Data Warehousing Assignment 4: Creating a simple ETL pipeline

Team Members:
Siby Abraham Kochuputhentharayil Chacko (50419148)
Sanjuna Senthil Kumar (50416860)
Viswanathan Jeyaraman Krishnamurthy (50428096)

Summary:

We first created the basic structure and the tables for the warehouse using the SQL server and SSMS. In the following report, we will be detailing the steps that we followed to fill the tables with the appropriate data to populate this data warehouse.

- In an effort to build the final production database for the organization, we tried importing a database in the form of a .bak file into the Server. However, we encountered errors owing to not storing the file in the appropriate file path.

  We then accessed the SSMS and accessed the locally running database server and restored the database provided. However, after restoring, we also faced two other challenges viz., the database diagrams

  - were missing relationships.
  - weren't accessible and required explicit public permission for access.

- Moving forward, we populated the Date dimension table for the organisation by writing and executing scripts, according to the defined caveats.

- We then manually added data by running an SQL Insert Query into the Customer dimension.

- We downloaded and installed Visual Studio along with the Data storage and processing capacity as this provides us with the inbuilt SSIS tool.

- After creating a SSIS project, we started our work on creating an ETL pipeline. The first step towards this involved creating a Data Flow task. This mandates the automated actions that needs to take place for the ETL pipeline.

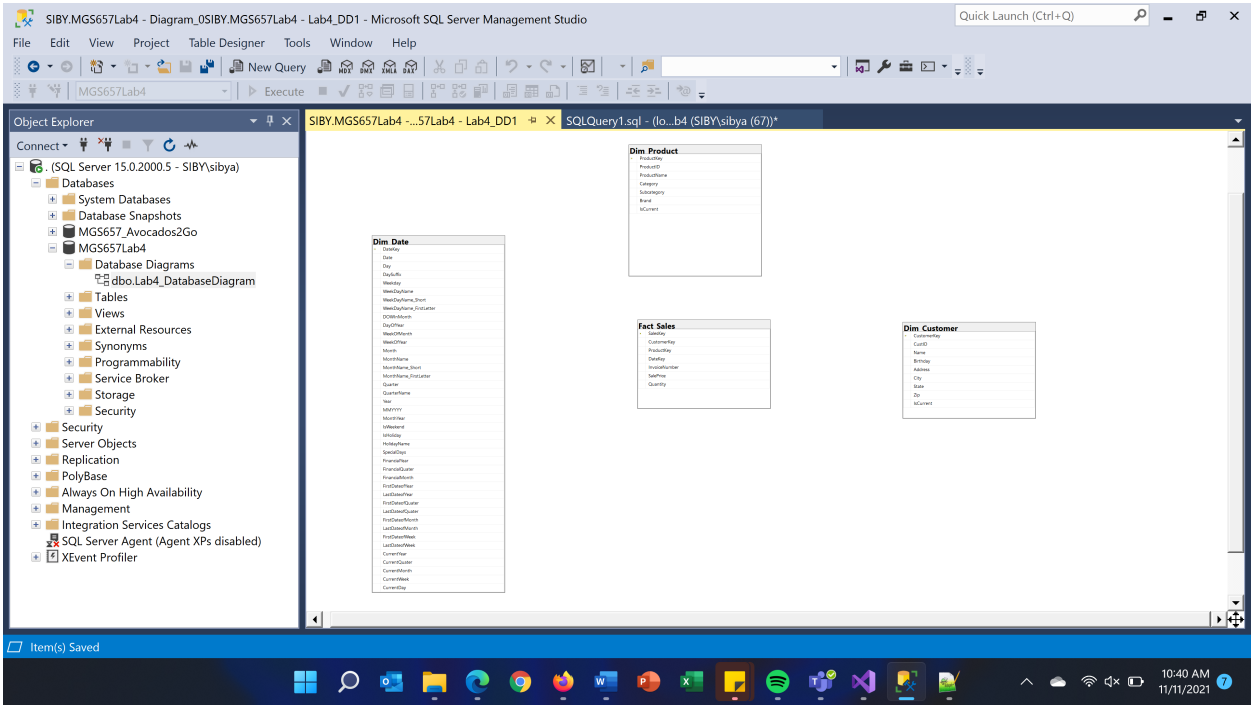
- We also set up the connection manager for setting the source and destination of the ETL pipeline as the Excel files and localhost server.

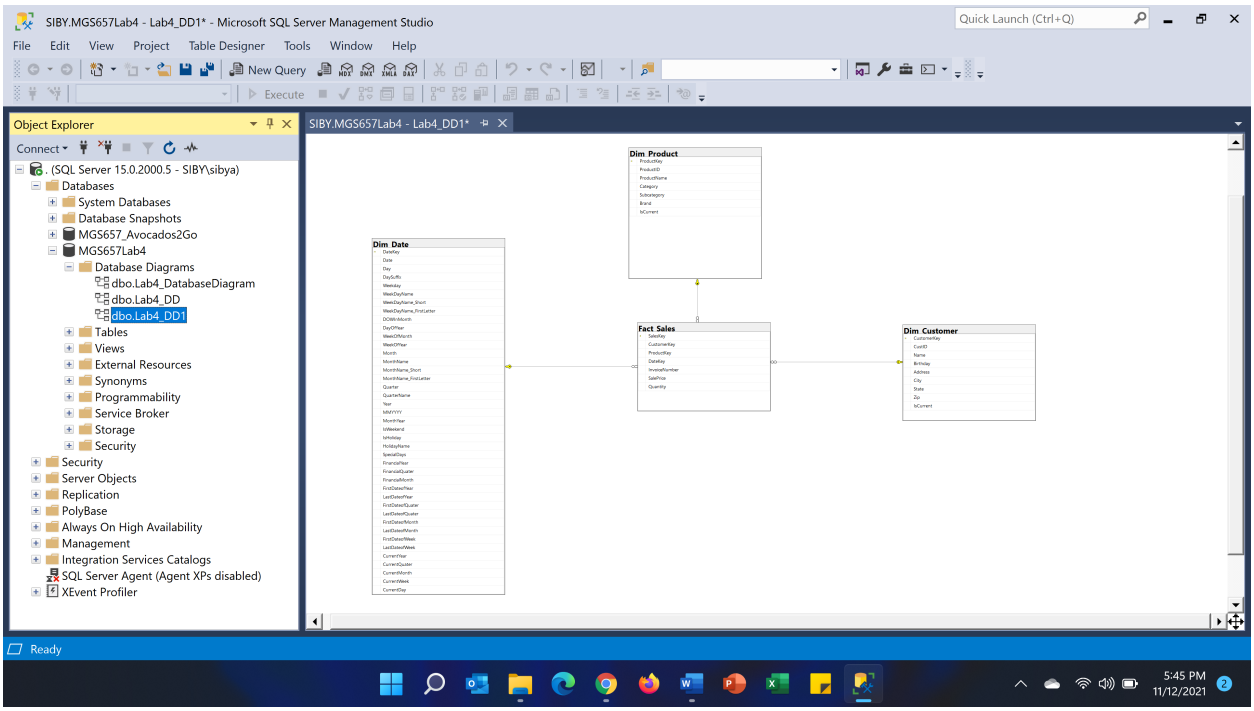
- We then standardized the data types by altering the Excel files' datatypes through Data Conversion.

- We then accessed the Slowly Changing Dimension Wizard to account for updation, deletion or creation scenarios to existing dimensions.

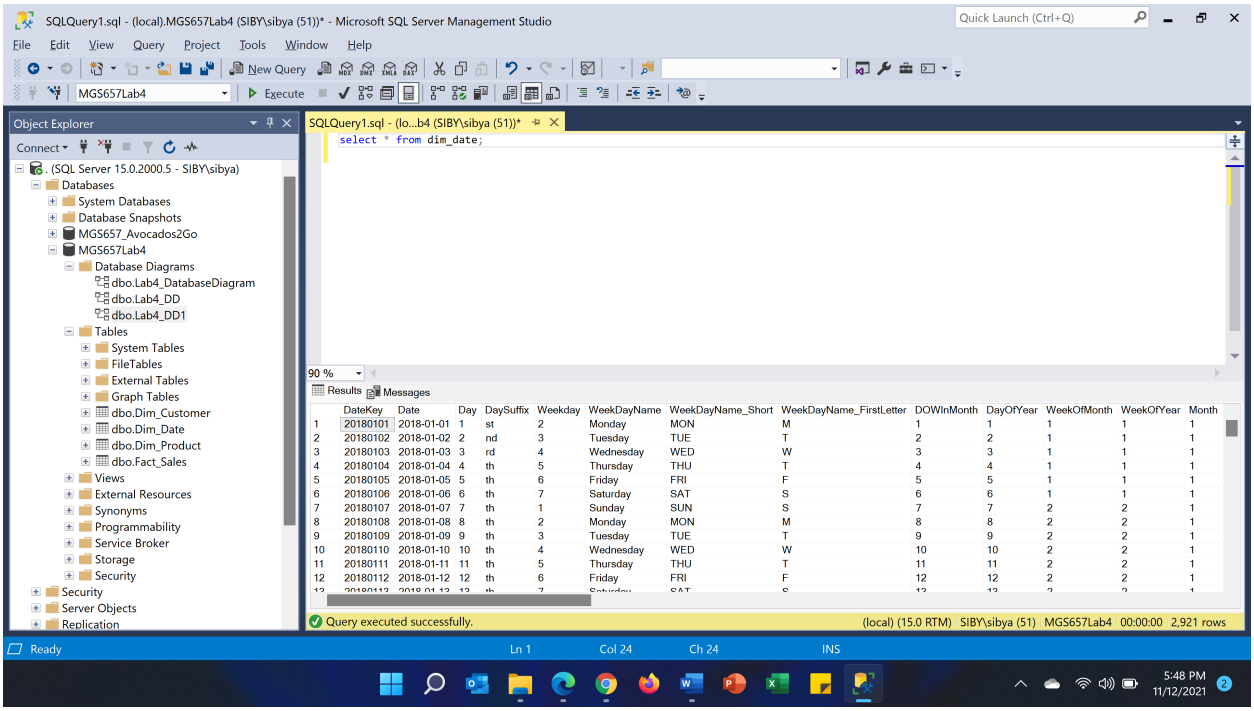
- Post this, we ran the ETL process successfully.

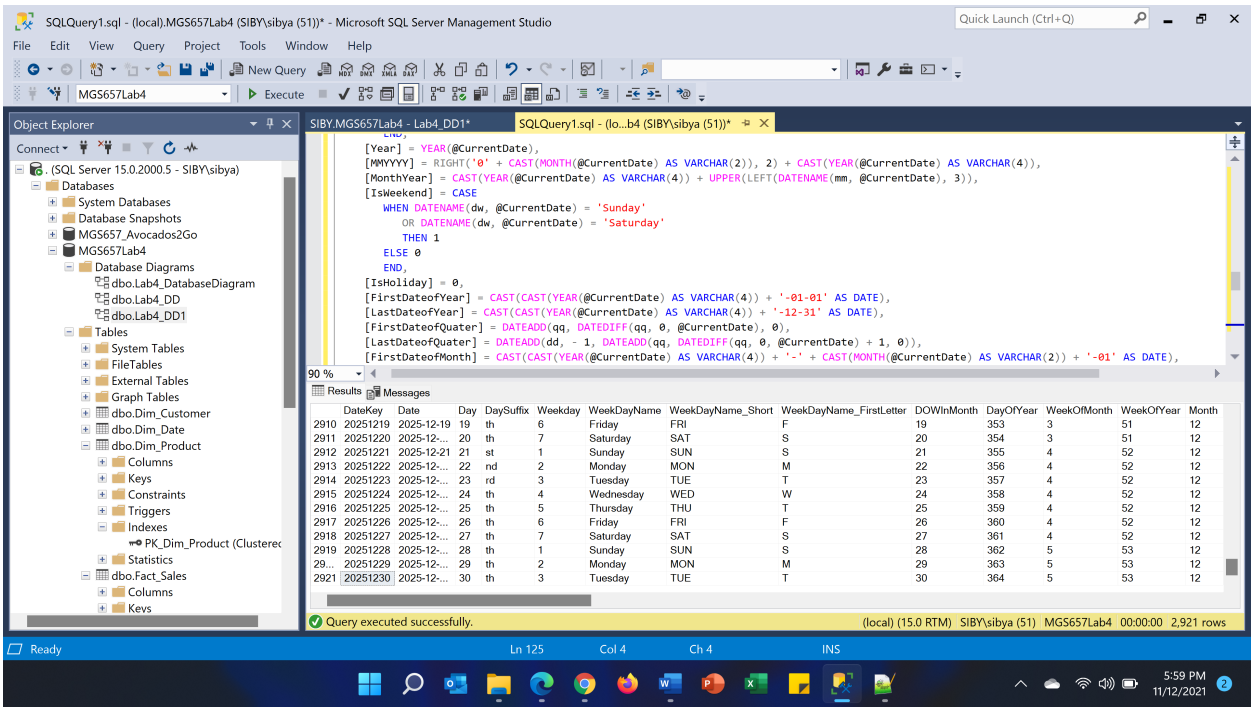
Database screenshot:

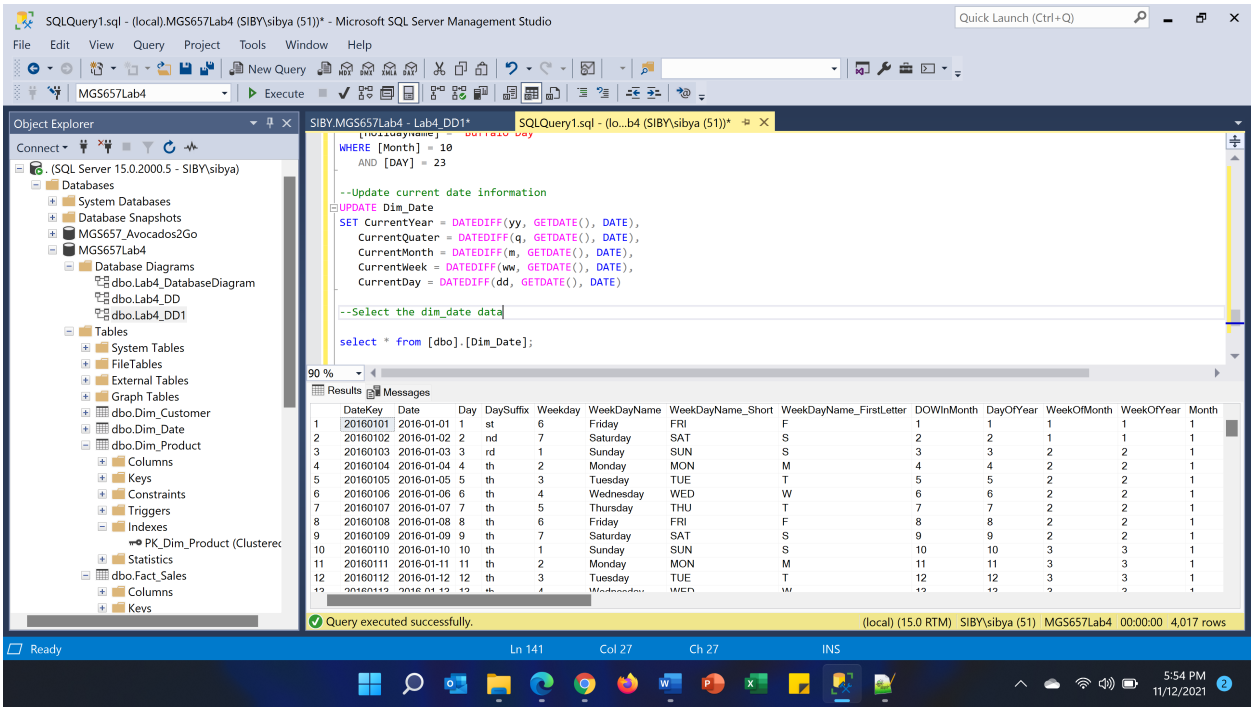


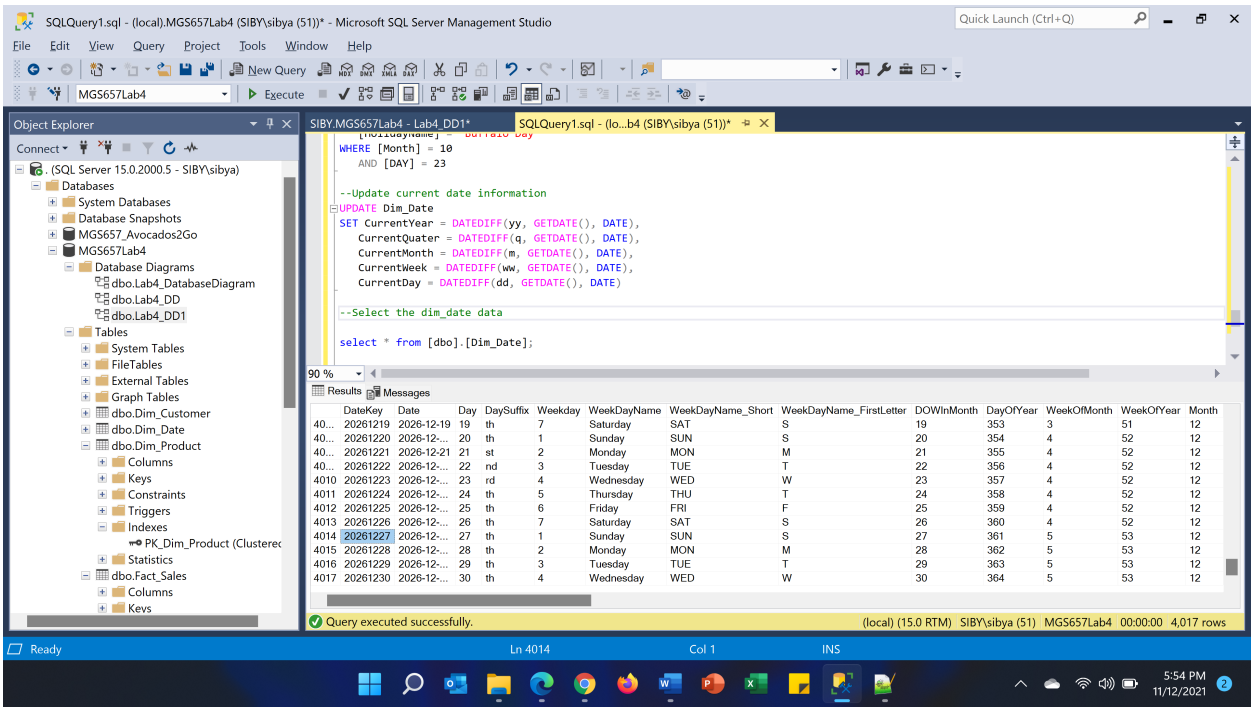

SIBY.MGS657Lab4 - Lab4_DD1* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Project Table Designer Tools Window Help
New Query
Execute
Object Explorer
Connect
. (SQL Server 15.0.2000.5 - SIBY\sibya)
Databases
System Databases
Database Snapshots
MGS657_Avocados2Go
MGS657Lab4
Database Diagrams
dbo.Lab4_DatabaseDiagram
dbo.Lab4_DD
dbo.Lab4_DD1
Tables
Views
External Resources
Synonyms
Programmability
Service Broker
Storage
Security
Security
Server Objects
Replication
PolyBase
Always On High Availability
Management
Integration Services Catalogs
SQL Server Agent (Agent XPs disabled)
XEvent Profiler
SIBY.MGS657Lab4 - Lab4_DD1*
Dim_Product
Dim_Date
Fact_Sales
Dim_Customer
Ready
5:45 PM
11/12/2021

Dim_Date:

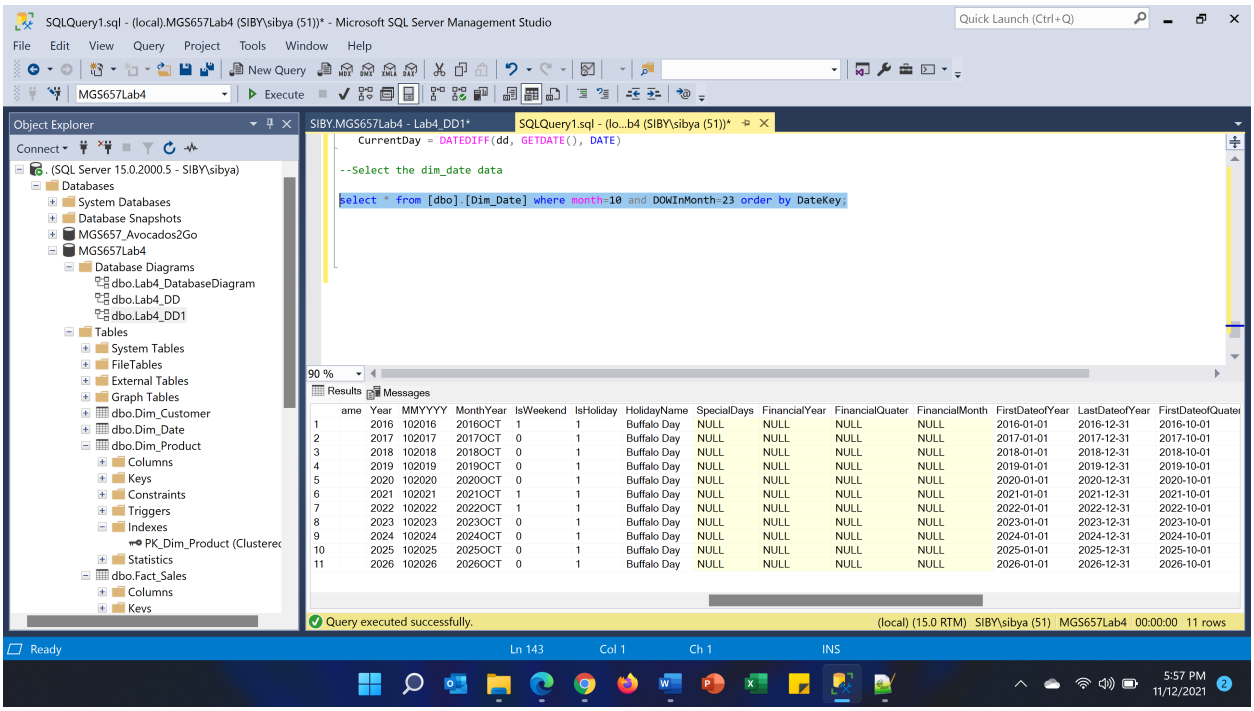

SQLQuery1.sql - (local).MGS657Lab4 (SIBY\sibya (51))* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Query Project Tools Window Help
New Query
MGS657Lab4
Execute
Object Explorer
Connect
. (SQL Server 15.0.2000.5 - SIBY\sibya)
Databases
System Databases
Database Snapshots
MGS657_Avocados2Go
MGS657Lab4
Database Diagrams
dbo.Lab4_DatabaseDiagram
dbo.Lab4_DD
dbo.Lab4_DD1
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.Dim_Customer
dbo.Dim_Date
dbo.Dim_Product
dbo.Fact_Sales
Views
External Resources
Synonyms
Programmability
Service Broker
Storage
Security
Security
Server Objects
Replication
SQLQuery1.sql - (lo...b4 (SIBY\sibya (51))*
select * from dim_date;
90 %
Results
Messages
DateKey Date Day DaySuffix Weekday WeekDayName WeekDayName_Short WeekDayName_FirstLetter DOWInMonth DayOfYear WeekOfMonth WeekOfYear Month
1 20180101 2018-01-01 1 st 2 Monday MON M 1 1 1 1 1
2 20180102 2018-01-02 2 nd 3 Tuesday TUE T 2 2 1 1 1
3 20180103 2018-01-03 3 rd 4 Wednesday WED W 3 3 1 1 1
4 20180104 2018-01-04 4 th 5 Thursday THU T 4 4 1 1 1
5 20180105 2018-01-05 5 th 6 Friday FRI F 5 5 1 1 1
6 20180106 2018-01-06 6 th 7 Saturday SAT S 6 6 1 1 1
7 20180107 2018-01-07 7 th 1 Sunday SUN S 7 7 2 2 1
8 20180108 2018-01-08 8 th 2 Monday MON M 8 8 2 2 1
9 20180109 2018-01-09 9 th 3 Tuesday TUE T 9 9 2 2 1
10 20180110 2018-01-10 10 th 4 Wednesday WED W 10 10 2 2 1
11 20180111 2018-01-11 11 th 5 Thursday THU T 11 11 2 2 1
12 20180112 2018-01-12 12 th 6 Friday FRI F 12 12 2 2 1
Query executed successfully.
(local) (15.0 RTM) SIBY\sibya (51) MGS657Lab4 00:00:00 2,921 rows
Ready
Ln 1 Col 24 Ch 24 INS
5:48 PM
11/12/2021


SQLQuery1.sql - (local).MGS657Lab4 (SIBY\sibya (51))* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Query Project Tools Window Help
New Query
Execute
MGS657Lab4
Object Explorer
Connect
. (SQL Server 15.0.2000.5 - SIBY\sibya)
Databases
System Databases
Database Snapshots
MGS657_Avocados2Go
MGS657Lab4
Database Diagrams
dbo.Lab4_DatabaseDiagram
dbo.Lab4_DD
dbo.Lab4_DD1
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.Dim_Customer
dbo.Dim_Date
dbo.Dim_Product
Columns
Keys
Constraints
Triggers
Indexes
PK_Dim_Product (Clustered
Statistics
dbo.Fact_Sales
Columns
Keys
SIBY.MGS657Lab4 - Lab4_DD1*
SQLQuery1.sql - (lo...b4 (SIBY\sibya (51))*
[Year] = YEAR(@CurrentDate),
[MMYYYY] = RIGHT('0' + CAST(MONTH(@CurrentDate) AS VARCHAR(2)), 2) + CAST(YEAR(@CurrentDate) AS VARCHAR(4)),
[MonthYear] = CAST(YEAR(@CurrentDate) AS VARCHAR(4)) + UPPER(LEFT(DATENAME(mm, @CurrentDate), 3)),
[IsWeekend] = CASE
WHEN DATENAME(dw, @CurrentDate) = 'Sunday'
OR DATENAME(dw, @CurrentDate) = 'Saturday'
THEN 1
ELSE 0
END,
[IsHoliday] = 0,
[FirstDateofYear] = CAST(CAST(YEAR(@CurrentDate) AS VARCHAR(4)) + '-01-01' AS DATE),
[LastDateofYear] = CAST(CAST(YEAR(@CurrentDate) AS VARCHAR(4)) + '-12-31' AS DATE),
[FirstDateofQuater] = DATEADD(qq, DATEDIFF(qq, 0, @CurrentDate), 0),
[LastDateofQuater] = DATEADD(dd, -1, DATEADD(qq, DATEDIFF(qq, 0, @CurrentDate) + 1, 0)),
[FirstDateofMonth] = CAST(CAST(YEAR(@CurrentDate) AS VARCHAR(4)) + '-' + CAST(MONTH(@CurrentDate) AS VARCHAR(2)) + '-01' AS DATE),
90 %
Results
Messages
DateKey Date Day DaySuffix Weekday WeekDayName WeekDayName_Short WeekDayName_FirstLetter DOWInMonth DayOfYear WeekOfMonth WeekOfYear Month
2910 20251219 2025-12-19 19 th 6 Friday FRI F 19 353 3 51 12
2911 20251220 2025-12-... 20 th 7 Saturday SAT S 20 354 3 51 12
2912 20251221 2025-12-21 21 st 1 Sunday SUN S 21 355 4 52 12
2913 20251222 2025-12-... 22 nd 2 Monday MON M 22 356 4 52 12
2914 20251223 2025-12-... 23 rd 3 Tuesday TUE T 23 357 4 52 12
2915 20251224 2025-12-... 24 th 4 Wednesday WED W 24 358 4 52 12
2916 20251225 2025-12-... 25 th 5 Thursday THU T 25 359 4 52 12
2917 20251226 2025-12-... 26 th 6 Friday FRI F 26 360 4 52 12
2918 20251227 2025-12-... 27 th 7 Saturday SAT S 27 361 4 52 12
2919 20251228 2025-12-... 28 th 1 Sunday SUN S 28 362 5 53 12
29... 20251229 2025-12-... 29 th 2 Monday MON M 29 363 5 53 12
2921 20251230 2025-12-... 30 th 3 Tuesday TUE T 30 364 5 53 12
Query executed successfully.
(local) (15.0 RTM) SIBY\sibya (51) MGS657Lab4 00:00:00 2,921 rows
Ready
Ln 125 Col 4 Ch 4 INS
5:59 PM
11/12/2021

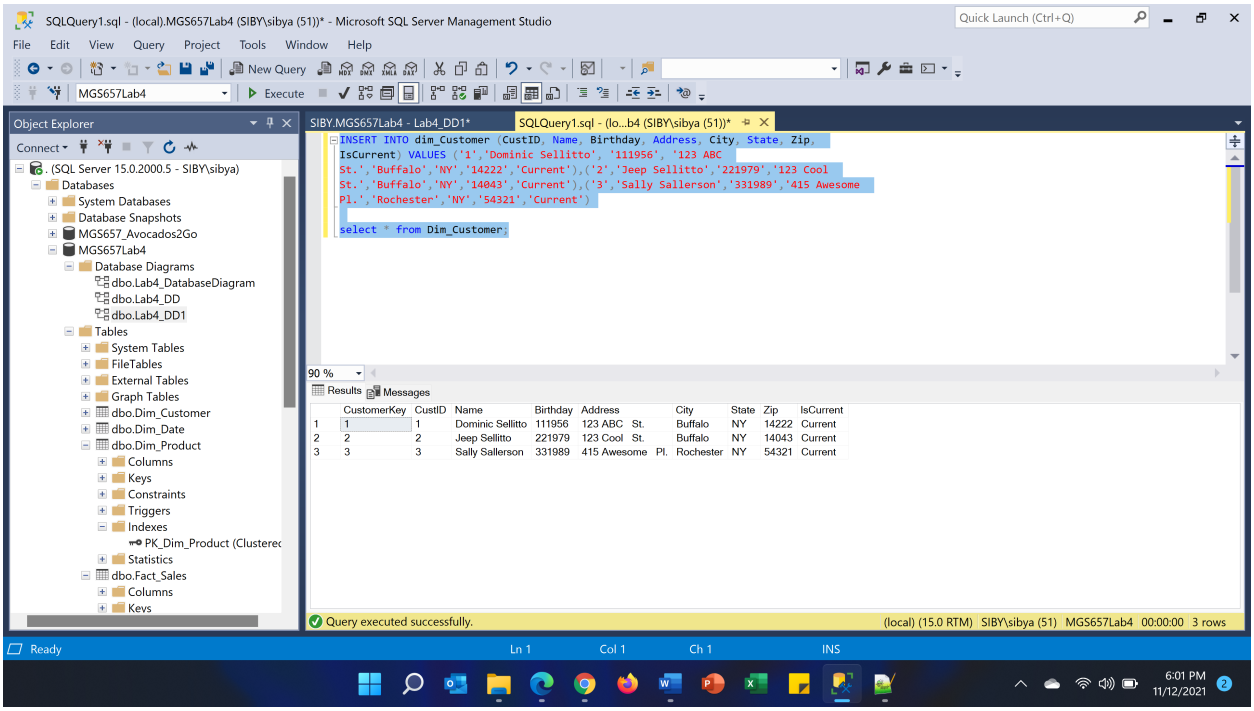

SQLQuery1.sql - (local).MGS657Lab4 (SIBY\sibya (51))* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Query Project Tools Window Help
New Query
Execute
MGS657Lab4
Object Explorer
Connect
. (SQL Server 15.0.2000.5 - SIBY\sibya)
Databases
System Databases
Database Snapshots
MGS657_Avocados2Go
MGS657Lab4
Database Diagrams
dbo.Lab4_DatabaseDiagram
dbo.Lab4_DD
dbo.Lab4_DD1
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.Dim_Customer
dbo.Dim_Date
dbo.Dim_Product
Columns
Keys
Constraints
Triggers
Indexes
PK_Dim_Product (Clustered
Statistics
dbo.Fact_Sales
Columns
Keys
SIBY.MGS657Lab4 - Lab4_DD1*
SQLQuery1.sql - (lo...b4 (SIBY\sibya (51))*
WHERE [Month] = 10
AND [DAY] = 23
--Update current date information
UPDATE Dim_Date
SET CurrentYear = DATEDIFF(yy, GETDATE(), DATE),
CurrentQuater = DATEDIFF(q, GETDATE(), DATE),
CurrentMonth = DATEDIFF(m, GETDATE(), DATE),
CurrentWeek = DATEDIFF(ww, GETDATE(), DATE),
CurrentDay = DATEDIFF(dd, GETDATE(), DATE)
--Select the dim_date data
select * from [dbo].[Dim_Date];
90 %
Results
Messages
DateKey Date Day DaySuffix Weekday WeekDayName WeekDayName_Short WeekDayName_FirstLetter DOWInMonth DayOfYear WeekOfMonth WeekOfYear Month
1 20160101 2016-01-01 1 st 6 Friday FRI F 1 1 1 1 1
2 20160102 2016-01-02 2 nd 7 Saturday SAT S 2 2 1 1 1
3 20160103 2016-01-03 3 rd 1 Sunday SUN S 3 3 2 2 1
4 20160104 2016-01-04 4 th 2 Monday MON M 4 4 2 2 1
5 20160105 2016-01-05 5 th 3 Tuesday TUE T 5 5 2 2 1
6 20160106 2016-01-06 6 th 4 Wednesday WED W 6 6 2 2 1
7 20160107 2016-01-07 7 th 5 Thursday THU T 7 7 2 2 1
8 20160108 2016-01-08 8 th 6 Friday FRI F 8 8 2 2 1
9 20160109 2016-01-09 9 th 7 Saturday SAT S 9 9 2 2 1
10 20160110 2016-01-10 10 th 1 Sunday SUN S 10 10 3 3 1
11 20160111 2016-01-11 11 th 2 Monday MON M 11 11 3 3 1
12 20160112 2016-01-12 12 th 3 Tuesday TUE T 12 12 3 3 1
Query executed successfully.
(local) (15.0 RTM) SIBY\sibya (51) MGS657Lab4 00:00:00 4,017 rows
Ready
Ln 141 Col 27 Ch 27 INS
5:54 PM
11/12/2021



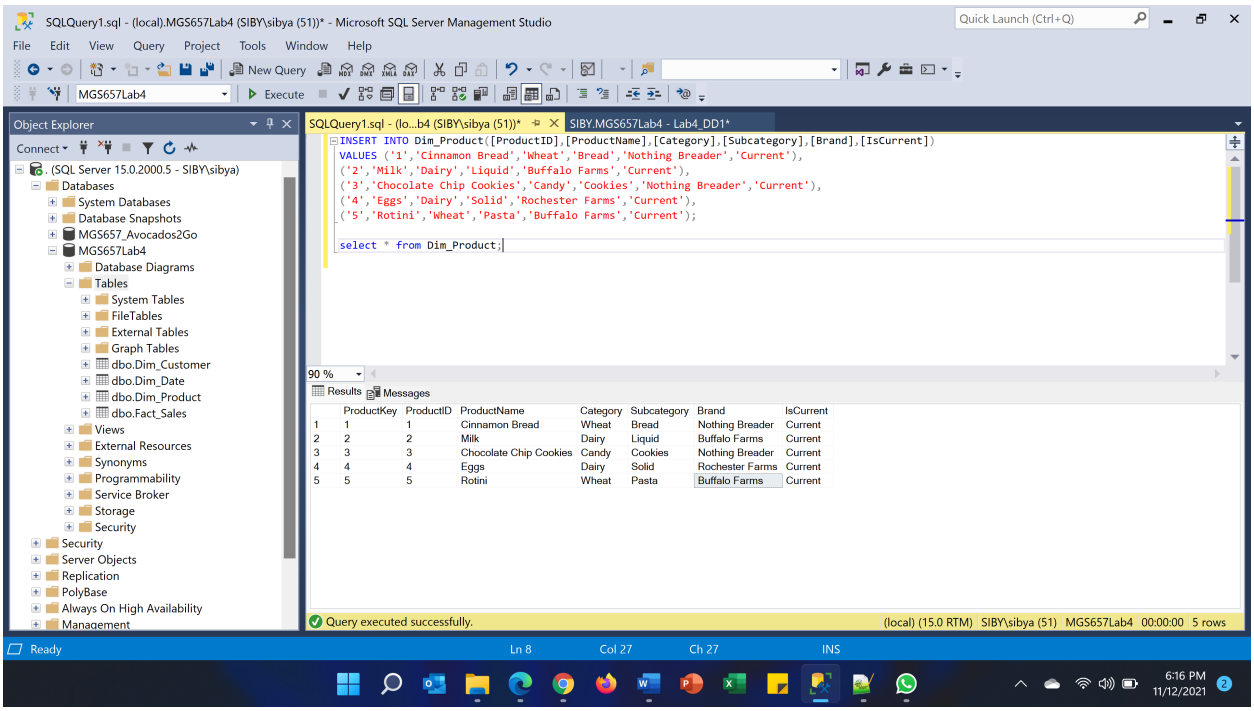
Dim_Date with Buffalo day:



Dim_Customer:

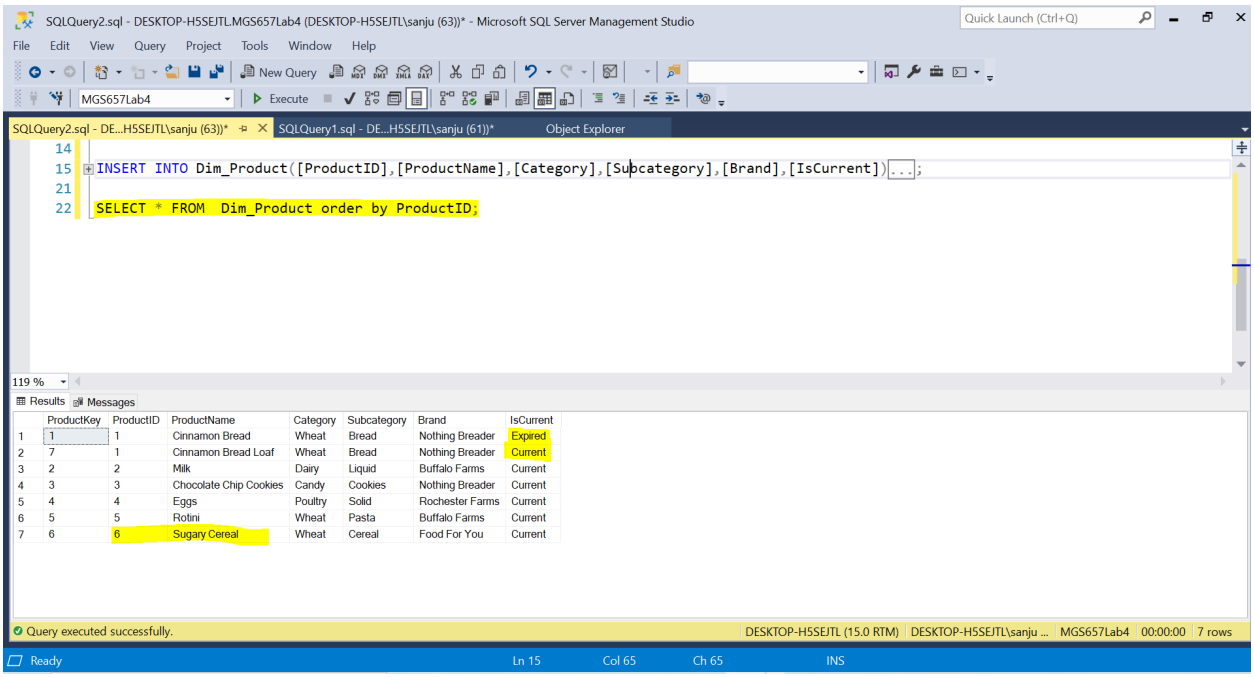


Dim_Product:



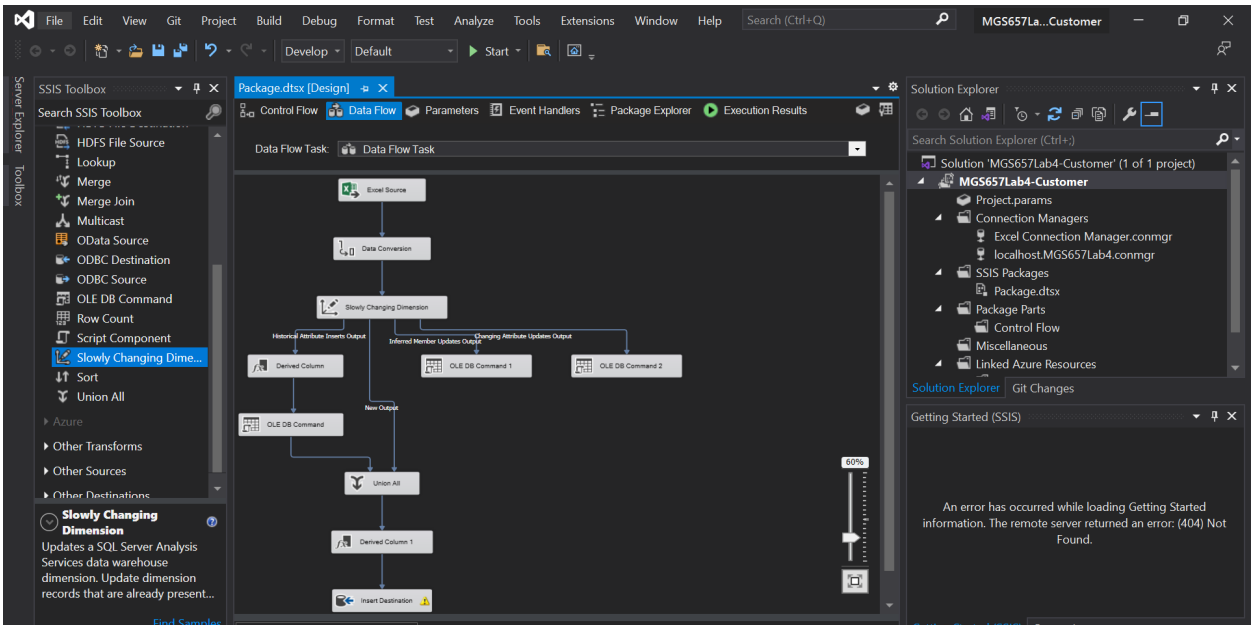
SSIS flow:

Dim_Product:

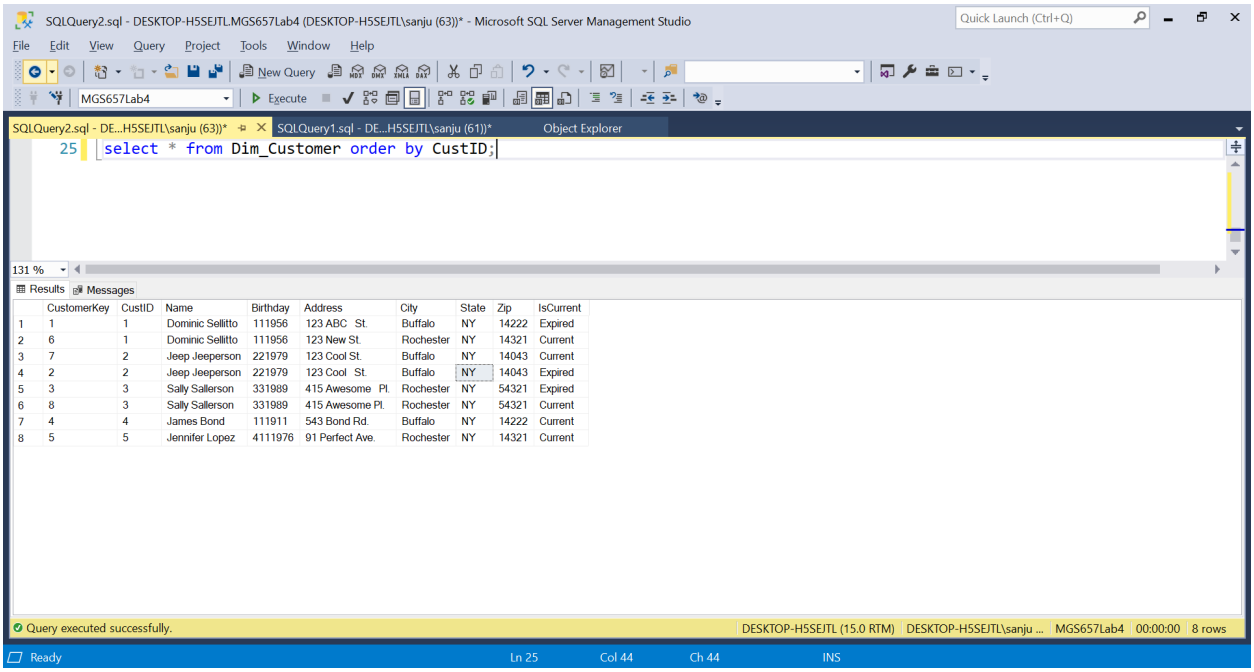


	ProductKey	ProductID	ProductName	Category	Subcategory	Brand	IsCurrent
1	1	1	Cinnamon Bread	Wheat	Bread	Nothing Breader	Expired
2	7	1	Cinnamon Bread Loaf	Wheat	Bread	Nothing Breader	Current
3	2	2	Milk	Dairy	Liquid	Buffalo Farms	Current
4	3	3	Chocolate Chip Cookies	Candy	Cookies	Nothing Breader	Current
5	4	4	Eggs	Poultry	Solid	Rochester Farms	Current
6	5	5	Rotini	Wheat	Pasta	Buffalo Farms	Current
7	6	6	Sugary Cereal	Wheat	Cereal	Food For You	Current

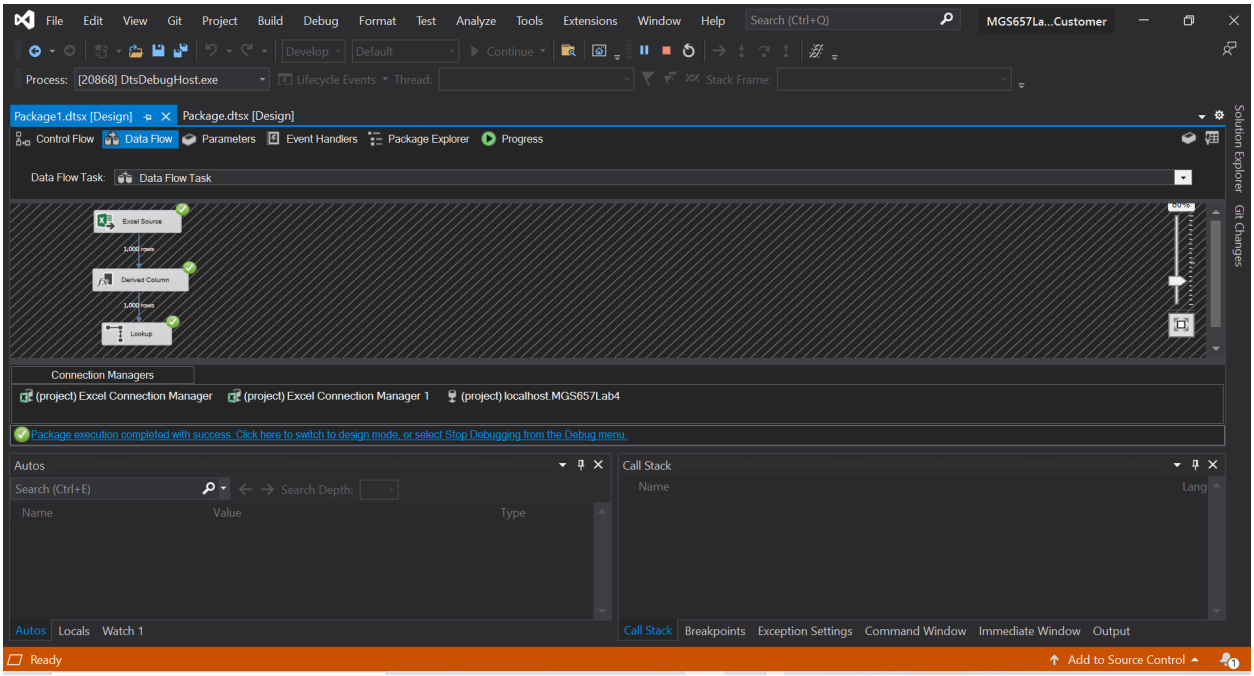
SSIS package flow for Customer Dimension:



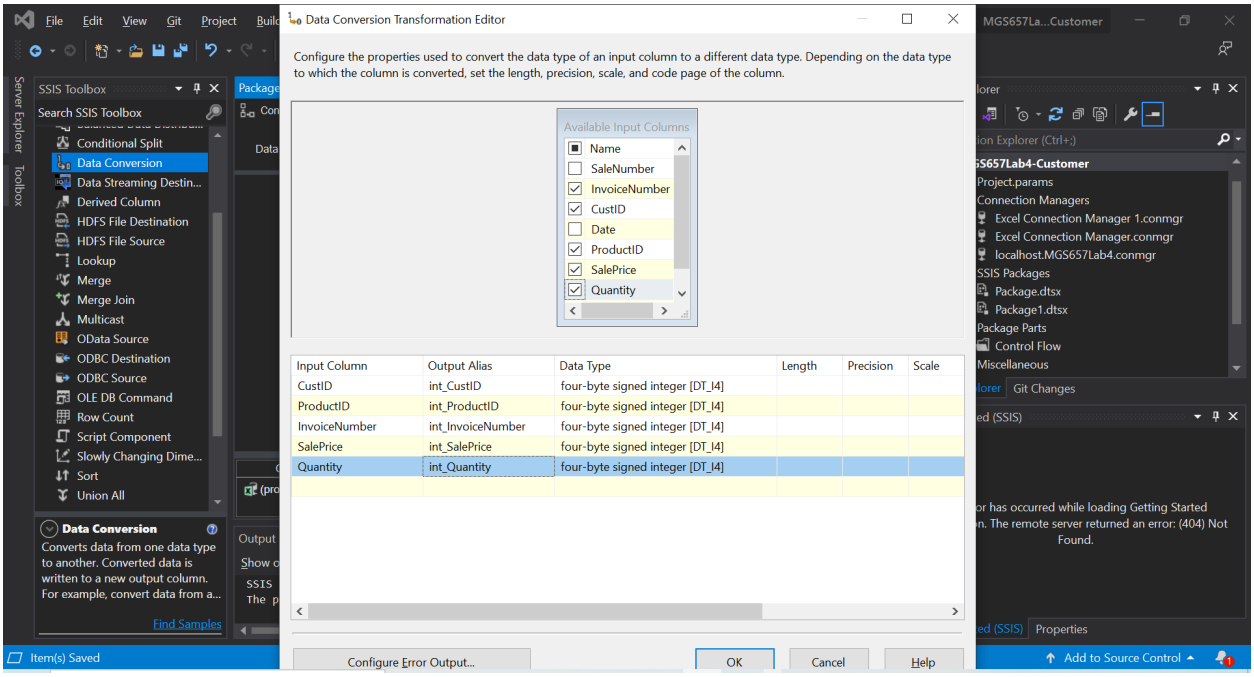
Dim_Customer:



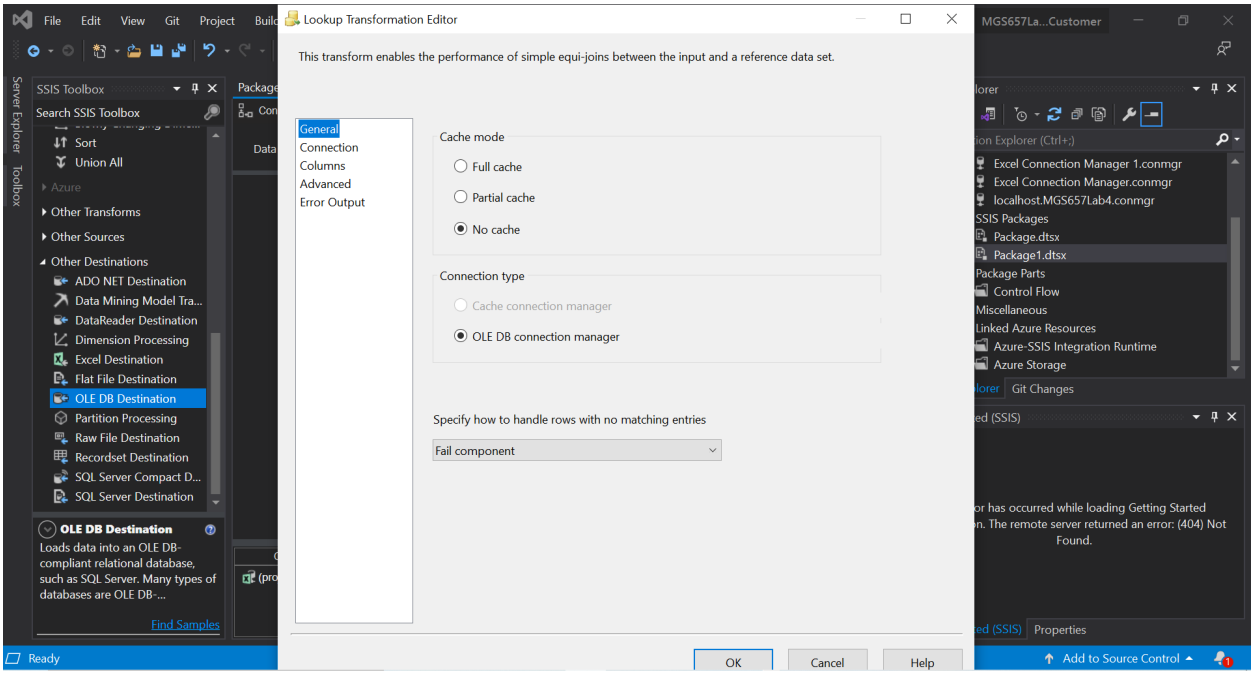
SalesUpdate_lookup_package:

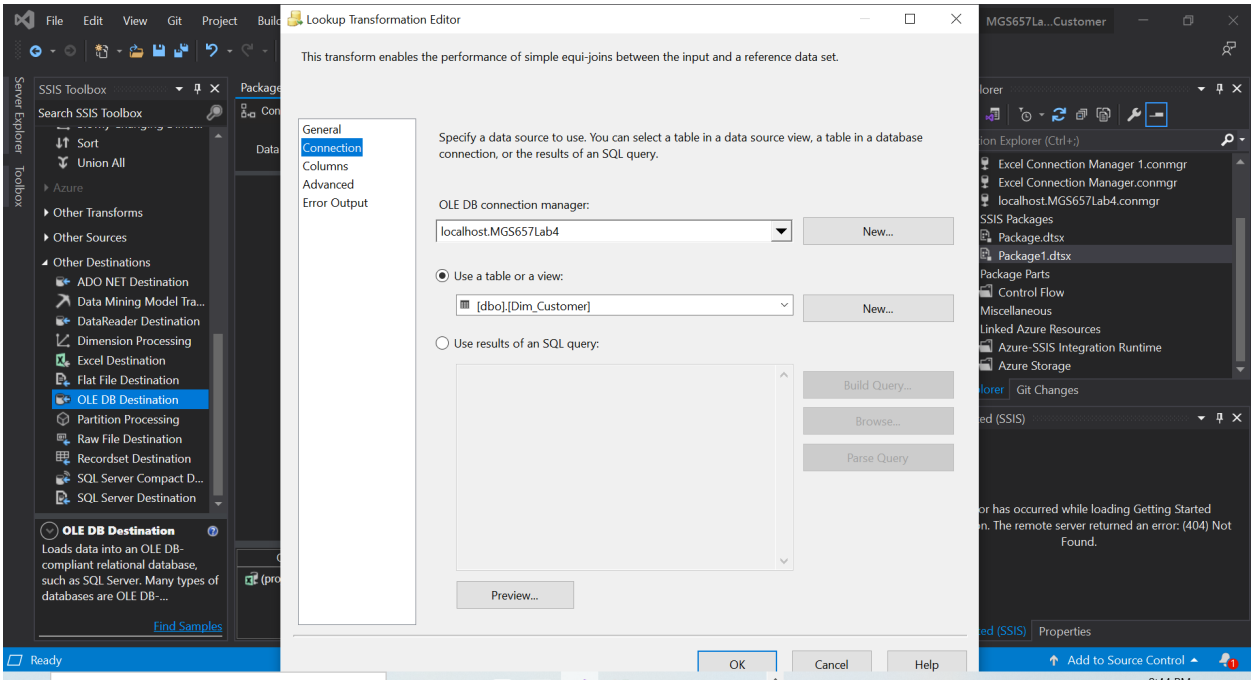


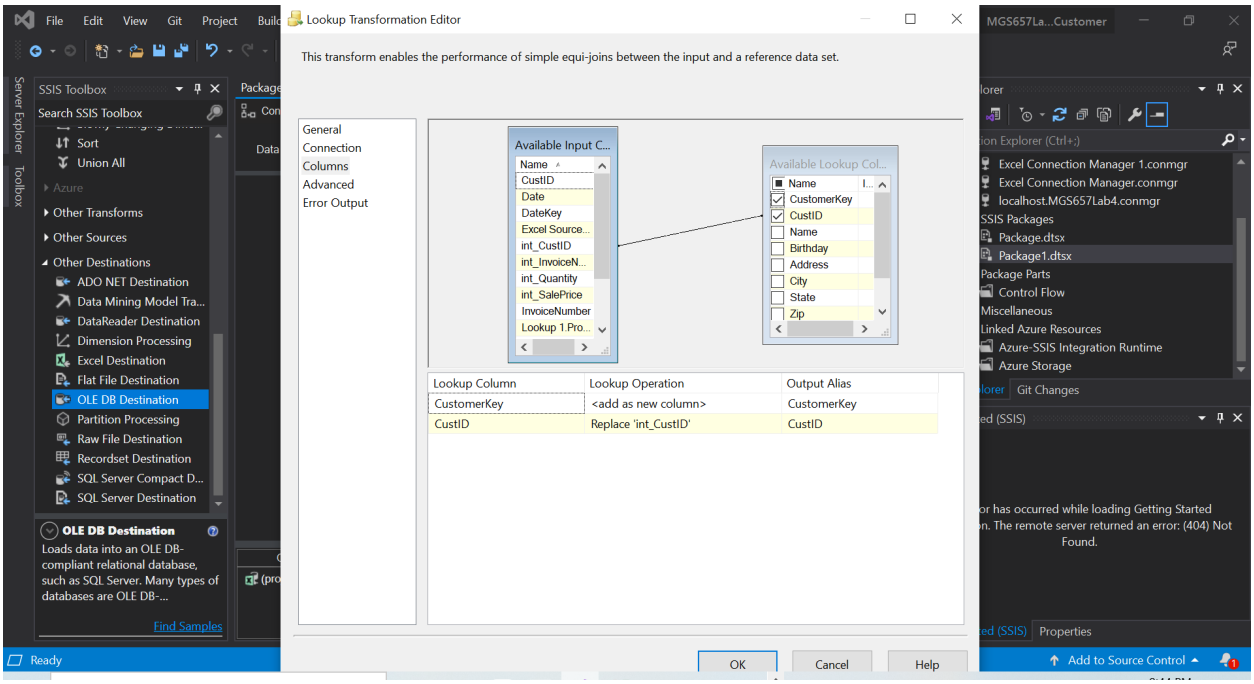
Data conversion transformation editor screenshot:

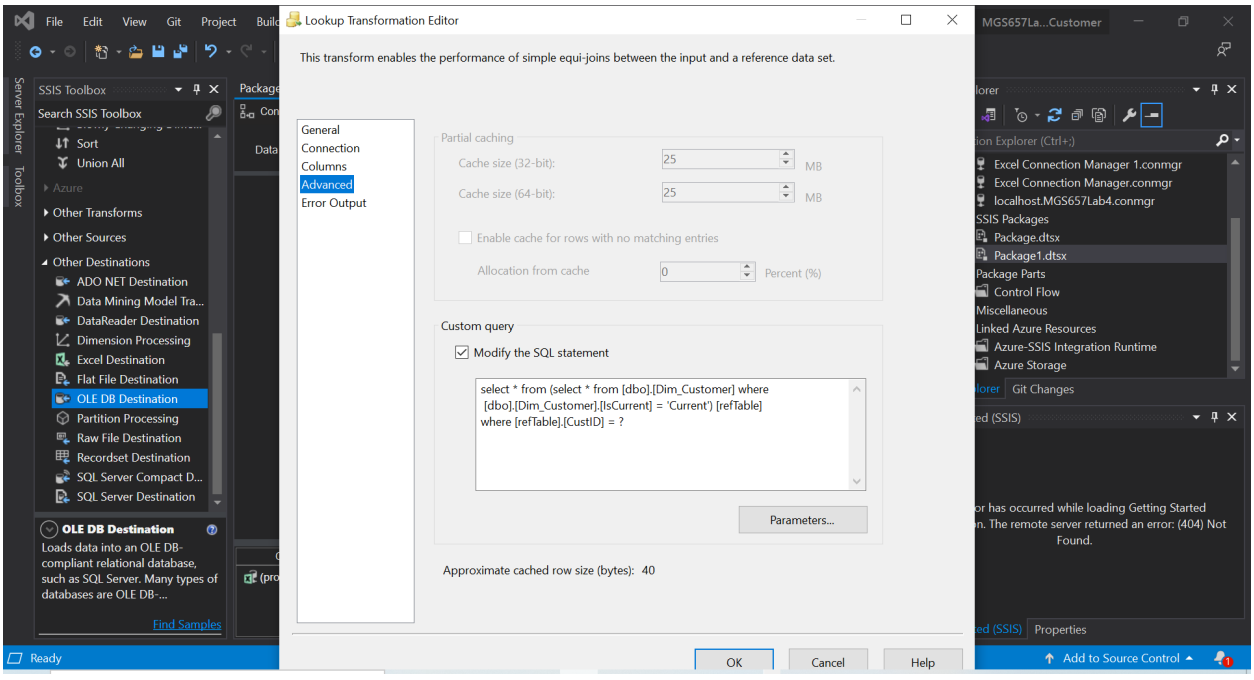


Lookup task for customer dimension:

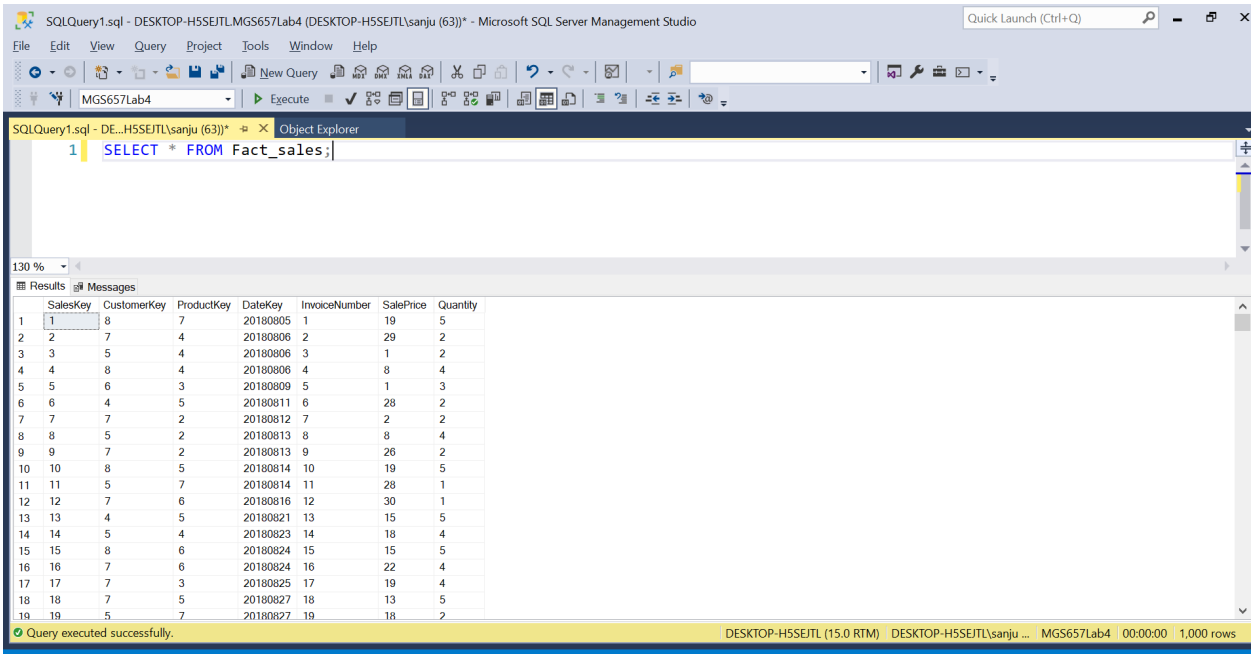





Lookup Transformation Editor
This transform enables the performance of simple equi-joins between the input and a reference data set.
General
Connection
Columns
Advanced
Error Output
Available Input C...
Name
CustID
Date
DateKey
Excel Source...
int_CustID
int_InvoiceN...
int_Quantity
int_SalePrice
InvoiceNumber
Lookup 1.Pro...
Available Lookup Col...
Name
CustomerKey
CustID
Name
Birthday
Address
City
State
Zip
Lookup Column
Lookup Operation
Output Alias
CustomerKey
<add as new column>
CustomerKey
CustID
Replace 'int_CustID'
CustID
OK
Cancel
Help
SSIS Toolbox
Search SSIS Toolbox
Sort
Union All
Other Transforms
Other Sources
Other Destinations
ADO NET Destination
Data Mining Model Tra...
DataReader Destination
Dimension Processing
Excel Destination
Flat File Destination
OLE DB Destination
Partition Processing
Raw File Destination
Recordset Destination
SQL Server Compact D...
SQL Server Destination
OLE DB Destination
Loads data into an OLE DB-compliant relational database, such as SQL Server. Many types of databases are OLE DB-...
Find Samples
Ready
Add to Source Control


Lookup Transformation Editor
This transform enables the performance of simple equi-joins between the input and a reference data set.
General
Connection
Columns
Advanced
Error Output
Partial caching
Cache size (32-bit):
25
MB
Cache size (64-bit):
25
MB
Enable cache for rows with no matching entries
Allocation from cache
0
Percent (%)
Custom query
Modify the SQL statement
select * from (select * from [dbo].[Dim_Customer] where [dbo].[Dim_Customer].[IsCurrent] = 'Current') [refTable] where [refTable].[CustID] = ?
Parameters...
Approximate cached row size (bytes): 40
OK
Cancel
Help
Ready
Add to Source Control

Fact_sales:


SQLQuery1.sql - DESKTOP-H5SEJTL.MGS657Lab4 (DESKTOP-H5SEJTL\sanju (63))* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Query Project Tools Window Help
New Query
MGS657Lab4
Execute
SQLQuery1.sql - DE...H5SEJTL\sanju (63))*
Object Explorer
1 SELECT * FROM Fact_sales;
130 %
Results
Messages
SalesKey CustomerKey ProductKey DateKey InvoiceNumber SalePrice Quantity
1 1 8 7 20180805 1 19 5
2 2 7 4 20180806 2 29 2
3 3 5 4 20180806 3 1 2
4 4 8 4 20180806 4 8 4
5 5 6 3 20180809 5 1 3
6 6 4 5 20180811 6 28 2
7 7 7 2 20180812 7 2 2
8 8 5 2 20180813 8 8 4
9 9 7 2 20180813 9 26 2
10 10 8 5 20180814 10 19 5
11 11 5 7 20180814 11 28 1
12 12 7 6 20180816 12 30 1
13 13 4 5 20180821 13 15 5
14 14 5 4 20180823 14 18 4
15 15 8 6 20180824 15 15 5
16 16 7 6 20180824 16 22 4
17 17 7 3 20180825 17 19 4
18 18 7 5 20180827 18 13 5
19 19 5 7 20180827 19 18 2
Query executed successfully.
DESKTOP-H5SEJTL (15.0 RTM)
DESKTOP-H5SEJTL\sanju ...
MGS657Lab4
00:00:00
1,000 rows