Assignment No. 1

Title:

Importing and loading of legacy data from different sources.

Learning Objectives:

Import the legacy data from different sources such as (Excel, OData feed Access, Web, SQL Server etc.) and load in the target system using Power BI tool.

Problem Statement:

Import the legacy data from different sources such as (SQL Server, Oracle etc.) and load in the target system. (You can download sample database such as Adventure works, Northwind, food mart etc.)

Theory Concepts:

Power BI

Microsoft Power BI is business intelligence (BI) platform that provides nontechnical business users with tools for aggregating, analyzing, visualizing, and sharing data. Power BI's user interface is fairly intuitive for users familiar with Excel, and its deep integration with other Microsoft products makes it a versatile self-service tool that requires little upfront training.

Users can download an application for Windows 10, called Power BI Desktop, and native mobile apps for Windows, Android and iOS devices. There is also Power BI Report Server for companies that must maintain their data and reports on premises. That version of Power BI requires a special version of the desktop app -- aptly called Power BI Desktop for Power BI Report Server.

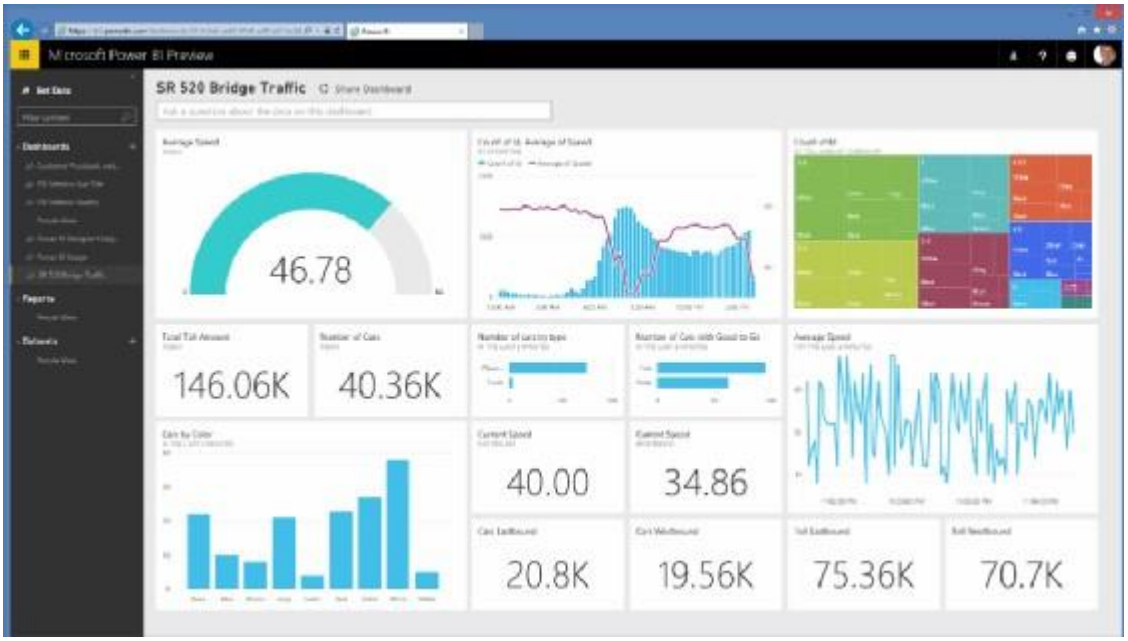
Common uses of Power BI

Microsoft Power BI is used to find insights within an organization's data. Power BI can help connect disparate data sets, transform, and clean the data into a data model and create charts or graphs to provide visuals of the data. All of this can be shared with other Power BI users within the organization.

The data models created from Power BI can be used in several ways for organizations, including the following:

- telling stories through charts and data visualizations;
- examining "what if" scenarios within the data; and
- creating reports that can answer questions in real time and help with forecasting to make sure departments meet business metrics.

  Power BI can also provide executive dashboards for administrators or managers, giving management more insight into how departments are doing.



MICROSOFT

This Power BI preview shows the reporting and dashboard capabilities that Power BI offers. **Who uses Power BI?**

Though Power BI is a self-service BI tool that brings data analytics to employees, it's mostly used by data analysts and BI professionals who create the data models before disseminating reports throughout the organization. However, those without an analytical background can still navigate Power BI and create reports.

Microsoft Power BI is used by both department reps and management, with reports and forecasts created to aid sales and marketing reps, while also providing data for management on how the department or individual employees are progressing toward their goals.

In addition, Power BI offers an admin portal for administrators to help configure the implementation of Power BI, as well as usage monitoring and licenses.

Power BI components

Microsoft Power BI works by connecting data sources and providing a dashboard of BI to the users. It can connect with just an Excel spreadsheet or bring together cloud-based and on-premises data warehouses. Data pulled from cloud-based sources, such as Salesforce CRM, is automatically refreshed.

With applications such as an Excel workbook or Power BI Desktop file connected to online or on-premises data sources, Power BI users must manually refresh or setup a refresh schedule to ensure the data in Power BI reports and dashboards use the most current data available.

Power BI consists of a collection of apps and can be used either on desktop, as a SaaS product or on a mobile device. Power BI Desktop is the on-premises version, Power BI Service is the cloud-based offering and mobile Power BI runs on mobile devices.

The different components of Power BI are meant to let users create and share business insights in a way that fits with their role.
Included within Power BI are several components that help users create and share data reports. Those are the following:

- **Power Query:** a data mashup and transformation tool
- **Power Pivot:** a memory tabular data modeling tool
- **Power View:** a data visualization tool
- **Power Map:** a 3D geospatial data visualization tool
- **Power Q&A:** a natural language question and answering engine

Additionally, there are dozens of data sources that connect into Power BI, ranging from files (Excel, PDF, SharePoint, XML), databases (SQL Server Database, Oracle Database, IBM databases, Amazon Redshift, Google Big Query), other Power BI data sets, Azure data connections and many online services (Dynamics 365, Salesforce Reports, Google Analytics, Adobe Analytics, Facebook and others).

How to use Power BI

Power BI Desktop is where analysts and other users can create data connections, data models and reports. The Power BI service is where those reports can be shared, so other users can view and interact with them.

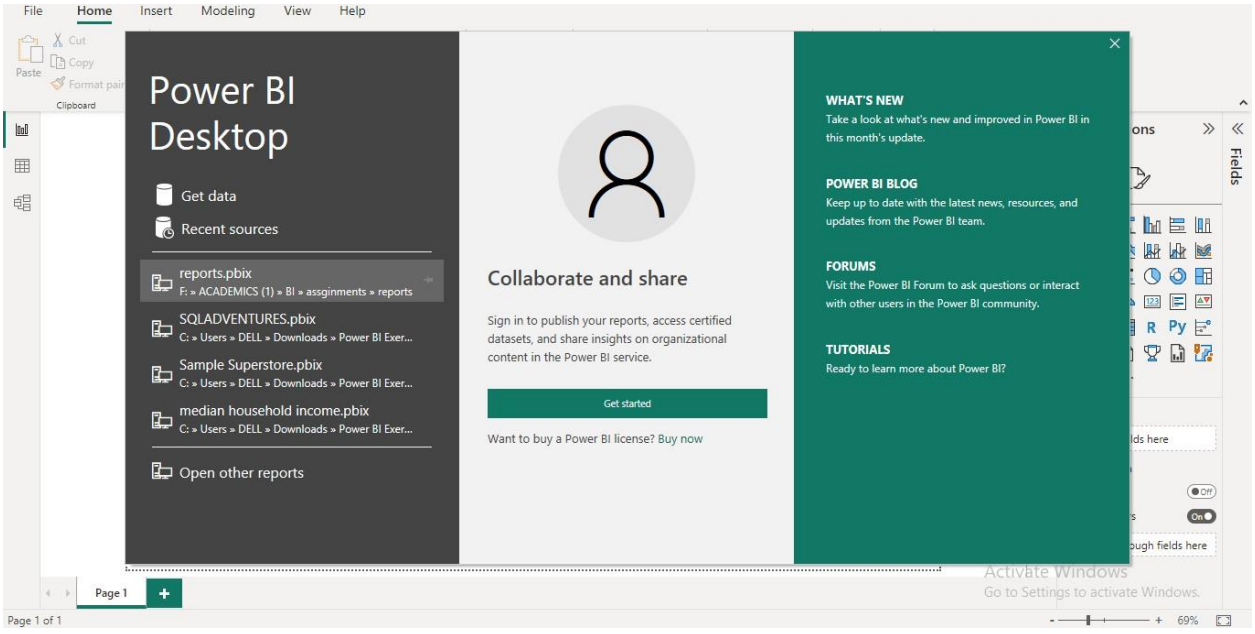
To build a Power BI report, take the following steps:

- Connect the data sources.
- Query the data to create reports based on user needs.
- Publish the report to the Power BI service.
- Share the report, so cloud and mobile users can see and interact with it.
- Add permissions to give colleagues the ability to edit reports or create dashboards or limit their ability to edit.

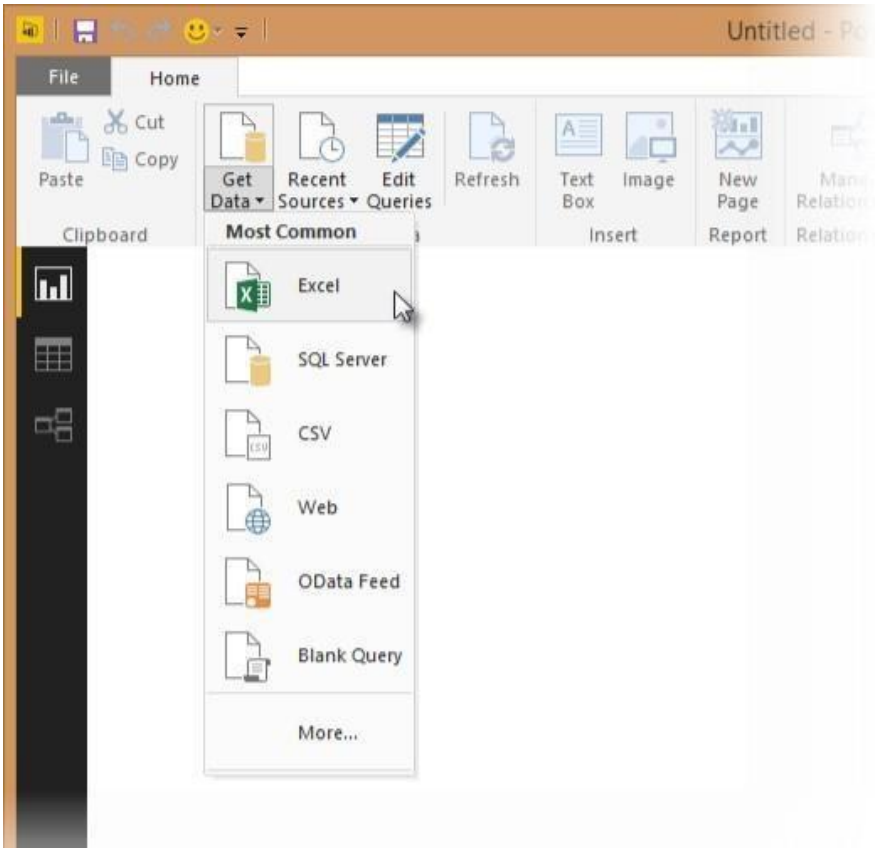
Import the legacy data from different sources such as (Excel, OData feed, Access, Web , SQL Server etc.) and load in the target system using Power BI tool.

Importing Excel Data

1) Launch Power BI Desktop.



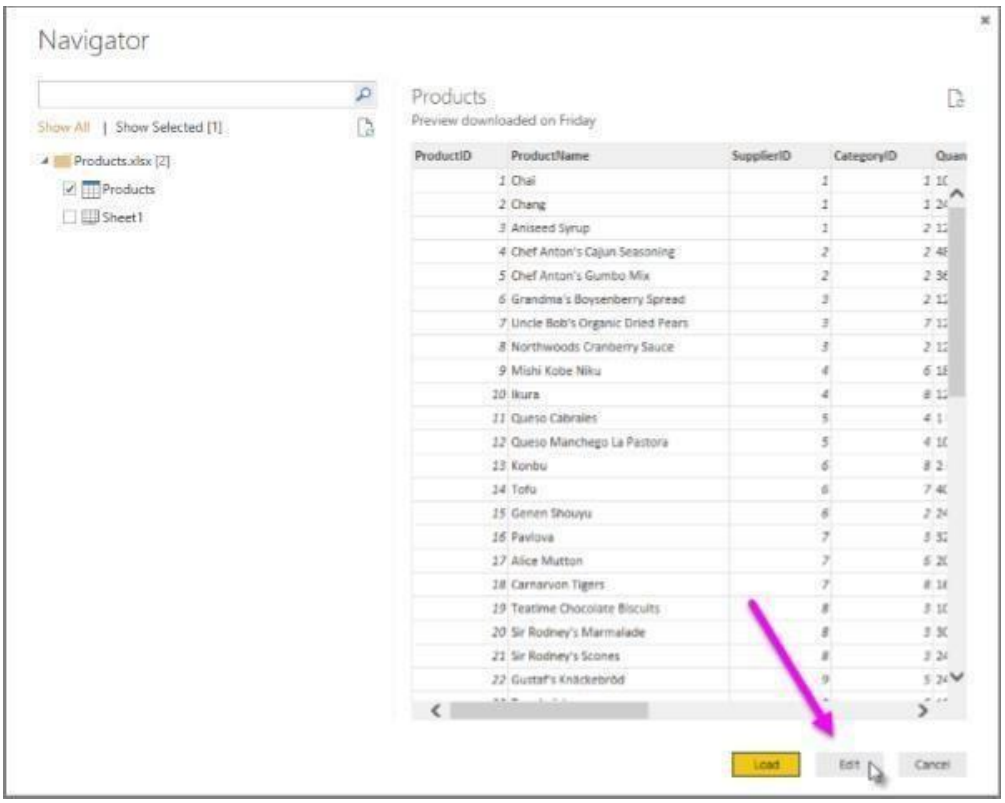
2) From the Home ribbon, select Get Data. Excel is one of the Most Common data connections, so you can select it directly from the Get Data menu.



3) If you select the Get Data button directly, you can also select FIle> Excel and select Connect.

4) In the Open File dialog box, select theProducts.xlsx file.

5) In the Navigator pane, select the Products table and then select Edit.



Importing Data from OData Feed

In this task, you will bring in order data. This step represents connecting to a sales system. You import data into Power BI Desktop from the sample Northwind OData feed at the following URL, which you can copy (and then paste) in the steps below: http://services.odata.org/V3/Northwind/Northwind.svc/
Connect to an OData feed:

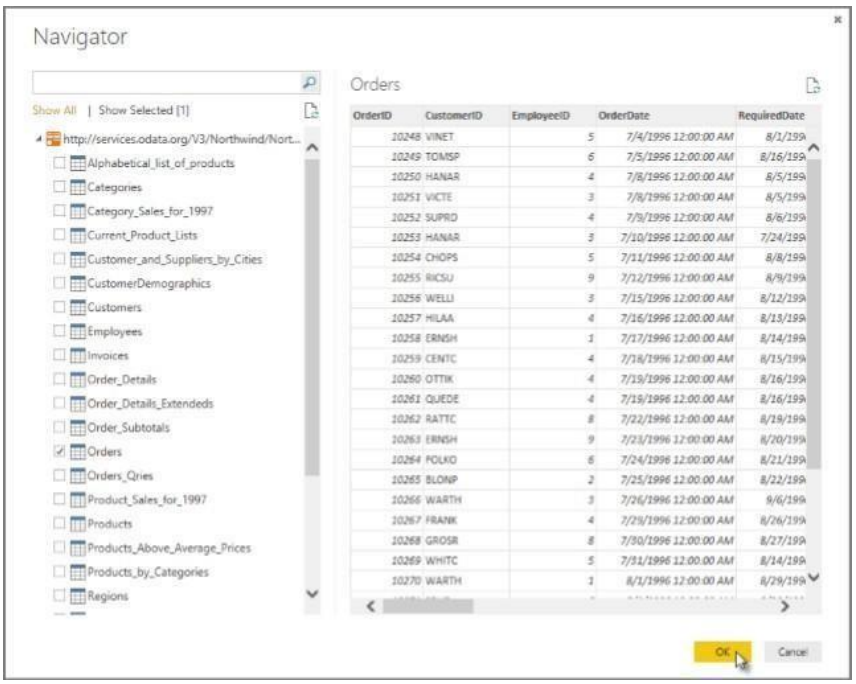
1) From the Home ribbon tab in Query Editor, select Get Data.

2) Browse to the OData Feed data source.

3) In the OData Feed dialog box, paste the URL for the Northwind OData feed.

4) Select OK.

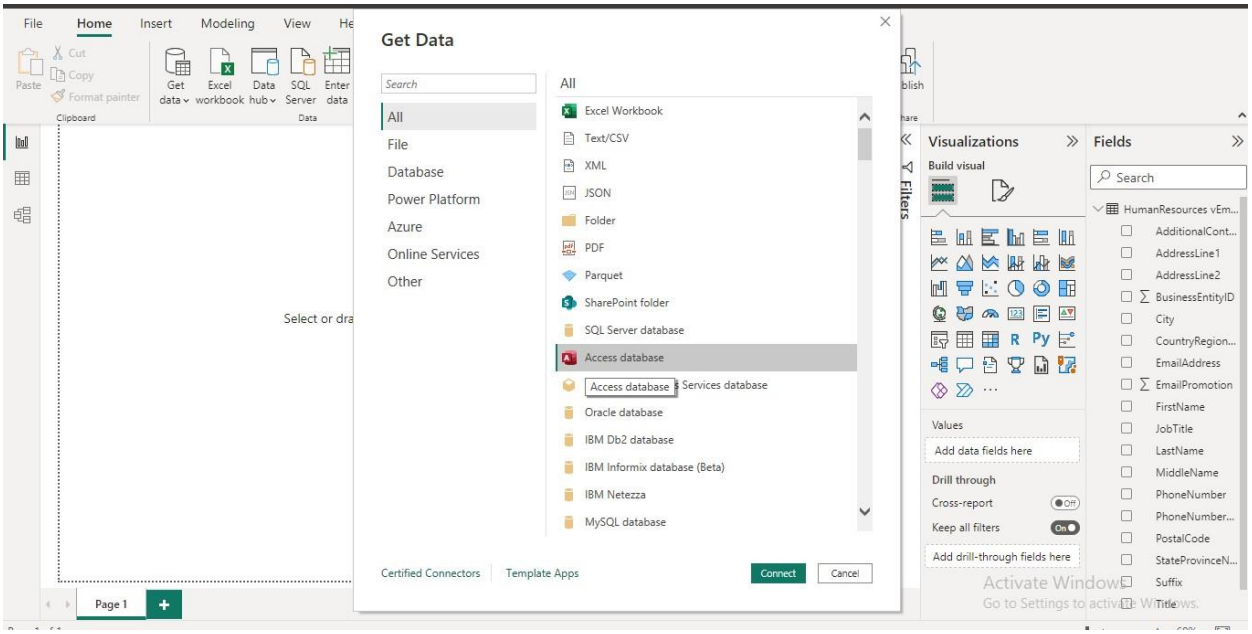
5) In the Navigator pane, select the Orders table, and then select Edit.



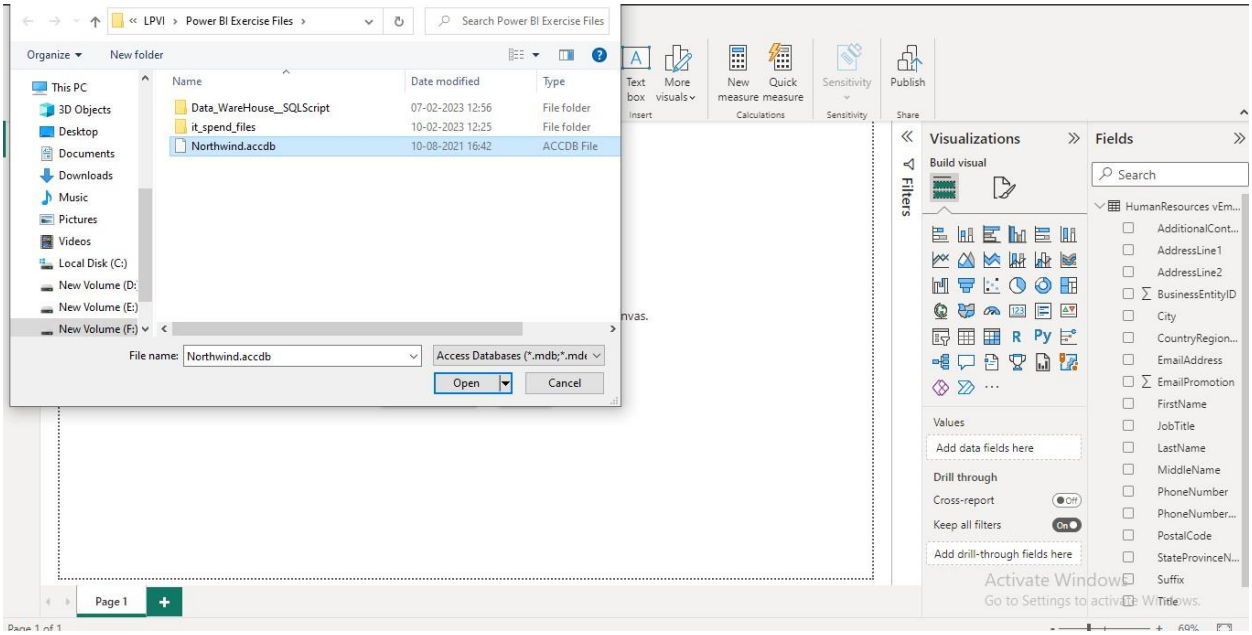
Note- You can click a table name, without selecting the checkbox, to see a preview.

Import Data from ACCESS database

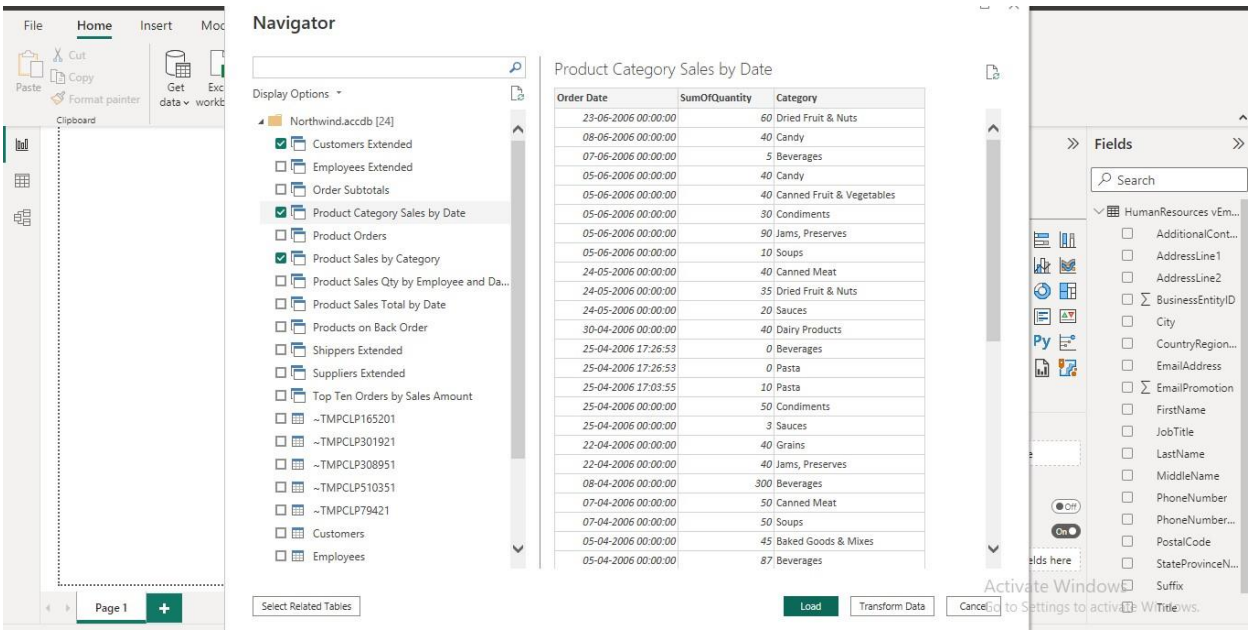
1. From the Home ribbon, select Get Data.
2. Get Data -> **Access database**

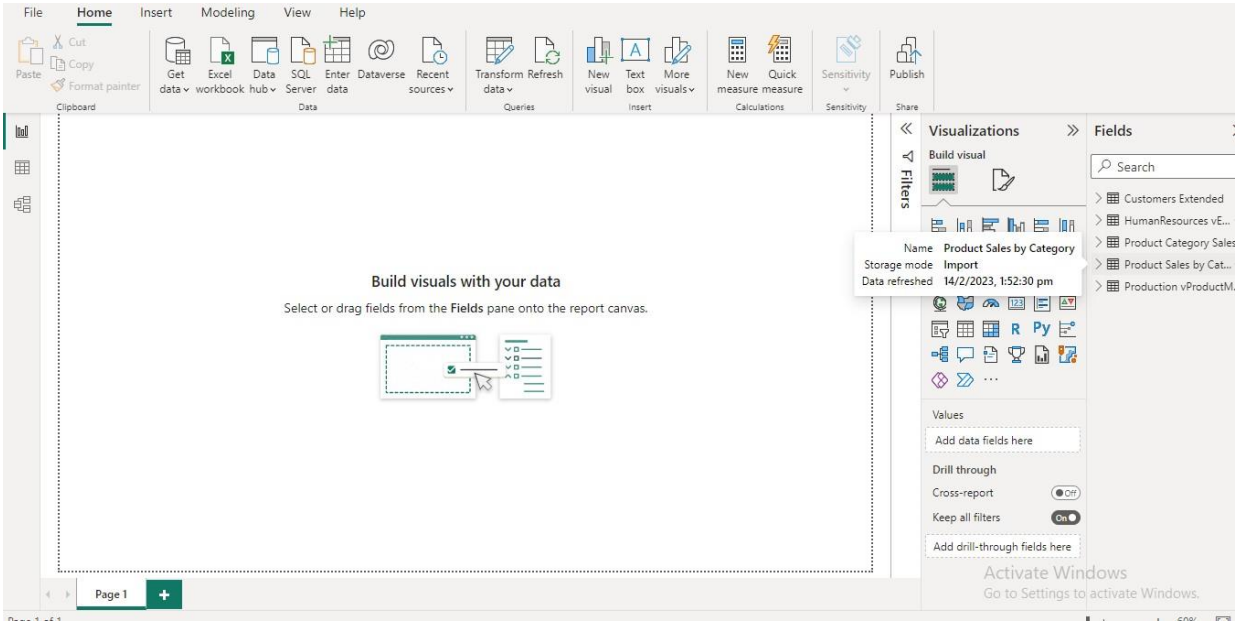


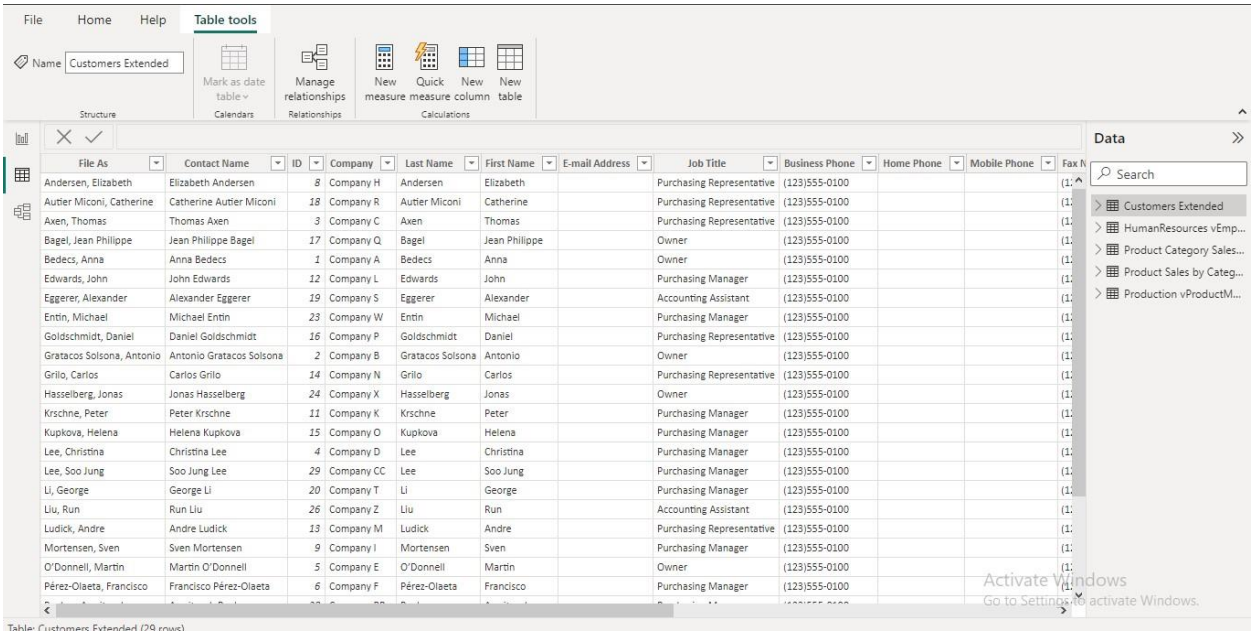
3. Select the access database



4. From the database select the table and click load





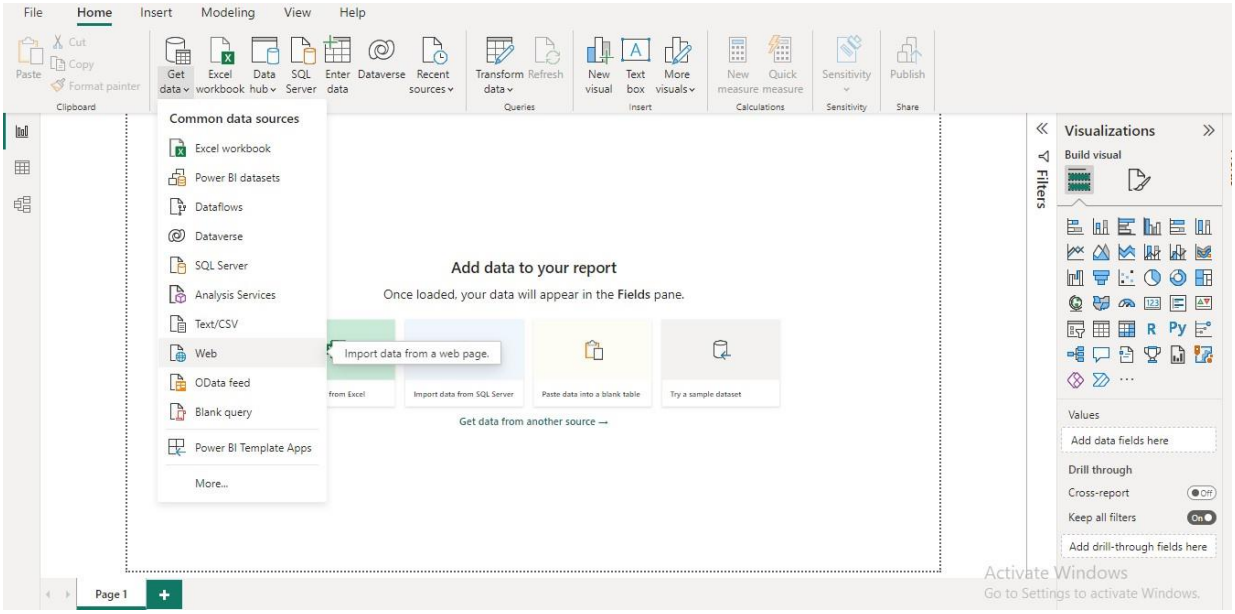


File As	Contact Name	ID	Company	Last Name	First Name	E-mail Address	Job Title	Business Phone	Home Phone	Mobile Phone
Andersen, Elizabeth	Elizabeth Andersen	8	Company H	Andersen	Elizabeth		Purchasing Representative	(123)555-0100		
Autier Miconi, Catherine	Catherine Autier Miconi	18	Company R	Autier Miconi	Catherine		Purchasing Representative	(123)555-0100		
Axen, Thomas	Thomas Axen	3	Company C	Axen	Thomas		Purchasing Representative	(123)555-0100		
Bagel, Jean Philippe	Jean Philippe Bagel	17	Company Q	Bagel	Jean Philippe		Owner	(123)555-0100		
Bedecs, Anna	Anna Bedecs	1	Company A	Bedecs	Anna		Owner	(123)555-0100		
Edwards, John	John Edwards	12	Company L	Edwards	John		Purchasing Manager	(123)555-0100		
Eggerer, Alexander	Alexander Eggerer	19	Company S	Eggerer	Alexander		Accounting Assistant	(123)555-0100		
Entin, Michael	Michael Entin	23	Company W	Entin	Michael		Purchasing Manager	(123)555-0100		
Goldschmidt, Daniel	Daniel Goldschmidt	16	Company P	Goldschmidt	Daniel		Purchasing Representative	(123)555-0100		
Gratacos Solsona, Antonio	Antonio Gratacos Solsona	2	Company B	Gratacos Solsona	Antonio		Owner	(123)555-0100		
Grilo, Carlos	Carlos Grilo	14	Company N	Grilo	Carlos		Purchasing Representative	(123)555-0100		
Hasselberg, Jonas	Jonas Hasselberg	24	Company X	Hasselberg	Jonas		Owner	(123)555-0100		
Krschne, Peter	Peter Krschne	11	Company K	Krschne	Peter		Purchasing Manager	(123)555-0100		
Kupkova, Helena	Helena Kupkova	15	Company O	Kupkova	Helena		Purchasing Manager	(123)555-0100		
Lee, Christina	Christina Lee	4	Company D	Lee	Christina		Purchasing Manager	(123)555-0100		
Lee, Soo Jung	Soo Jung Lee	29	Company CC	Lee	Soo Jung		Purchasing Manager	(123)555-0100		
Li, George	George Li	20	Company T	Li	George		Purchasing Manager	(123)555-0100		
Liu, Run	Run Liu	26	Company Z	Liu	Run		Accounting Assistant	(123)555-0100		
Ludick, Andre	Andre Ludick	13	Company M	Ludick	Andre		Purchasing Representative	(123)555-0100		
Mortensen, Sven	Sven Mortensen	9	Company I	Mortensen	Sven		Purchasing Manager	(123)555-0100		
O'Donnell, Martin	Martin O'Donnell	5	Company E	O'Donnell	Martin		Owner	(123)555-0100		
Pérez-Olaeta, Francisco	Francisco Pérez-Olaeta	6	Company F	Pérez-Olaeta	Francisco		Purchasing Manager	(123)555-0100		

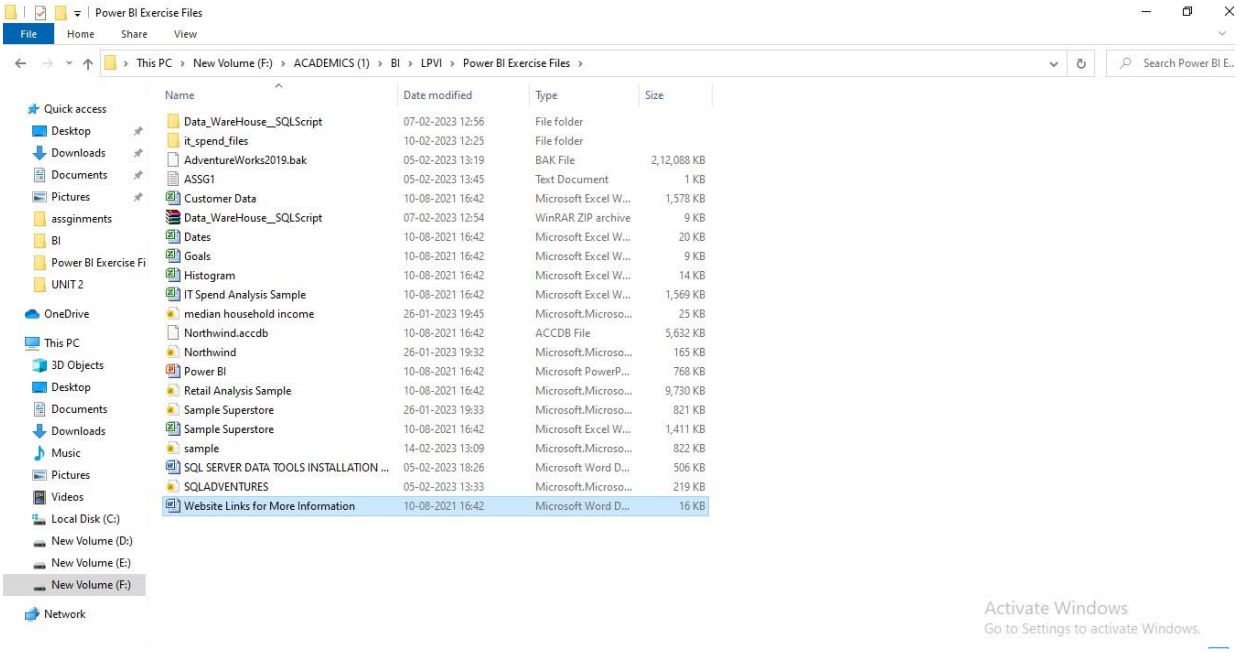
Table: Customers Extended (29 rows)

Import Data from WEB

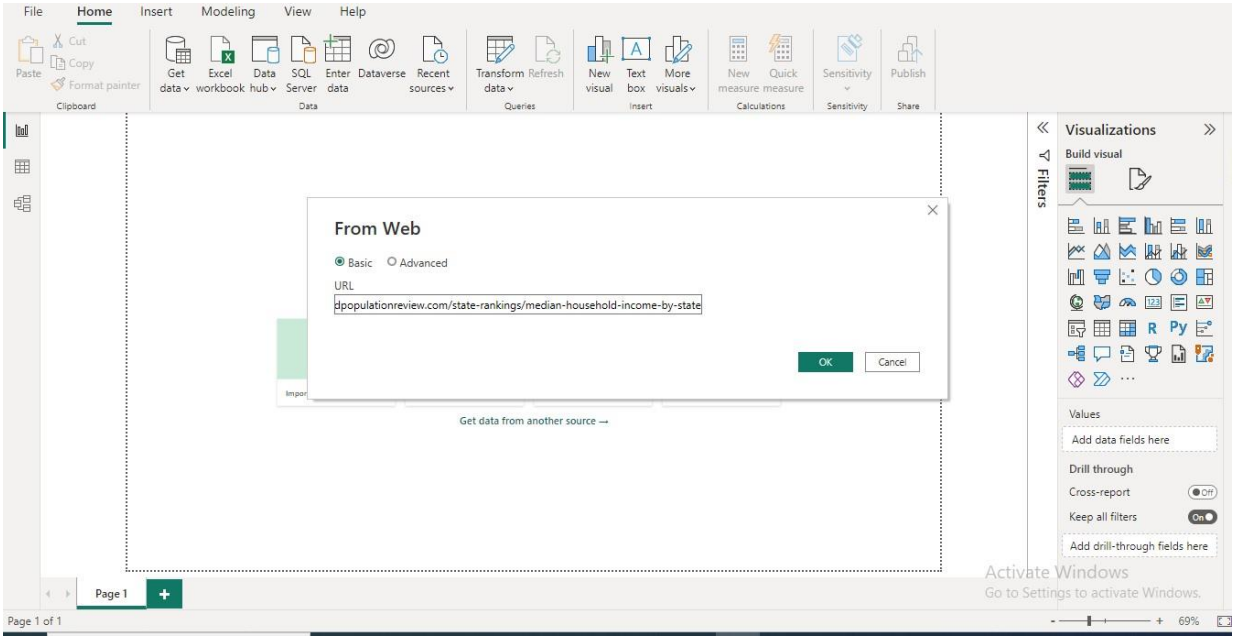
1. From the Home ribbon, select Get Data.
2. Get Data -> **web**



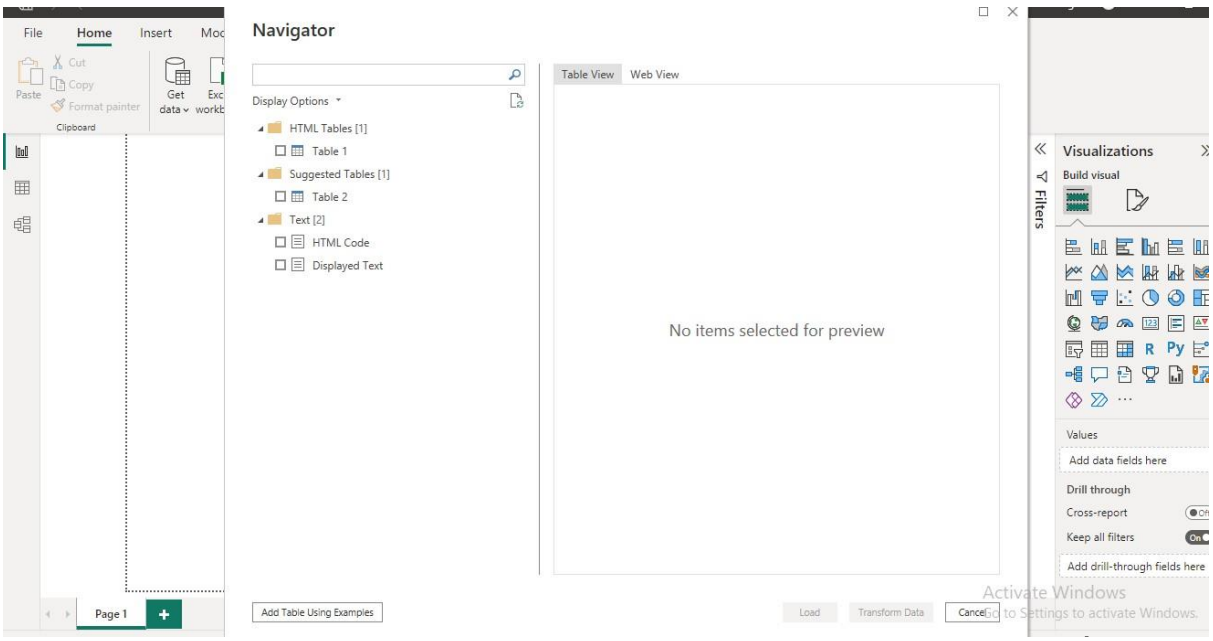
3. From the file given select any URL

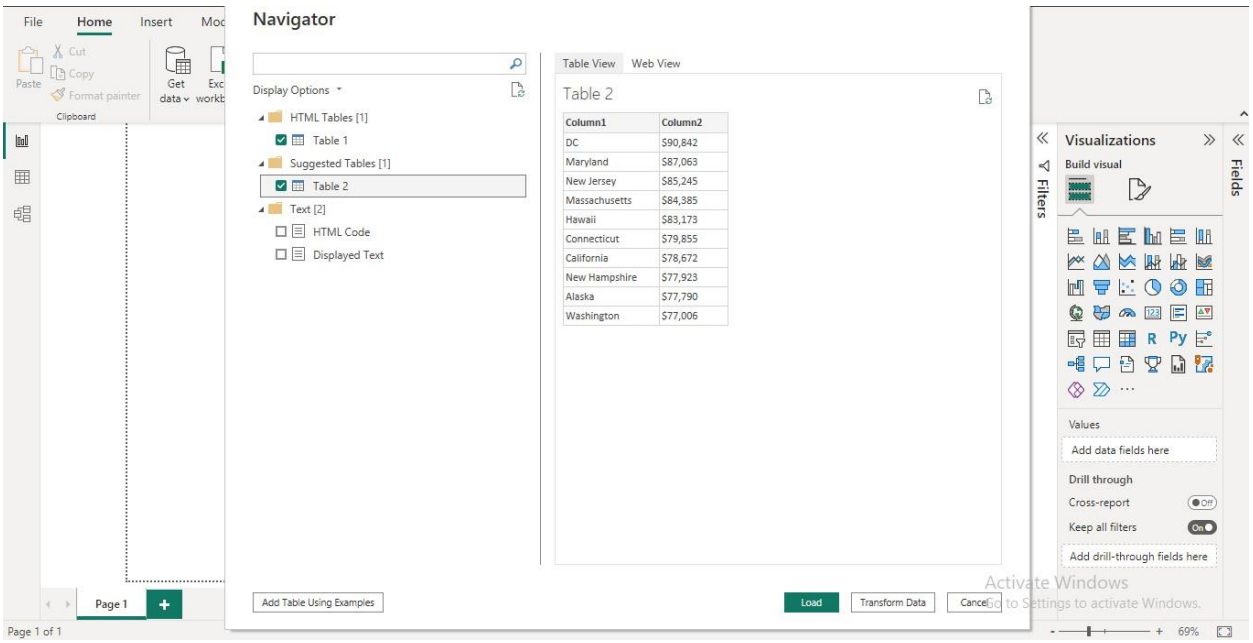


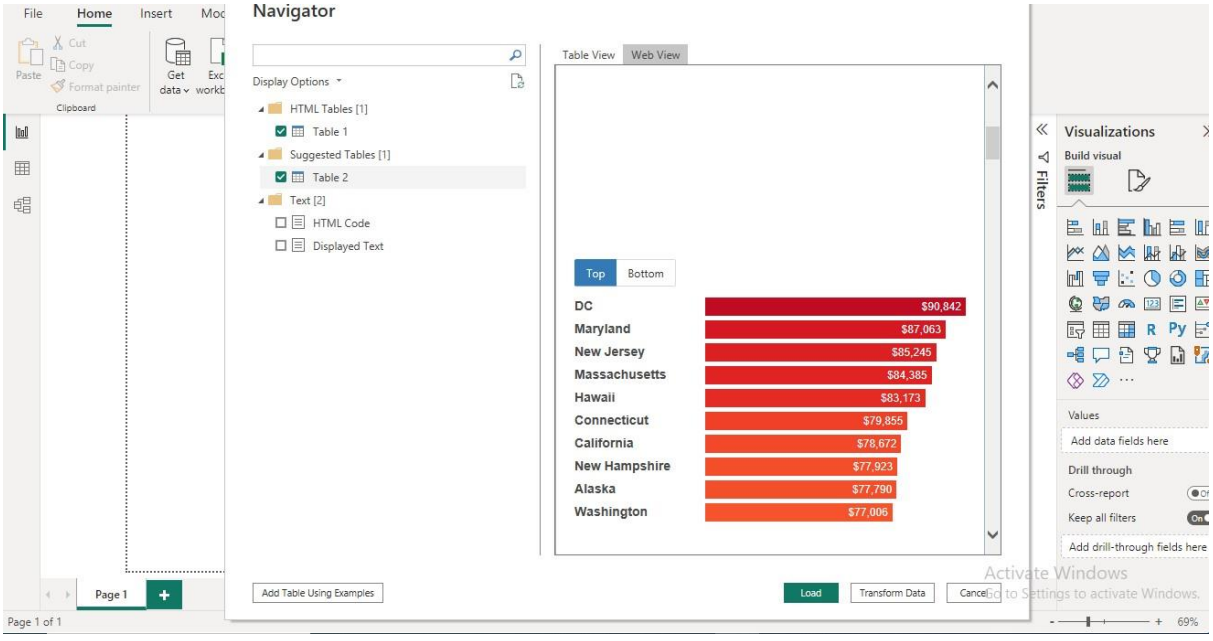
4. Enter the URL

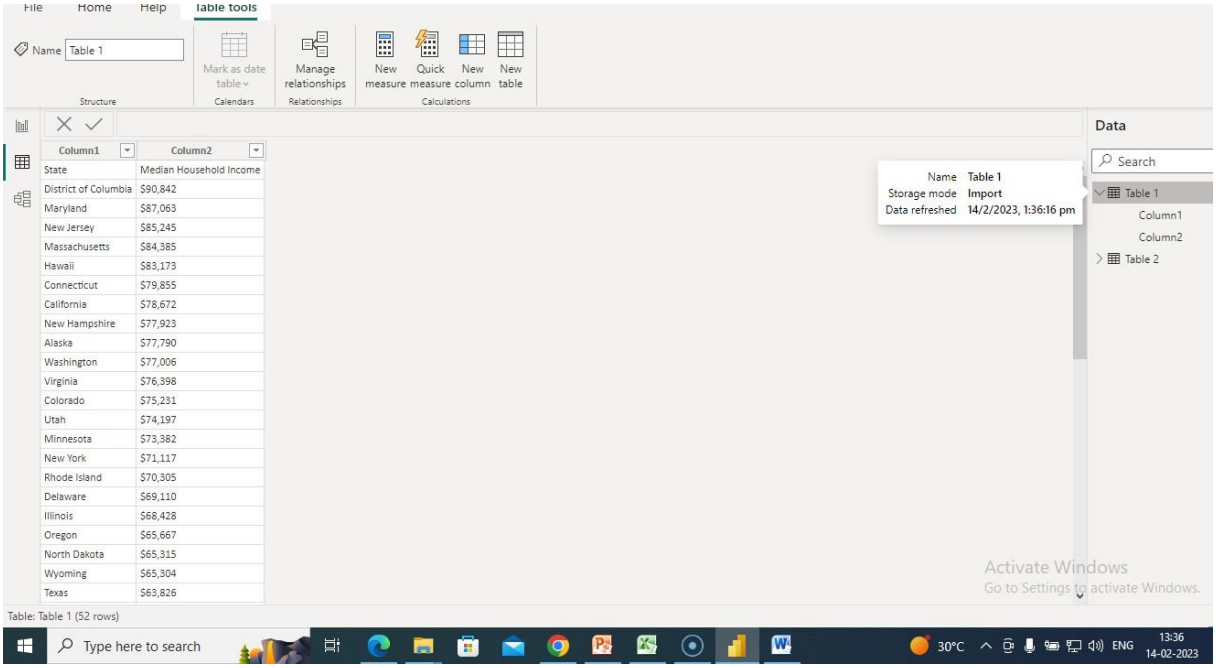


5. In navigator select the tables or code you want to display



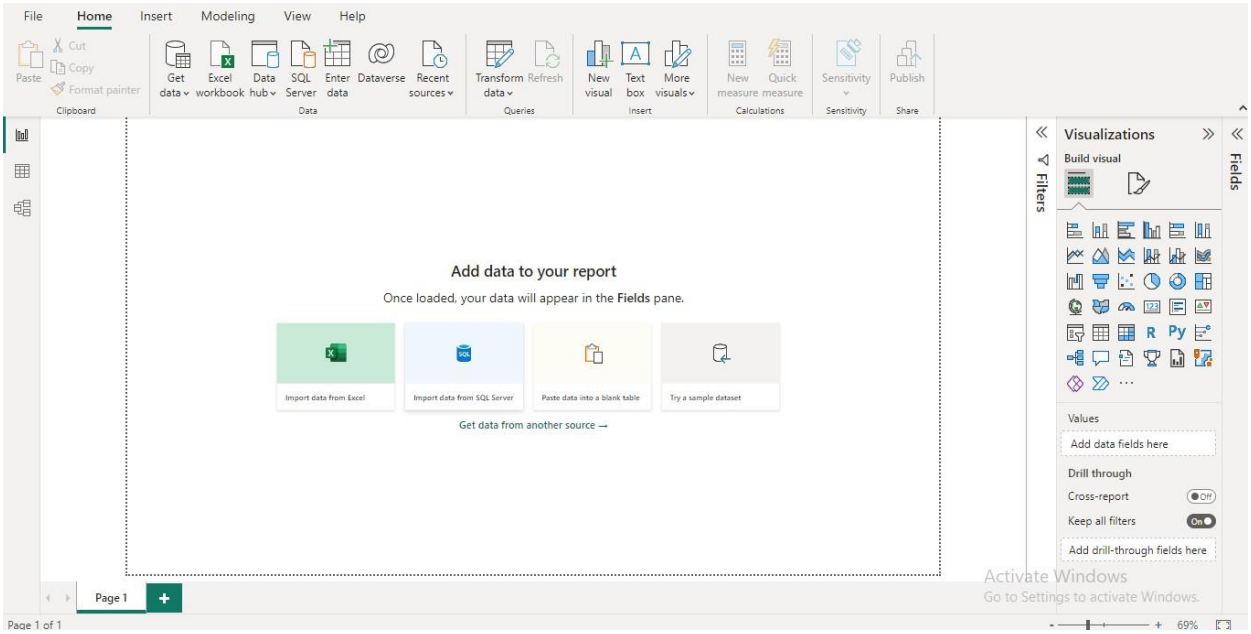

File
Home
Insert
Navigator
Display Options
HTML Tables [1]
Table 1
Suggested Tables [1]
Table 2
Text [2]
HTML Code
Displayed Text
Table View
Web View
Table 2
Column1
Column2
DC
$90,842
Maryland
$87,063
New Jersey
$85,245
Massachusetts
$84,385
Hawaii
$83,173
Connecticut
$79,855
California
$78,672
New Hampshire
$77,923
Alaska
$77,790
Washington
$77,006
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Keep all filters
Add drill-through fields here
Add Table Using Examples
Load
Transform Data
Page 1
Page 1 of 1


File
Home
Insert
Navigator
Display Options
HTML Tables [1]
Table 1
Suggested Tables [1]
Table 2
Text [2]
HTML Code
Displayed Text
Table View
Web View
Top
Bottom
DC
$90,842
Maryland
$87,063
New Jersey
$85,245
Massachusetts
$84,385
Hawaii
$83,173
Connecticut
$79,855
California
$78,672
New Hampshire
$77,923
Alaska
$77,790
Washington
$77,006
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Keep all filters
Add drill-through fields here
Add Table Using Examples
Load
Transform Data
Page 1
Page 1 of 1

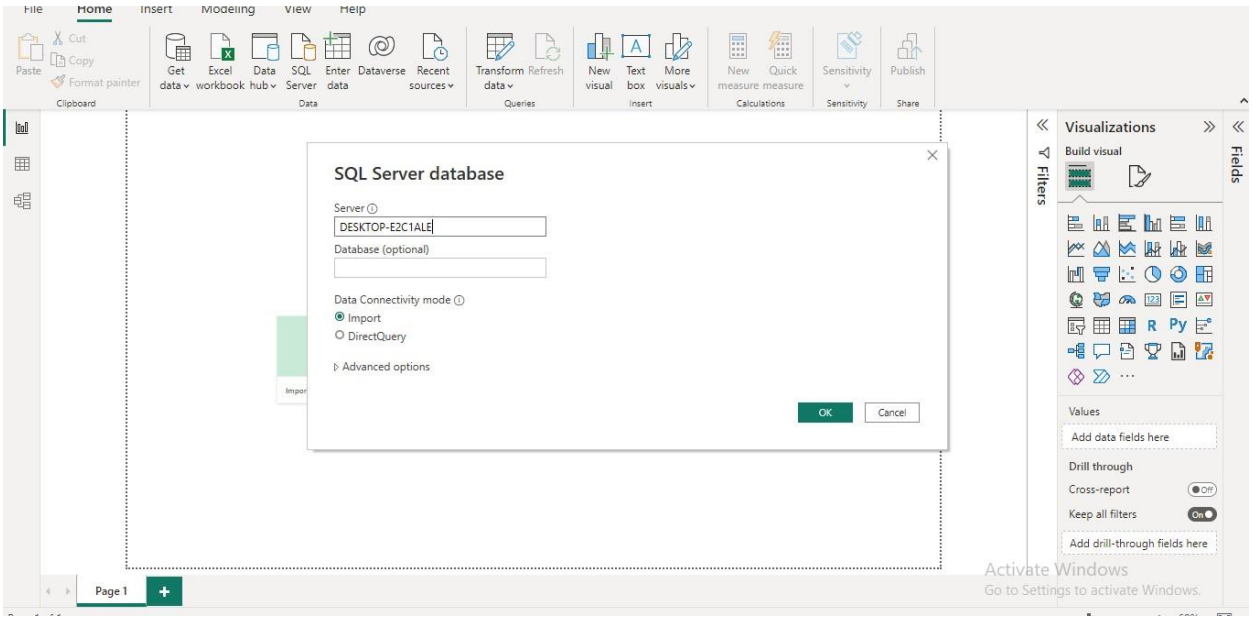


Import data from SQL SERVER 1. From the Home ribbon, select Get Data.

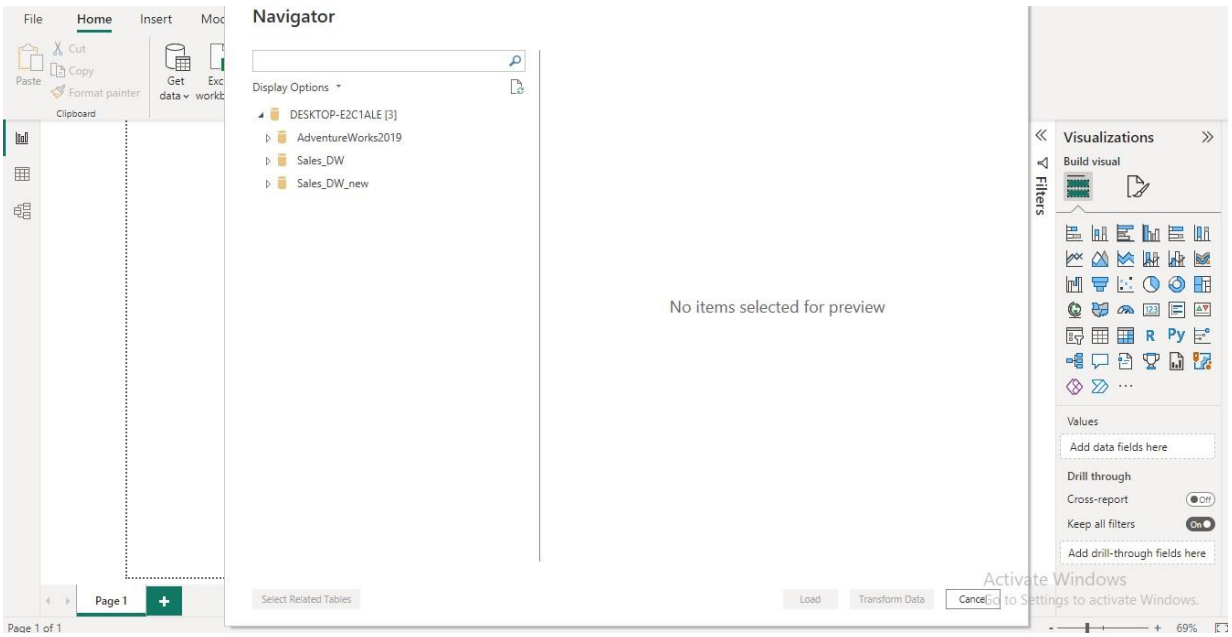
2. Get Data -> **SQL server**

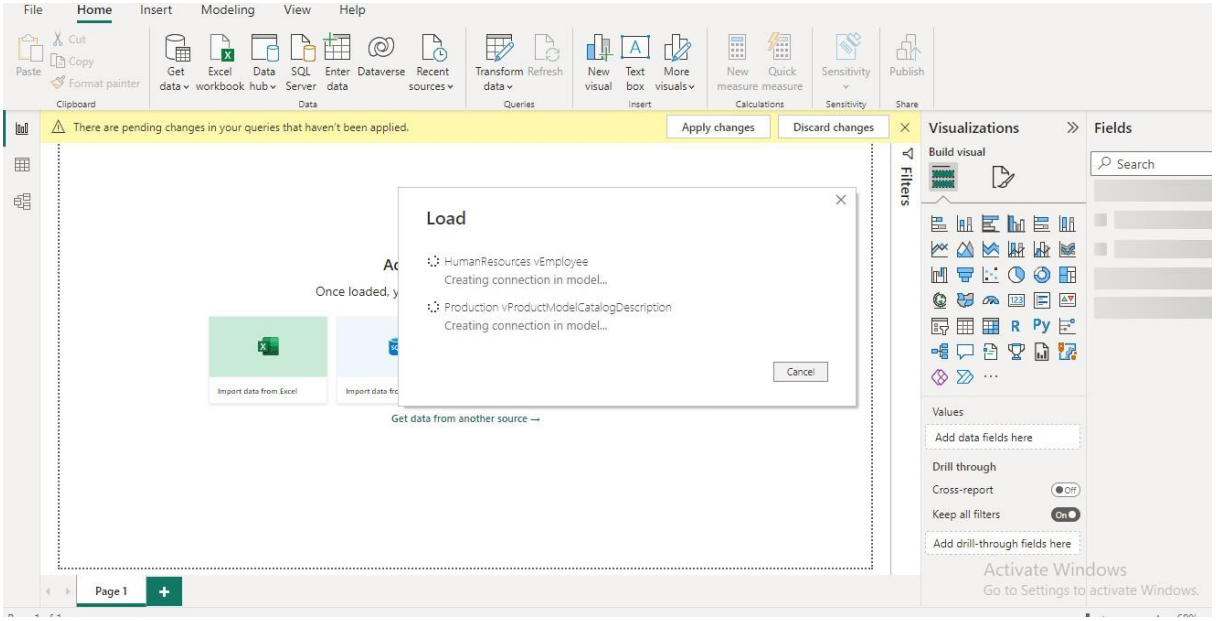


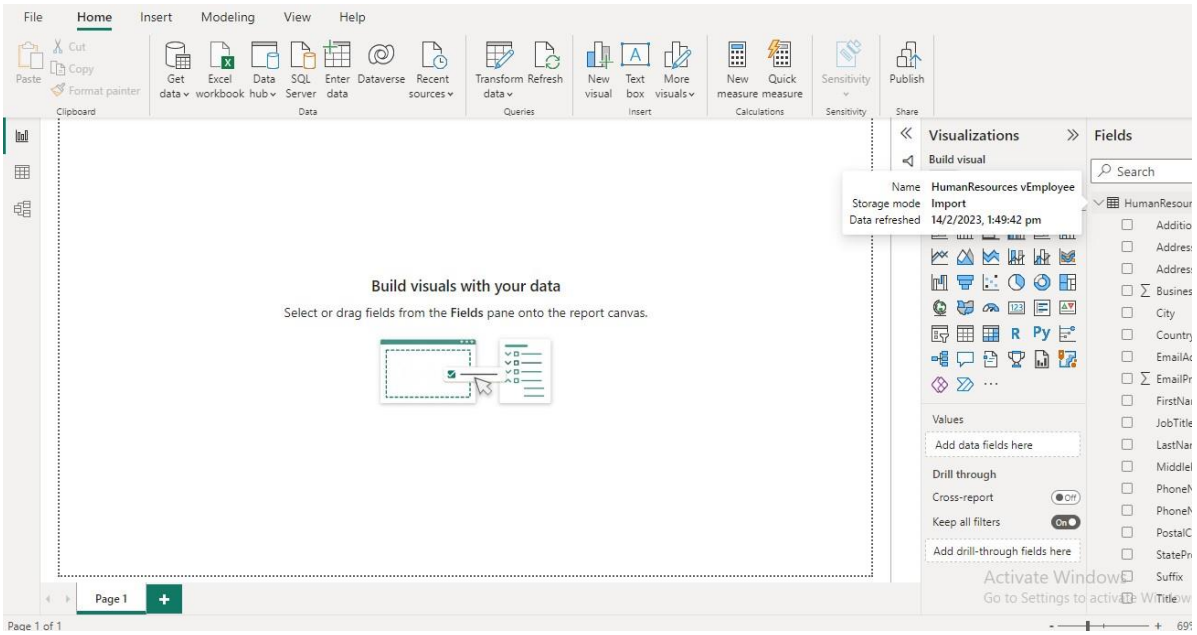
3. Enter server name



4. Select database and the tables from the database







Conclusion:

With the help of PowerBI the legacy data from different sources such as (Excel, OData feed, Access, Web , SQL Server etc.) was imported and loaded in the target system successfully.

Assignment No. 2

Title: Extraction Transformation and Loading (ETL) process

Learning Objectives:

Perform the Extraction Transformation and Loading (ETL) process to construct the database in the SQL server

Problem Statement:

Perform the Extraction Transformation and Loading (ETL) process to construct the database in the SQL server.

Theory Concepts:

Extraction, Transformation and Load

Step 1: Data Extraction:

The data extraction is first step of ETL. There are 2 Types of Data Extraction

1. Full Extraction: All the data from source systems or operational systems gets extracted to staging area. (Initial Load)

2. Partial Extraction: Sometimes we get notification from the source system to update specific date. It is called as Delta load.

Source System Performance: The Extraction strategies should not affect source system performance.

Step 2: Data Transformation:

The data transformation is second step. After extracting the data there is big need to do the transformation as per the target system. I would like to give you some bullet points of Data Transformation.

- Data Extracted from source system is in to Raw format. We need to transform it before loading in to target server.
- Data has to be cleaned, mapped and transformed
- There are following important steps of Data Transformation:

1. **Selection:** Select data to load in target
2. **Matching:** Match the data with target system
3. **Data Transforming:** We need to change data as per target table structures

Real life examples of Data Transformation:

- Standardizing data: Data is fetched from multiple sources so it needs to be standardized as per the target system.

- Character set conversion: Need to transform the character sets as per the target systems. (First name and last name example)
- Calculated and derived values: In source system there is first val and second val and in target we need the calculation of first val and second val.
- Data Conversion in different formats: If in source system date in in DDMMYY format and in target the date is in DDMONYYYY format then this transformation needs to be done at transformation phase.

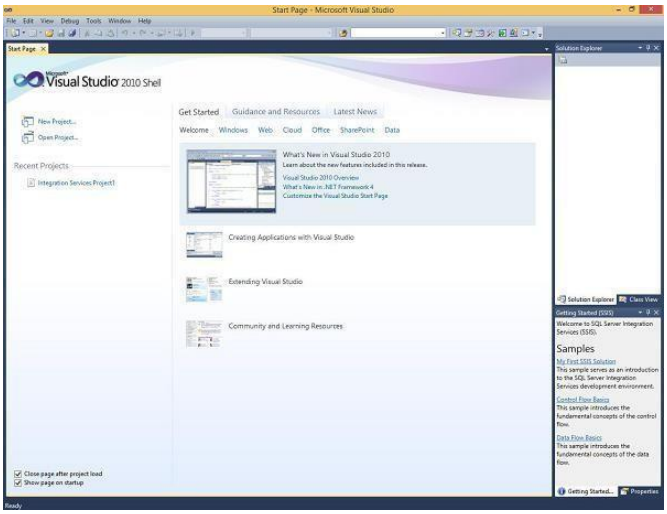
Step 3: Data Loading

- Data loading phase loads the prepared data from staging tables to main tables.

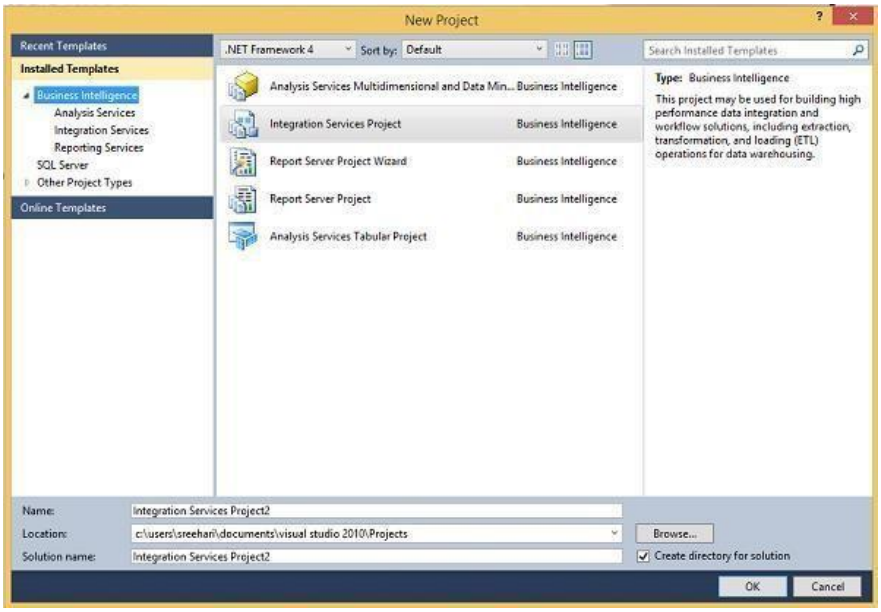
ETL Process in SQL SERVER

Following are the steps to open BIDS\SSDT.

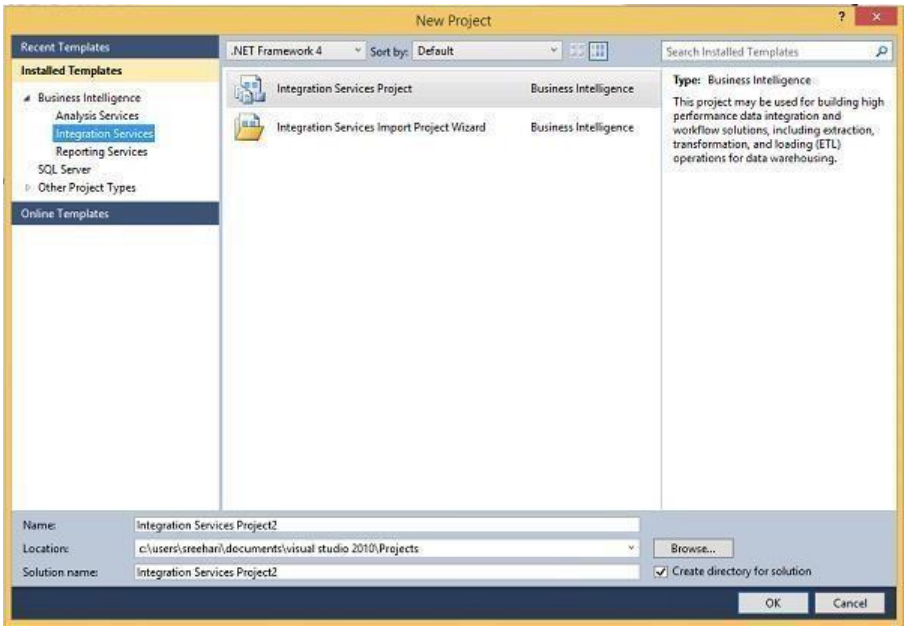
Step 1 − Open either BIDS\SSDT based on the version from the Microsoft SQL Server programs group. The following screen appears.



Step 2 − The above screen shows SSDT has opened. Go to file at the top left corner in the above image and click New. Select project and the following screen opens.



Step 3 – Select Integration Services under Business Intelligence on the top left corner in the above screen to get the following screen.



Step 4 – In the above screen, select either Integration Services Project or Integration Services Import Project Wizard based on your requirement to develop\create the package.

Modes

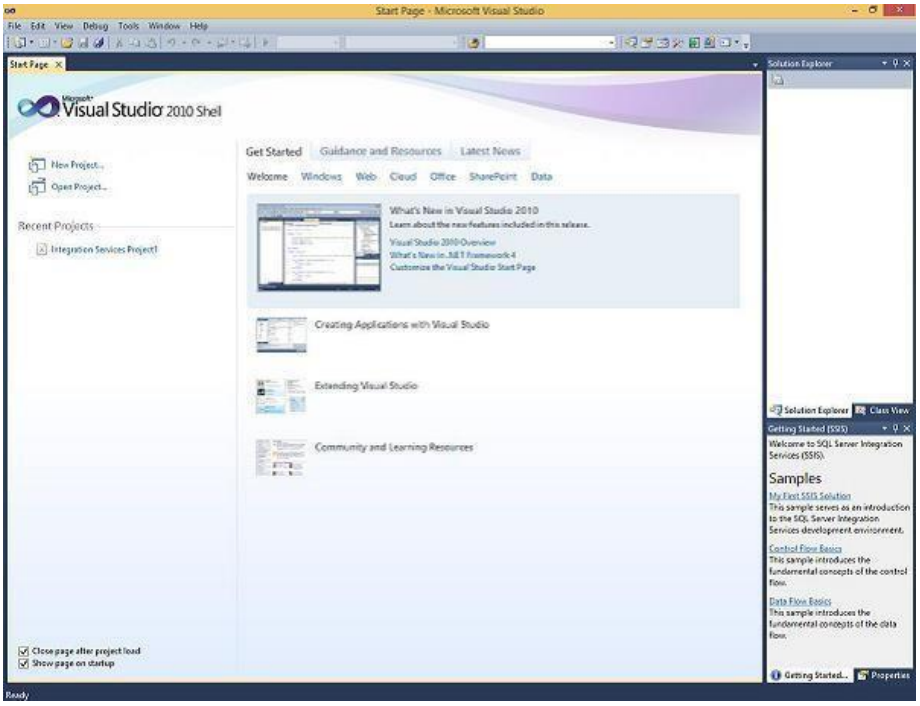
There are two modes – Native Mode (SQL Server Mode) and Share Point Mode.

Models

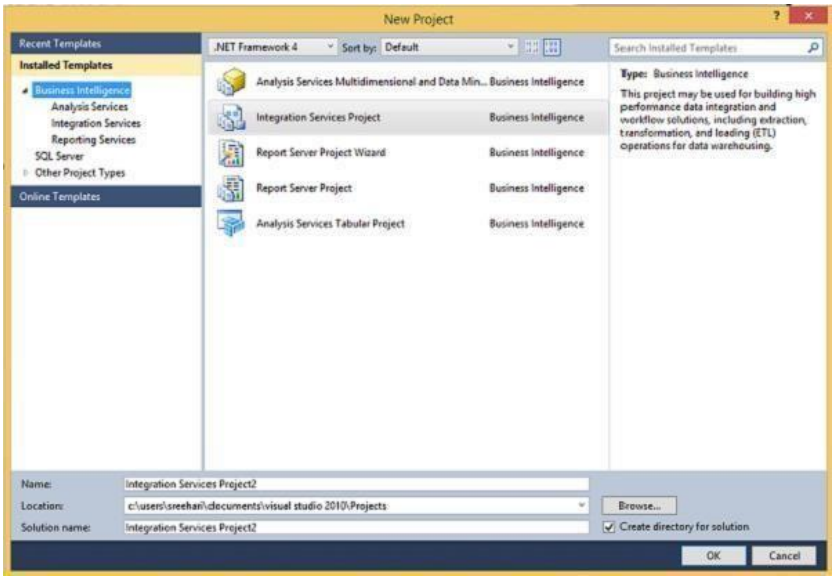
There are two models – Tabular Model (For Team and Personal Analysis) and Multi Dimensions Model (For Corporate Analysis).

The BIDS (Business Intelligence Studio till 2008 R2) and SSDT (SQL Server Data Tools from 2012) are environments to work with SSAS.

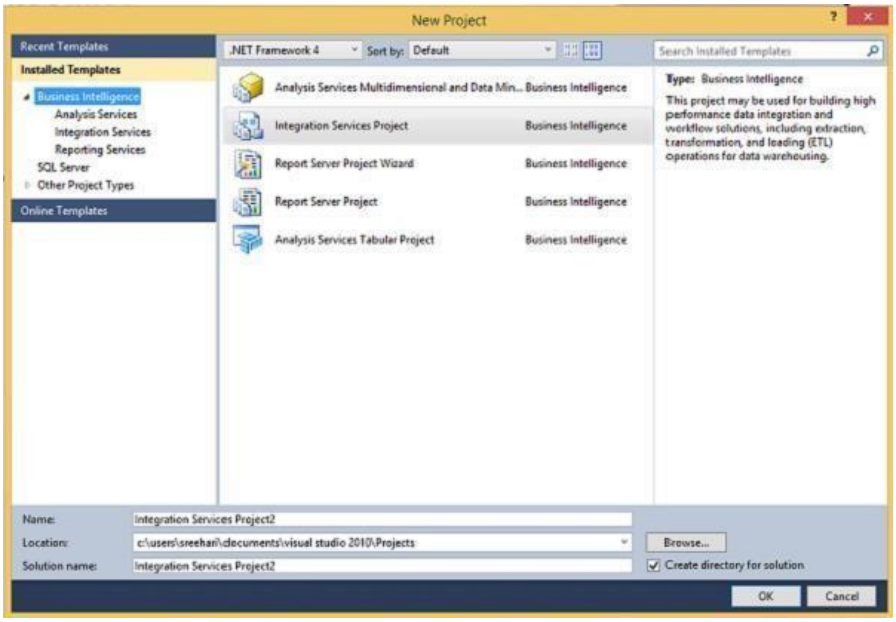
Step 1 – Open either BIDS\SSDT based on the version from the Microsoft SQL Server programs group. The following screen will appear.



Step 2 – The above screen shows SSDT has opened. Go to file on the top left corner in the above image and click New. Select project and the following screen opens.

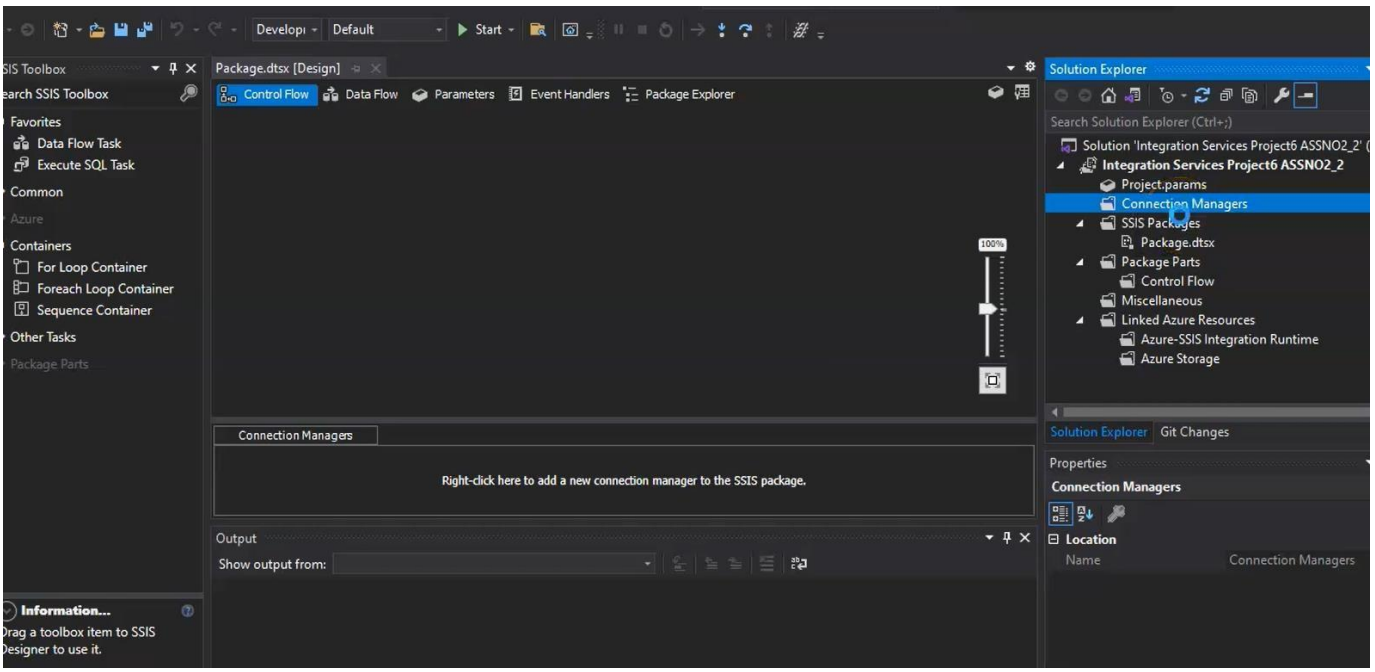


Step 3 – Select Analysis Services in the above screen under Business Intelligence as seen on the top left corner. The following screen pops up.

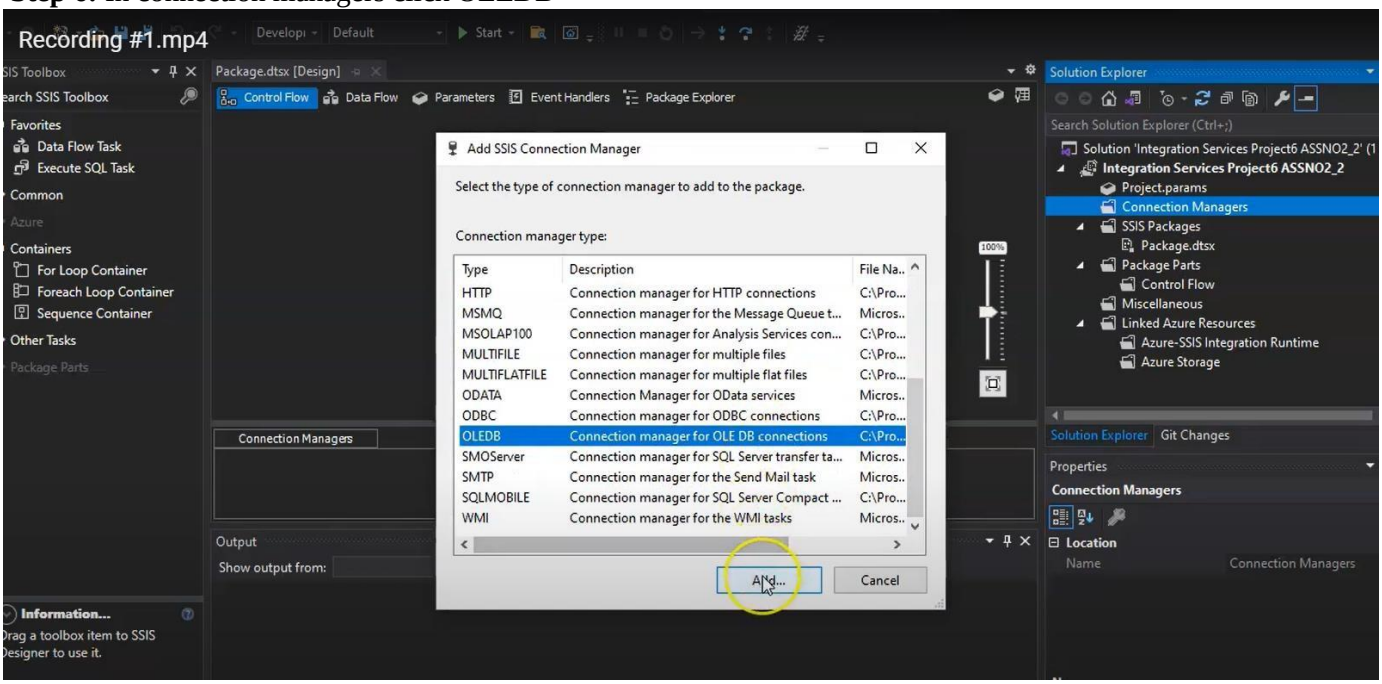


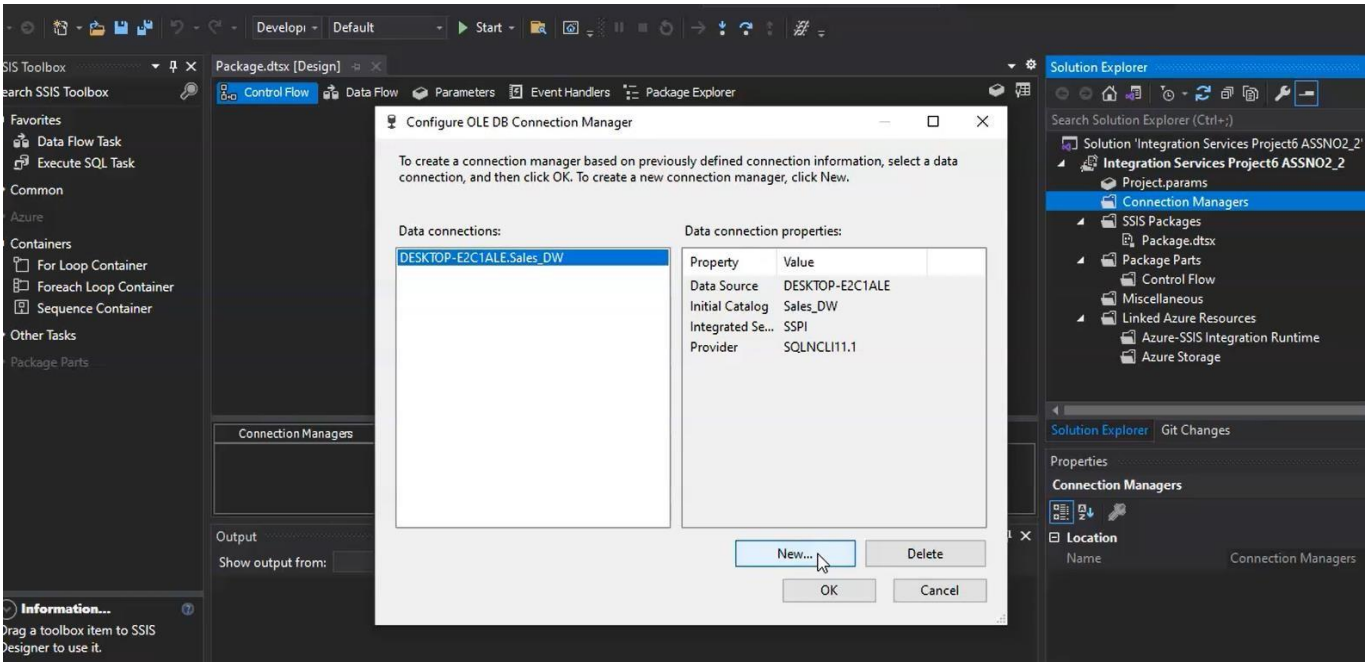
Step 4 – In the above screen, select any one option from the listed five options based on your requirement to work with Analysis services.

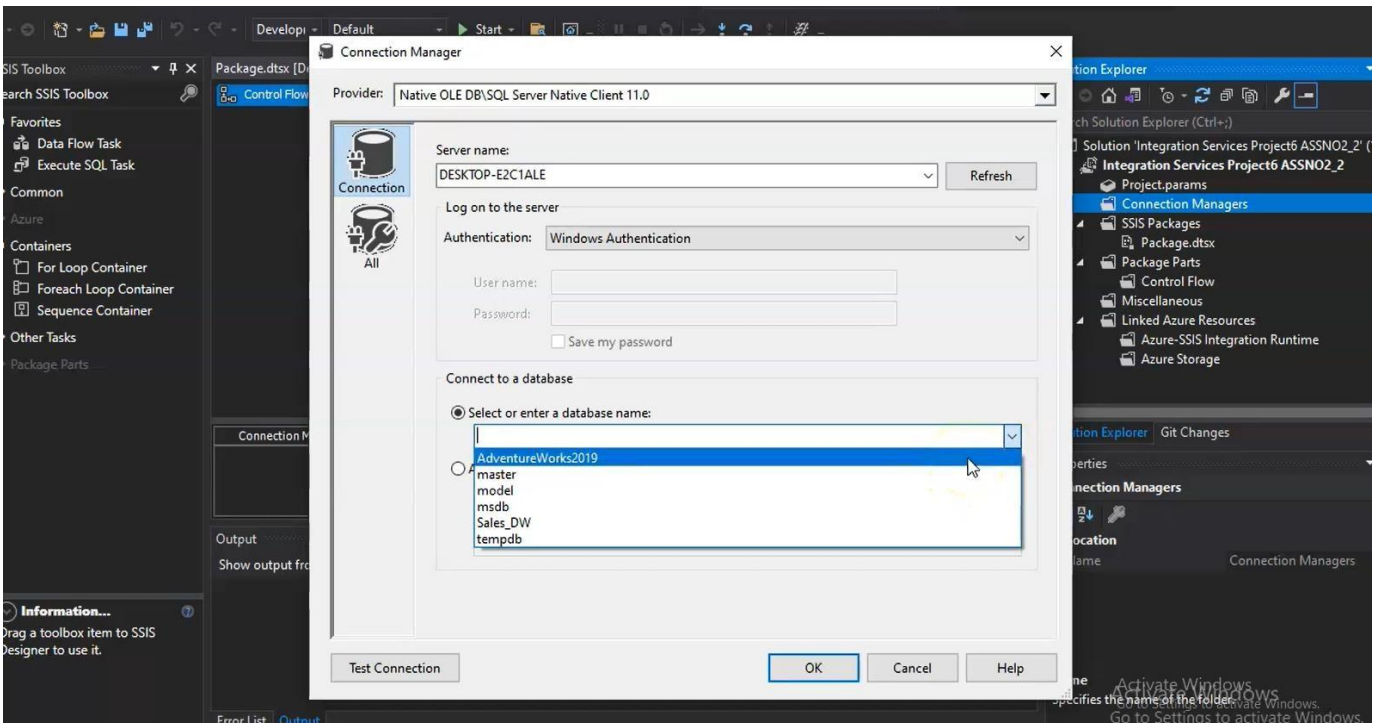
Step 5: Open Connection managers

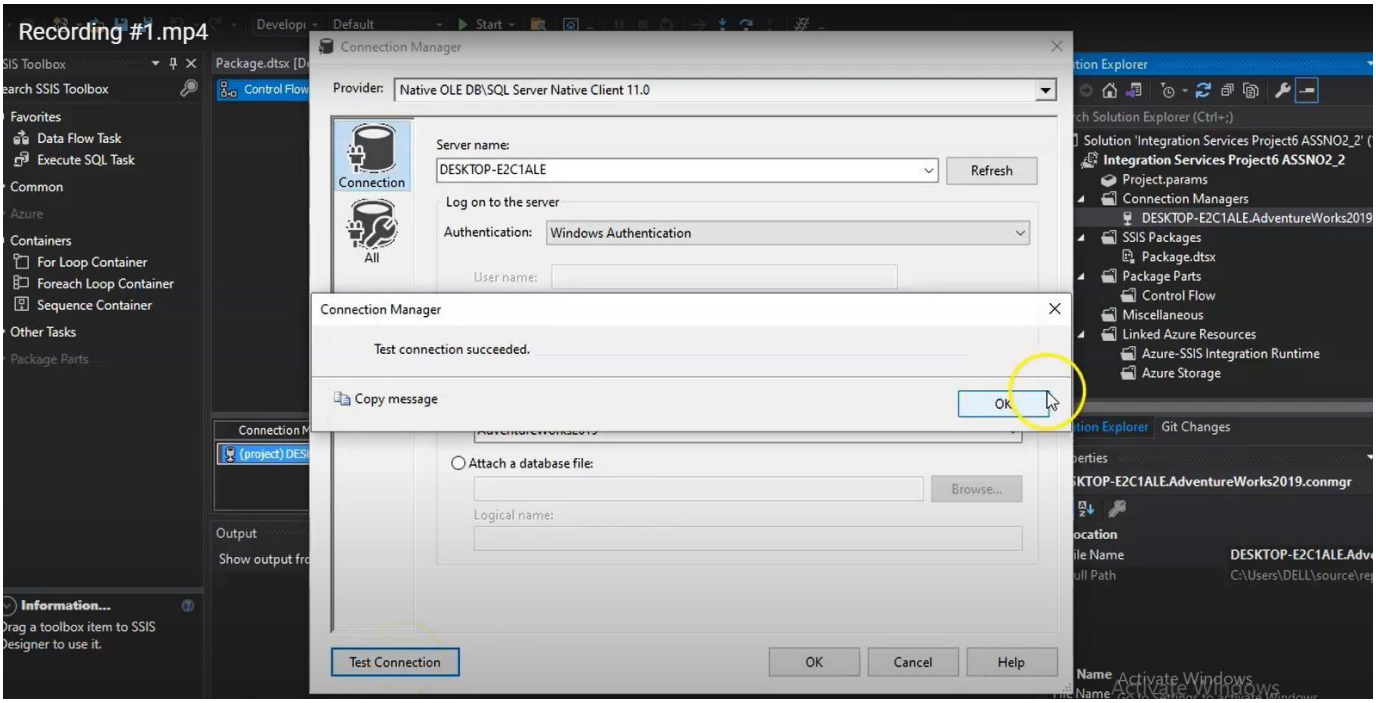


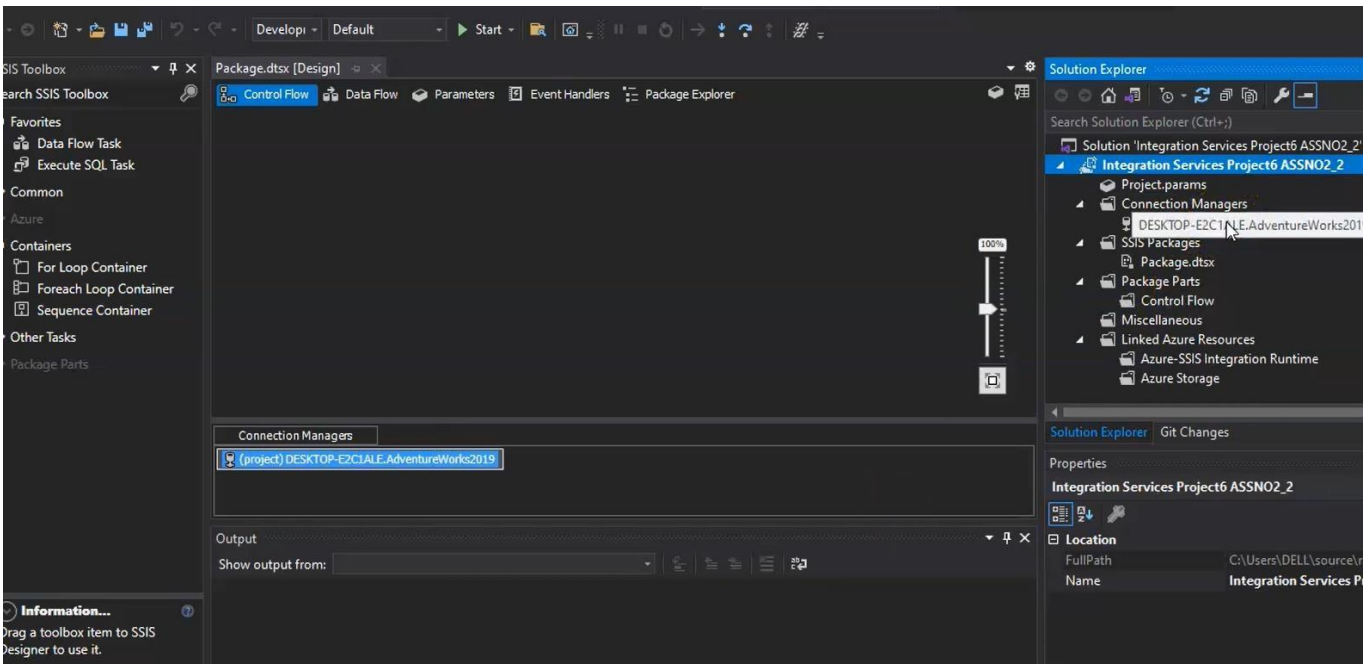
Step 6: In connection managers click OLEDB

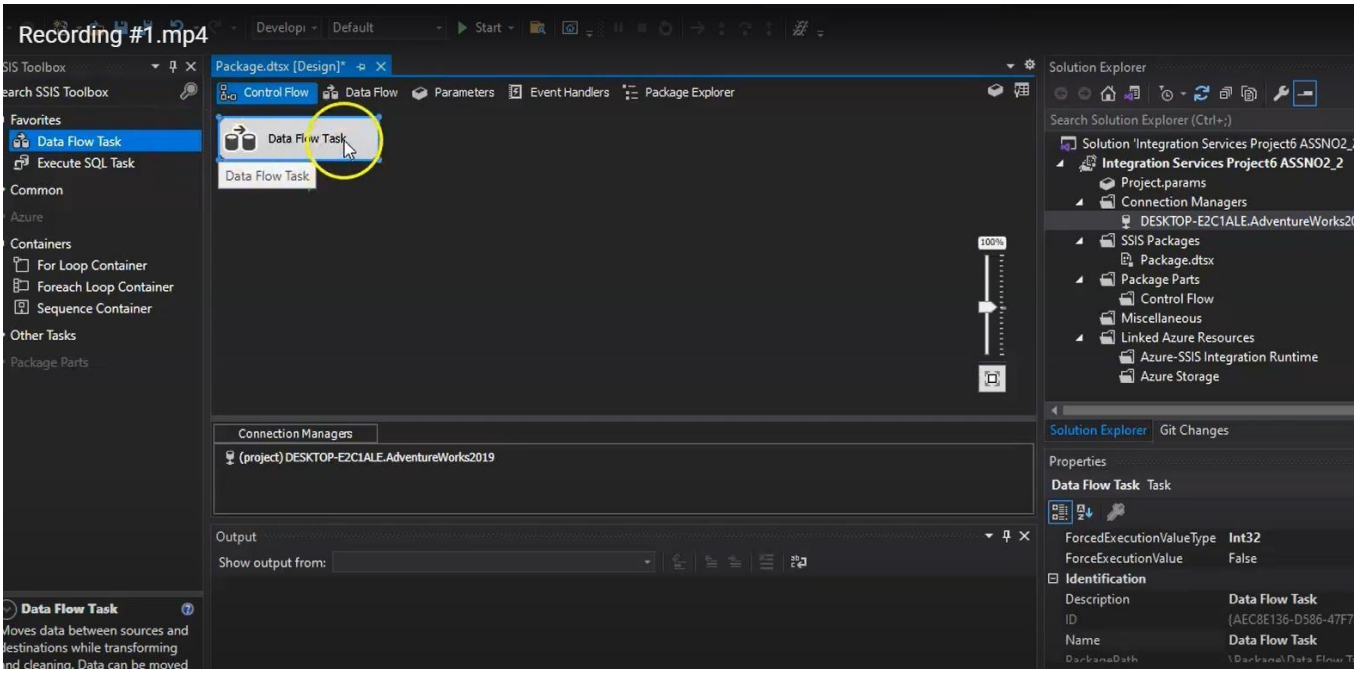


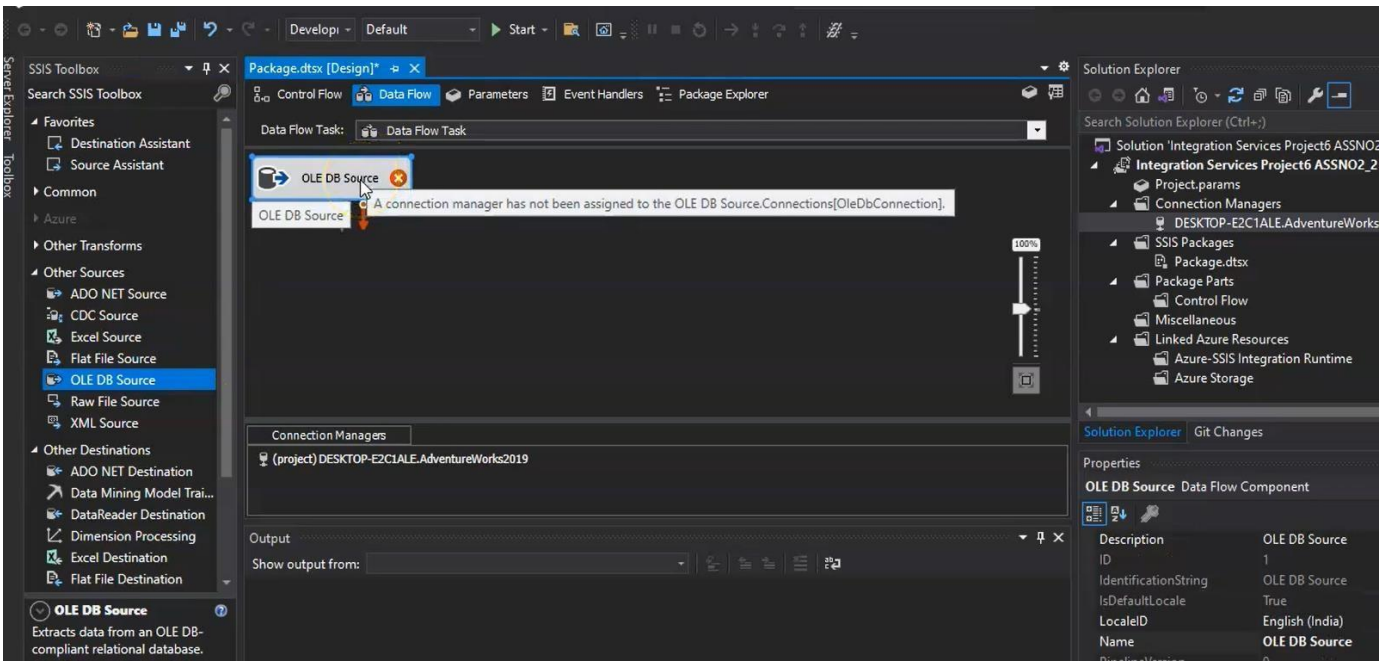


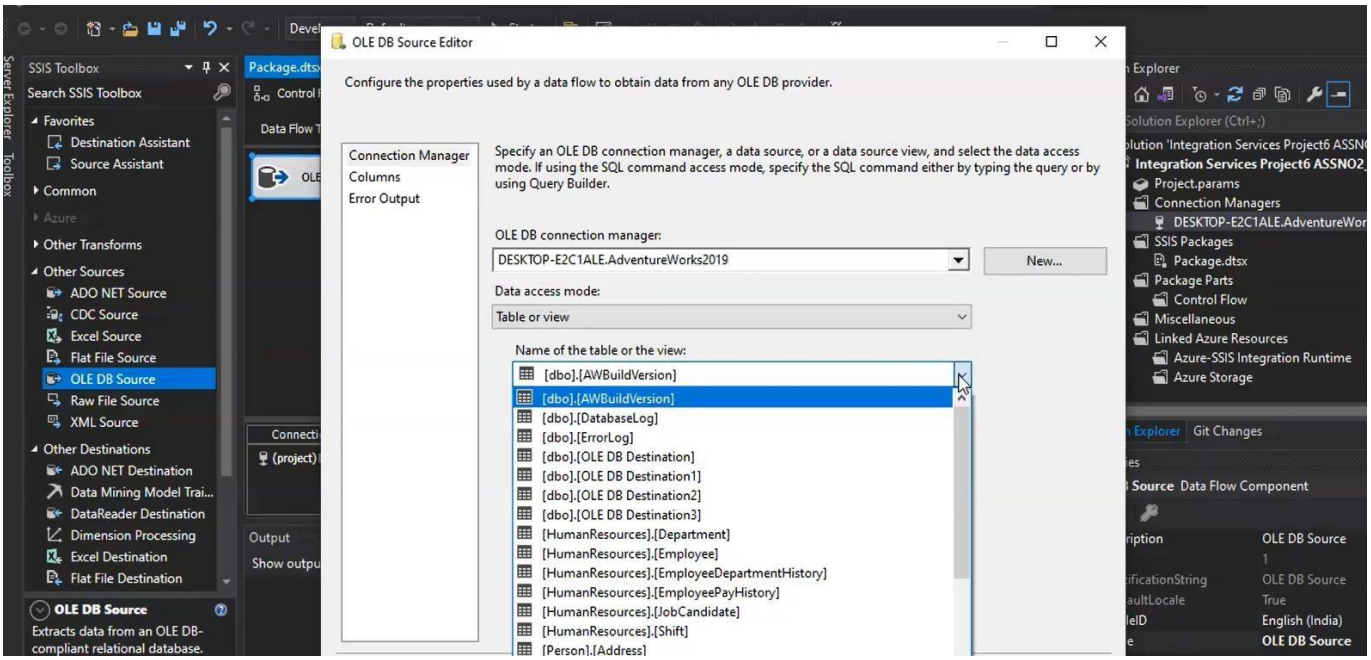

Connection Manager
Provider:
Native OLE DB\SQL Server Native Client 11.0
Connection
All
Server name:
DESKTOP-E2C1ALE
Refresh
Log on to the server
Authentication:
Windows Authentication
User name:
Password:
Save my password
Connect to a database
Select or enter a database name:
AdventureWorks2019
master
model
msdb
Sales_DW
tempdb
Test Connection
OK
Cancel
Help
Integration Services Project6 ASSNO2_2
Project.params
Connection Managers
SSIS Packages
Package.dtsx
Package Parts
Control Flow
Miscellaneous
Linked Azure Resources
Azure-SSIS Integration Runtime
Azure Storage

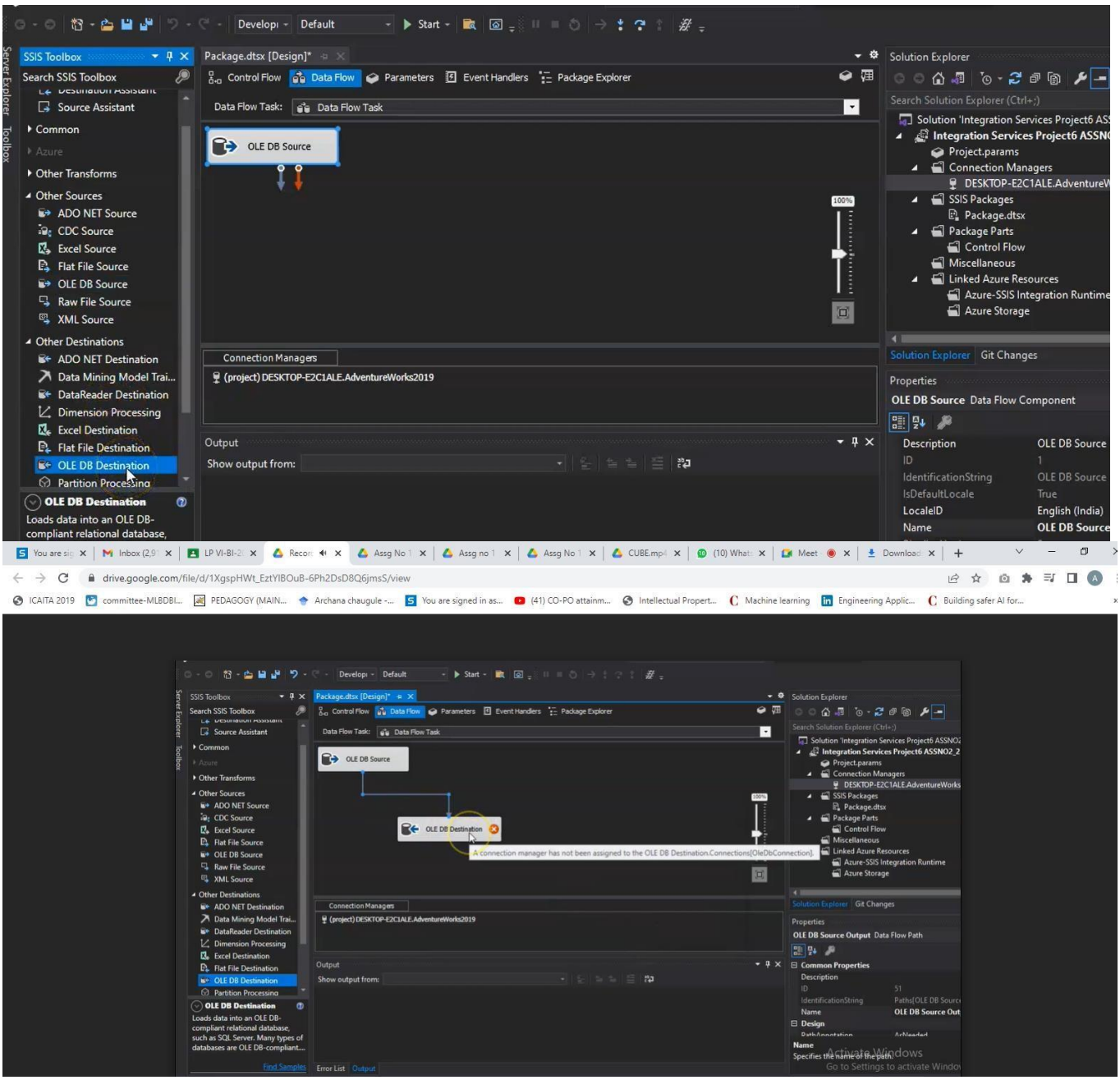

Recording #1.mp4
Connection Manager
Provider:
Native OLE DB\SQL Server Native Client 11.0
Server name:
DESKTOP-E2C1ALE
Refresh
Log on to the server
Authentication:
Windows Authentication
Connection Manager
Test connection succeeded.
Copy message
OK
Attach a database file:
Browse...
Logical name:
Test Connection
OK
Cancel
Help
DESKTOP-E2C1ALE.AdventureWorks2019.conmgr


Package.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
SSIS Toolbox
Favorites
Data Flow Task
Execute SQL Task
Common
Containers
For Loop Container
Foreach Loop Container
Sequence Container
Other Tasks
Connection Managers
(project) DESKTOP-E2C1ALE.AdventureWorks2019
Output
Show output from:
Solution Explorer
Integration Services Project6 ASSNO2_2
Project.params
Connection Managers
SSIS Packages
Package.dtsx
Package Parts
Control Flow
Miscellaneous
Linked Azure Resources
Azure-SSIS Integration Runtime
Azure Storage
Properties
Location

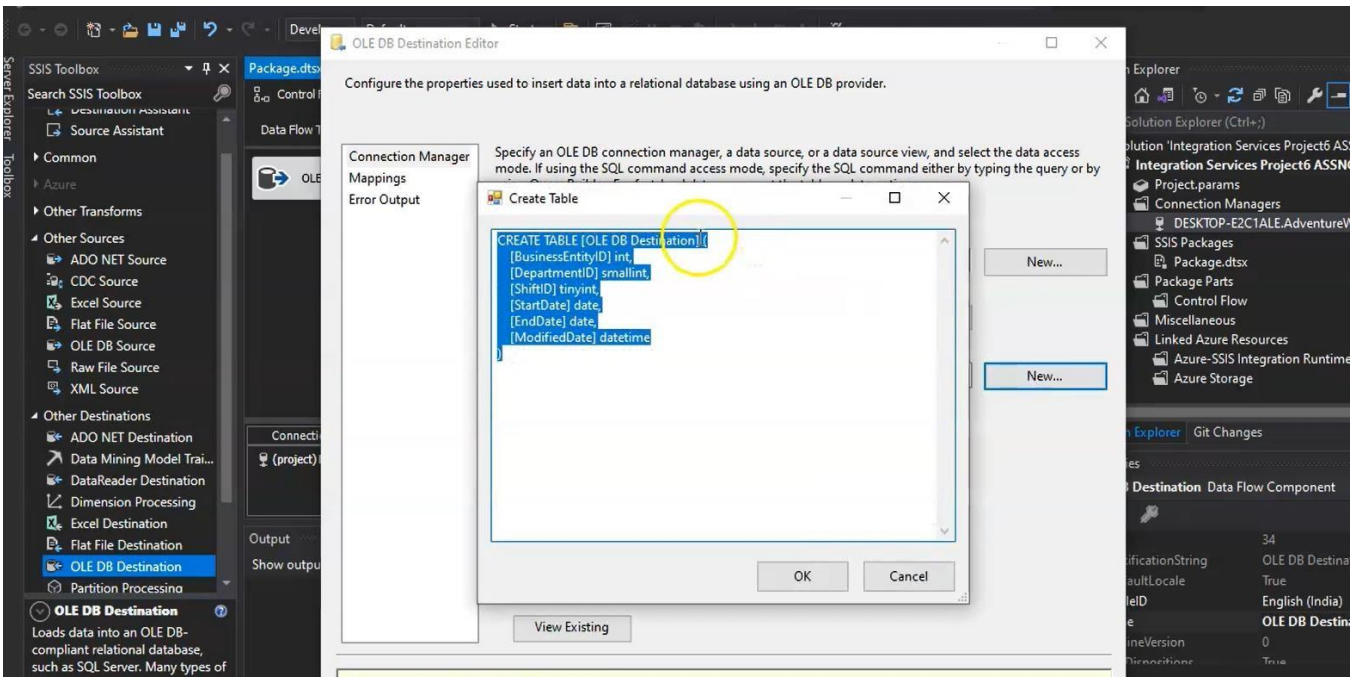

Recording #1.mp4
Package.dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Data Flow Task
Execute SQL Task
Common
Containers
For Loop Container
Foreach Loop Container
Sequence Container
Other Tasks
Connection Managers
(project) DESKTOP-E2C1ALE.AdventureWorks2019
Output
Show output from:
Solution Explorer
Integration Services Project6 ASSNO2_2
Project.params
Connection Managers
SSIS Packages
Package.dtsx
Package Parts
Control Flow
Miscellaneous
Linked Azure Resources
Azure-SSIS Integration Runtime
Azure Storage
Properties
Data Flow Task Task
ForcedExecutionValueType Int32
ForceExecutionValue False
Identification
Description Data Flow Task
Name Data Flow Task

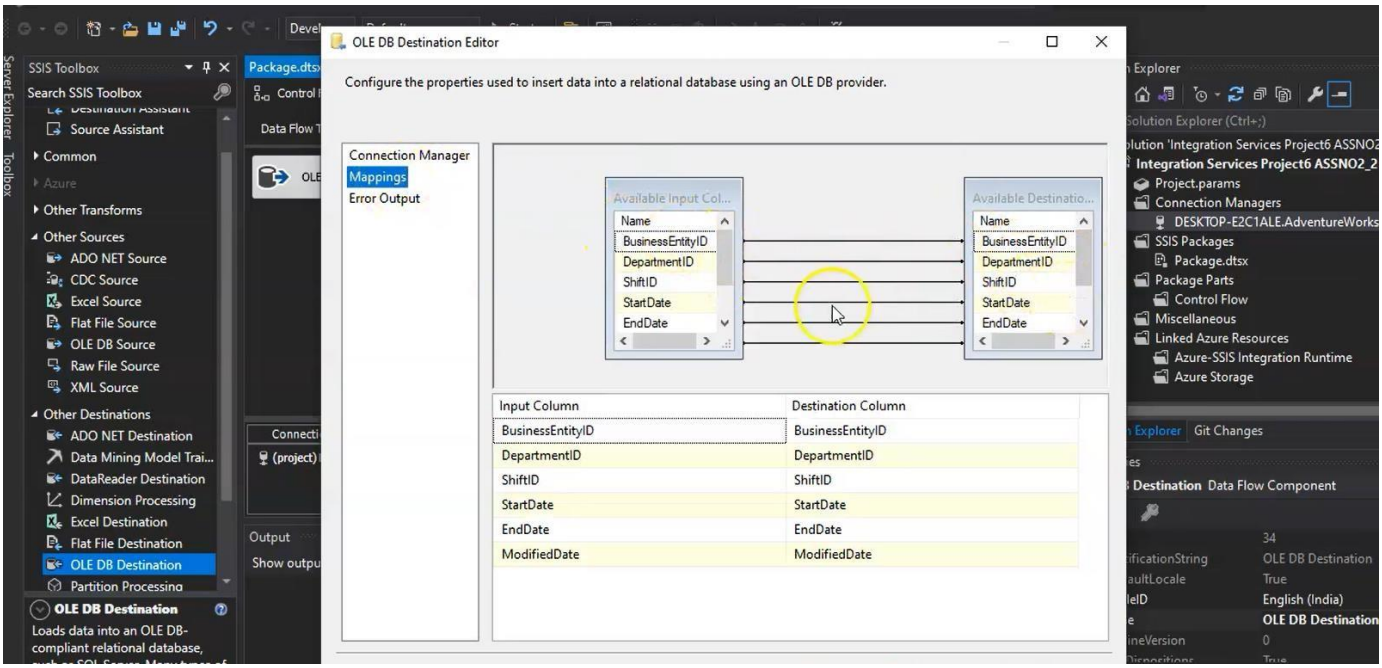

SSIS Toolbox
Search SSIS Toolbox
Favorites
Destination Assistant
Source Assistant
Common
Other Transforms
Other Sources
ADO NET Source
CDC Source
Excel Source
Flat File Source
OLE DB Source
Raw File Source
XML Source
Other Destinations
ADO NET Destination
Data Mining Model Trai...
DataReader Destination
Dimension Processing
Excel Destination
Flat File Destination
OLE DB Source
Extracts data from an OLE DB-compliant relational database.
Package.dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Data Flow Task:
Data Flow Task
OLE DB Source
A connection manager has not been assigned to the OLE DB Source.Connections[OleDbConnection].
Connection Managers
(project) DESKTOP-E2C1ALE.AdventureWorks2019
Output
Show output from:
Solution Explorer
Search Solution Explorer (Ctrl+;)
Integration Services Project6 ASSNO2_2
Project.params
Connection Managers
SSIS Packages
Package.dtsx
Package Parts
Control Flow
Miscellaneous
Linked Azure Resources
Azure-SSIS Integration Runtime
Azure Storage
Solution Explorer
Git Changes
Properties
OLE DB Source Data Flow Component
Description
OLE DB Source
ID
1
IdentificationString
OLE DB Source
IsDefaultLocale
True
LocaleID
English (India)
Name
OLE DB Source

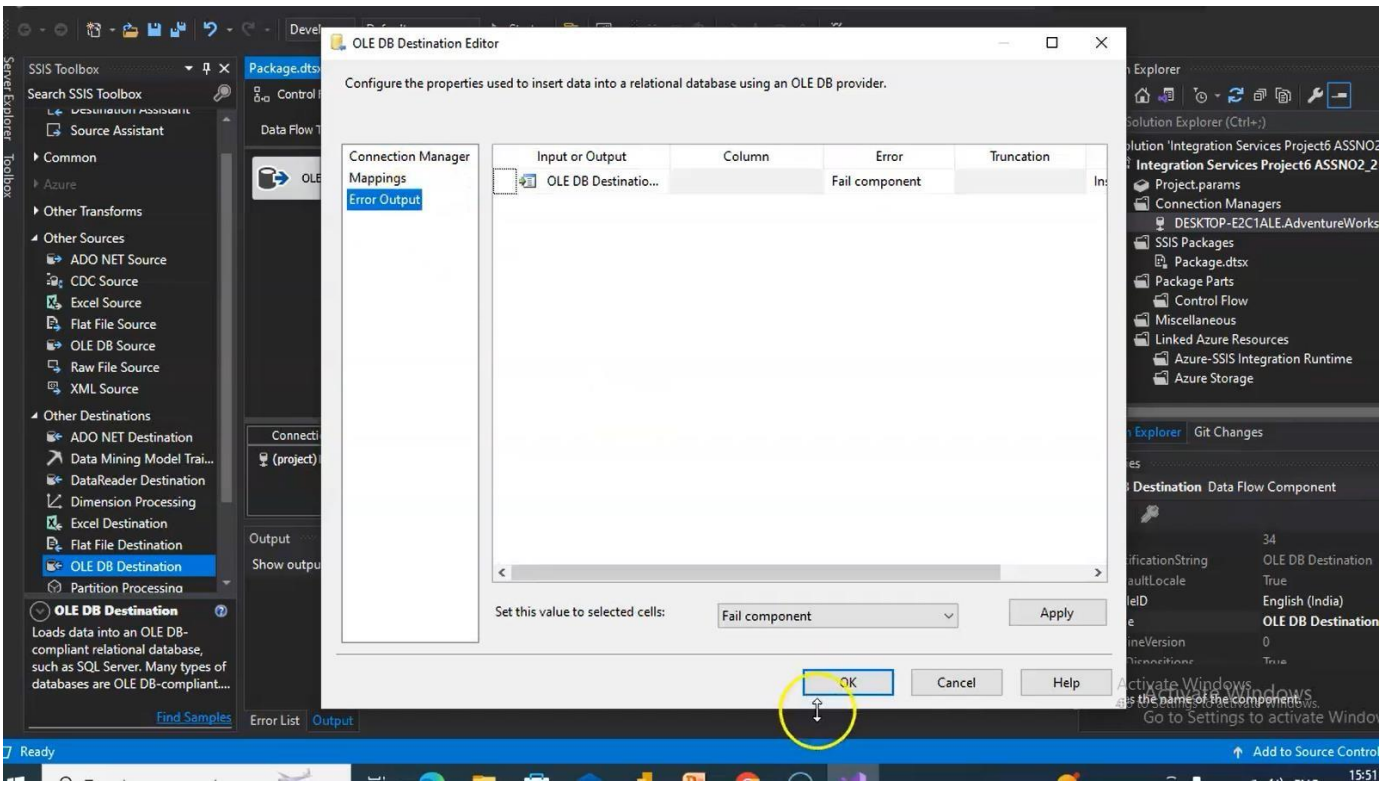

OLE DB Source Editor
Configure the properties used by a data flow to obtain data from any OLE DB provider.
Connection Manager
Columns
Error Output
Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder.
OLE DB connection manager:
DESKTOP-E2C1ALE.AdventureWorks2019
New...
Data access mode:
Table or view
Name of the table or the view:
[dbo].[AWBuildVersion]
[dbo].[AWBuildVersion]
[dbo].[DatabaseLog]
[dbo].[ErrorLog]
[dbo].[OLE DB Destination]
[dbo].[OLE DB Destination1]
[dbo].[OLE DB Destination2]
[dbo].[OLE DB Destination3]
[HumanResources].[Department]
[HumanResources].[Employee]
[HumanResources].[EmployeeDepartmentHistory]
[HumanResources].[EmployeePayHistory]
[HumanResources].[JobCandidate]
[HumanResources].[Shift]
[Person].[Address]

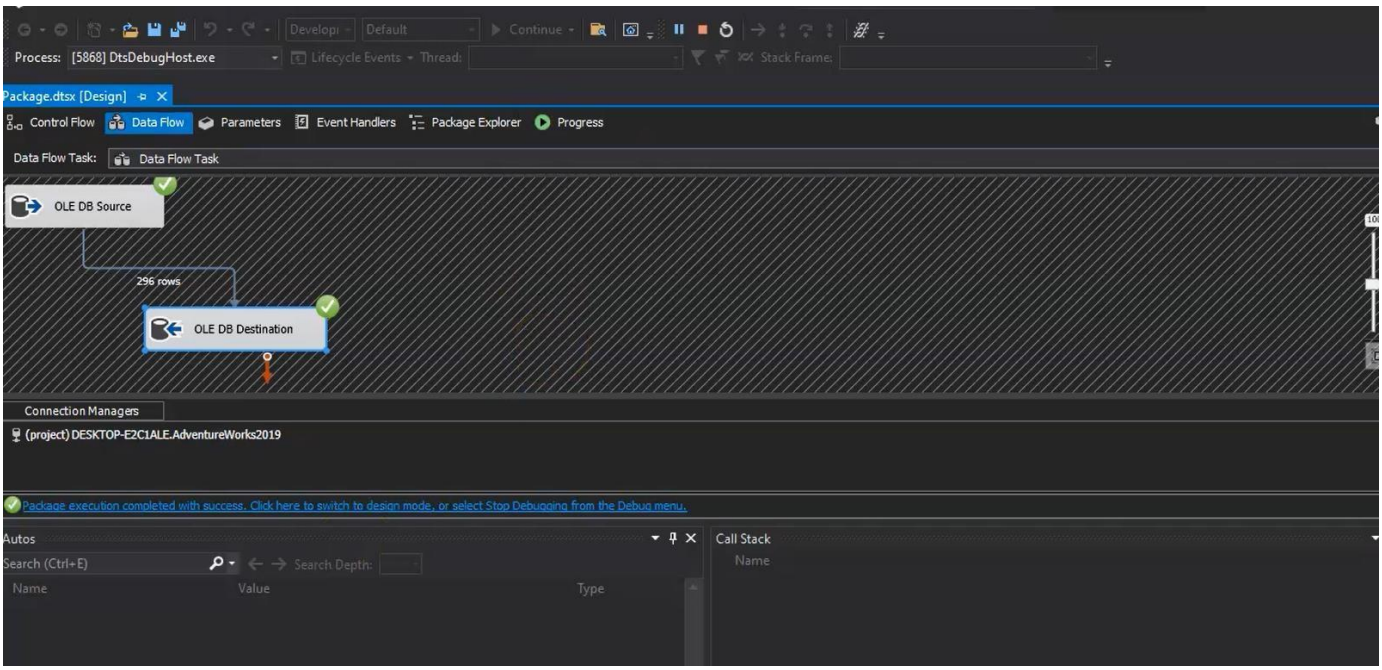

SSIS Toolbox
Search SSIS Toolbox
Source Assistant
Common
Other Transforms
Other Sources
ADO NET Source
CDC Source
Excel Source
Flat File Source
OLE DB Source
Raw File Source
XML Source
Other Destinations
ADO NET Destination
Data Mining Model Trai...
DataReader Destination
Dimension Processing
Excel Destination
Flat File Destination
OLE DB Destination
Partition Processing
OLE DB Destination
Loads data into an OLE DB-compliant relational database,
Package.dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Data Flow Task:
Data Flow Task
OLE DB Source
Connection Managers
(project) DESKTOP-E2C1ALE.AdventureWorks2019
Output
Show output from:
Solution Explorer
Search Solution Explorer (Ctrl+;)
Integration Services Project6 ASSN
Project.params
Connection Managers
SSIS Packages
Package.dtsx
Package Parts
Control Flow
Miscellaneous
Linked Azure Resources
Azure-SSIS Integration Runtime
Azure Storage
Solution Explorer
Git Changes
Properties
OLE DB Source Data Flow Component
Description
ID
IdentificationString
IsDefaultLocale
LocaleID
Name
OLE DB Source
English (India)

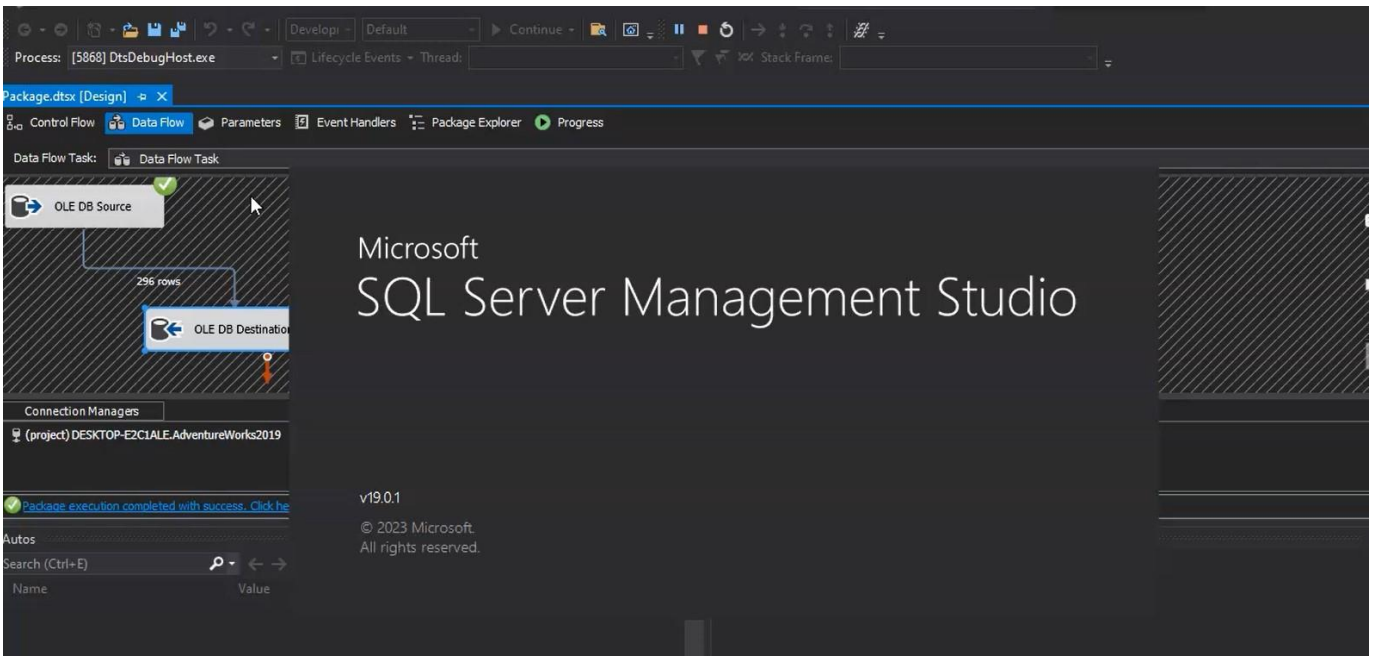
drive.google.com/file/d/1XgspHWt_EztYlBOuB-6Ph2DsD8Q6jmsS/view
OLE DB Source
OLE DB Destination
A connection manager has not been assigned to the OLE DB Destination.Connections[OleDbConnection].
Connection Managers
(project) DESKTOP-E2C1ALE.AdventureWorks2019
Output
Show output from:
OLE DB Source Output Data Flow Path
Common Properties
Design
Activate Windows

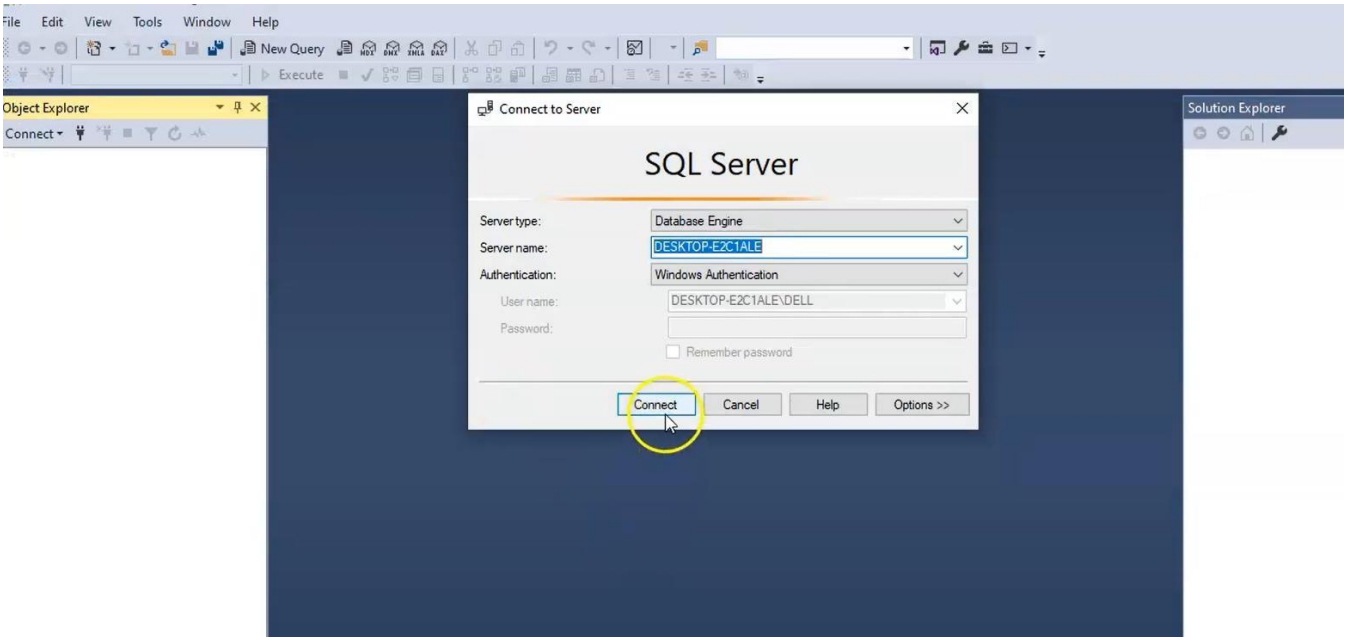

OLE DB Destination Editor
Configure the properties used to insert data into a relational database using an OLE DB provider.
Connection Manager
Mappings
Error Output
Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by
Create Table
CREATE TABLE [OLE DB Destination] (
[BusinessEntityID] int,
[DepartmentID] smallint,
[ShiftID] tinyint,
[StartDate] date,
[EndDate] date,
[ModifiedDate] datetime
)
New...
New...
OK
Cancel
View Existing
SSIS Toolbox
Search SSIS Toolbox
Source Assistant
Common
Other Transforms
Other Sources
ADO NET Source
CDC Source
Excel Source
Flat File Source
OLE DB Source
Raw File Source
XML Source
Other Destinations
ADO NET Destination
Data Mining Model Trai...
DataReader Destination
Dimension Processing
Excel Destination
Flat File Destination
OLE DB Destination
Partition Processing
OLE DB Destination
Loads data into an OLE DB-compliant relational database, such as SQL Server. Many types of
Solution Explorer
Integration Services Project6 ASSN
Project.params
Connection Managers
SSIS Packages
Package.dtsx
Package Parts
Control Flow
Miscellaneous
Linked Azure Resources
Azure-SSIS Integration Runtime
Azure Storage
Git Changes
Destination Data Flow Component

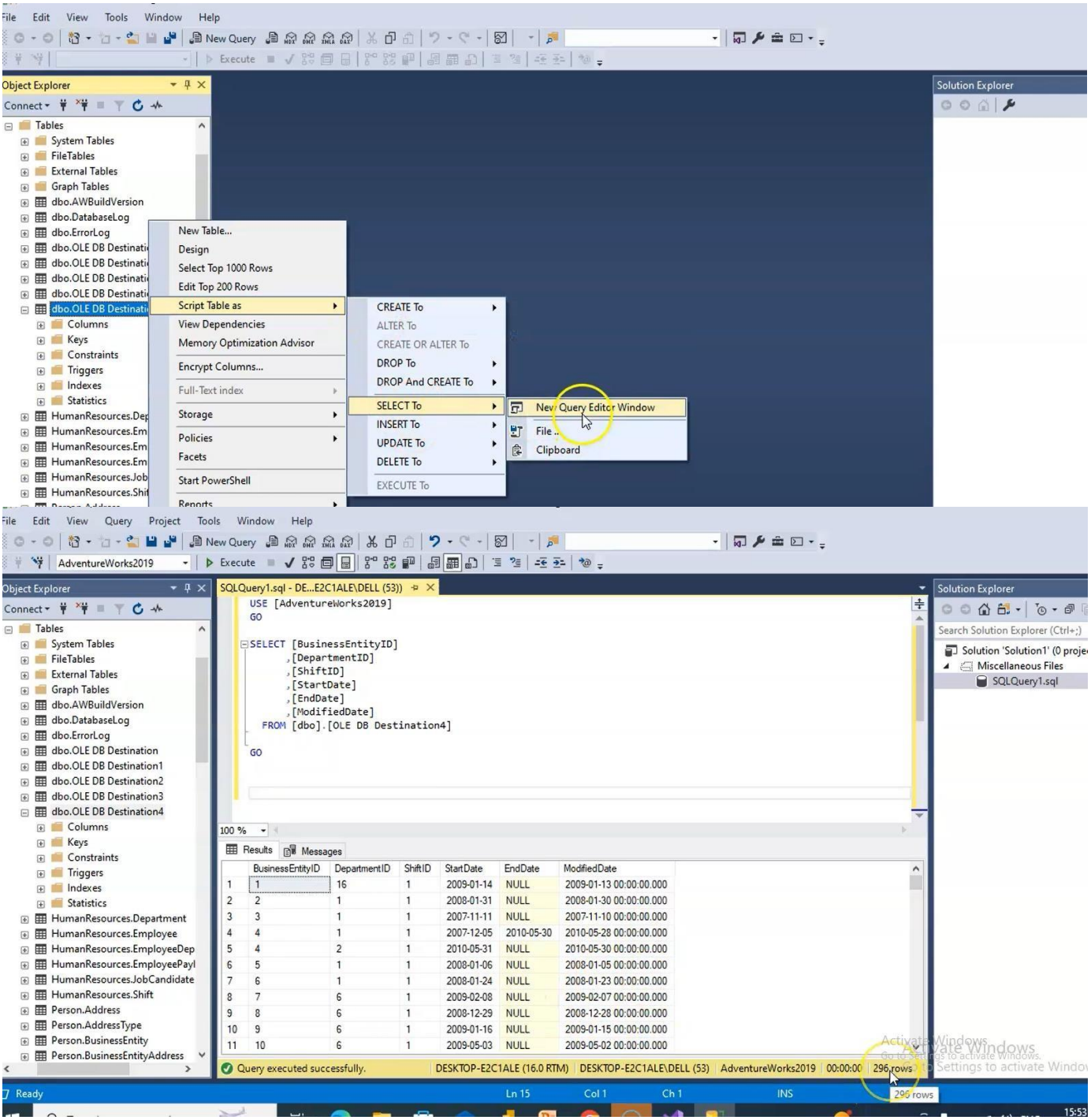

OLE DB Destination Editor
Configure the properties used to insert data into a relational database using an OLE DB provider.
Connection Manager
Mappings
Error Output
Available Input Col...
Name
BusinessEntityID
DepartmentID
ShiftID
StartDate
EndDate
Available Destinatio...
Name
BusinessEntityID
DepartmentID
ShiftID
StartDate
EndDate
Input Column
Destination Column
BusinessEntityID
BusinessEntityID
DepartmentID
DepartmentID
ShiftID
ShiftID
StartDate
StartDate
EndDate
EndDate
ModifiedDate
ModifiedDate
SSIS Toolbox
Search SSIS Toolbox
Source Assistant
Common
Other Transforms
Other Sources
ADO NET Source
CDC Source
Excel Source
Flat File Source
OLE DB Source
Raw File Source
XML Source
Other Destinations
ADO NET Destination
Data Mining Model Trai...
DataReader Destination
Dimension Processing
Excel Destination
Flat File Destination
OLE DB Destination
Partition Processing
OLE DB Destination
Loads data into an OLE DB-compliant relational database,
Solution Explorer
Integration Services Project6 ASSNO2_2
Project.params
Connection Managers
SSIS Packages
Package.dtsx
Package Parts
Control Flow
Miscellaneous
Linked Azure Resources
Azure-SSIS Integration Runtime
Azure Storage
Git Changes
Destination Data Flow Component


OLE DB Destination Editor
Configure the properties used to insert data into a relational database using an OLE DB provider.
Connection Manager
Mappings
Error Output
Input or Output
Column
Error
Truncation
OLE DB Destinatio...
Fail component
Set this value to selected cells:
Fail component
Apply
OK
Cancel
Help
SSIS Toolbox
Search SSIS Toolbox
Other Transforms
Other Sources
ADO NET Source
CDC Source
Excel Source
Flat File Source
OLE DB Source
Raw File Source
XML Source
Other Destinations
ADO NET Destination
Data Mining Model Trai...
DataReader Destination
Dimension Processing
Excel Destination
Flat File Destination
OLE DB Destination
Partition Processing
OLE DB Destination
Loads data into an OLE DB-compliant relational database, such as SQL Server. Many types of databases are OLE DB-compliant....
Find Samples
Error List
Output
Ready
Add to Source Control


Process: [5868] DtsDebugHost.exe
Package.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task:
Data Flow Task
OLE DB Source
296 rows
OLE DB Destination
Connection Managers
(project) DESKTOP-E2C1ALE.AdventureWorks2019
Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.
Autos
Search (Ctrl+E)
Name
Value
Type
Call Stack
Name


Process: [5868] DtsDebugHost.exe
Package.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task:
Data Flow Task
OLE DB Source
296 rows
OLE DB Destinatio
Connection Managers
(project) DESKTOP-E2C1ALE.AdventureWorks2019
Package execution completed with success. Click he
Autos
Search (Ctrl+E)
Name
Value
Microsoft
SQL Server Management Studio
v19.0.1
© 2023 Microsoft.
All rights reserved.


File
Edit
View
Tools
Window
Help
New Query
Execute
Object Explorer
Connect
Connect to Server
SQL Server
Server type:
Database Engine
Server name:
DESKTOP-E2C1ALE
Authentication:
Windows Authentication
User name:
DESKTOP-E2C1ALE\DELL
Password:
Remember password
Connect
Cancel
Help
Options >>
Solution Explorer


File Edit View Tools Window Help
New Query
Execute
Object Explorer
Connect
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.AWBuildVersion
dbo.DatabaseLog
dbo.ErrorLog
dbo.OLE DB Destinati
Columns
Keys
Constraints
Triggers
Indexes
Statistics
HumanResources.Dep
HumanResources.Em
HumanResources.Em
HumanResources.Em
HumanResources.Job
HumanResources.Shif
New Table...
Design
Select Top 1000 Rows
Edit Top 200 Rows
Script Table as
View Dependencies
Memory Optimization Advisor
Encrypt Columns...
Full-Text index
Storage
Policies
Facets
Start PowerShell
Reports
CREATE To
ALTER To
CREATE OR ALTER To
DROP To
DROP And CREATE To
SELECT To
INSERT To
UPDATE To
DELETE To
EXECUTE To
New Query Editor Window
File...
Clipboard
Solution Explorer
File Edit View Query Project Tools Window Help
New Query
AdventureWorks2019
Execute
Object Explorer
Connect
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.AWBuildVersion
dbo.DatabaseLog
dbo.ErrorLog
dbo.OLE DB Destination
dbo.OLE DB Destination1
dbo.OLE DB Destination2
dbo.OLE DB Destination3
dbo.OLE DB Destination4
Columns
Keys
Constraints
Triggers
Indexes
Statistics
HumanResources.Department
HumanResources.Employee
HumanResources.EmployeeDep
HumanResources.EmployeePayl
HumanResources.JobCandidate
HumanResources.Shift
Person.Address
Person.AddressType
Person.BusinessEntity
Person.BusinessEntityAddress
SQLQuery1.sql - DE...E2C1ALE\DELL (53))
USE [AdventureWorks2019]
GO
SELECT [BusinessEntityID]
,[DepartmentID]
,[ShiftID]
,[StartDate]
,[EndDate]
,[ModifiedDate]
FROM [dbo].[OLE DB Destination4]
GO
100 %
Results
Messages
BusinessEntityID DepartmentID ShiftID StartDate EndDate ModifiedDate
1 1 16 1 2009-01-14 NULL 2009-01-13 00:00:00.000
2 2 1 1 2008-01-31 NULL 2008-01-30 00:00:00.000
3 3 1 1 2007-11-11 NULL 2007-11-10 00:00:00.000
4 4 1 1 2007-12-05 2010-05-30 2010-05-28 00:00:00.000
5 4 2 1 2010-05-31 NULL 2010-05-30 00:00:00.000
6 5 1 1 2008-01-06 NULL 2008-01-05 00:00:00.000
7 6 1 1 2008-01-24 NULL 2008-01-23 00:00:00.000
8 7 6 1 2009-02-08 NULL 2009-02-07 00:00:00.000
9 8 6 1 2008-12-29 NULL 2008-12-28 00:00:00.000
10 9 6 1 2009-01-16 NULL 2009-01-15 00:00:00.000
11 10 6 1 2009-05-03 NULL 2009-05-02 00:00:00.000
Query executed successfully.
DESKTOP-E2C1ALE (16.0 RTM)
DESKTOP-E2C1ALE\DELL (53)
AdventureWorks2019
00:00:00
296 rows
Solution Explorer
Search Solution Explorer (Ctrl+;)
Solution 'Solution1' (0 proje
Miscellaneous Files
SQLQuery1.sql
Ready
Ln 15
Col 1
Ch 1
INS
296 rows

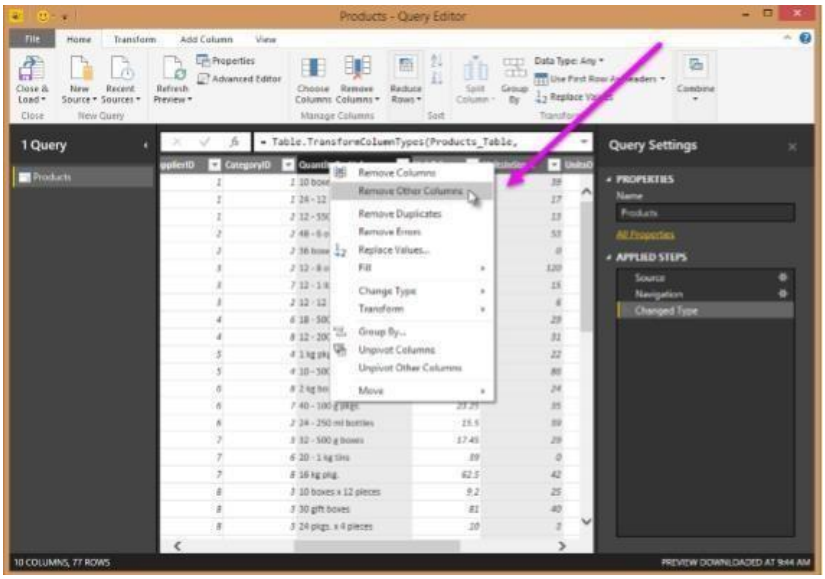
ETL Process in Power BI

1) Remove other columns to only display columns of interest

In this step you remove all columns except **ProductID**, **ProductName**, **UnitsInStock**, and **QuantityPerUnit**

Power BI Desktop includes Query Editor, which is where you shape and transform your data connections. Query Editor opens automatically when you select **Edit** from Navigator. You can also open the Query Editor by selecting Edit Queries from the Home ribbon in Power BI Desktop. The following steps are performed in Query Editor.

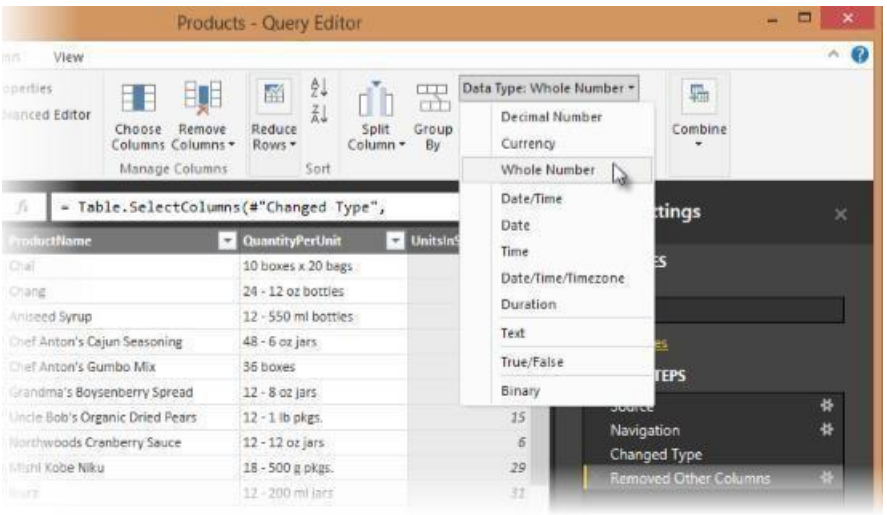
1. In **Query Editor**, select the **ProductID, ProductName**, **QuantityPerUnit,** and **UnitsInStock** columns (use **Ctrl+Click** to select more than one column, or **Shift+Click** to select columns that are beside each other).
2. Select **Remove Columns > Remove** Other Columns from the ribbon, or right-click on a column header and click Remove Other Columns.



3. **Change the data type of the UnitsInStock column**

When Query Editor connects to data, it reviews each field and to determine the best data type. For the Excel workbook, products in stock will always be a whole number, so in this step you confirm the **UnitsInStock** column's datatype is Whole Number.

1. Select the **UnitsInStock** column.
2. Select the **Data Type drop-down button** in the **Home ribbon**.
3. If not already a Whole Number, select **Whole Number** for data type from the drop down (the Data Type: button also displays the data type for the current selection).



3. **Expand the Order_Details table**

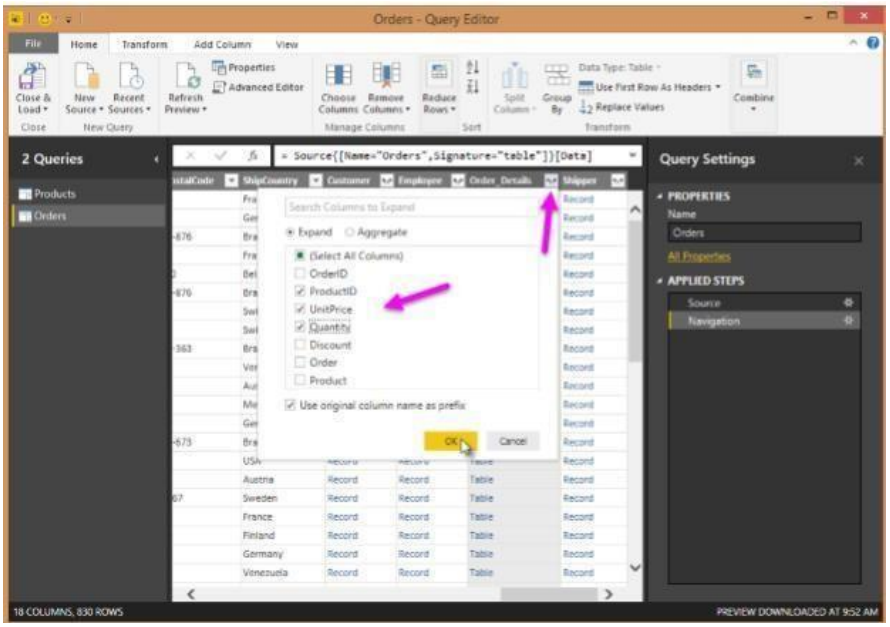
The Orders table contains a reference to a Details table, which contains the individual products that were included in each Order. When you connect to data sources with multiples tables (such as a relational database) you can use these references to build up your query.

In this step, you expand the **Order_Details** table that is related to the Orders table, to combine the **ProductID**, **UnitPrice**, and **Quantity** columns from **Order_Details** into the **Orders table**. This is a representation of the data in these tables:

The Expand operation combines columns from a related table into a subject table. When the query runs, rows from the related table (**Order_Details**) are combined into rows from the subject table (**Orders**).

After you expand the Order_Details table, three new columns and additional rows are added to the Orders table, one for each row in the nested or related table.

1. In the Query View, scroll to the Order_Details column.
2. In the Order_Details column, select the expand icon ( ).
3. In the Expand drop-down:
    a. Select (Select All Columns) to clear all columns.
    b. Select ProductID, UnitPrice, and Quantity.
    c. Click OK.

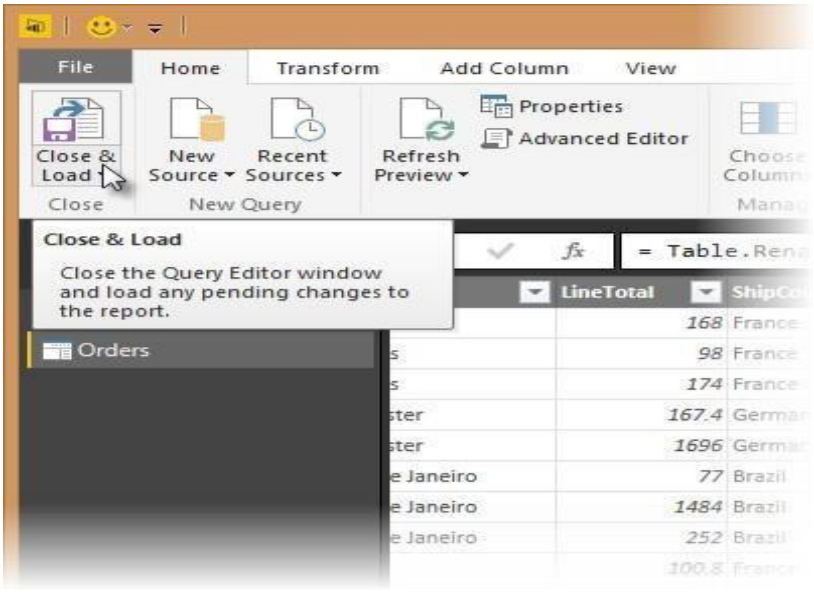


4. **Calculate the line total for each Order_Details row**

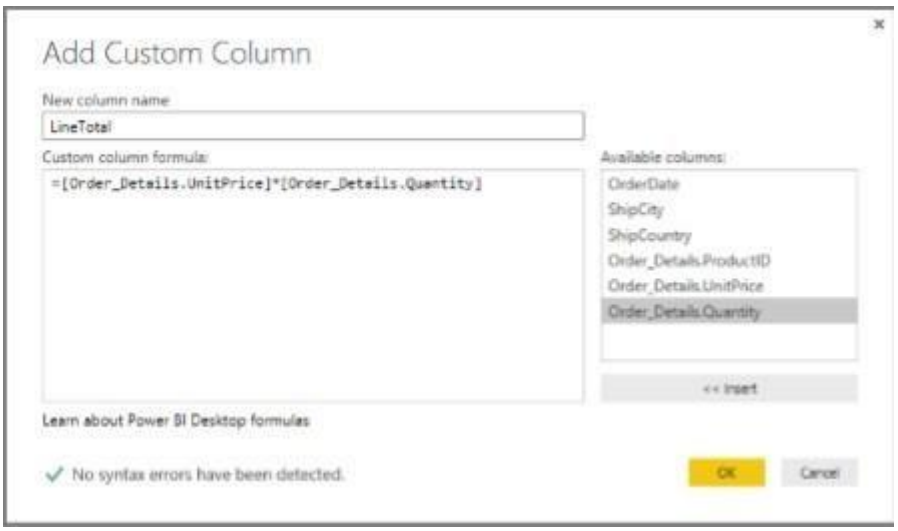
Power BI Desktop lets you to create calculations based on the columns you are importing, so you can enrich the data that you connect to. In this step, you create a Custom Column to calculate the line total for each Order_Details row.

Calculate the line total for each Order_Details row:

1. In the Add Column ribbon tab, click Add Custom Column.



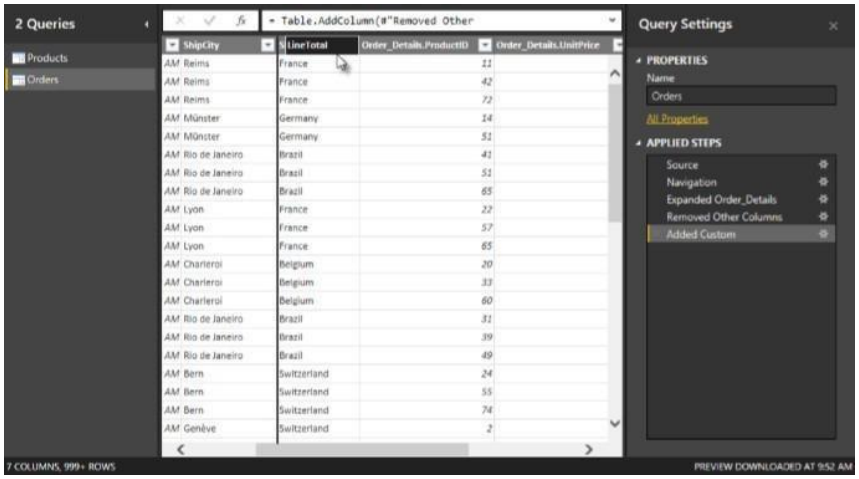
2. In the Add Custom Column dialog box, in the Custom Column Formula textbox, enter [Order_Details.UnitPrice] * [Order_Details.Quantity].
3. In the New column name textbox, enter LineTotal.
4. Click OK.



5. **Rename and reorder columns in the query**

In this step you finish making the model easy to work with when creating reports, by renaming the final columns and changing their order.

1. In Query Editor, drag the LineTotal column to the left, after Ship Country.



2.Remove the Order_Details. prefix from the Order_Details.ProductID, Order_Details.UnitPrice and Order_Details.Quantity columns, by double-clicking on each column header, and then deleting that text from the column name.

6. **Combine the Products and Total Sales queries**

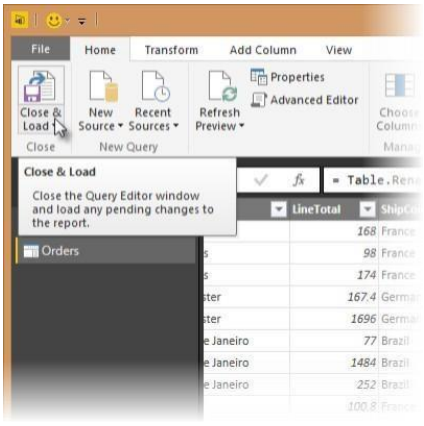
Power BI Desktop does not require you to combine queries to report on them. Instead, you can create Relationships between datasets. These relationships can be created on any column that is common to your datasets

we have Orders and Products data that share a common 'ProductID' field, so we need to ensure there's a relationship between them in the model we're using with Power BI Desktop. Simply specify in Power BI Desktop that the columns from each table are related (i.e. columns that have the same values). Power BI Desktop works out the direction and cardinality of the relationship for you. In some cases, it will even detect the relationships automatically.

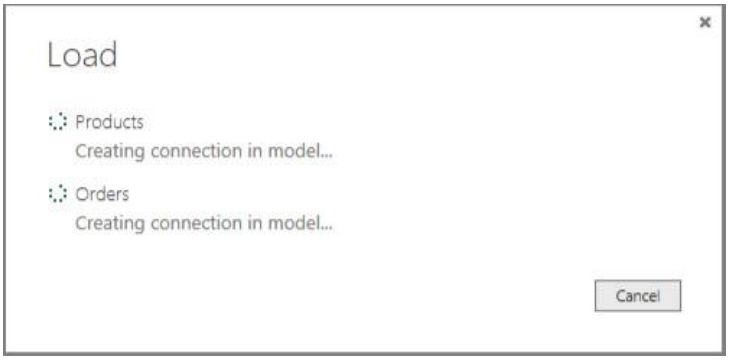
In this task, you confirm that a relationship is established in Power BI Desktop between the Products and Total Sales queries

Step 1: Confirm the relationship between Products and Total Sales

1. First, we need to load the model that we created in Query Editor into Power BI Desktop. From the Home ribbon of Query Editor, select Close & Load.

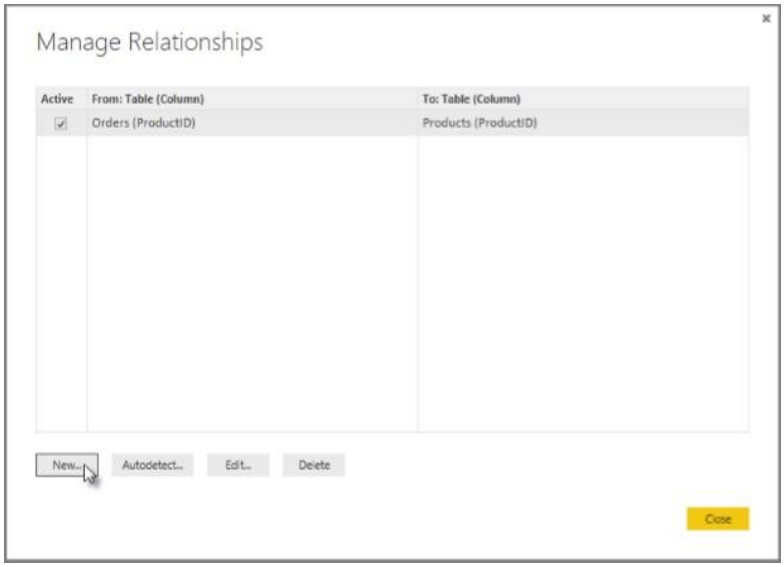


2. Power BI Desktop loads the data from the two queries.

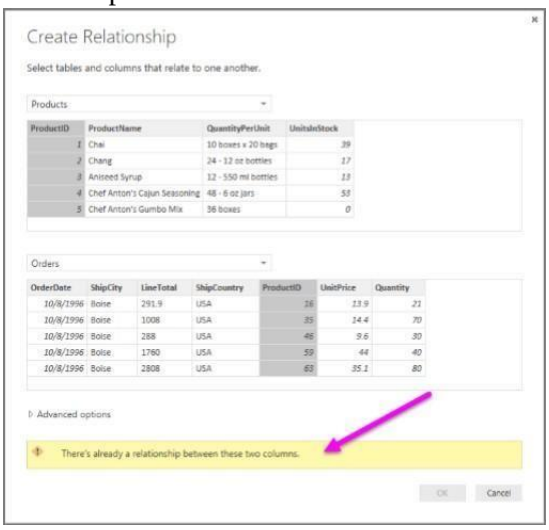


3. Once the data is loaded, select the Manage Relationships button Home ribbon.

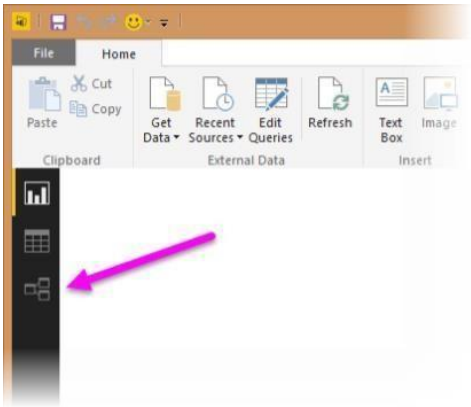
4. Select the New… button



5. When we attempt to create the relationship, we see that one already exists! As shown in the Create Relationship dialog (by the shaded columns), the ProductsID fields in each query already have an established relationship.

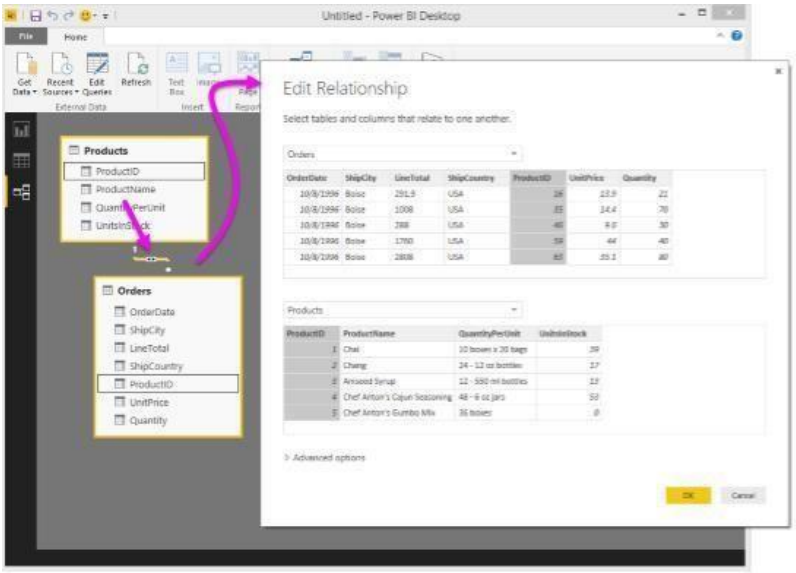


5. Select Cancel, and then select Relationship view in Power BI Desktop.



6. We see the following, which visualizes the relationship between the queries.

7. When you double-click the arrow on the line that connects the to queries, an Edit Relationship dialog appears.



8. No need to make any changes, so we'll just select Cancel to close the Edit Relationship dialog.

Conclusion:

With the help of SQL server and PowerBI the Extraction Transformation and Loading (ETL) process was constructed in the SQL server and Power BI successfully.

Assignment No. 3

Title: Cube in SQL server

Learning Objectives:

Creating a Cube in SQL server

Problem Statement:

Create the cube with suitable dimension and fact tables based on ROLAP, MOLAP and HOLAP model.

Theory Concepts:

Creating Data Warehouse

Let us execute our T-SQL Script to create data warehouse with fact tables, dimensions and populate them with appropriate test values.

Download T-SQL script attached with this article for creation of Sales Data Warehouse or download from this article "**Create First Data Warehouse**" and run it in your SQL Server.

Follow the given steps to run the query in **SSMS** (SQL Server Management Studio).

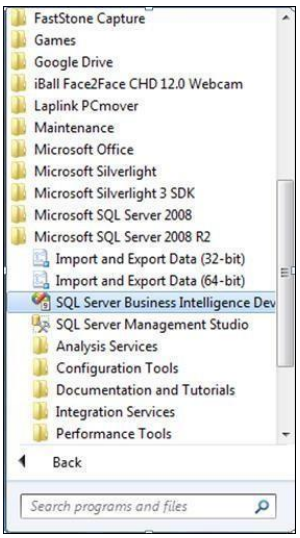
1. Open SQL Server Management Studio 2008
2. Connect Database Engine
3. Open **New Query** editor
4. Copy paste Scripts given below in various steps in new query editor window one by one
5. To run the given SQL Script, press **F5**
6. It will create and populate "Sales_DW" database on your SQL Server

Developing an OLAP Cube

For creation of OLAP Cube in Microsoft BIDS Environment, follow the 10 easy steps given below.

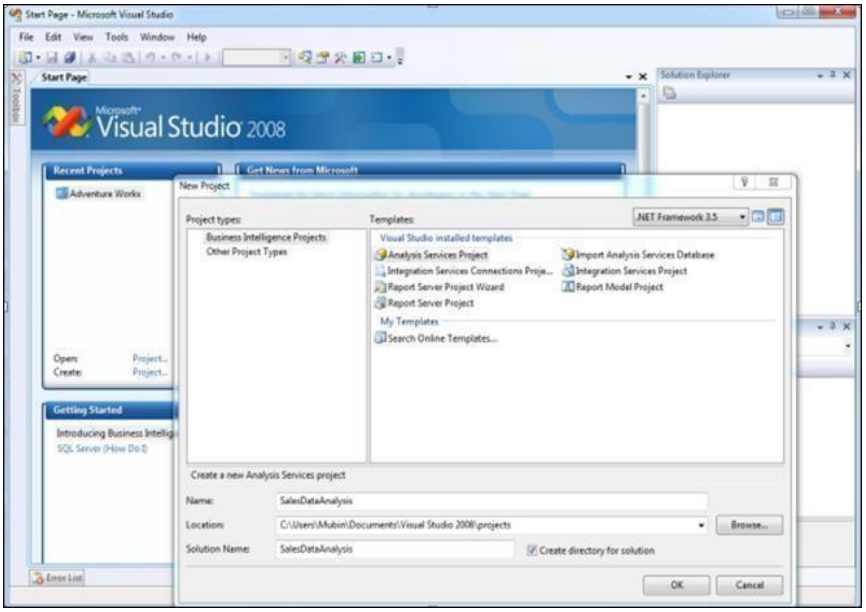
Step 1: Start BIDS Environment

Click on **Start Menu** -> Microsoft **SQL Server 2008 R2** -> Click **SQL Server Business Intelligence Development Studio.**



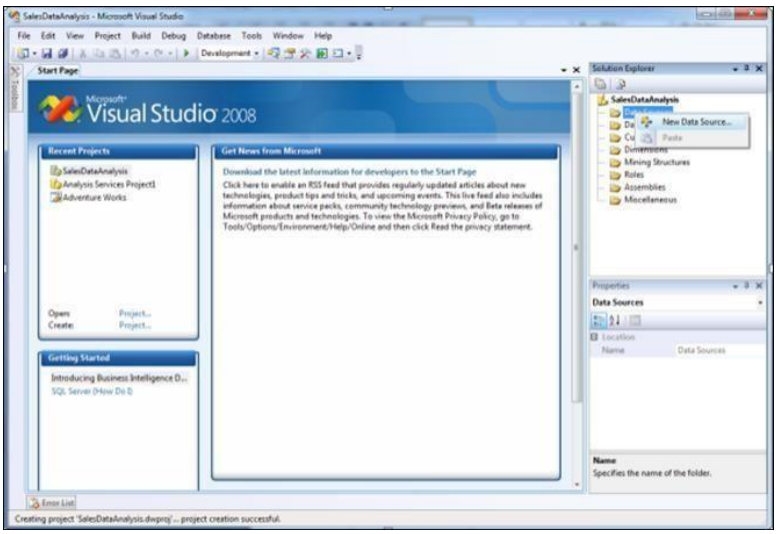
Step 2: Start Analysis Services Project

Click **File** -> **New** -> **Project** ->**Business Intelligence Projects** ->select **Analysis Services Project**-> Assign Project **Name** -> Click **OK**



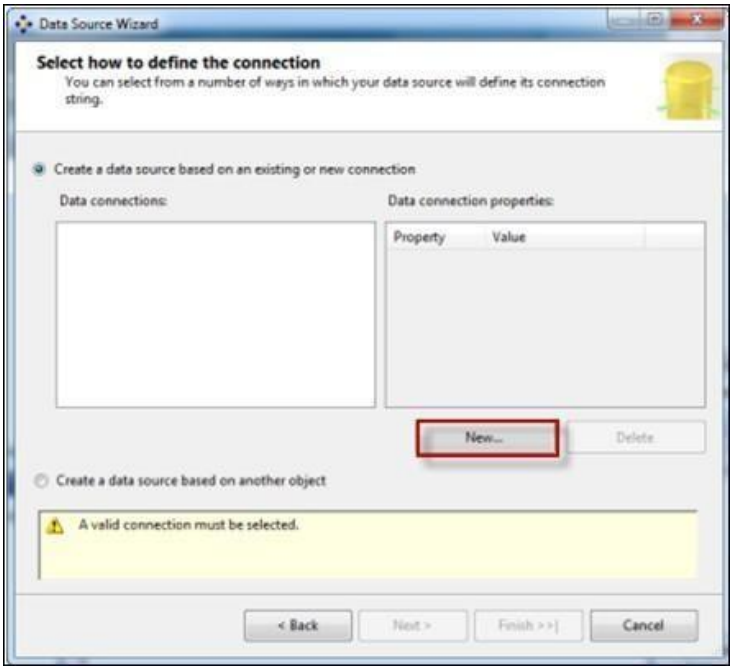
Step 3: Creating New Data Source

In Solution Explorer, Right click on **Data Source** -> Click **New Data Source**



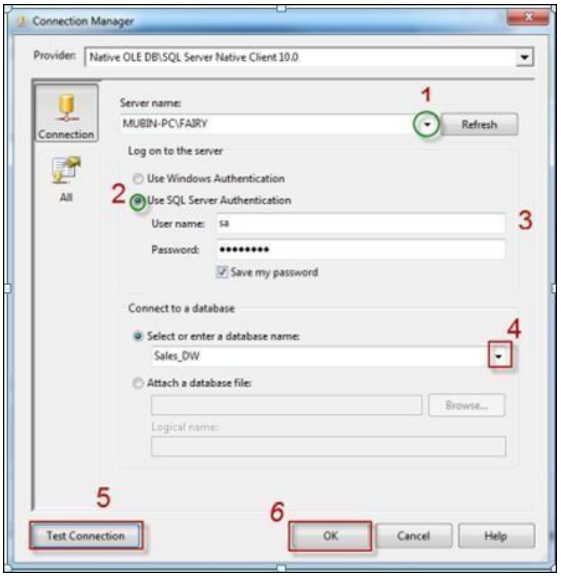
Click on **Next**

Click on New **Button**

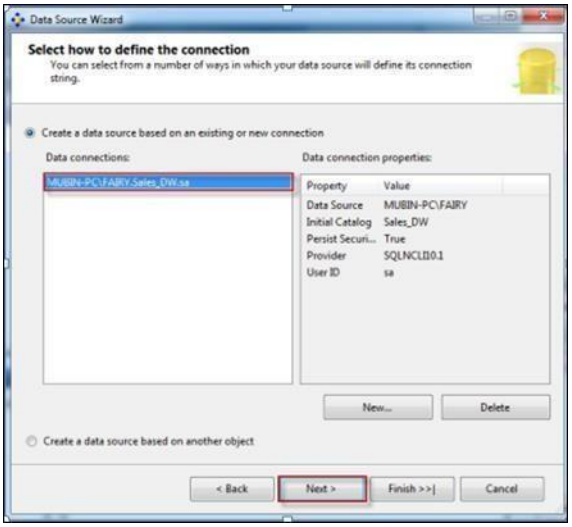


Creating **New connection**

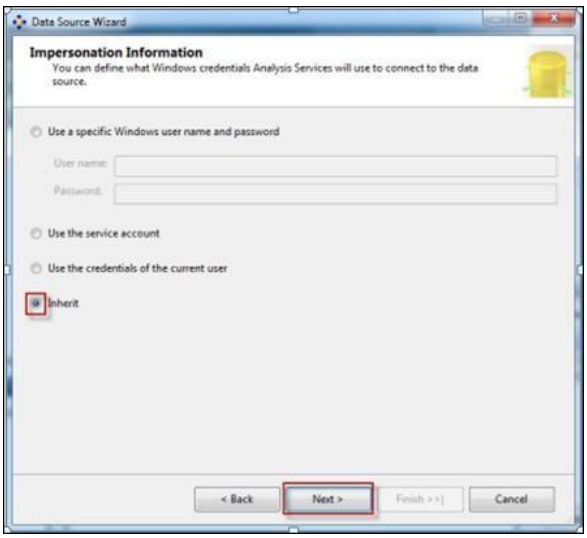
1. Specify Your SQL **Server Name** where your Data Warehouse was created
2. Select Radio Button according to your **SQL Server Authentication** mode 3. Specify your **Credentials** using which you can connect to your SQL Server
4. Select database Sales_DW.
5. Click on **Test Connection** and verify for its success
6. Click **OK**.



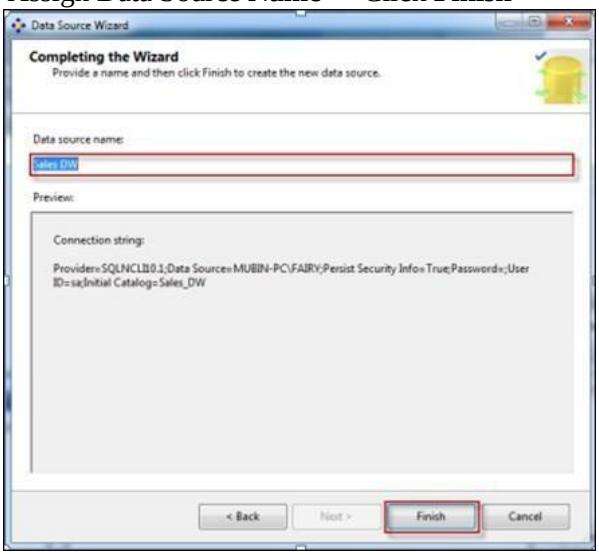
Select Connection created in **Data Connections**-> Click **Next**

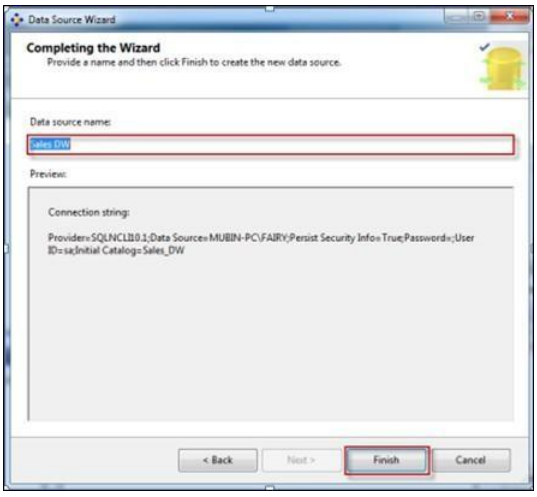


Select Option **Inherit**



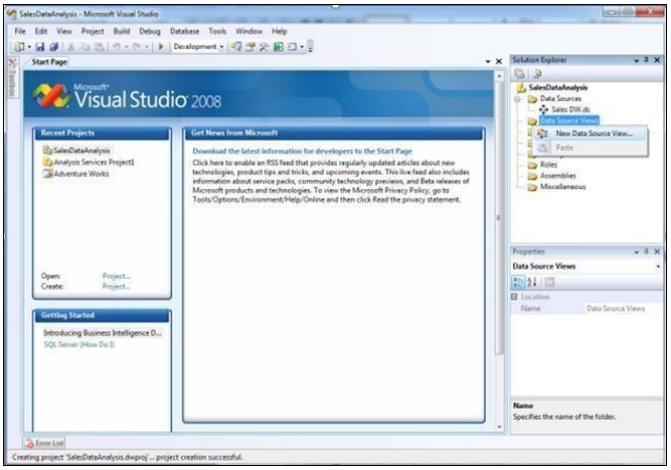
Assign Data Source **Name** -> Click **Finish**



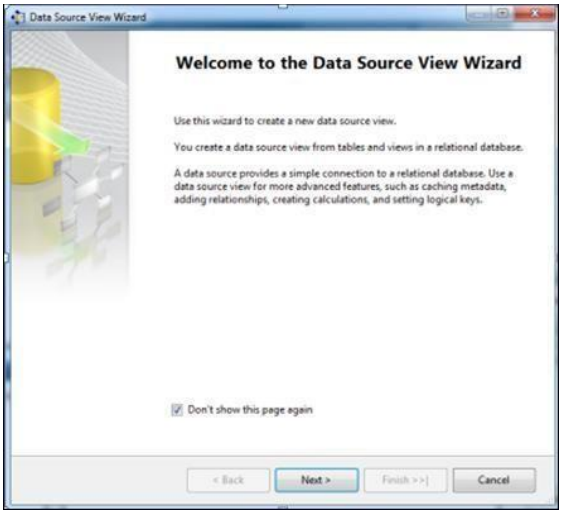


Step 4: Creating New Data Source View

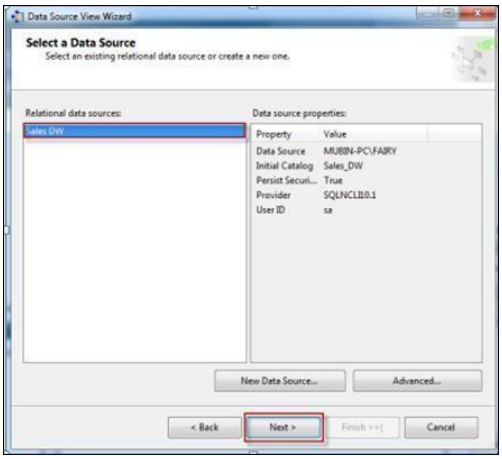
In the Solution Explorer, Right Click on **Data Source View** -> Click on **New Data Source View**



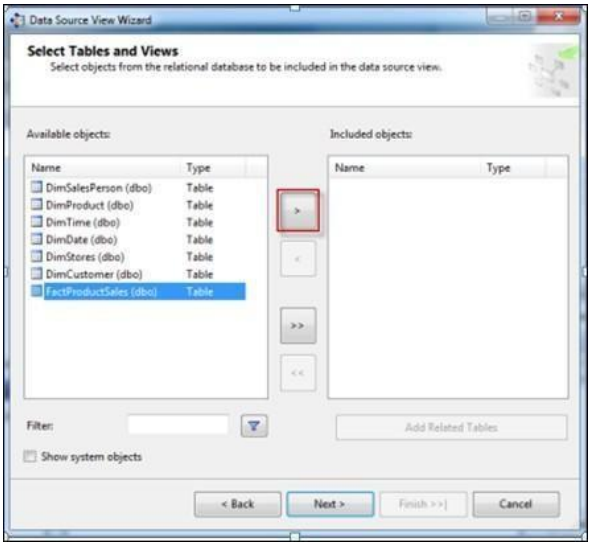
Click **Next**



Select **Relational Data Source** we have created previously (Sales_DW)-> Click **Next**

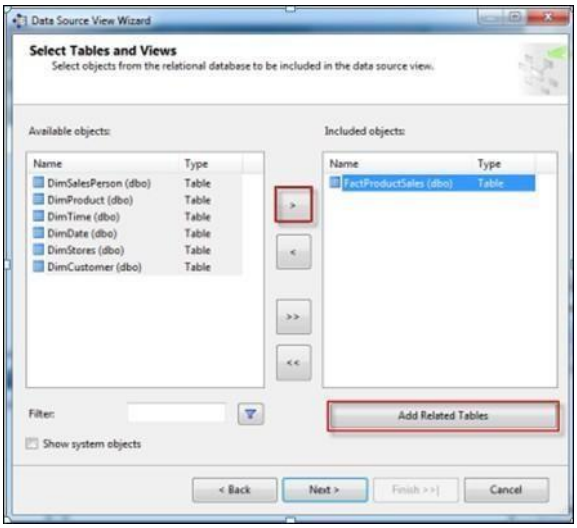


First move your **Fact Table** to the right side to include in object list.



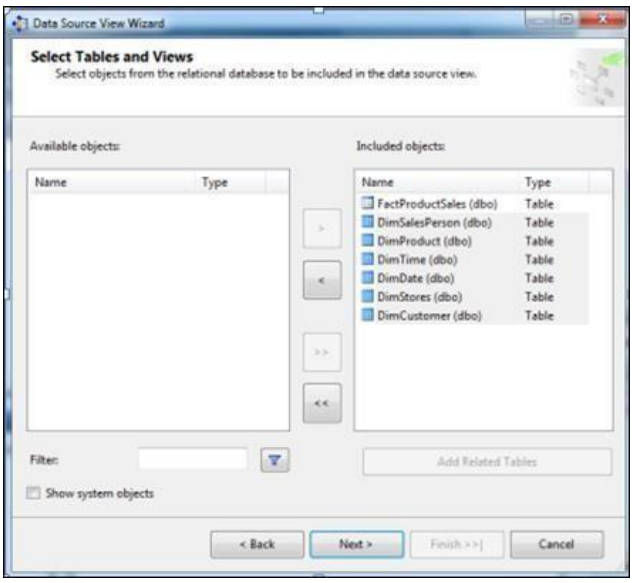
Select FactProductSales Table -> Click on Arrow Button to move the selected object to Right Pane.

Now to **add dimensions** which are **related** to your **Fact Table**, follow the given steps: Select **Fact Table** in Right Pane (Fact product Sales) -> Click on **Add Related Tables**

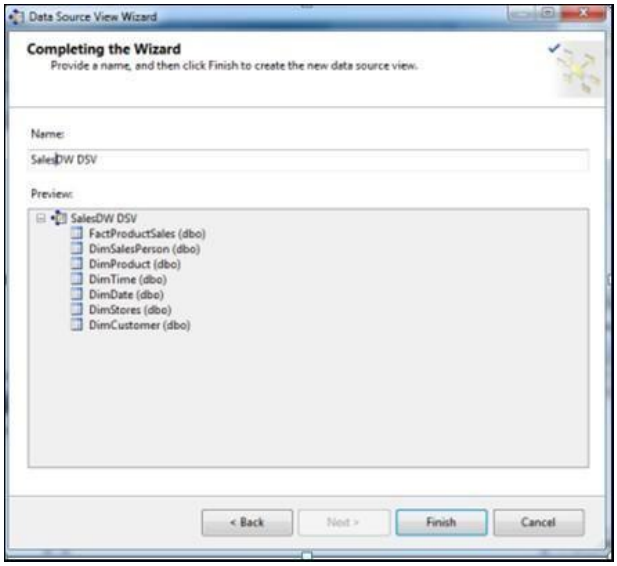


It will add all associated dimensions to your Fact table as per relationship specified in your SQL DW (Sales_DW).

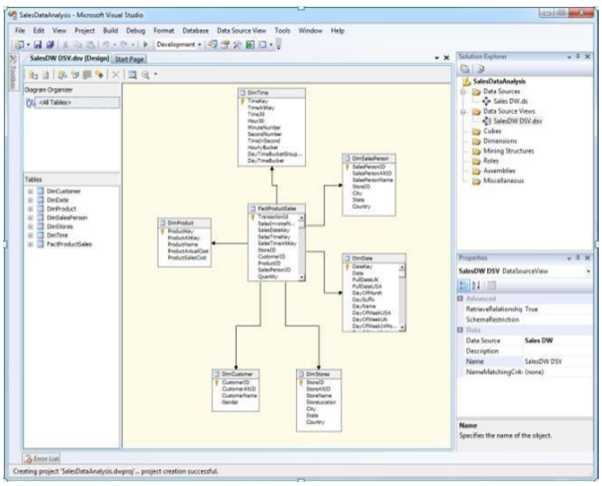
Click **Next**.



Assign **Name (SalesDW DSV)->** Click **Finish**

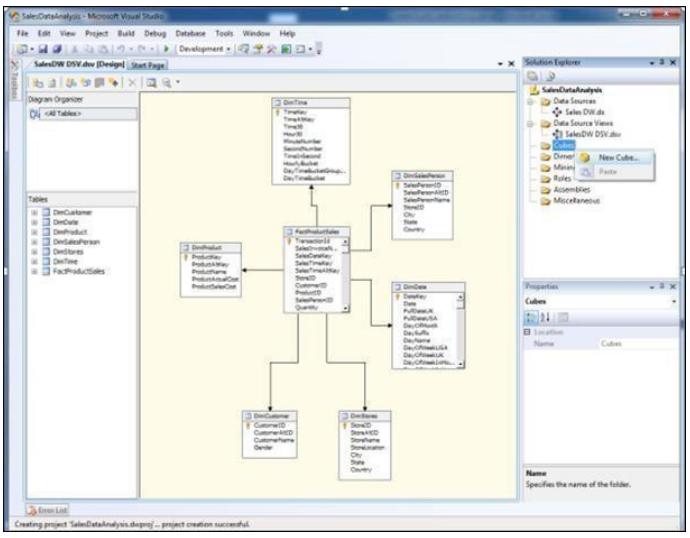


Now Data Source View is ready to use.

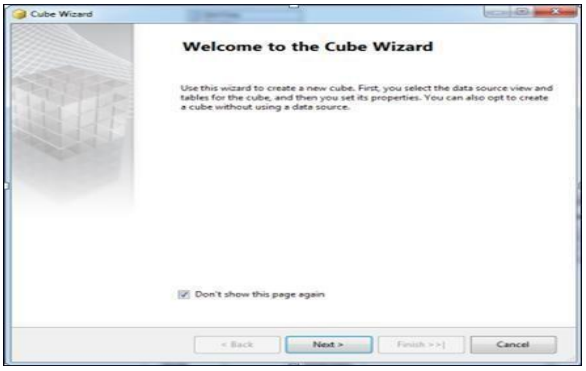


Step 5: Creating New Cube

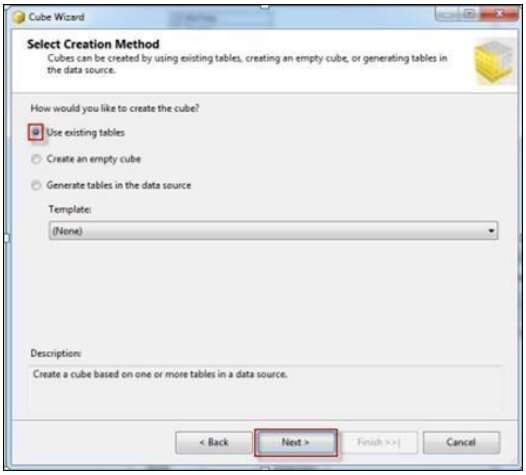
In Solution Explorer -> Right Click on **Cube->** Click **New Cube**



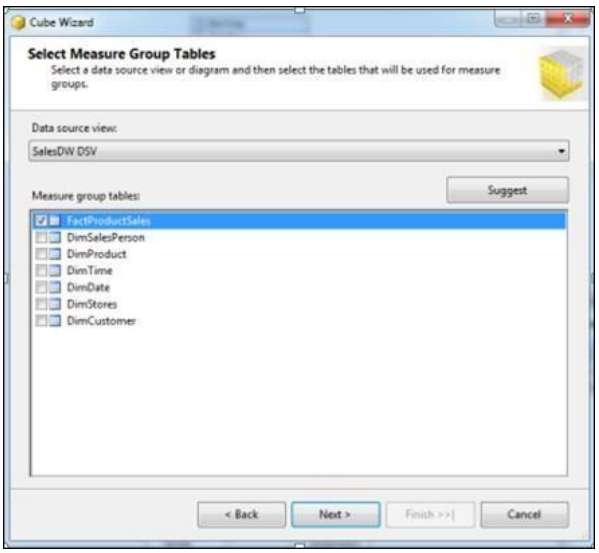
Click **Next**



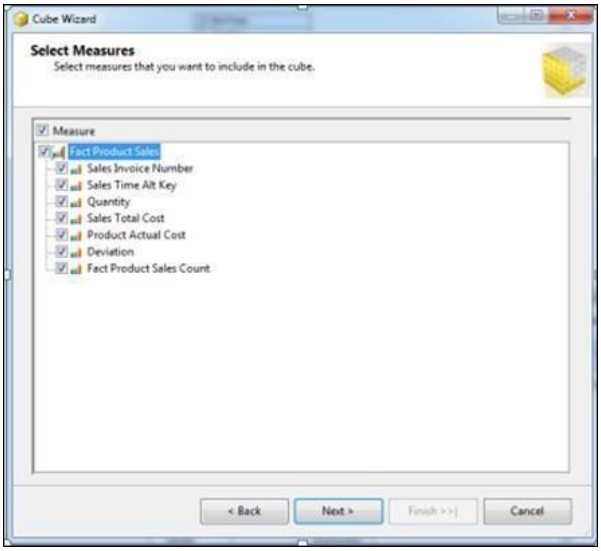
Select Option **Use Existing Tables ->** Click **Next**



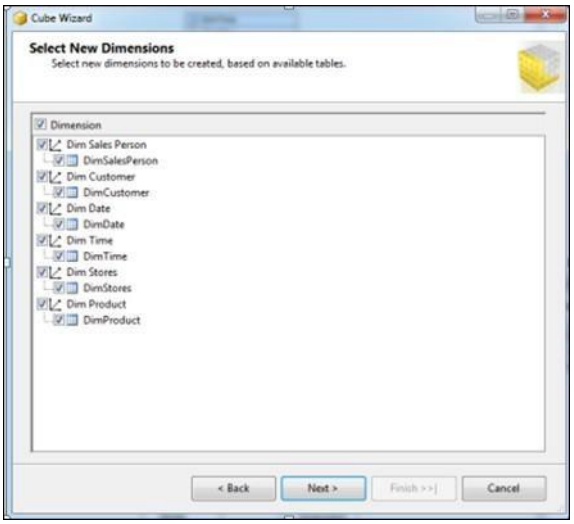
Select Fact Table Name from **Measure Group Tables (FactProductSales) ->** Click **Next**



Choose **Measures** from the List which you want to place in your Cube --> Click **Next**

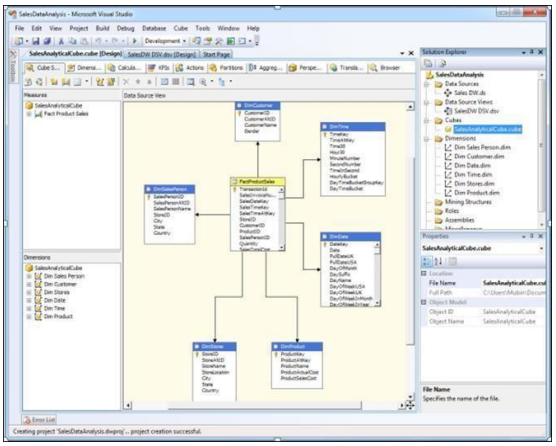


Select All **Dimensions** here which are associated with your Fact Table-> Click **Next**



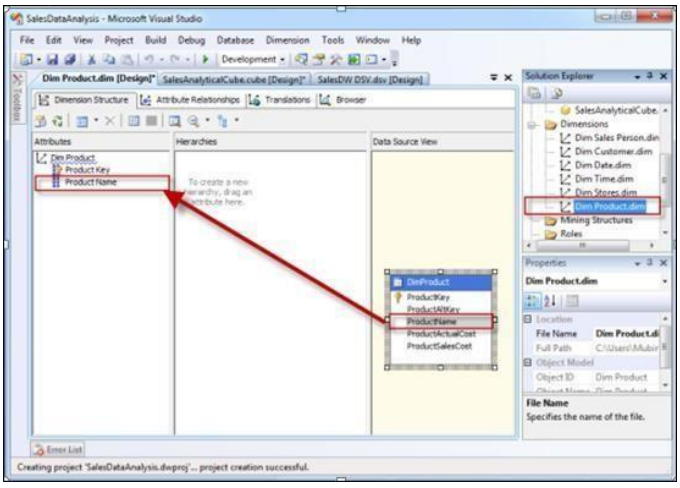
Assign **Cube Name** (SalesAnalyticalCube) -> Click **Finish**

Now your Cube is ready, you can see the newly created cube and dimensions added in your solution explorer.



Step 6: Dimension Modification

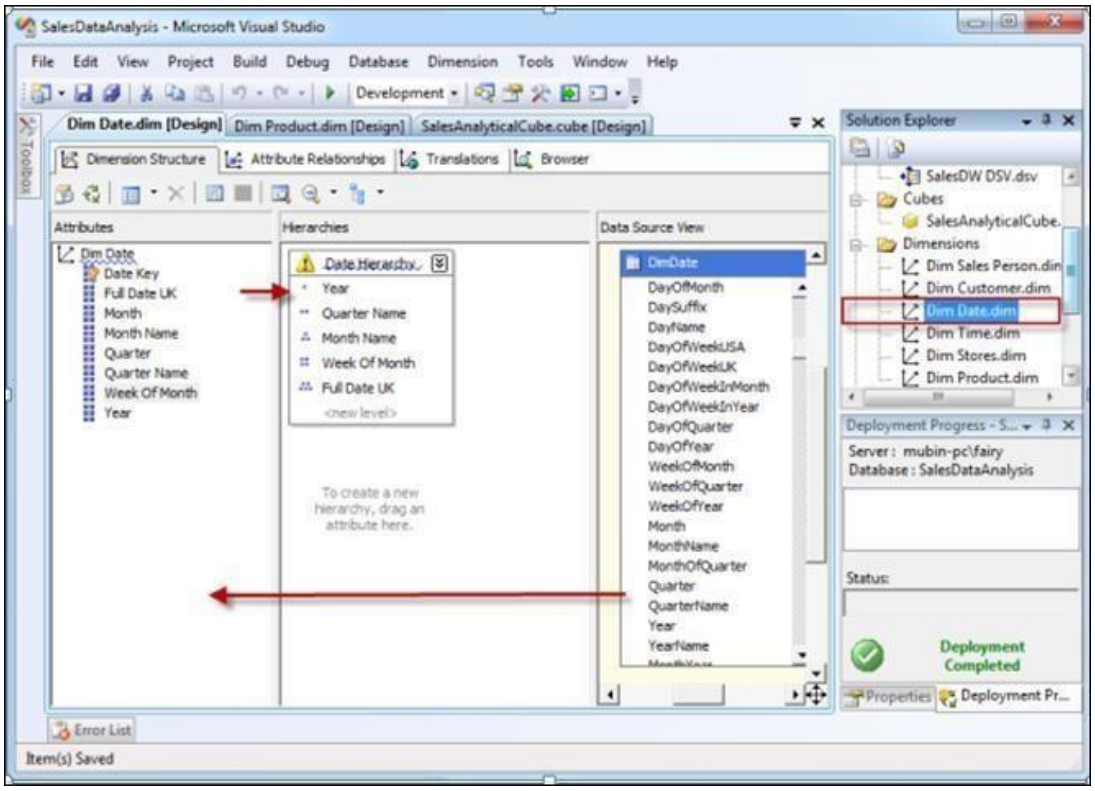
In Solution Explorer, double click on dimension **Dim Product ->** Drag and Drop Product Name from Table in Data Source View and Add in Attribute Pane at left side.



Step 7: Creating Attribute Hierarchy In Date Dimension

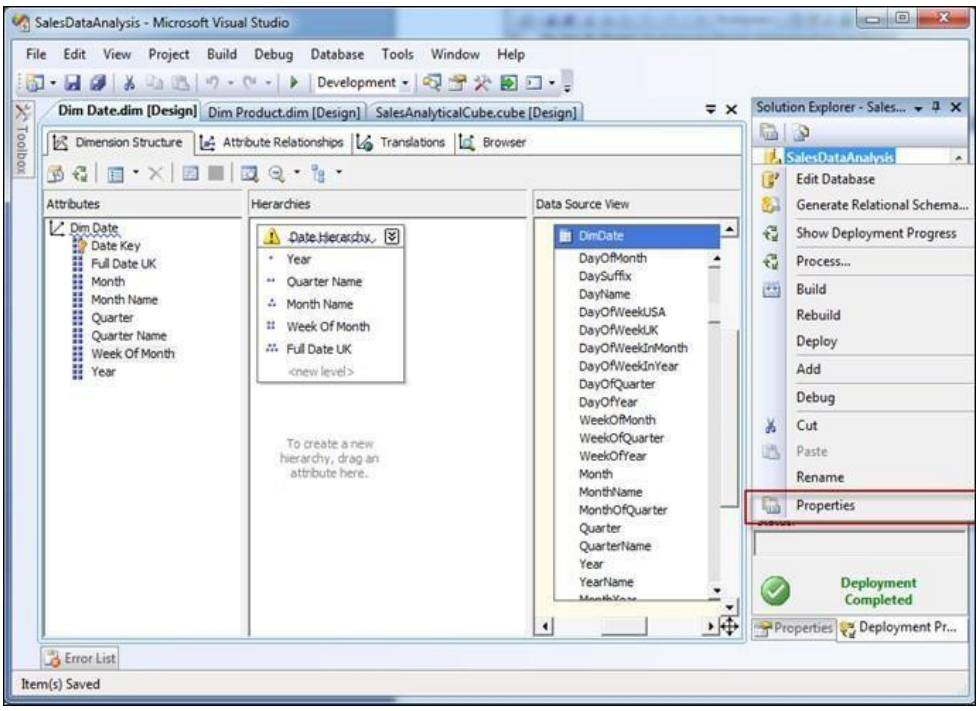
Double click On **Dim Date** dimension -> Drag and Drop Fields from Table shown in Data Source View to Attributes-> Drag and Drop attributes from leftmost pane of attributes to middle pane of Hierarchy.

Drag fields in sequence from Attributes to Hierarchy window (Year, Quarter Name, Month Name, Week of the Month, Full Date UK),



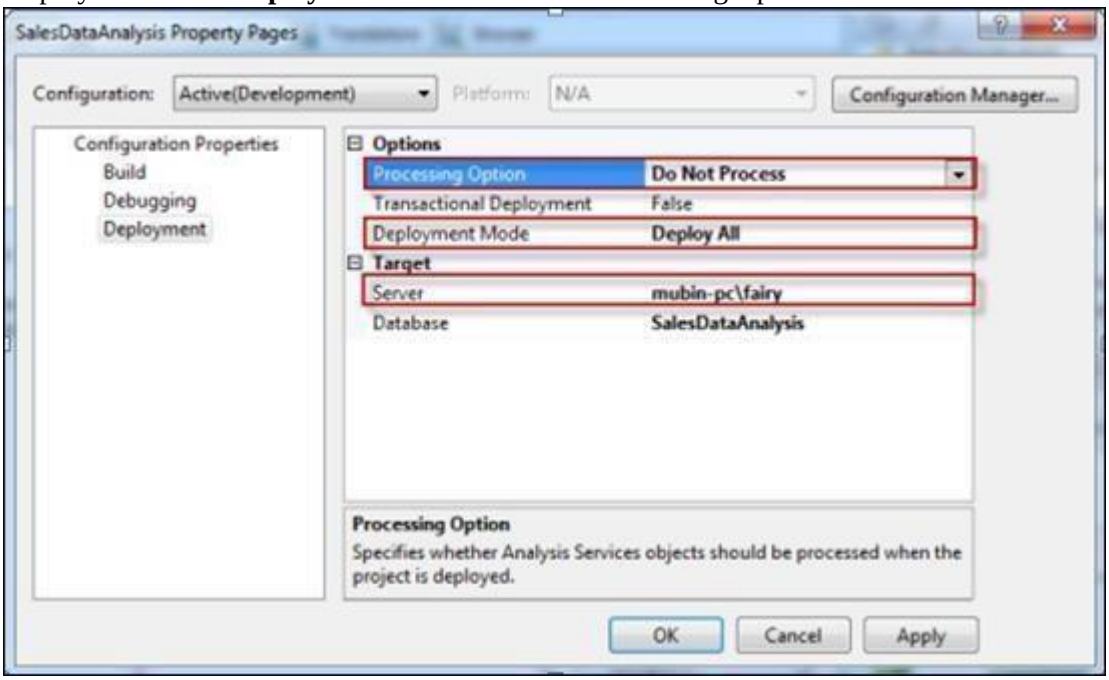
Step 8: Deploy the Cube

In Solution Explorer, right click on Project Name (SalesDataAnalysis) -- > Click **Properties**

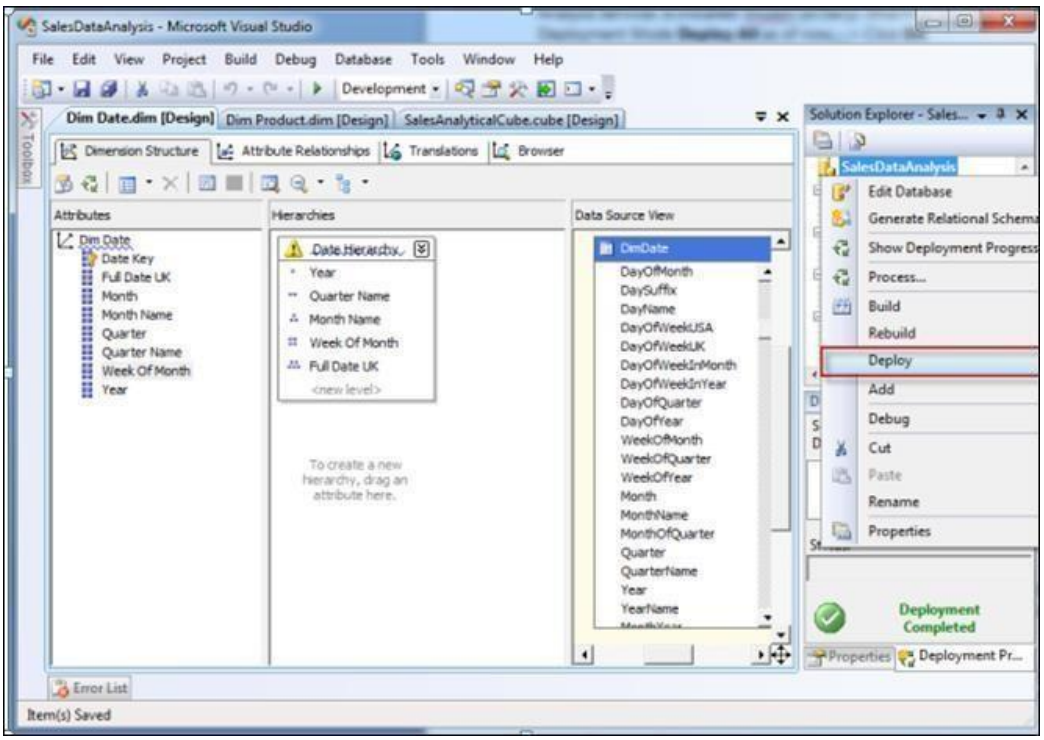


Set **Deployment Properties** First

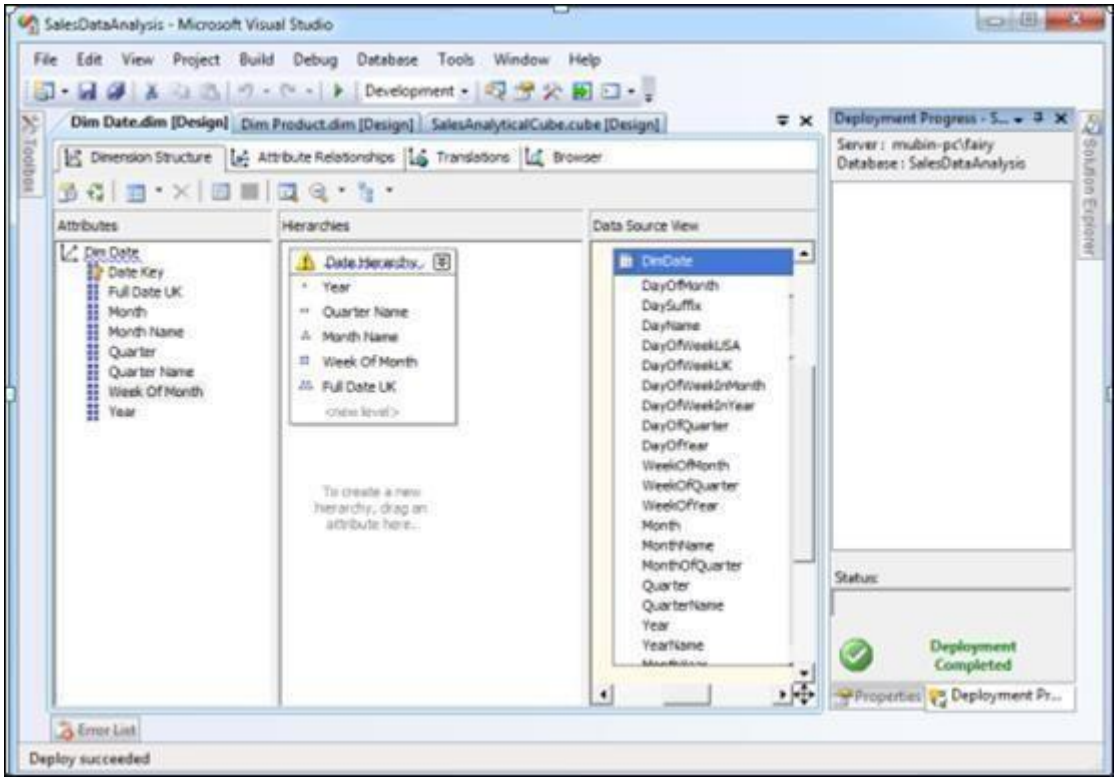
In Configuration Properties, Select Deployment-> Assign Your SQL Server Instance Name Where Analysis Services Is Installed (*mubin-pc\fairy*) (*Machine Name\Instance Name*) -> Choose Deployment Mode **Deploy All** as of now ->Select Processing Option **Do Not Process** -> Click **OK**



In Solution Explorer, right click on **Project Name** (SalesDataAnalysis) -- > Click **Deploy**

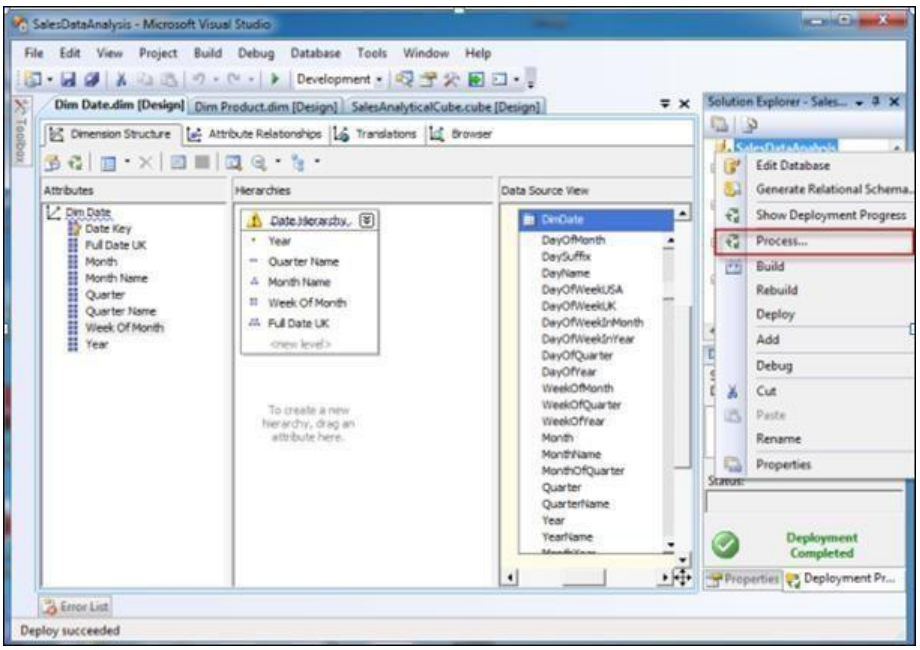


Once Deployment will finish, you can see the message **Deployment Completed** in deployment Properties.

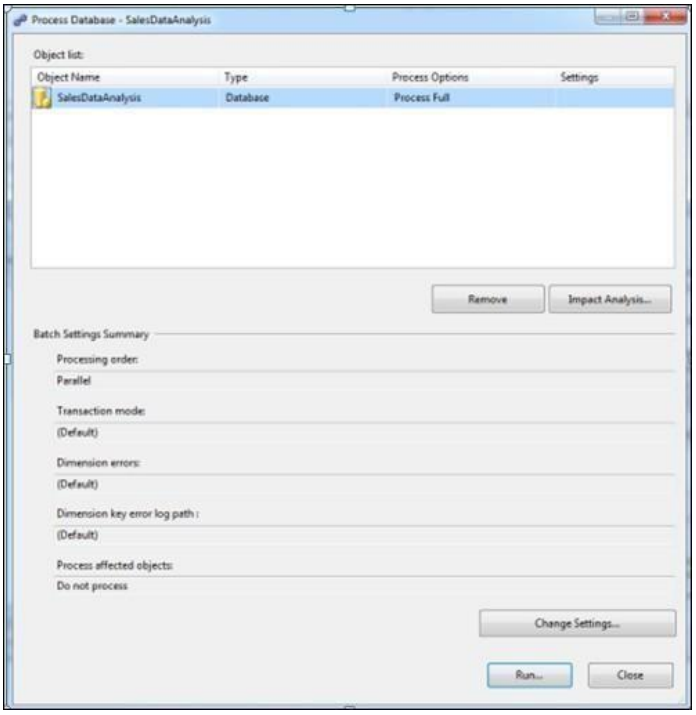


Step 9: Process the Cube

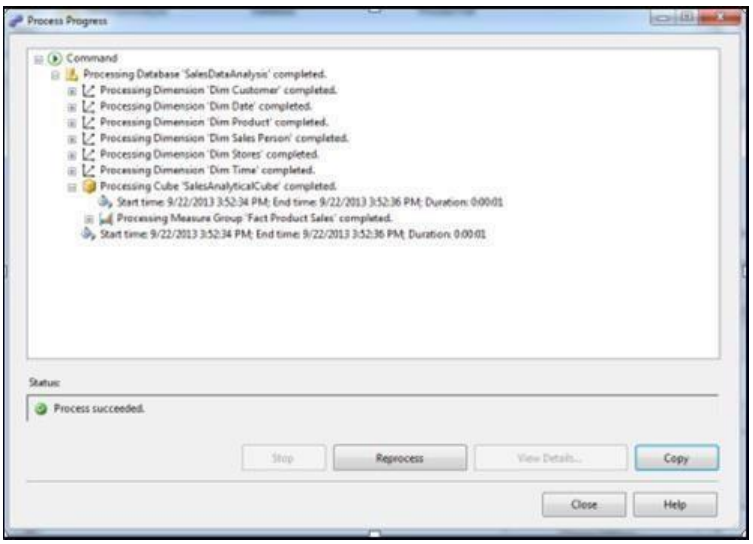
In Solution Explorer, right click on Project Name (SalesDataAnalysis) -- > Click **Process**



Click on **Run** button to process the Cube

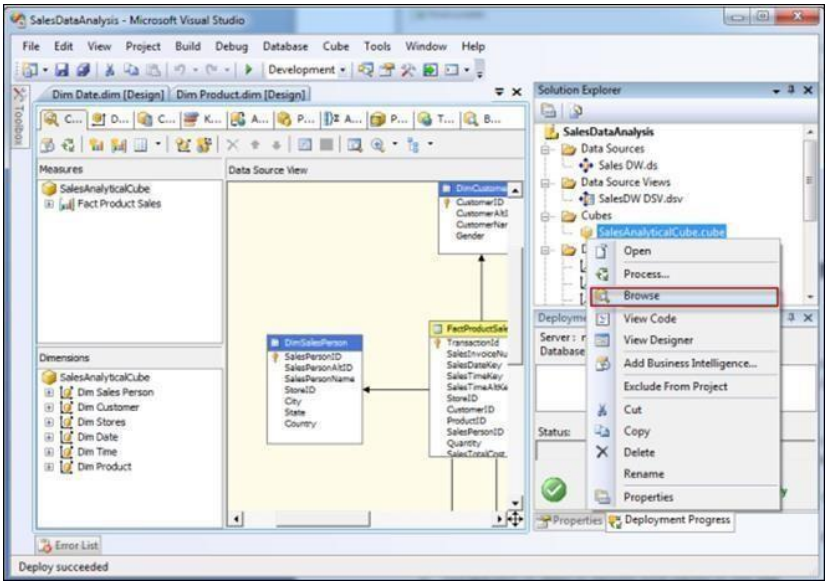


Once processing is complete, you can see **Status** as **Process Succeeded** -->Click **Close** to close both the open windows for processing one after the other.



Step 10: Browse the Cube for Analysis

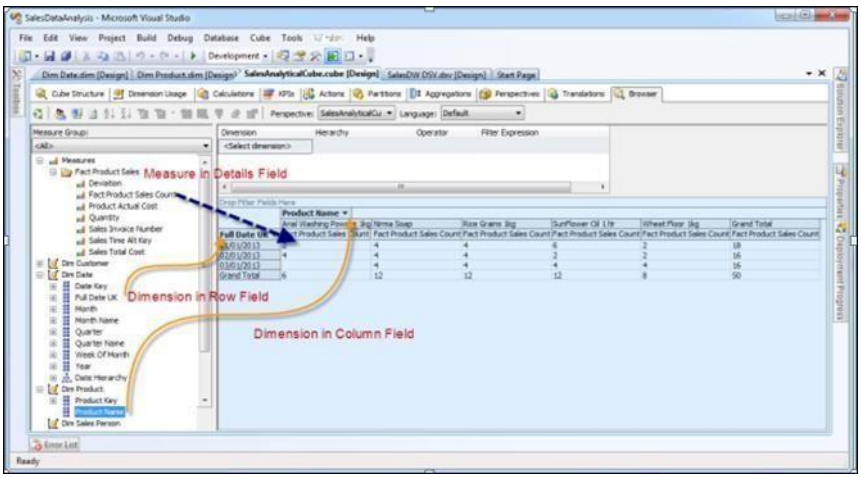
In Solution Explorer, right click on Cube Name (SalesDataAnalysisCube) -- > Click **Browse**



Drag and drop measures in to Detail fields, & Drag and Drop Dimension Attributes in Row Field or Column fields.

Now to **Browse Our Cube**

1. Product Name Drag & Drop into Column
2. Full Date UK Drag & Drop into Row Field
3. FactProductSalesCount Drop this measure in Detail area



Conclusion:

With the help of visual studio, the cube was created successfully.

Assignment No. 4

Title: Pivot table and Pivot Chart

Learning Objectives:

Learn to import the data warehouse data in Microsoft Excel and create the Pivot table and Pivot Chart

Problem Statement:

Import the data warehouse data in Microsoft Excel and create the Pivot table and Pivot Chart

Theory Concepts:

Pivot table

A PivotTable is a powerful tool to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data. A pivot table is a table of grouped values that aggregates the individual items of a more extensive table within one or more discrete categories. This summary might include sums, averages, or other statistics, which the pivot table groups together using a chosen aggregation function applied to the grouped values.

Pivot chart

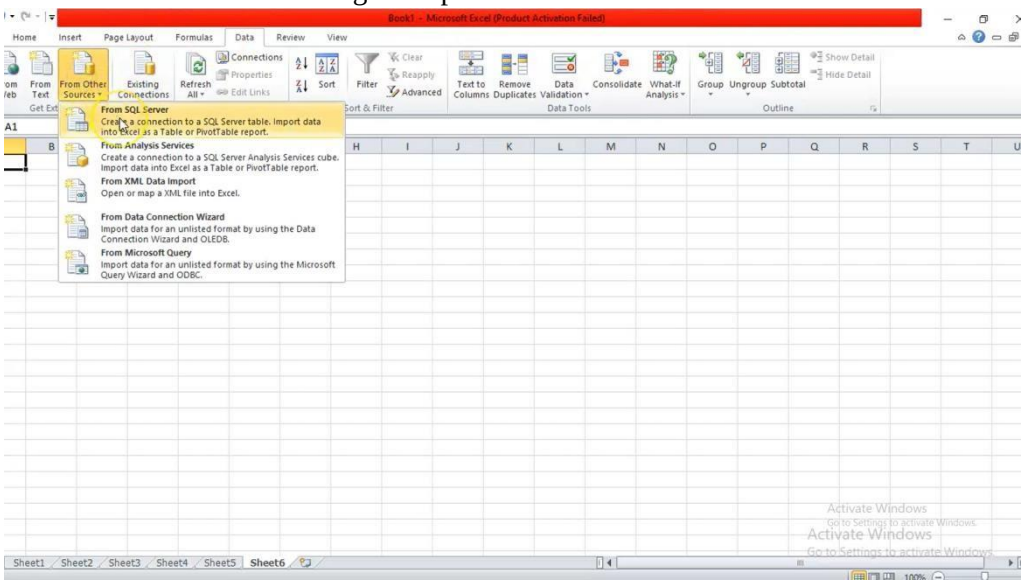
A pivot chart is the visual representation of a pivot table in Excel. Pivot charts and pivot tables are connected with each other.

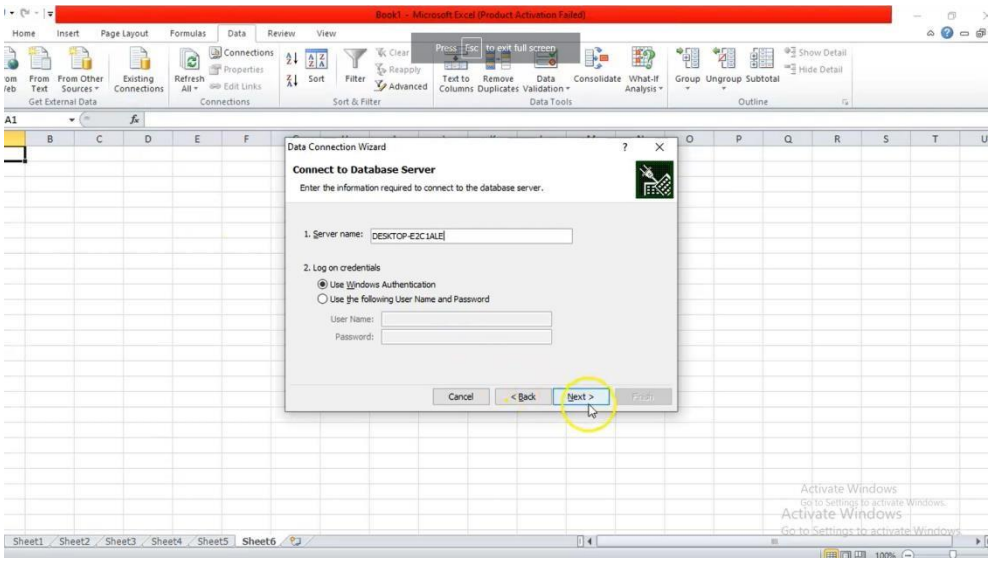
Create a Pivot Chart and Table to analyze worksheet data

Step 1 – Open a new blank Workbook in Excel.

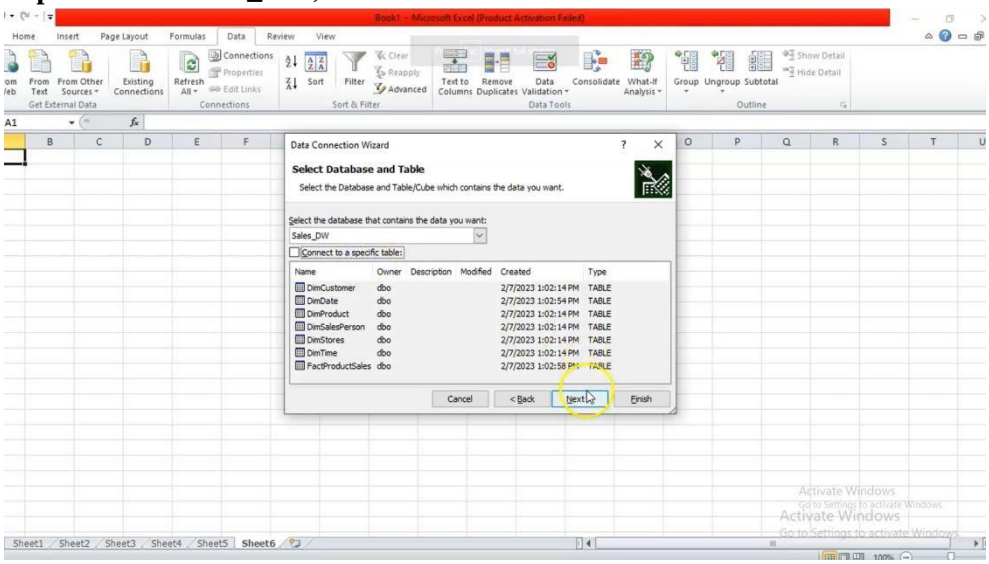
Step 2 – Click on the **DATA** tab.

Step 3 – In the **Get External Data** group, click on the option **From SQL SERVER**. The **Select Data Source** dialog box opens.

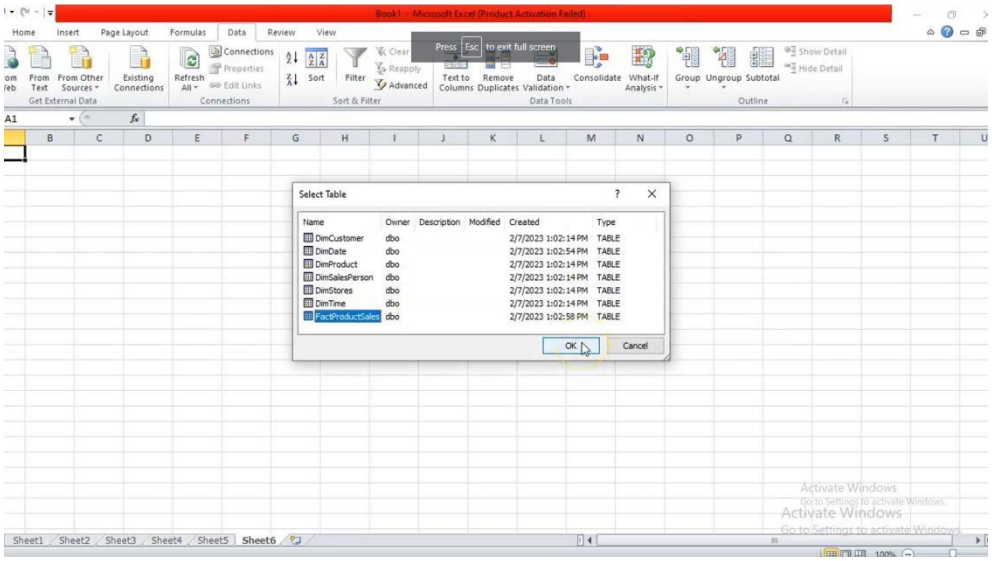




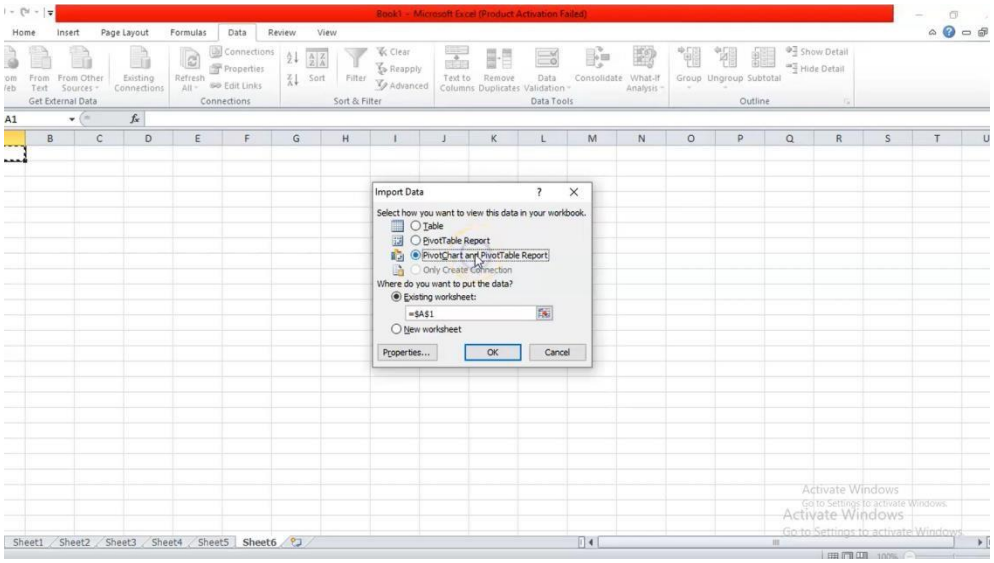
Step 4 − Select **Sales_DW**, Database file.



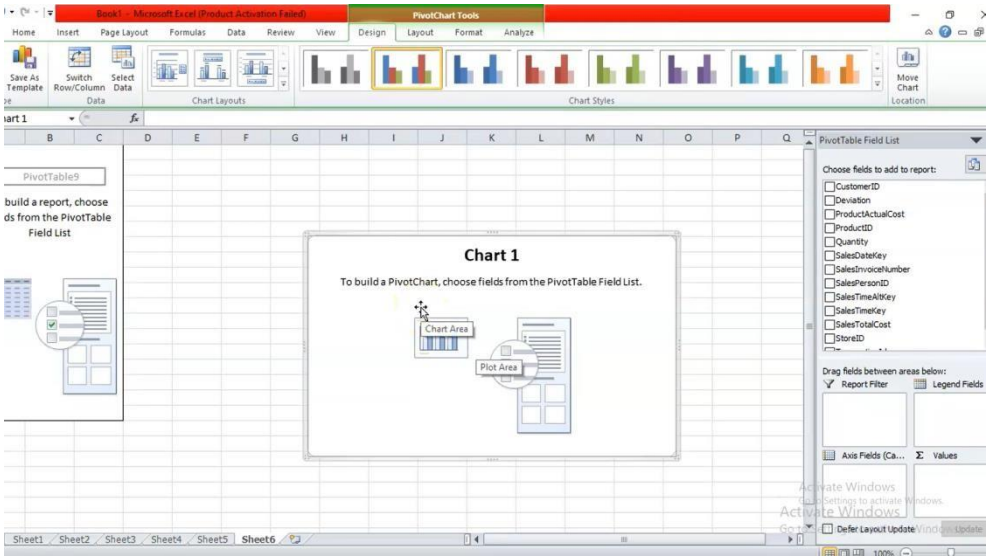
Step 5 − The **Select Table** window, displaying all the **tables** found in the database, appears.



Step 6 – From Tables in a database select the fact table. Then click **OK**.



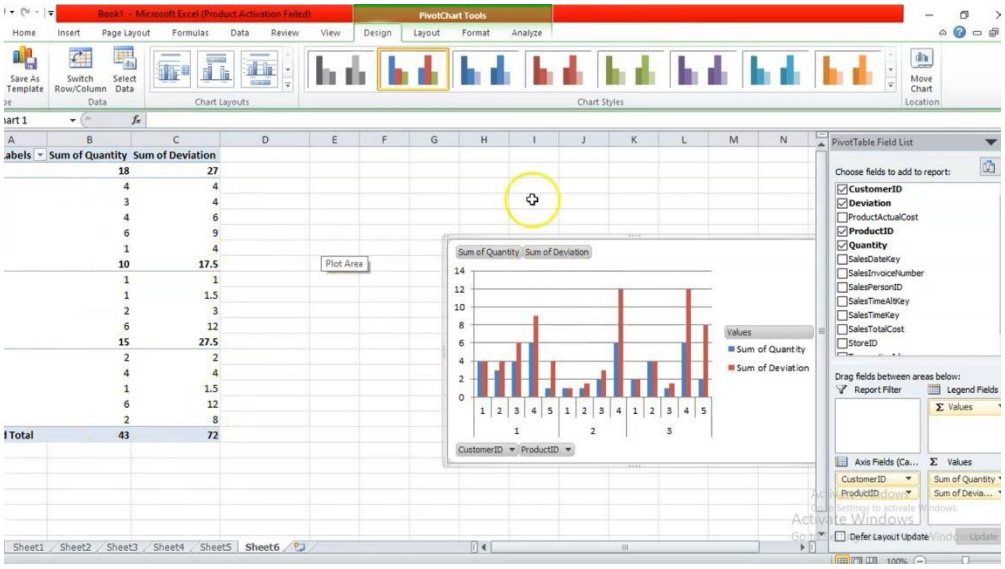
Step 7 − The **Import Data** window appears. Select the **PivotChart and PivotTable Report** option. This option imports the tables into Excel and prepares a Pivot Chart and Table for analyzing the imported tables.

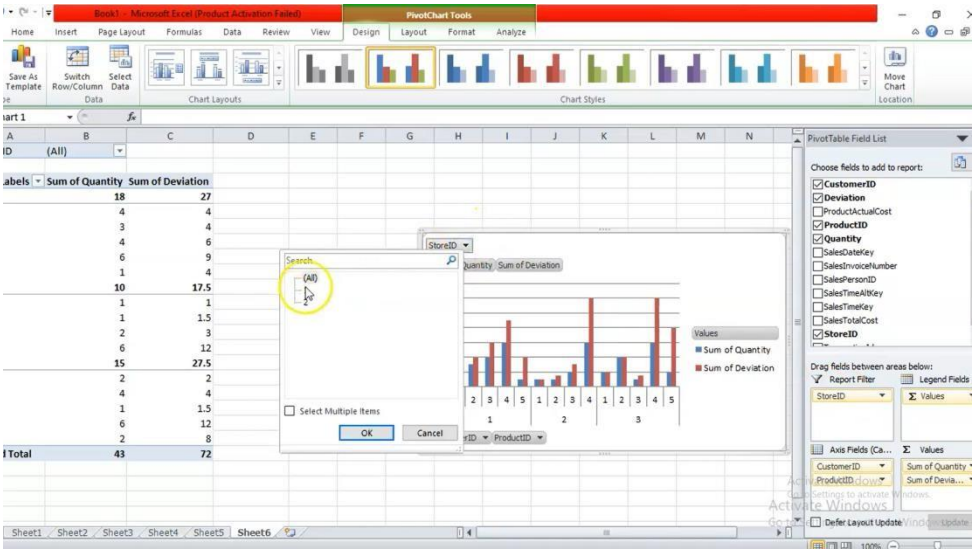


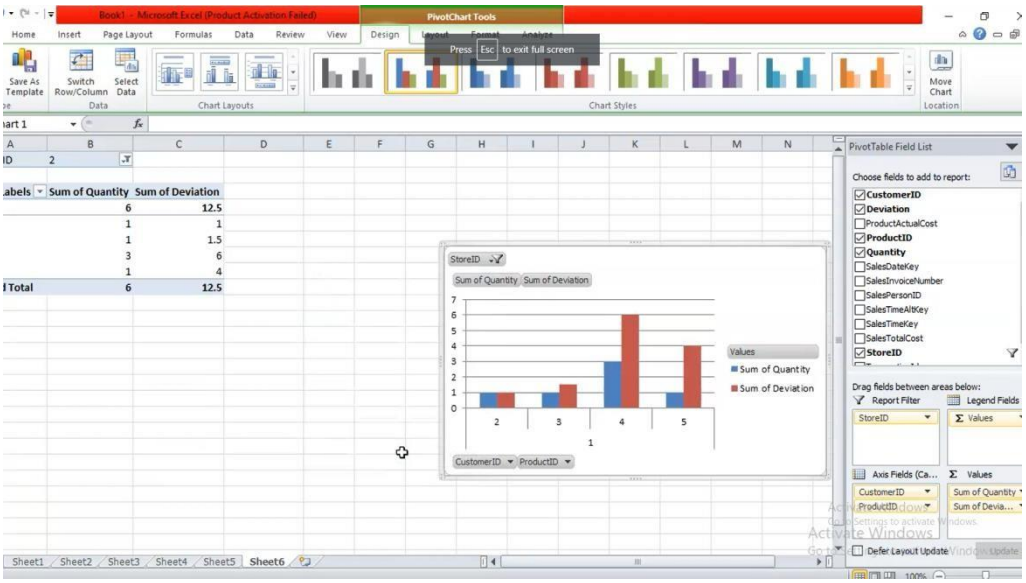
Step 8 – The data is imported, and a **Pivot Chart and Table** is created using the imported tables.

Explore Data Using PivotTable

Step 1 – You know how to add fields to PivotTable and drag fields across areas. Even if you are not sure of the final report that you want, you can play with the data and choose the best- suited report.







Conclusion

With the help of Microsoft Excel the data warehouse data was Imported and the Pivot table and Pivot Chart was created successfully.

Assignment No. 5

Title: Classification and Clustering Algorithms

Learning Objectives:

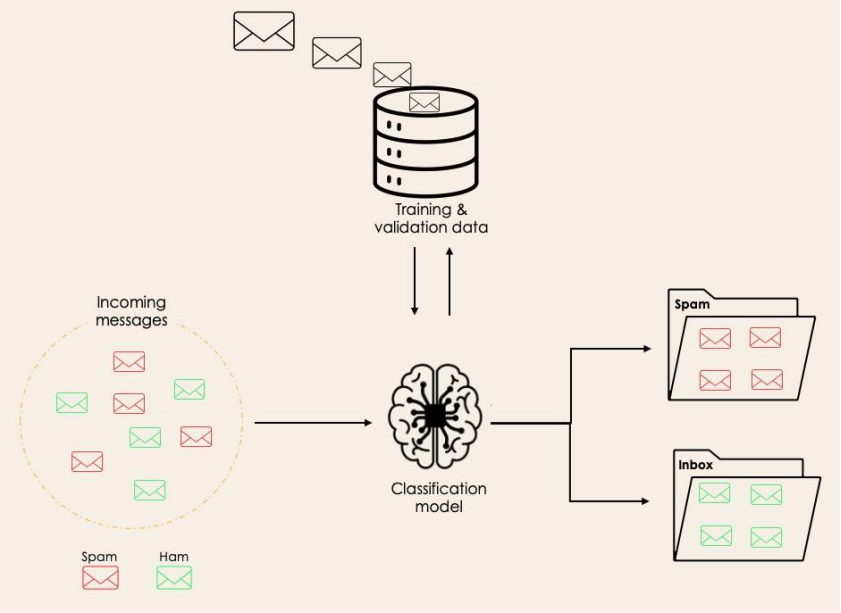
Learn to perform the data classification using classification algorithm or perform the data clustering using clustering algorithm.

Problem Statement: Perform the data classification using classification algorithm. Or Perform the data clustering using clustering algorithm.

Theory Concepts:

A) **Classification:** Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data.

For instance, an algorithm can learn to predict whether a given email is spam or ham (no spam), as illustrated below.



Before diving into the classification concept, we will first understand the difference between the two types of learners in classification: lazy and eager learners. Then we will clarify the misconception between classification and regression.

Lazy Learners Vs. Eager Learners

There are two types of learners in machine learning classification: lazy and eager learners.

Eager learners are machine learning algorithms that first build a model from the training dataset before making any prediction on future datasets. They spend more time during the training process because of their eagerness to have a better generalization during the training from learning the weights, but they require less time to make predictions.

Most machine learning algorithms are eager learners, and below are some examples:

- Logistic Regression.
- Support Vector Machine.
- Decision Trees.
- Artificial Neural Networks.

Lazy learners or instance-based learners, on the other hand, do not create any model immediately from the training data, and this is where the lazy aspect comes from. They just memorize the training data, and each time there is a need to make a prediction, they search for the nearest neighbor from the whole training data, which makes them very slow during prediction. Some examples of this kind are:

K-Nearest Neighbor.

- Case-based reasoning.

However, some algorithms, such as **BallTrees** and **KDTrees**, can be used to improve the prediction latency.

Machine Learning Classification Vs. Regression

There are four main categories of Machine Learning algorithms: supervised, unsupervised, semi supervised, and reinforcement learning.

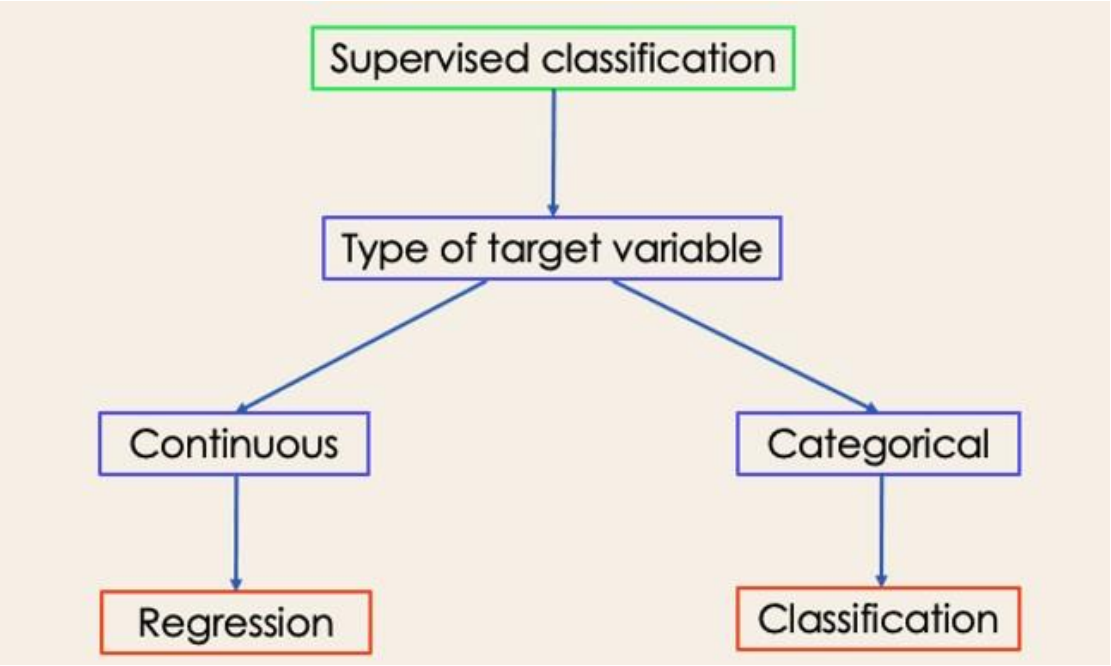
Even though classification and regression are both from the category of supervised learning, they are not the same.

- The prediction task is a **classification** when the target variable is discrete. An application is the identification of the underlying sentiment of a piece of text.

- The prediction task is a **regression** when the target variable is continuous. An example can be the prediction of the salary of a person given their education degree, previous work experience, geographical location, and level of seniority.

If you are interested in knowing more about classification, courses on **Supervised Learning with scikit-learn** and **Supervised Learning in R might be helpful**. They provide you with a better understanding of how each algorithm approaches tasks and the Python and R functions required to implement them.

Regarding regression, **Introduction to Regression in R** and **Introduction to Regression with stats models in Python** will help you explore different types of regression models as well as their implementation in R and Python.



Examples of Machine Learning Classification in Real Life

Supervised Machine Learning Classification has different applications in multiple domains of our day-to-day life. Below are some examples.

Healthcare

Training a machine learning model on historical patient data can help healthcare specialists accurately analyze their diagnoses:

-
- During the COVID-19 pandemic, machine learning models were implemented to efficiently predict whether a person had COVID-19 or not.
- Researchers can use machine learning models to predict new diseases that are more likely to emerge in the future.

Education

Education is one of the domains dealing with the most textual, video, and audio data. This unstructured information can be analyzed with the help of Natural Language technologies to perform different tasks such as:

- The classification of documents per category.
- Automatic identification of the underlying language of students' documents during their application. Analysis of students' feedback sentiments about a professor.

Transportation

Transportation is the key component of many countries' economic development. As a result, industries are using machine and deep learning models:

- To predict which geographical location will have a rise in traffic volume.
- Predict potential issues that may occur in specific locations due to weather conditions.

Sustainable agriculture

Agriculture is one of the most valuable pillars of human survival. Introducing sustainability can help improve farmers' productivity at a different level without damaging the environment:

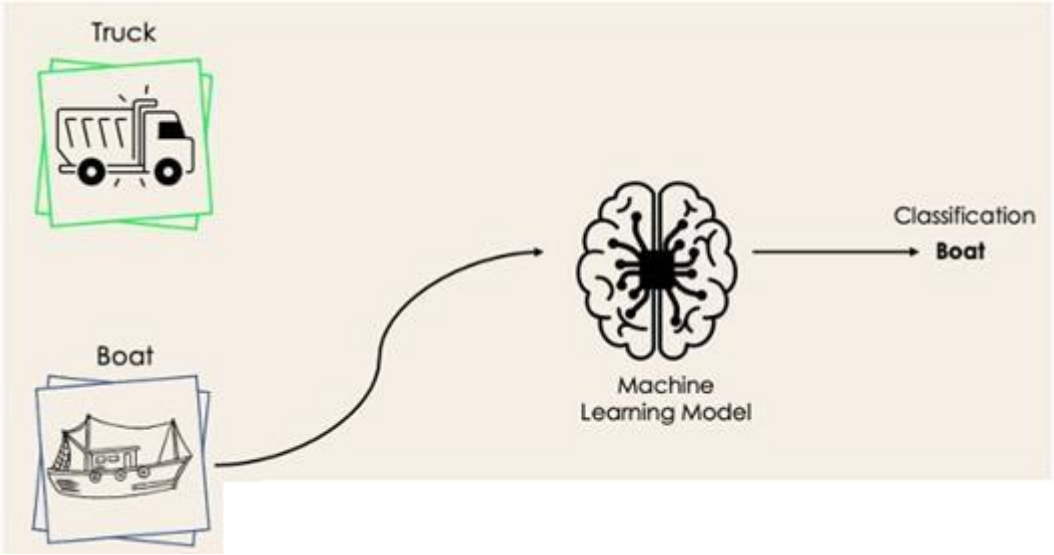
- By using classification models to predict which type of land is suitable for a given type of seed.
- Predict the weather to help them take proper preventive measures.

Different Types of Classification Tasks in Machine Learning

There are four main classification tasks in Machine learning: binary, multi-class, multi-label, and imbalanced classifications.

Binary Classification

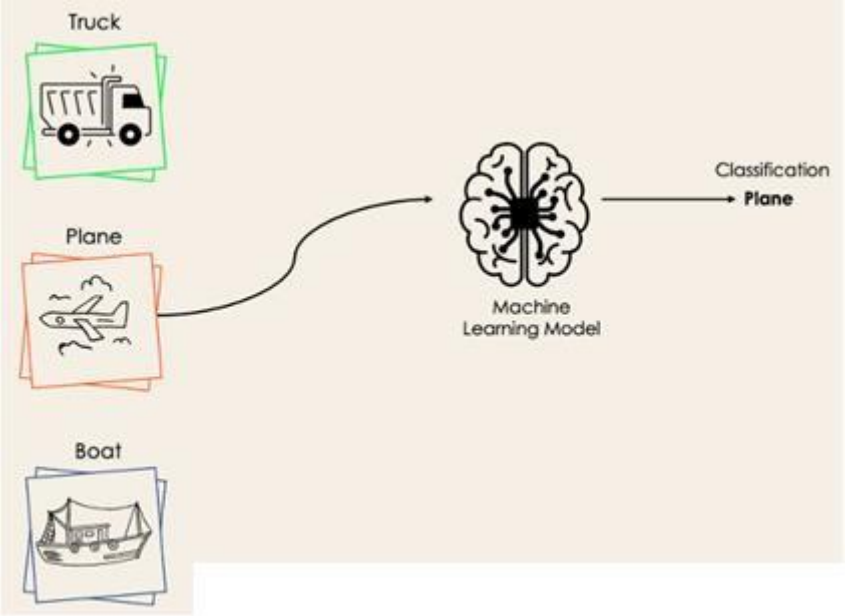
In a binary classification task, the goal is to classify the input data into two mutually exclusive categories. The training data in such a situation is labeled in a binary format: true and false; positive and negative; O and 1; spam and not spam, etc. depending on the problem being tackled. For instance, we might want to detect whether a given image is a truck or a boat.



Logistic Regression and Support Vector Machines algorithms are natively designed for binary classifications. However, other algorithms such as K-Nearest Neighbors and Decision Trees can also be used for binary classification.

Multi-Class Classification

The multi-class classification, on the other hand, has at least two mutually exclusive class labels, where the goal is to predict to which class a given input example belongs to. In the following case, the model correctly classified the image to be a plane.

- 

Most of the binary classification algorithms can be also used for multi-class classification. These algorithms include but are not limited to:

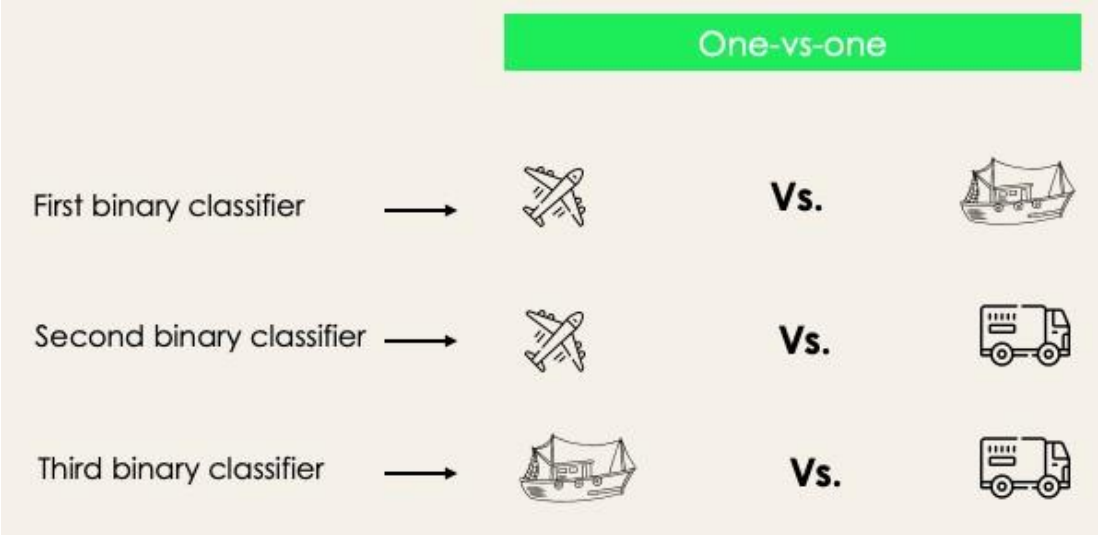
Random Forest

- Naive Bayes
- K-Nearest Neighbors
- Gradient Boosting
- SVM
- Logistic Regression.

But wait! Didn't you say that SVM and Logistic Regression do not support multi-class classification by default?

→ That's correct. However, we can apply binary transformation approaches such as one-versus-one and one-versus-all to adapt native binary classification algorithms for multi-class classification tasks.

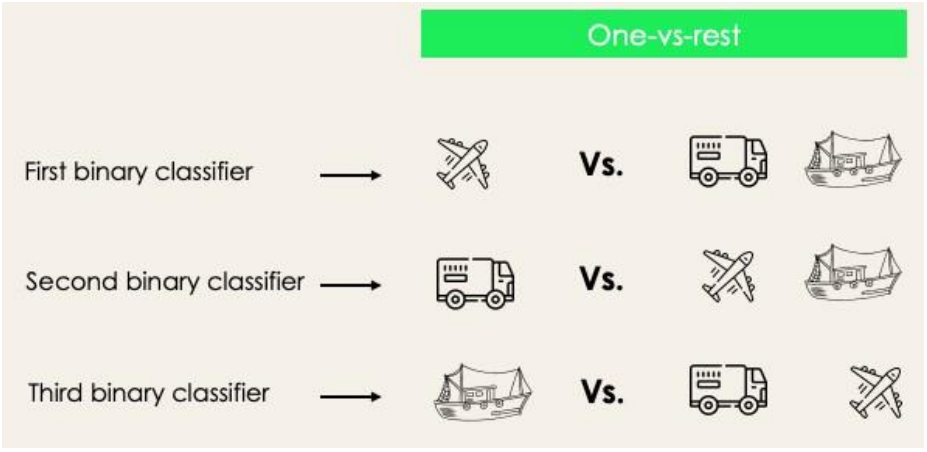
One-versus-one: this strategy trains as many classifiers as there are pairs of labels. If we have a 3class classification, we will have three pairs of labels, thus three classifiers, as shown below.



In general, for N labels, we will have Nx(N-1)/2 classifiers. Each classifier is trained on a single binary dataset, and the final class is predicted by a majority vote between all the classifiers. One-vsone approach works best for SVM and other kernel-based algorithms.

One-versus-rest: at this stage, we start by considering each label as an independent label and consider the rest combined as only one label. With 3-classes, we will have three classifiers.

In general, for N labels, we will have **N** binary classifiers.

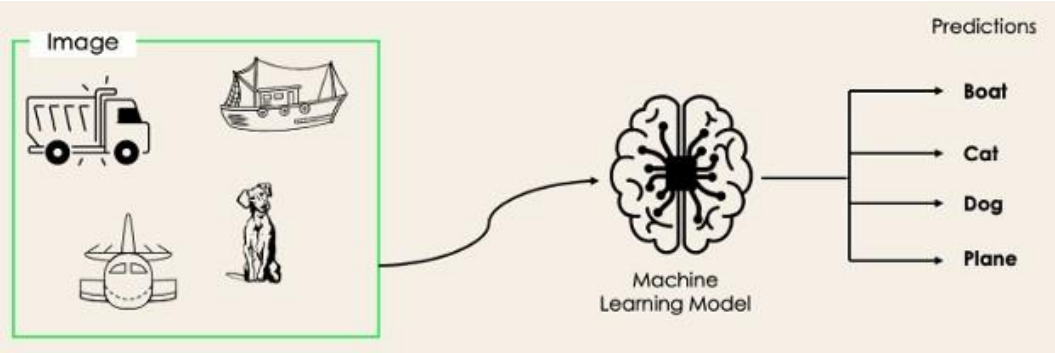


Multi-Label Classification

In multi-label classification tasks, we try to predict 0 or more classes for each input example. In this case, there is no mutual exclusion because the input example can have more than one label.

Such a scenario can be observed in different domains, such as auto-tagging in Natural Language Processing, where a given text can contain multiple topics. Similarly to computer vision, an image can

- contain multiple objects, as illustrated below: the model predicted that the image contains: a plane, a boat, a truck, and a dog.

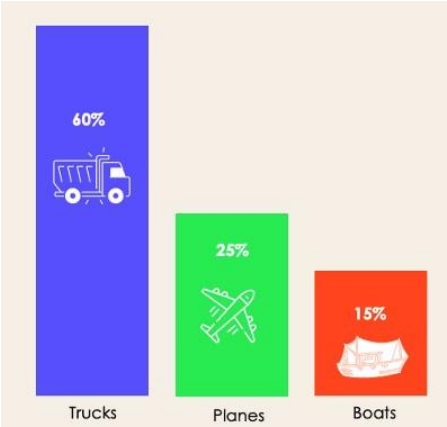


It is not possible to use multi-class or binary classification models to perform multi-label classification. However, most algorithms used for those standard classification tasks have their specialized versions for multi-label classification. We can cite:

- Multi-label Decision Trees
- Multi-label Gradient Boosting
- Multi-label Random Forests

Imbalanced Classification

For the imbalanced classification, the number of examples is unevenly distributed in each class, meaning that we can have more of one class than the others in the training data. Let's consider the following 3-class classification scenario where the training data contains: 60% of trucks, 25% of planes, and 15% of boats.



The imbalanced classification problem could occur in the following scenario:

- Fraudulent transaction detections in financial industries
- Rare disease diagnosis
- Customer churn analysis

Using conventional predictive models such as Decision Trees, Logistic Regression, etc. could not be effective when dealing with an imbalanced dataset, because they might be biased toward predicting the class with the highest number of observations, and considering those with fewer numbers as noise.

So, does that mean that such problems are left behind?

Of course not! We can use multiple approaches to tackle the imbalance problem in a dataset. The most commonly used approaches include sampling techniques or harnessing the power of cost-sensitive algorithms.

Sampling Techniques

These techniques aim to balance the distribution of the original by:

- Cluster-based Oversampling:

- Random under sampling: random elimination of examples from the majority class.
- SMOTE Oversampling: random replication of examples from the minority class.

Cost-Sensitive Algorithms

These algorithms take into consideration the cost of misclassification. They aim to minimize the total cost generated by the models.

- Cost-sensitive Decision Trees.
- Cost-sensitive Logistic Regression.
- Cost-sensitive Support Vector Machines.

Metrics to Evaluate Machine Learning Classification Algorithms

Now that we have an idea of the different types of classification models, it is crucial to choose the right evaluation metrics for those models. In this section, we will cover the most commonly used metrics: accuracy, precision, recall, F1 score, and area under the ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve).

A bit of context

Imagine you are a healthcare startup, and want an AI assistant able to predict whether a given patient has a heart disease or not based on its health record. This is a binary classification problem where the model will predict

- 1, True or Yes if the patient has heart disease
- 0, False or No otherwise

1 Confusion matrix

A 2X2 matrix that nicely summarizes the number of correct predictions of the model. It also helps in computing different other performance metrics.

Predicti / Reality	Yes	No
Yes	True Positives (TP)	False Negatives (FN)
No	False Positives(FP)	True Negatives (TN)

Type I Error

Type II Error

Type I & II Errors can be used interchangeably when referring to False Positives and False negatives respectively

2 Accuracy

We get accuracy by answering this question: **"out of the predictions made by the model, what percentage is correct?"**

$$\text{Accuracy} = \frac{TP + TN}{\text{Total number observation}}$$

3 Precision

We get *precision* by answering this question: **"out of all the YES predictions, how many of them were correct?"**

$$Precision = \frac{TP}{TP + FP}$$

4 Recall / Sensitivity

It aims to answer this question: **"how good was the model at predicting real Yes events?"**, which can be considered as the flip of the precision.

$$Sensitivity = \frac{TP}{TP + FN}$$

5 Recall / Specificity

It aims to answer this question: **"how good was the model at predicting real No events?".**

$$Specificity = \frac{TN}{TN + FP}$$

6 F1 Score

Sometimes used when dealing with imbalanced data set, meaning that there are more of one class/label than there are of the other. It corresponds to the harmonic mean of the precision and recall.

$$F1\ Score = 2 . \frac{Precision \times Recall}{Precision + Recall}$$

7 AUC – ROC Curve

AUC– ROC generates probability values instead of binary 0/1 values. It should be used when your data set is roughly balanced.

Using ROC for imbalanced data sets lead to incorrect interpretation.

ROC curves provide good overview of trade-off between the TP rate and FP rate for binary classifier using different probability thresholds.

- A value below 0.5 indicates a poor classifier
- A value of 0.5 means random classifier
- Value over 0.7 corresponds to a good classifier
- 0.8 indicates a strong classifier
- We have 1 when the classifier perfectly predicts everything.

Strategies to choose the right metric

Choose accuracy
- The cost of FP and FN are roughly equal.
- The benefit of TP and TN are roughly equal.

Choose Precision
- The cost of FP is much higher than a FN.
- The benefit of a TP is much higher than a TN.

Choose recall
- The cost of FN is much higher than a FP.
- The cost of a TN is much higher than a TP.

ROC AUC & Precision – Recall curves
- Use ROC when the dealing with balanced data sets.
- Use precision-recall for imbalanced data sets.

B) Clustering: Clustering is the act of organizing similar objects into groups within a machine learning algorithm. Assigning related objects into clusters is beneficial for AI models. Clustering has many uses in data science, like image processing, knowledge discovery in data, unsupervised learning, and various other applications. Cluster analysis, or clustering, is done by scanning the unlabeled datasets in a machine learning model and setting measurements for specific data point features. The cluster analysis will then classify and place the data points in a group with matching features. Once data has been grouped together, it will be assigned a cluster ID number to help identify the cluster characteristics. Breaking down large, intricate datasets in a machine learning model using the clustering technique can alleviate stress when deciphering complex data.

Examples of clustering Instances **that benefit from data cluster analysis:**

- Optimizing city planning
- Customizing training sets for professional athletes
- Detecting spam threats and criminal activity

- Identifying misinformation
- Analyzing documents
- Personalizing advertisements to customers
- Tracking online business traffic

The capabilities of AI utilizing cluster analysis are expansive. Large machine learning datasets can be compacted and numbered to simplify data tracking. Cluster IDs can transform minute data points into data mining tools that streamline machine learning trend prediction.

Why is clustering important? When clustering is utilized in AI, scalability increases and automates mundane tasks in data science.

Documenting datasets natural grouping patterns can simplify data collection and application. Through identifying and organizing similar data, companies can optimize research and provide more efficient products.

Should every machine learning model use clustering?

While clustering is not required to filter and organize data, it could provide previously unidentified data pattern information. Data clustering algorithms can increase a machine learning model's value by automating data organization. Clustering is recommended, but not mandatory.

What are clustering changes based on?

Clustering updates and changes are based on the most recent documents of an algorithm. If a recluster is needed, previous clusters will need to be redocumented, placing them back into their clusters with any new documents.

When should re-clustering take place?

Any removal or addition of documents requires a re-cluster. Algorithms on the latest documents within a given system. Re-clustering and assigning updated documentation will lead to better data clusters.

Clustering vs Other Technologies & Methodologies

There are various clustering methods, including, but not limited to:

● **K-Means Clustering** is an iterative algorithm that scans across all datasets to derive a consensus of the available data, consuming less power and having a faster turnaround than other clustering methods.

- **Mean-Shift Clustering** uses a moving sensor, called a sliding window, to detect dense data point areas. Within the data area, mean-shift clustering will locate its center cluster and clean up surrounding data until an acceptable cluster is formed.
- **Density-Based Clustering** identifies large data areas and eliminates small ones. A valuable aspect of density-based clustering lies in going beyond cluster identification. It detects other data points outside of clusters and recognizes them as noise.
- **Hierarchical Clustering** organizes and ranks multiple clusters. There are two categories: Agglomerative and Divisive. Agglomerative considers each data point as its own cluster and merges them at each iteration to create the optimal clustering. The Divisive technique is inverse to Agglomerative. Divisive clusters all data points from the start of each iteration, removing irrelevant data points from the cluster.

Example: K means clustering algorithm

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

What is K-Means Algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

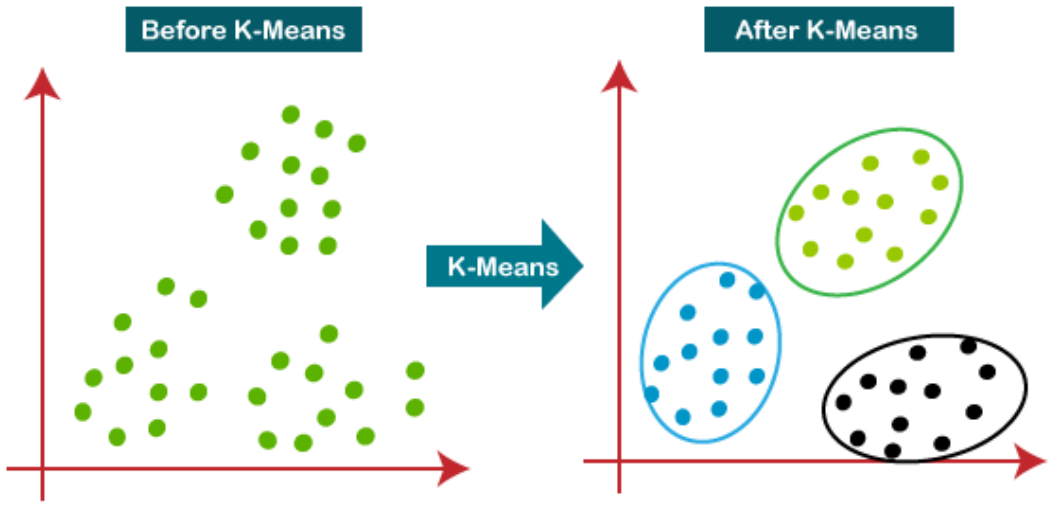
It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



How does the K-Means Algorithm Work?

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

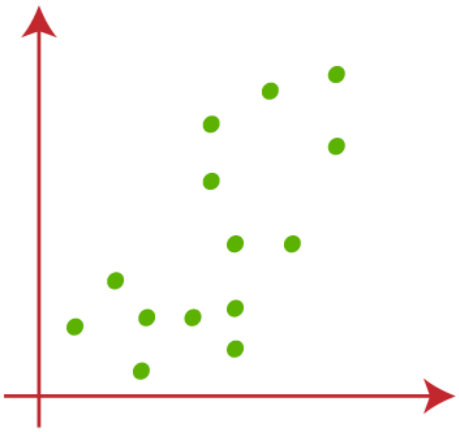
Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

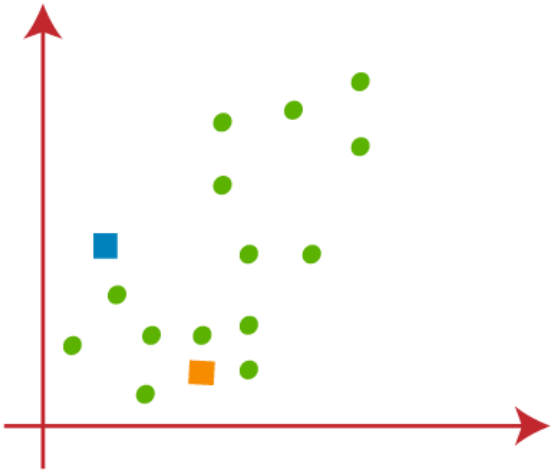
Step-7: The model is ready.

Let's understand the above steps by considering the visual plots:

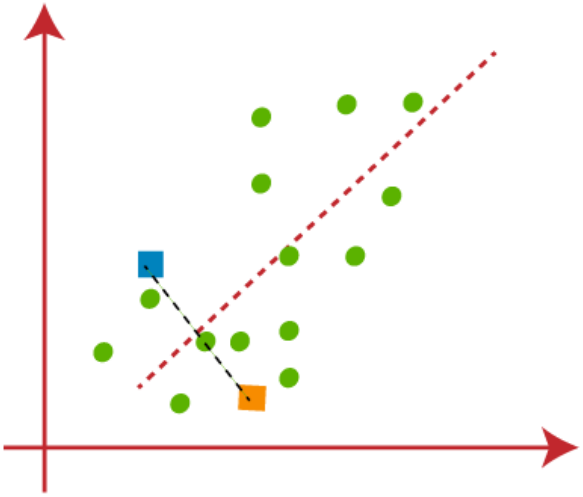
Suppose we have two variables M1 and M2. The x-y axis scatter plot of these two variables is given below:



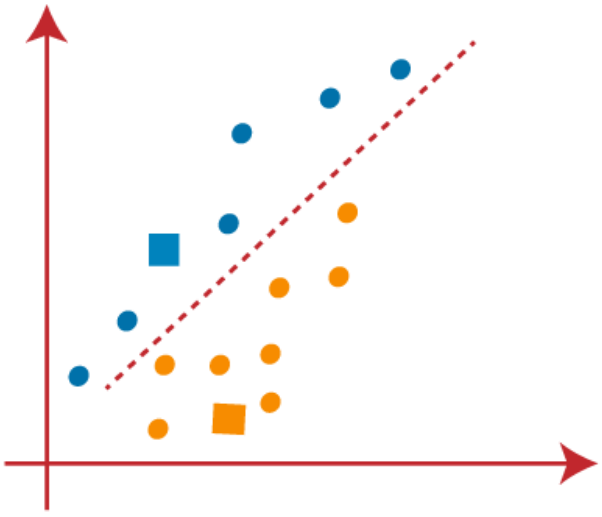
- Let's take number k of clusters, i.e., K=2, to identify the dataset and to put them into different clusters. It means here we will try to group these datasets into two different clusters.
- We need to choose some random k points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here we are selecting the below two points as k points, which are not the part of our dataset. Consider the below image:



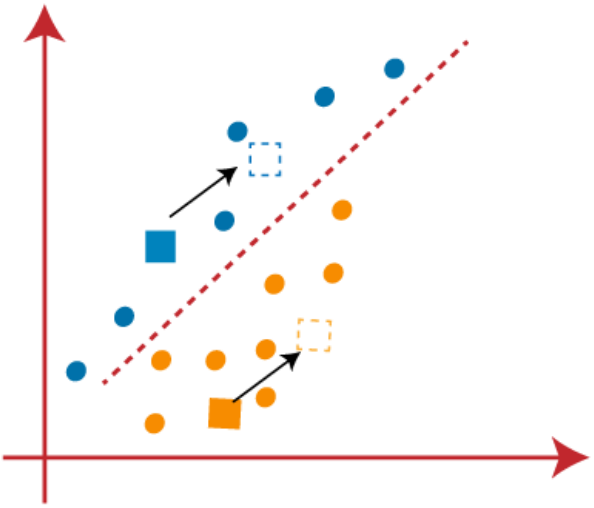
- Now we will assign each data point of the scatter plot to its closest K-point or centroid. We will compute it by applying some mathematics that we have studied to calculate the distance between two points. So, we will draw a median between both the centroids. Consider the below image:



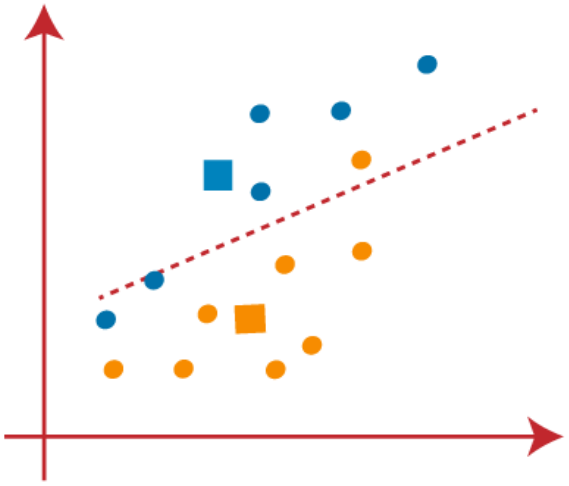
From the above image, it is clear that points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid. Let's color them as blue and yellow for clear visualization.



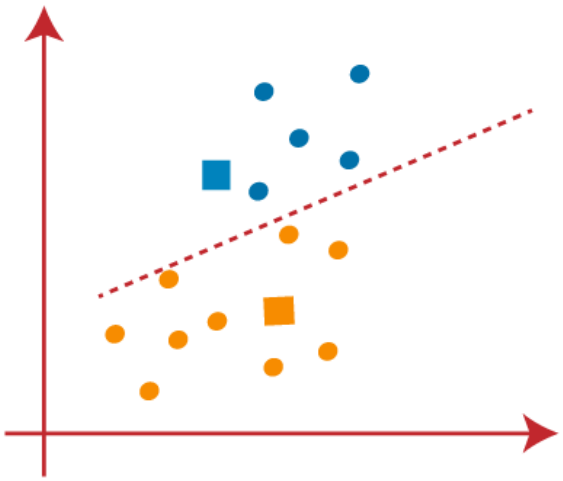
- As we need to find the closest cluster, so we will repeat the process by choosing **a new centroid**. To choose the new centroids, we will compute the center of gravity of these centroids, and will find new centroids as below:



- Next, we will reassign each datapoint to the new centroid. For this, we will repeat the same process of finding a median line. The median will be like below image:



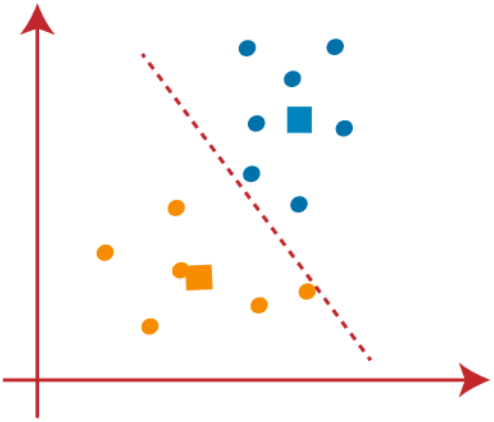
From the above image, we can see, one yellow point is on the left side of the line, and two blue points are right to the line. So, these three points will be assigned to new centroids.



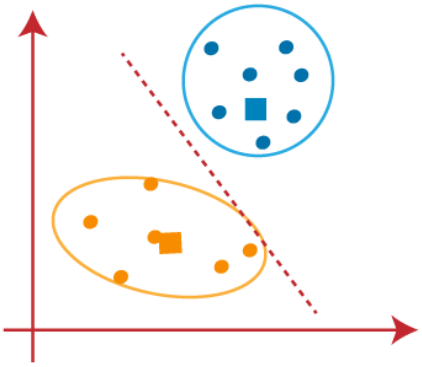
As reassignment has taken place, so we will again go to the step-4, which is finding new centroids or K-points.

- We will repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the below image:

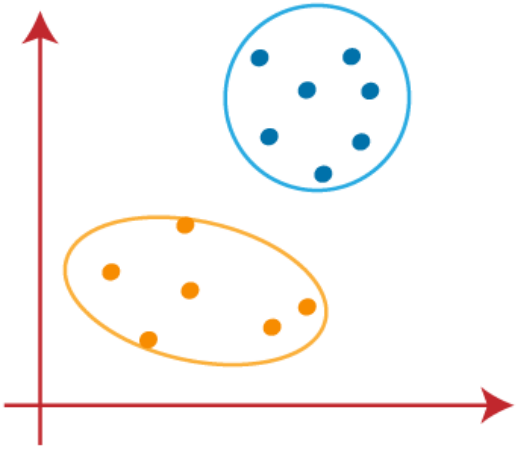
- As we got the new centroids so again will draw the median line and reassign the data points. So, the image will be:



- e can see in the above image; there are no dissimilar data points on either side of the line, which means our model is formed. Consider the below image:



As our model is ready, so we can now remove the assumed centroids, and the two final clusters will be as shown in the below image:



How to choose the value of "K number of clusters" in K-means Clustering?

The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal number of clusters, but here we are discussing the most appropriate method to find the number of clusters or value of K. The method is given below:

Elbow Method: The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$WCSS= \sum_{Pi \text{ in Cluster1}} distance\ (P_i\ C_1)_2 + \sum_{Pi \text{ in Cluster2}} distance\ (P_i\ C_2)_2 + \sum_{Pi \text{ in CLuster3}} distance(P_i\ C_3)_2$$

In the above formula of WCSS,

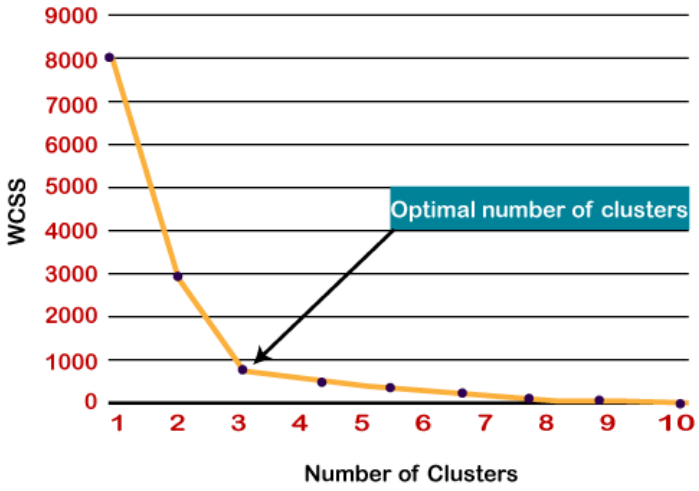
$\sum_{Pi \text{ in Cluster1}}$ distance $(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 110).
- For each value of K, calculates the WCSS value. ○ Plots a curve between calculated WCSS values and the number of clusters K. ○ The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method. The graph for the elbow method looks like the below image:



5.8 Sample of **K-means Clustering :**

R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications. Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

#Perform the clustering using clustering algorithm
#k-means clustering using R
#Apply K means to iris and store result
>newiris<-iris
>newiris$Species<-NULL
> (kc<-kmeans(newiris,3))

K-means clustering with 3 clusters of sizes 96, 21, 33

Cluster means:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	6.314583	2.895833	4.973958	1.7031250
2	4.738095	2.904762	1.790476	0.3523810
3	5.175758	3.624242	1.472727	0.2727273

Clustering vector:

[1] 3 2 2 2 3 3 3 3 2 2 3 3 2 2 3 3 3 3 3 3 3 3 3 3 2 2 3 3 3 2 2 3 3 3 2 3 3
[38] 3 2 3 3 2 2 3 3 2 3 2 3 3 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1
[75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1
[112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[149] 1 1

Within cluster sum of squares by cluster:

[1] 118.651875 17.669524 6.432121
(between_SS / total_SS = 79.0 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss"
[6] "betweenss" "size" "iter" "ifault"

#Compare the species label with the clustering result

>table(iris$Species,kc$cluster)

1 2 3 setosa 0
17 33 versicolor
46 4 0
virginica 50 0 0

#Plot the clusters and their centers
>plot (newiris[c ("Sepal.Length","Sepal.Width")], col=kc$cluster)
dev.off()

Conclusion:

Clustering algorithm is implemented using R programming