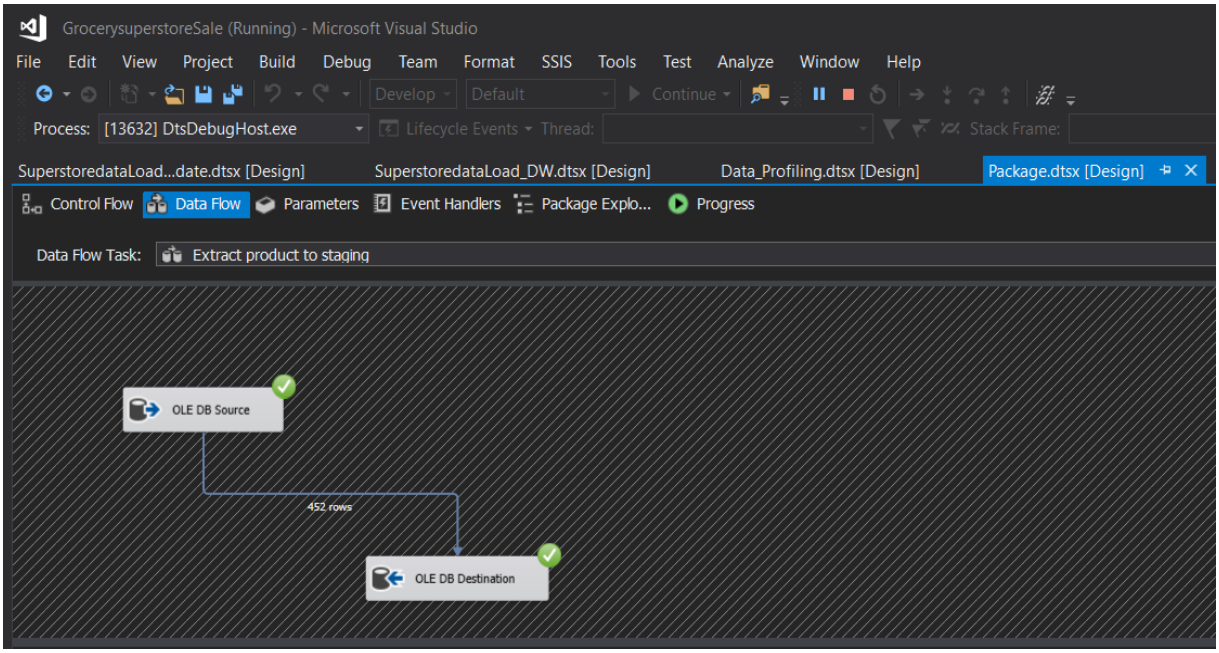


Screenshots of all the data sources that were staged and truncate tables created are attached below

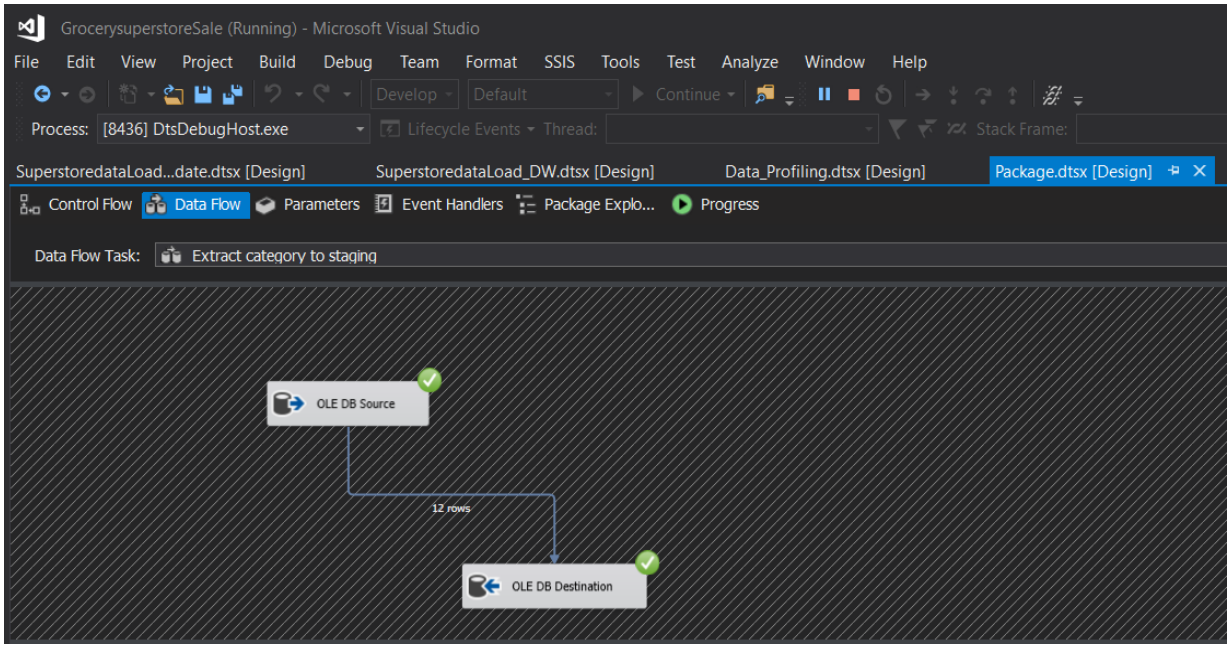
Load data Customer Staging



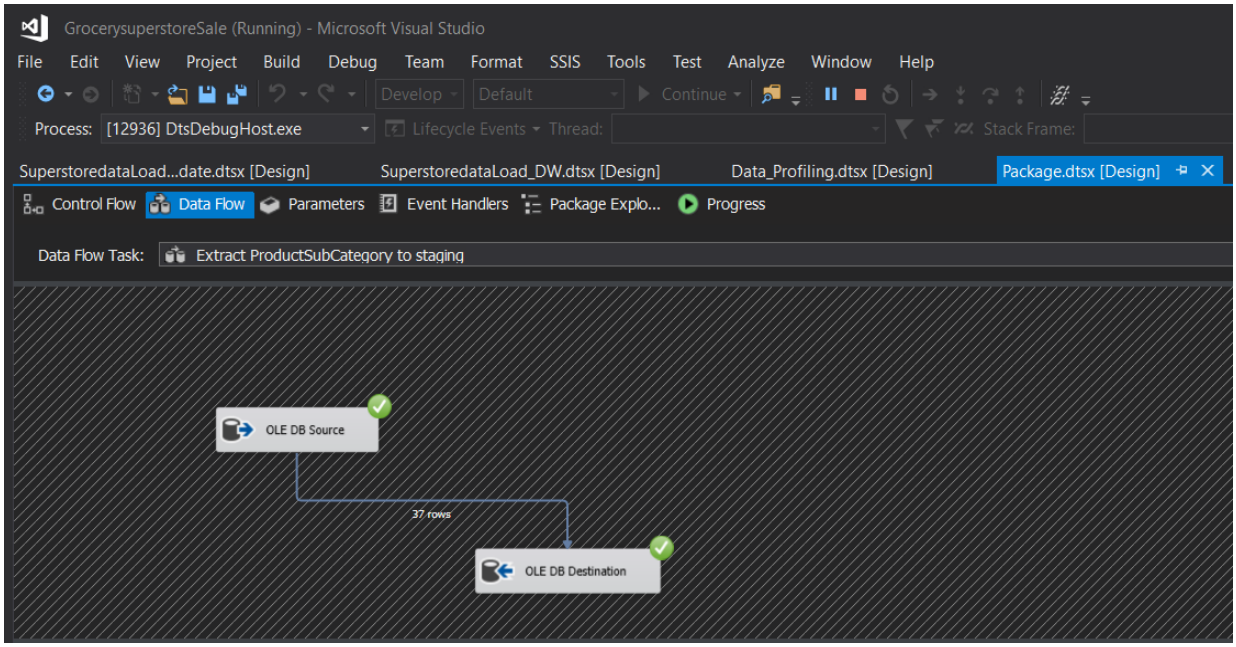
Load data Product Staging



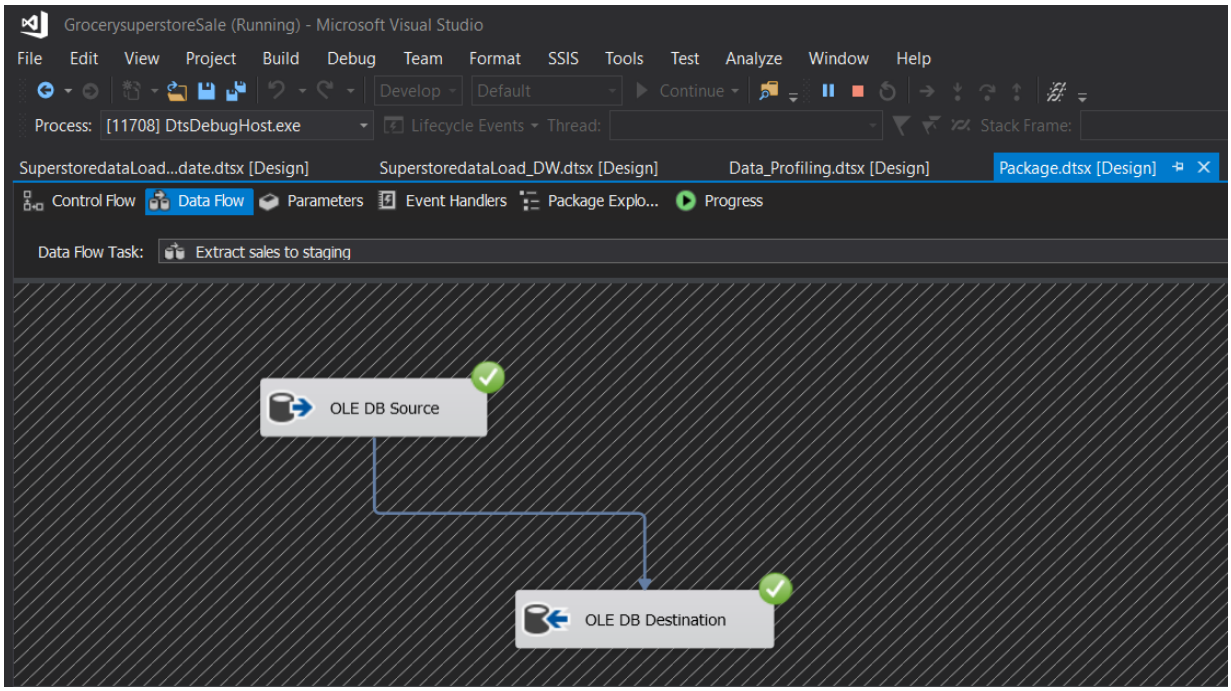
Load data Category Staging



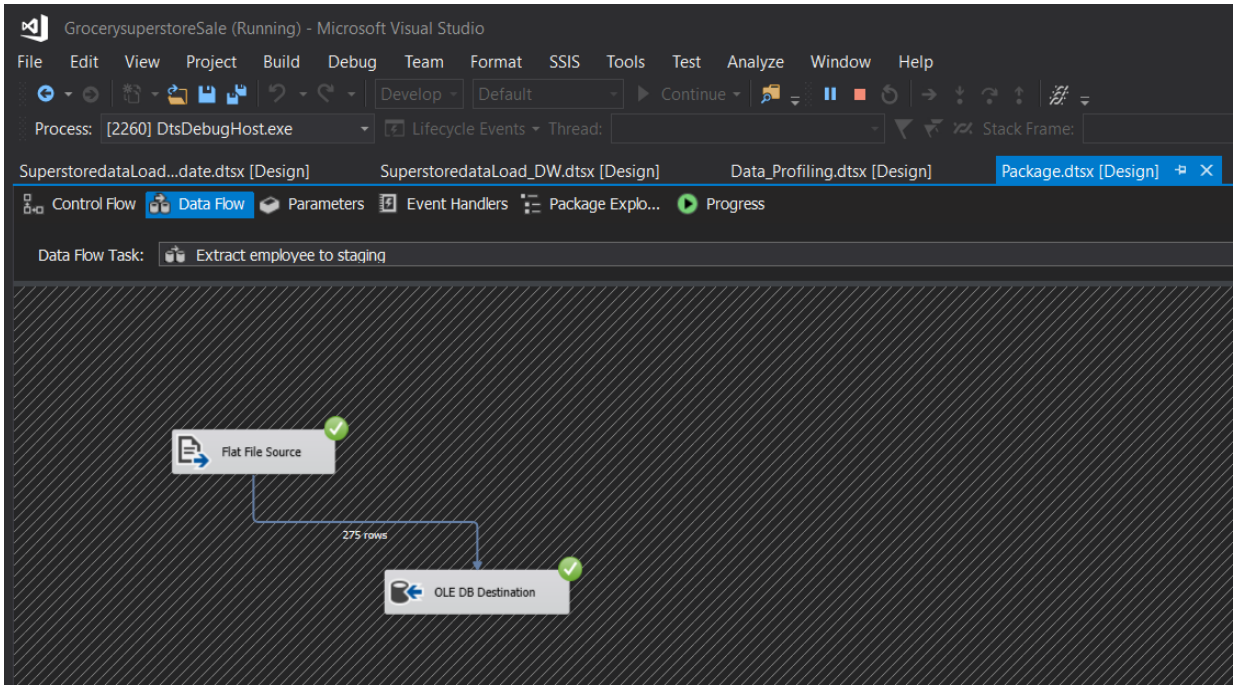
Load data ProductSubCategory Staging



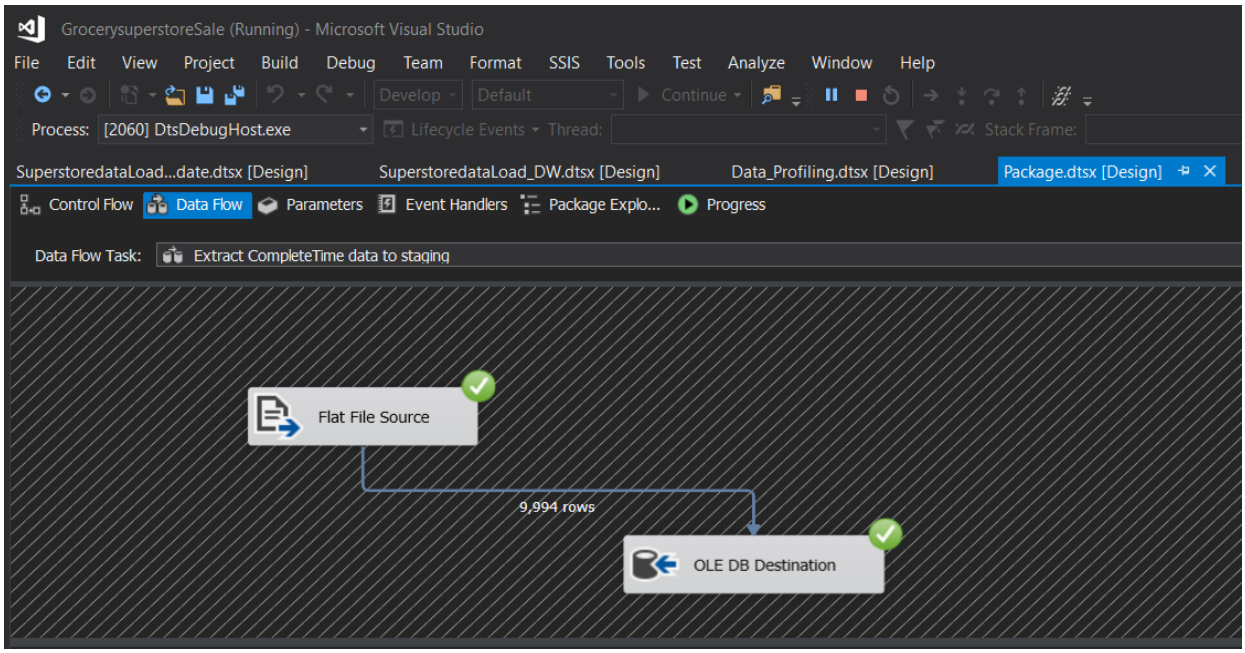
Load data Sales Staging



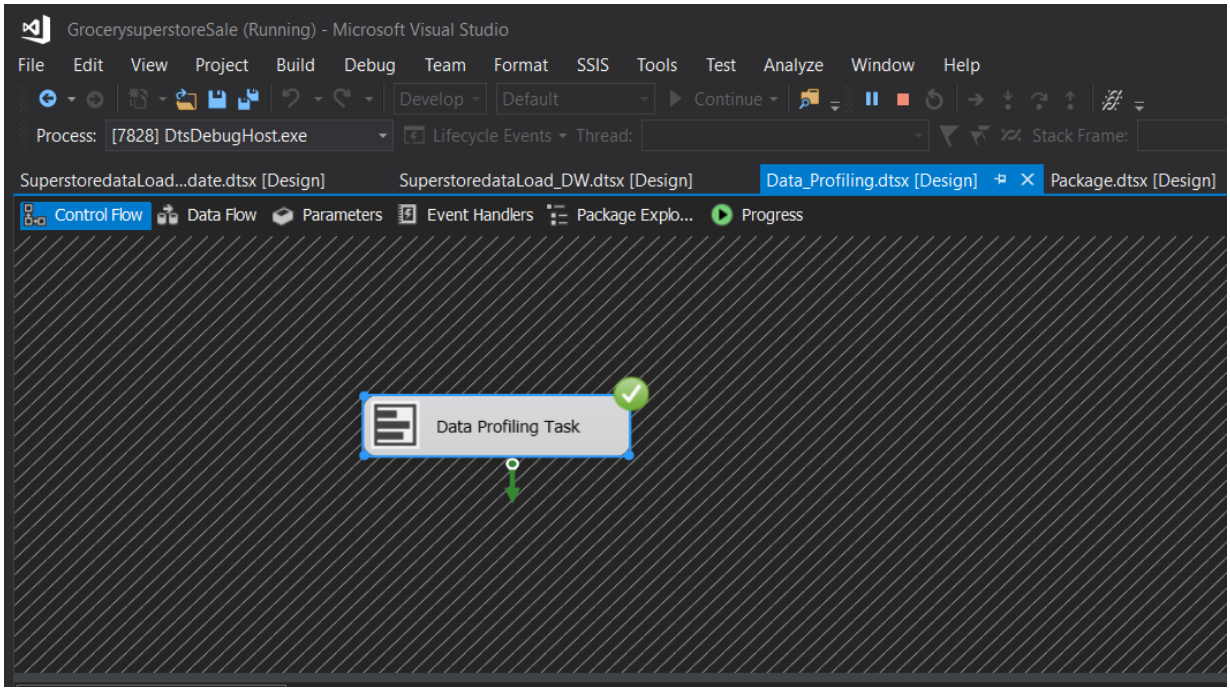
Load data Employee Staging

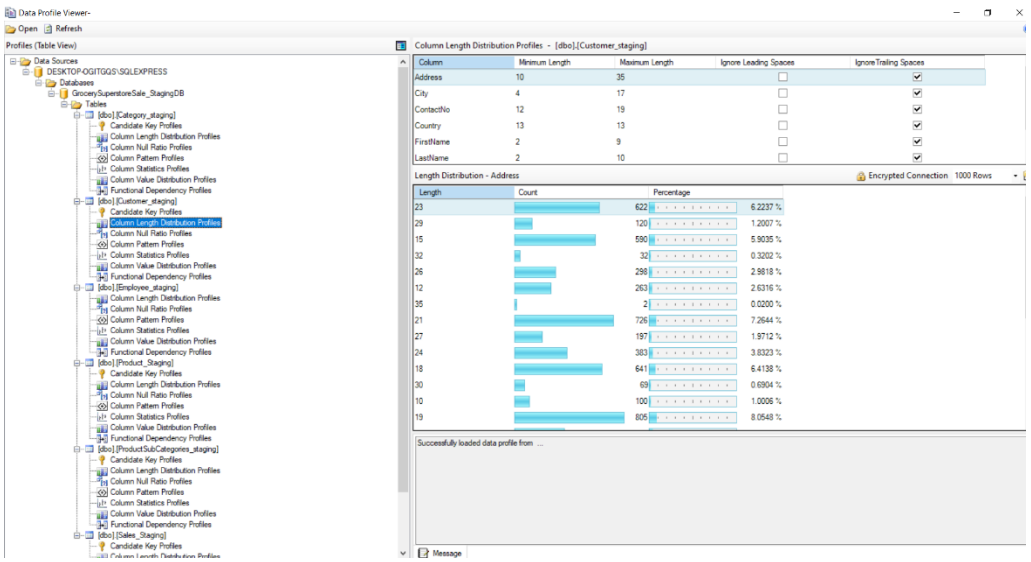


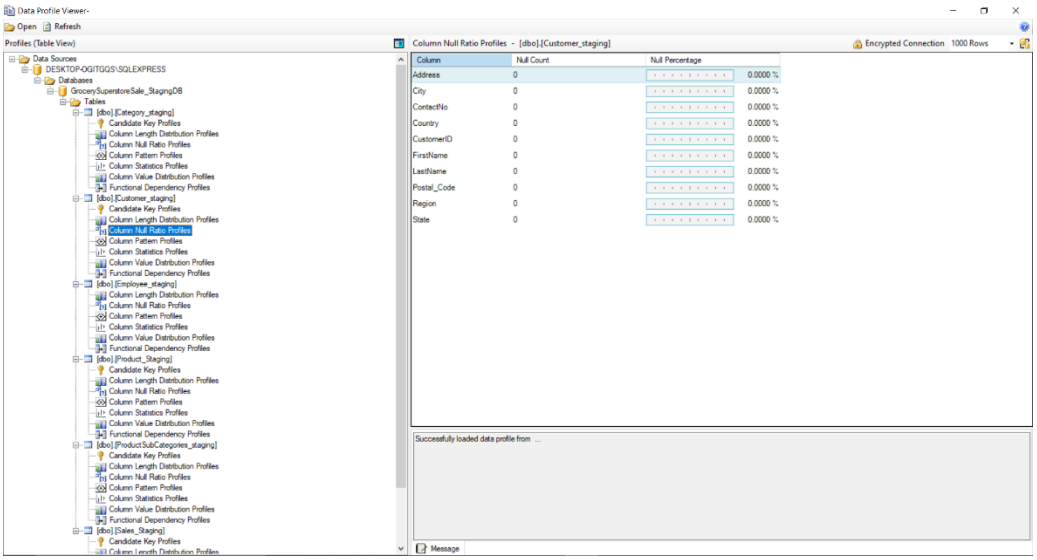
Load data CompleteTime Staging

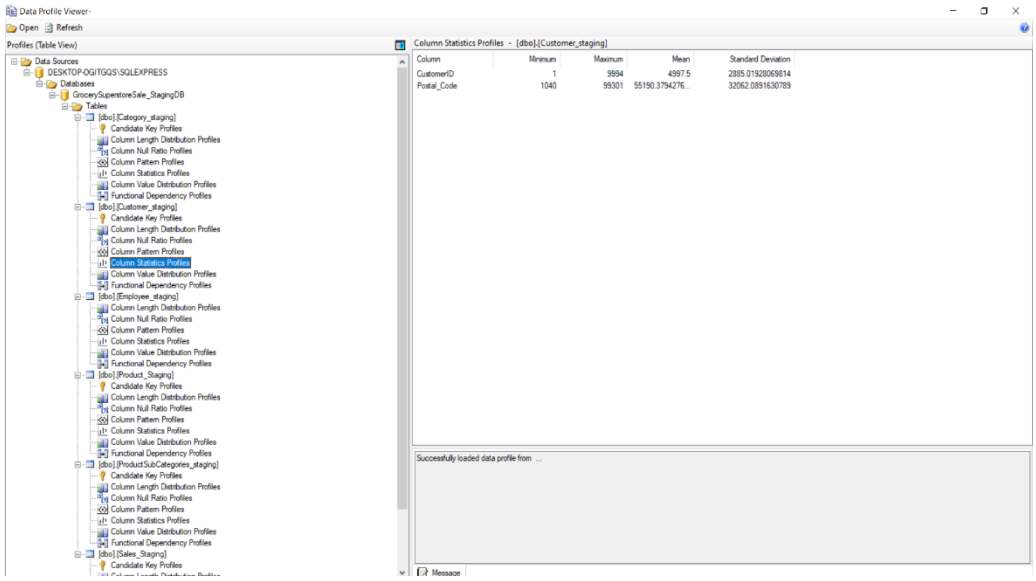


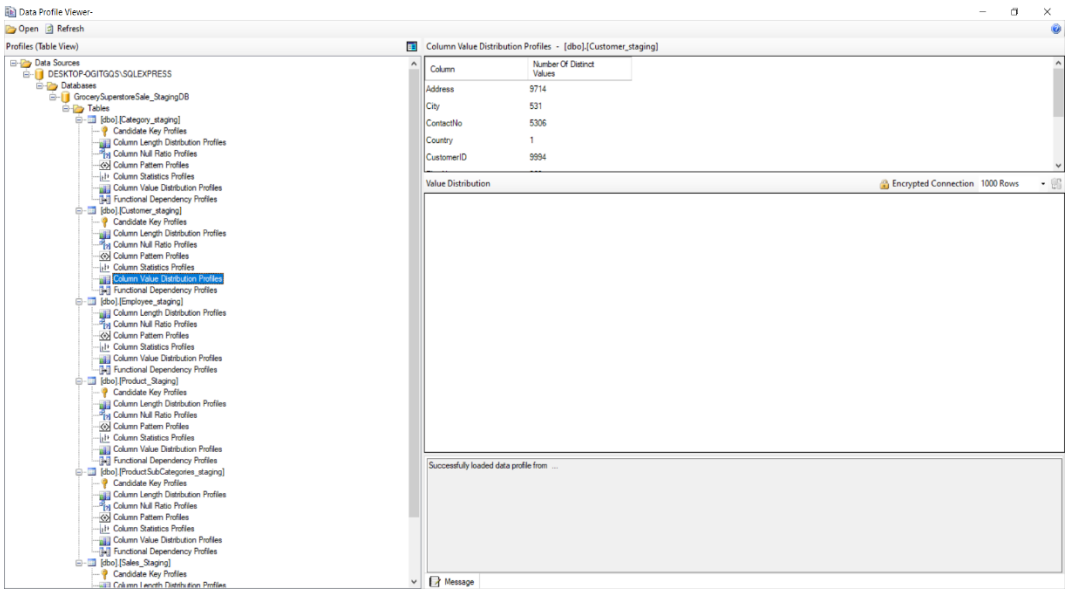
Data Profiling-Customer Data Profiling

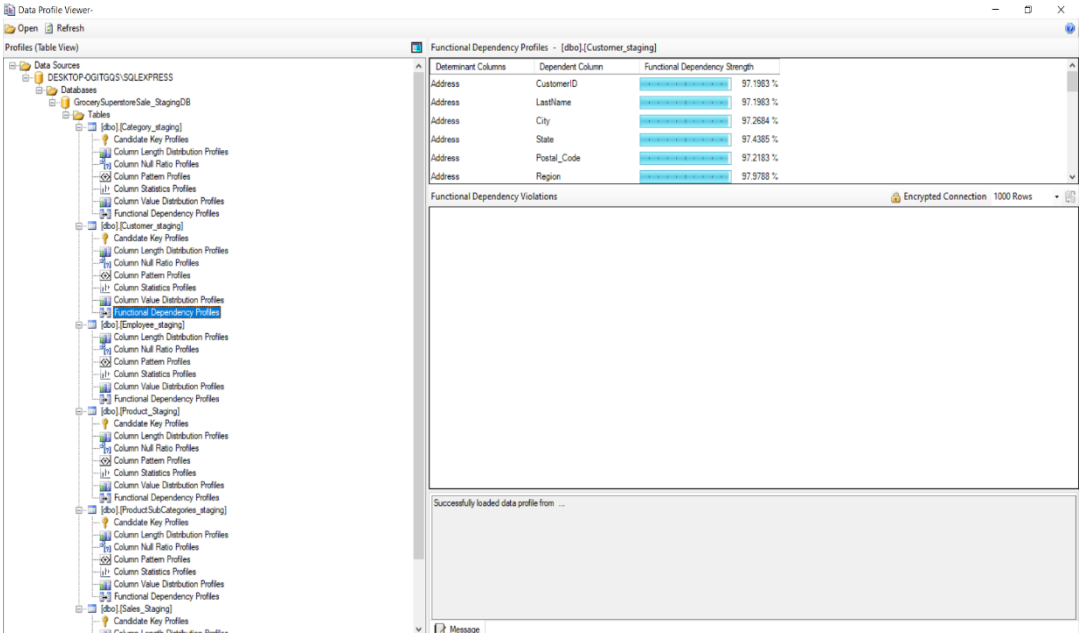



Data Profile Viewer-
Open
Refresh
Profiles (Table View)
Column Length Distribution Profiles - [dbo].[Customer_staging]
Column
Minimum Length
Maximum Length
Ignore Leading Spaces
IgnoreTrailing Spaces
Address 10 35
City 4 17
ContactNo 12 19
Country 13 13
FirstName 2 9
LastName 2 10
Length Distribution - Address
Encrypted Connection
1000 Rows
Length
Count
Percentage
23 622 6.2237 %
29 120 1.2007 %
15 590 5.9035 %
32 32 0.3202 %
26 298 2.9818 %
12 263 2.6316 %
35 2 0.0200 %
21 726 7.2644 %
27 197 1.9712 %
24 383 3.8323 %
18 641 6.4138 %
30 69 0.6904 %
10 100 1.0006 %
19 805 8.0548 %
Successfully loaded data profile from ...
Message

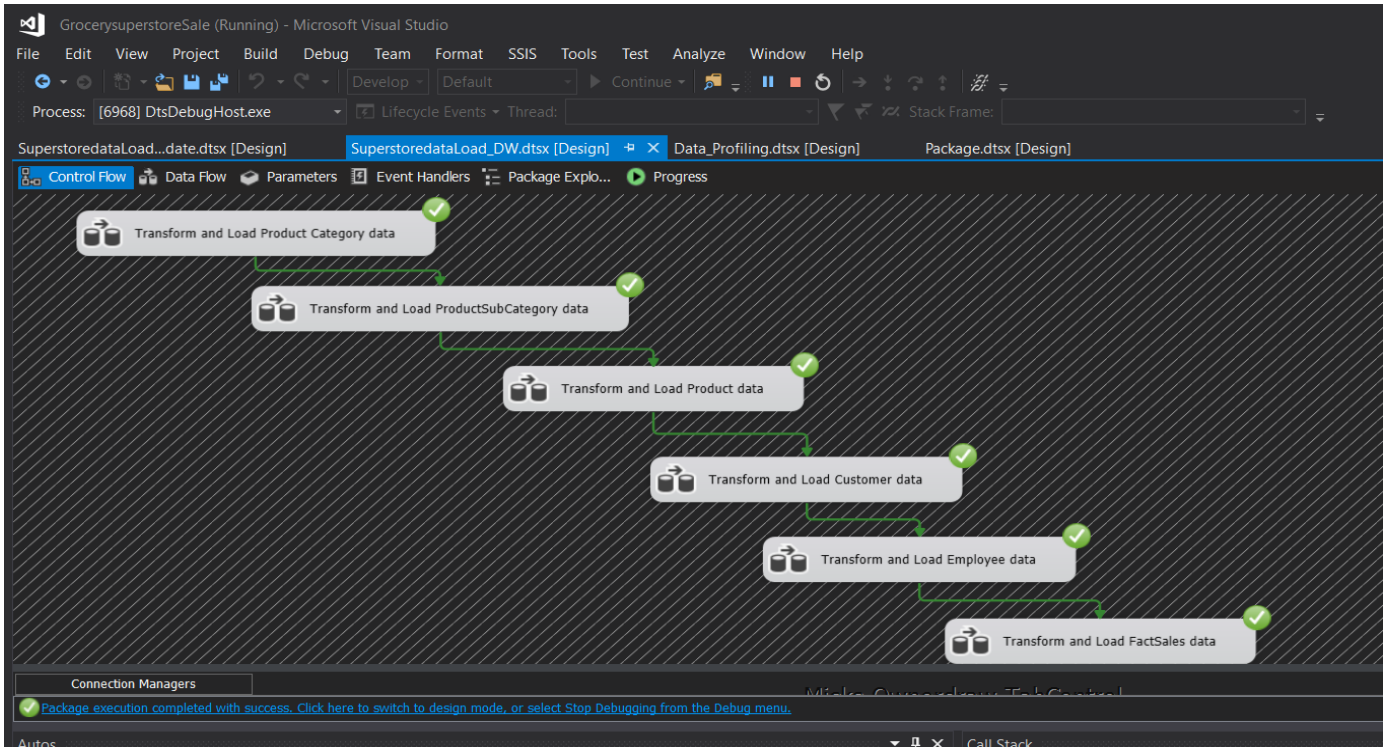

Data Profile Viewer-
Open
Refresh
Profiles (Table View)
Column Null Ratio Profiles - [dbo].[Customer_staging]
Encrypted Connection
1000 Rows
Column
Null Count
Null Percentage
Address 0 0.0000 %
City 0 0.0000 %
ContactNo 0 0.0000 %
Country 0 0.0000 %
CustomerID 0 0.0000 %
FirstName 0 0.0000 %
LastName 0 0.0000 %
Postal_Code 0 0.0000 %
Region 0 0.0000 %
State 0 0.0000 %
Successfully loaded data profile from ...
Message


Data Profile Viewer-
Open
Refresh
Profiles (Table View)
Column Statistics Profiles - [dbo].[Customer_staging]
Column
Minimum
Maximum
Mean
Standard Deviation
CustomerID 1 9994 4997.5 2885.01928069814
Postal_Code 1040 99301 55190.3794276... 32062.0891630789
Successfully loaded data profile from ...
Message

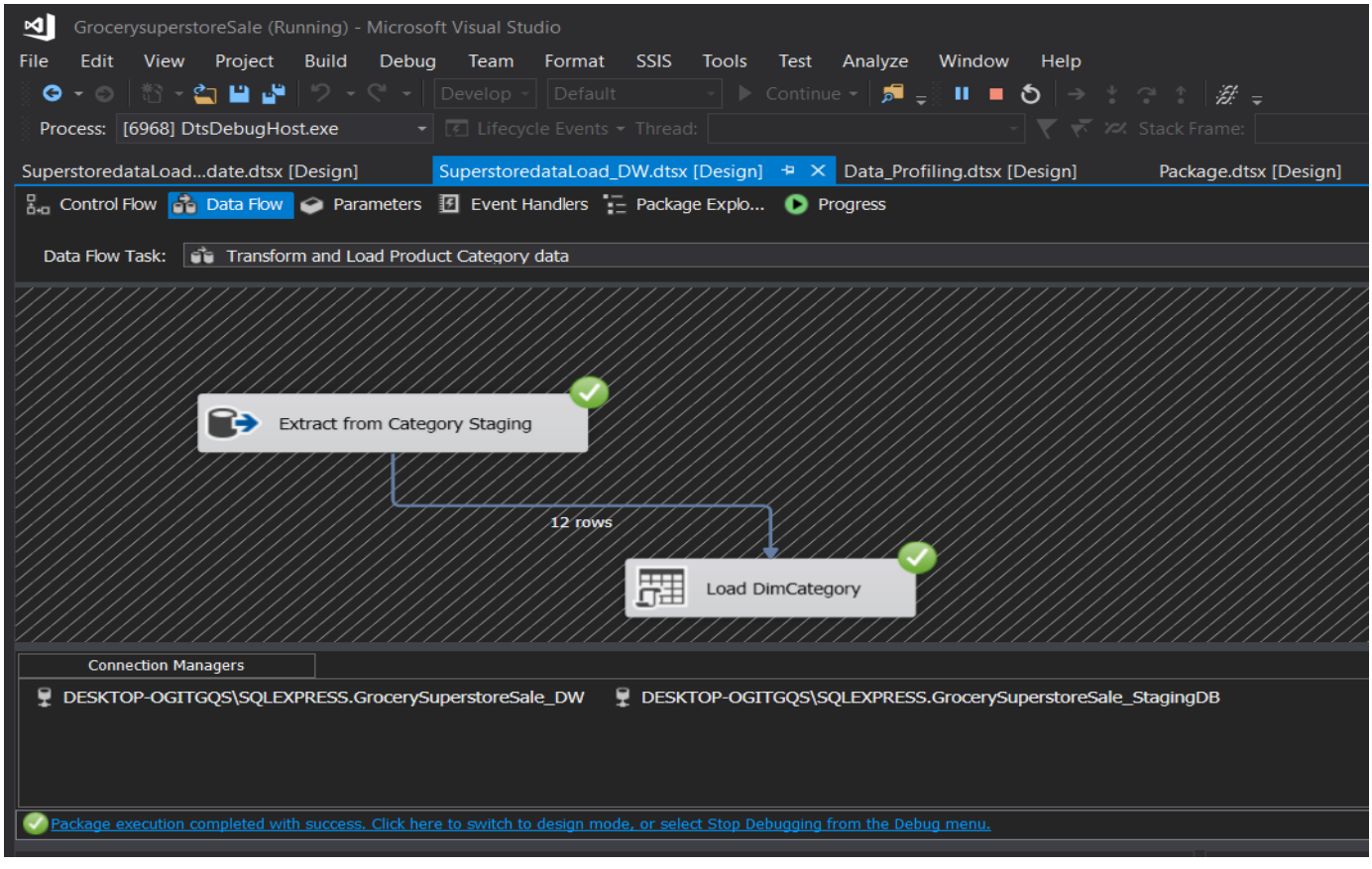

Data Profile Viewer-
Open
Refresh
Profiles (Table View)
Data Sources
DESKTOP-OGITGQS\SQLEXPRESS
Databases
GroceryStoreSale_StagingDB
Tables
[dbo].[Category_staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[Customer_staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[Employee_staging]
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[Product_Staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[ProductSubCategories_staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[Sales_Staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Value Distribution Profiles - [dbo].[Customer_staging]
Column
Number Of Distinct Values
Address 9714
City 531
ContactNo 5306
Country 1
CustomerID 9994
Value Distribution
Encrypted Connection
1000 Rows
Successfully loaded data profile from ...
Message


Data Profile Viewer-
Open
Refresh
Profiles (Table View)
Data Sources
DESKTOP-OGITGQS\SQLEXPRESS
Databases
GroceryStoreSale_StagingDB
Tables
[dbo].[Category_staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[Customer_staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[Employee_staging]
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[Product_Staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[ProductSubCategories_staging]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[Sales_Staging]
Candidate Key Profiles
Functional Dependency Profiles - [dbo].[Customer_staging]
Determinant Columns
Dependent Column
Functional Dependency Strength
Address CustomerID 97.1983 %
Address LastName 97.1983 %
Address City 97.2684 %
Address State 97.4385 %
Address Postal_Code 97.2183 %
Address Region 97.9788 %
Functional Dependency Violations
Encrypted Connection
1000 Rows
Successfully loaded data profile from ...
Message

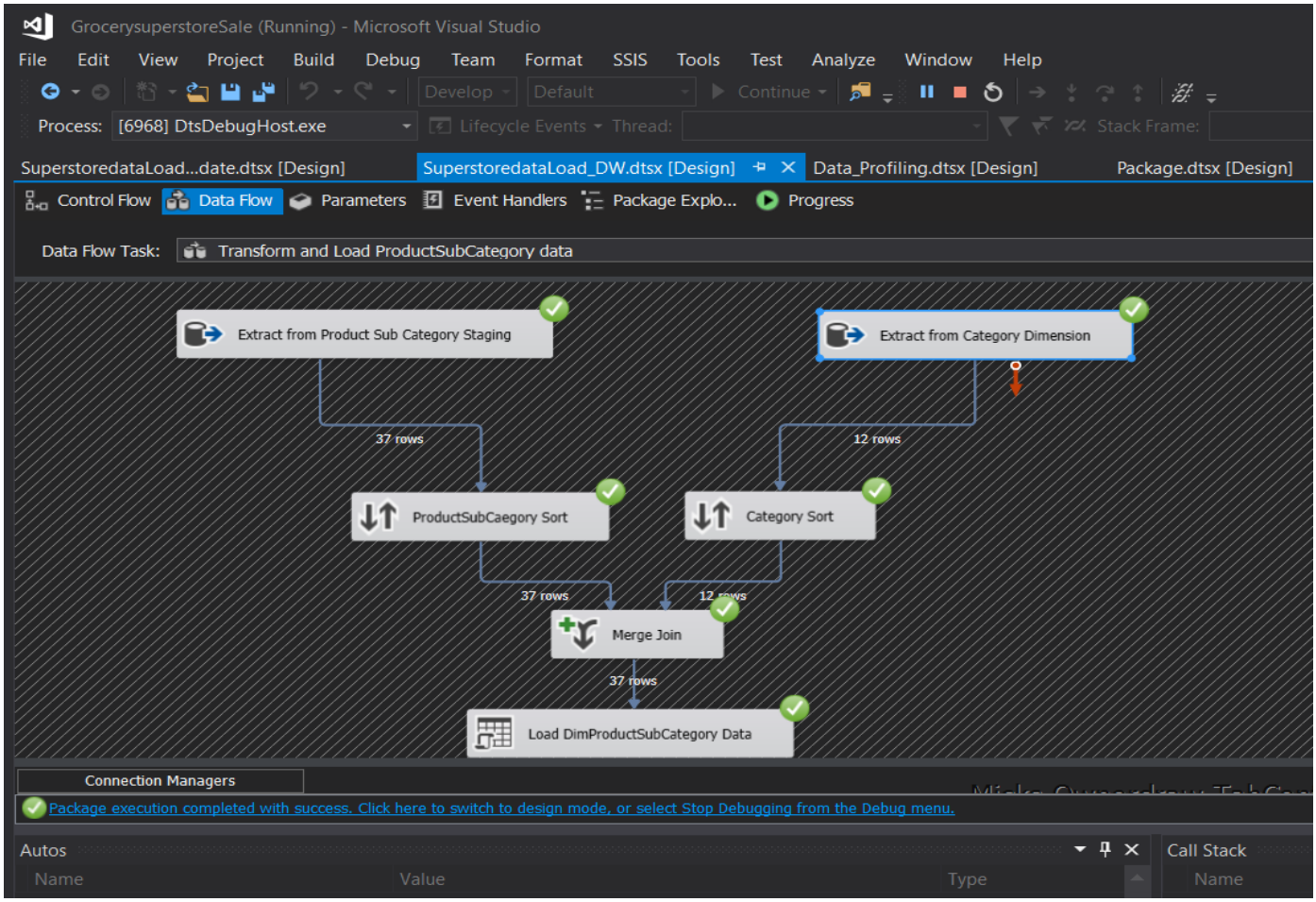
Staging To DW.



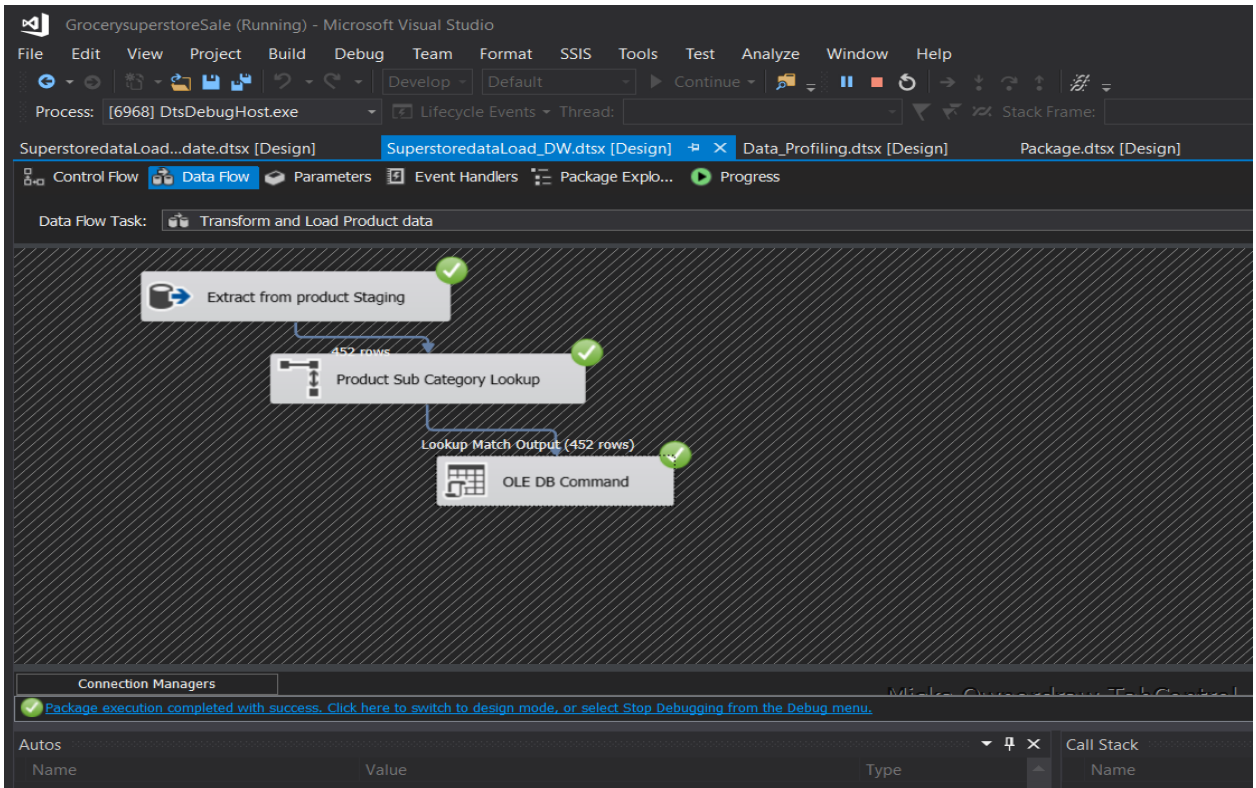
Load DimCategory data



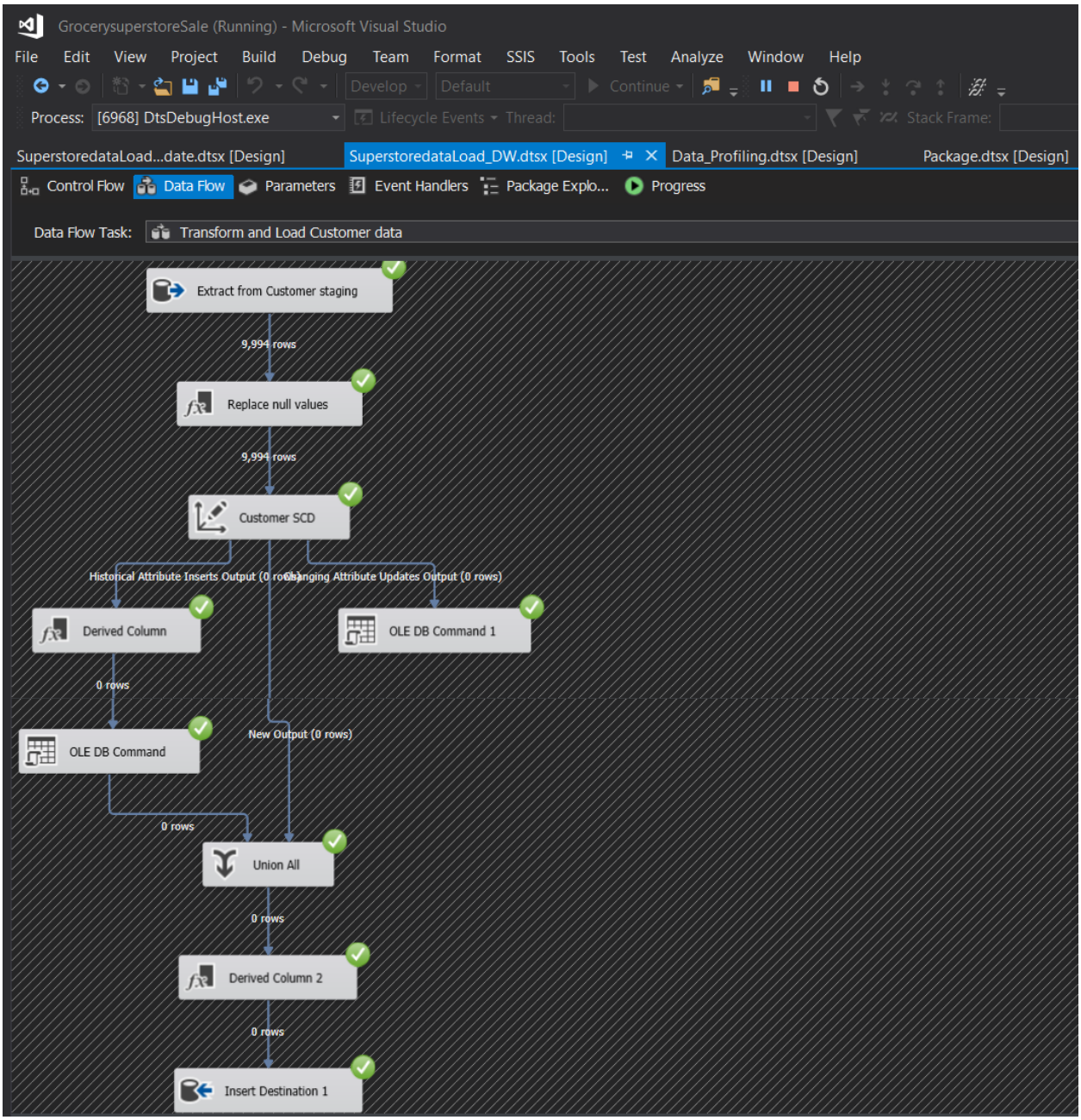
Load DimProductSubCategory data



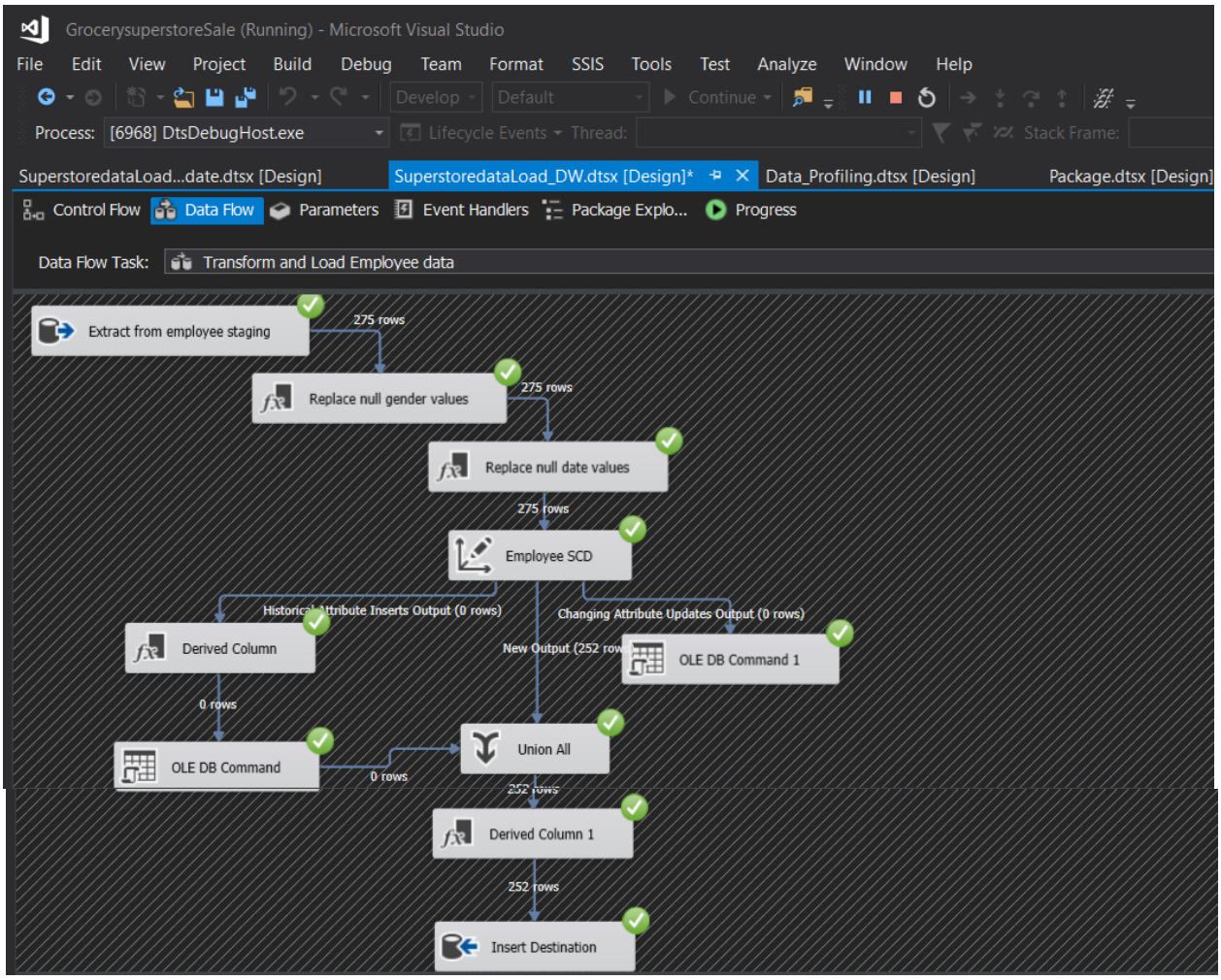
Load DimProduct data



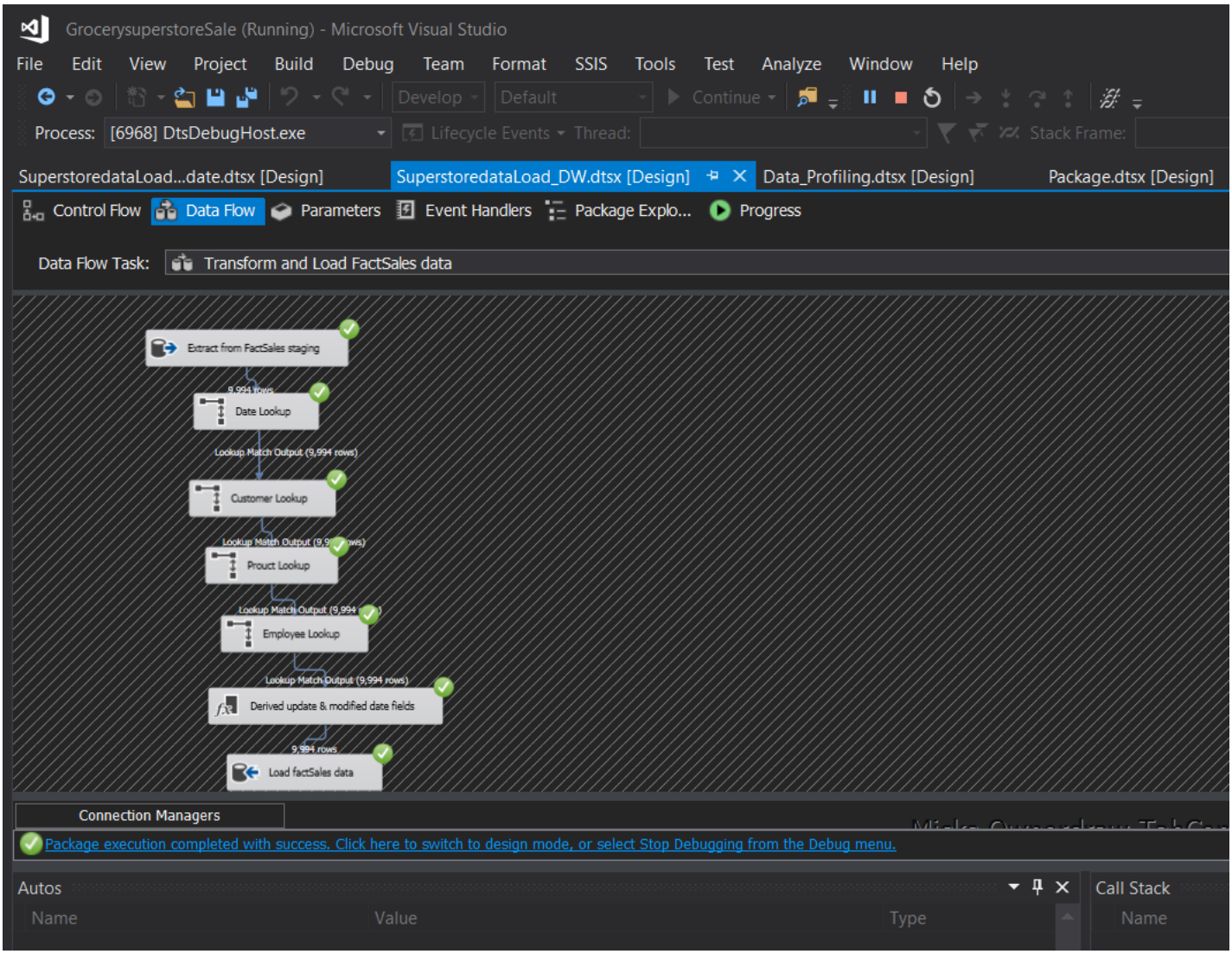
Load DimCustomer data



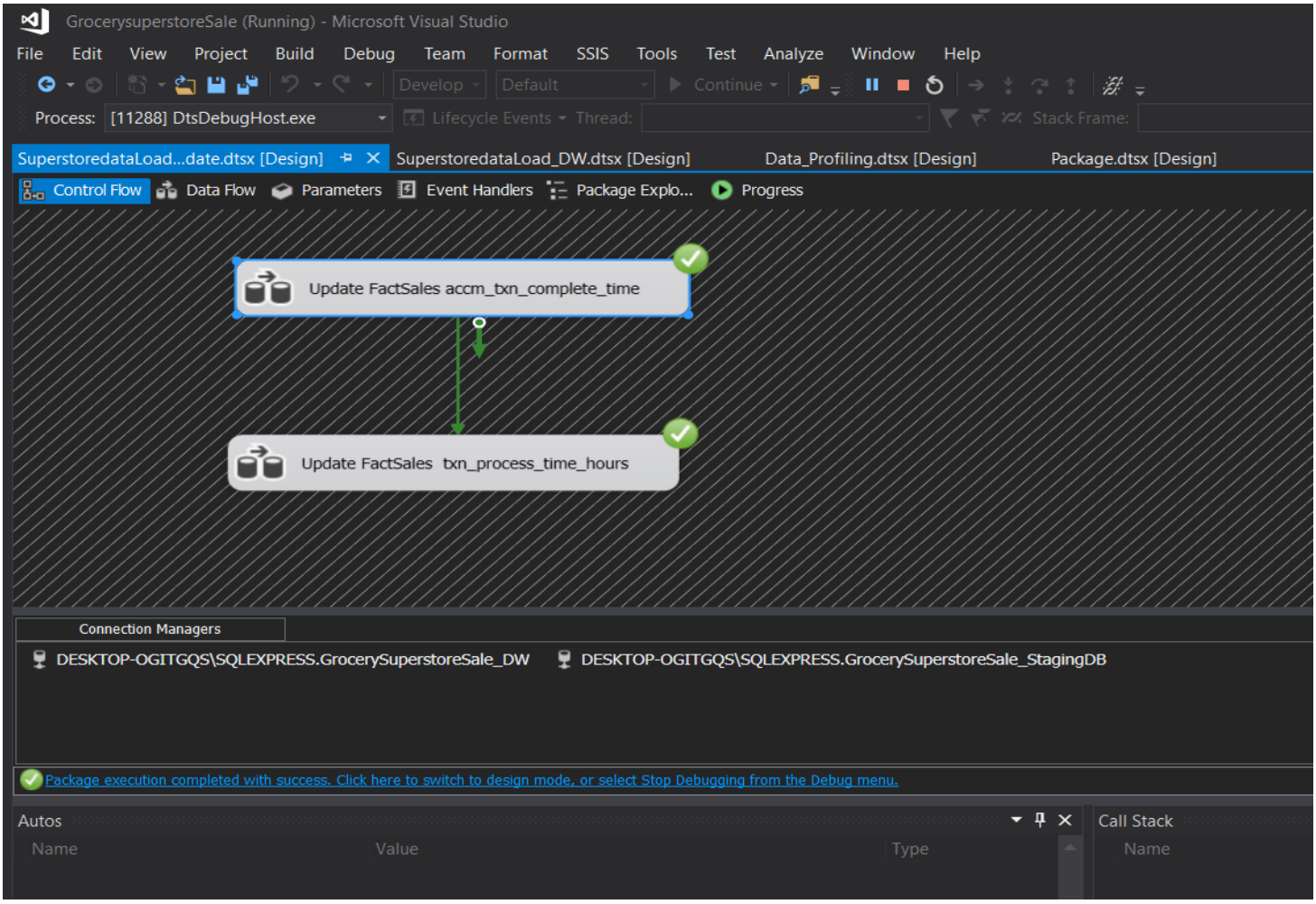
Load DimEmployee data

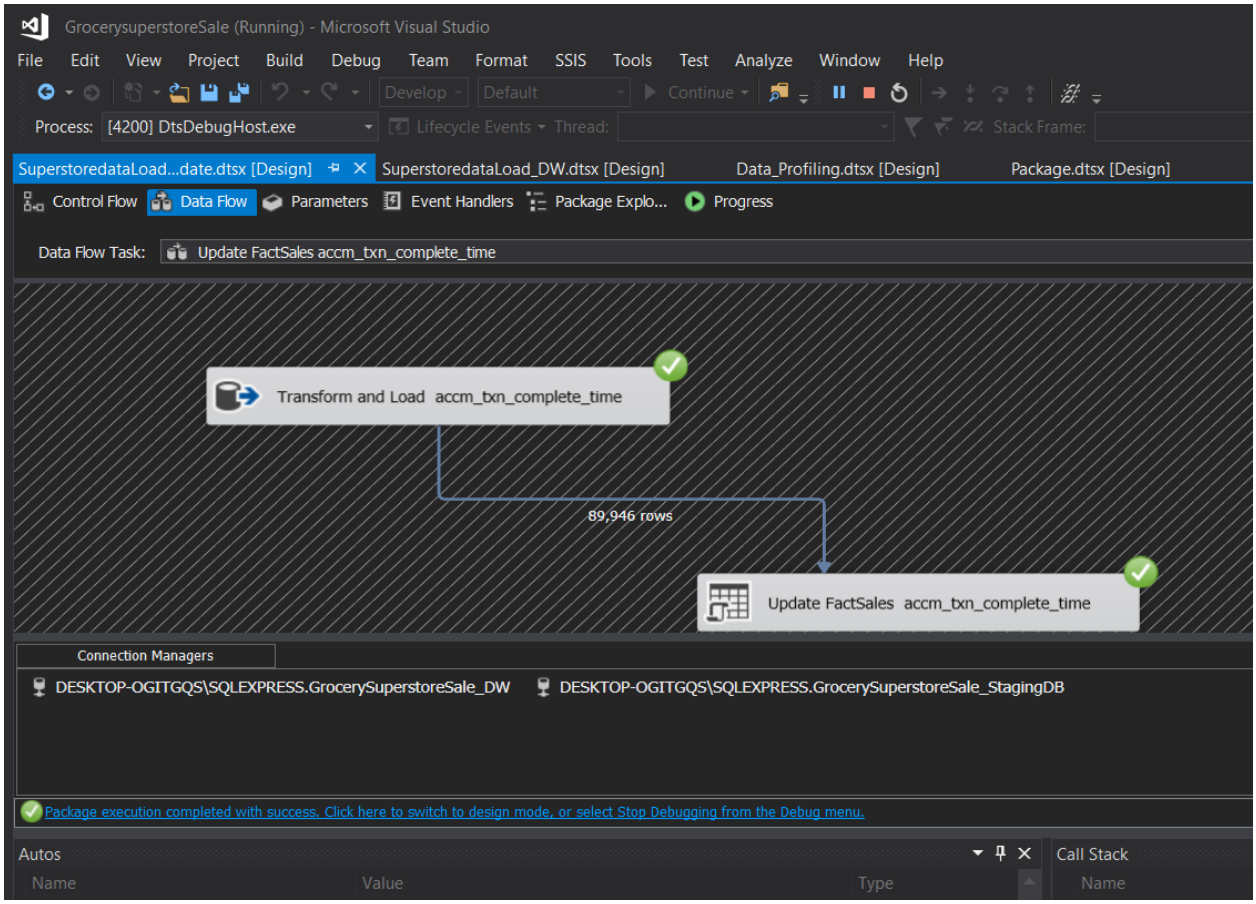


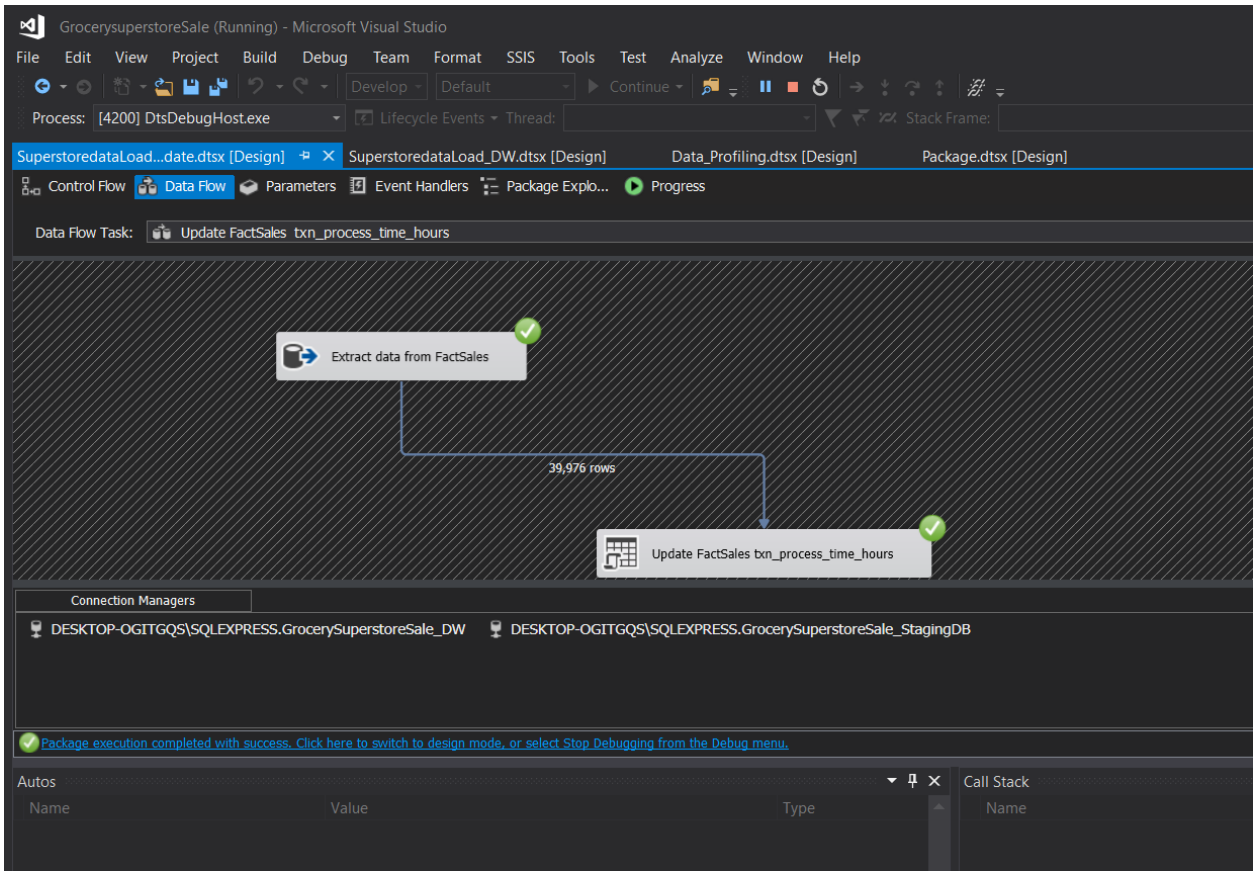
Load FactSales data



Accumulated fact table




GrocerysuperstoreSale (Running) - Microsoft Visual Studio
File Edit View Project Build Debug Team Format SSIS Tools Test Analyze Window Help
Develop Default Continue
Process: [4200] DtsDebugHost.exe
Lifecycle Events Thread:
Stack Frame:
SuperstoredataLoad...date.dtsx [Design]
SuperstoredataLoad_DW.dtsx [Design]
Data_Profiling.dtsx [Design]
Package.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Data Flow Task:
Update FactSales accm_txn_complete_time
Transform and Load accm_txn_complete_time
89,946 rows
Update FactSales accm_txn_complete_time
Connection Managers
DESKTOP-OGITGQS\SQLEXPRESS.GrocerySuperstoreSale_DW
DESKTOP-OGITGQS\SQLEXPRESS.GrocerySuperstoreSale_StagingDB
Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.
Autos
Name
Value
Type
Call Stack
Name


GrocerysuperstoreSale (Running) - Microsoft Visual Studio
File Edit View Project Build Debug Team Format SSIS Tools Test Analyze Window Help
Develop Default Continue
Process: [4200] DtsDebugHost.exe
Lifecycle Events Thread:
Stack Frame:
SuperstoredataLoad...date.dtsx [Design]
SuperstoredataLoad_DW.dtsx [Design]
Data_Profiling.dtsx [Design]
Package.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Data Flow Task:
Update FactSales txn_process_time_hours
Extract data from FactSales
39,976 rows
Update FactSales txn_process_time_hours
Connection Managers
DESKTOP-OGITGQS\SQLEXPRESS.GrocerySuperstoreSale_DW
DESKTOP-OGITGQS\SQLEXPRESS.GrocerySuperstoreSale_StagingDB
Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.
Autos
Name
Value
Type
Call Stack
Name