



**SHRI G.P.M. DEGREE COLLEGE OF  
SCIENCE & COMMERCE.**



**SHRI G.P.M. DEGREE COLLEGE OF  
SCIENCE & COMMERCE**

(COMMITTED TO EXCELLENCE IN EDUCATION)

**CERTIFICATE**

This is to certify that Mr/Ms \_\_\_\_\_

Student of class BSc-IT [ Roll No: \_\_\_\_\_ ] has completed the required number of practical's in the subject of Business-Intelligence as prescribed by the University of Mumbai under my supervision during the academic year 2023-2024.

---

Prof. In Charge

---

Course Co-coordinator

---

External Examiner

---

Principal

---

Date: \_\_\_\_\_

---

College Seal

<b>Prof. Name :</b>	<b>Class /SEM : T.Y. B.Sc. - IT / Sem – VI (2023-2024)</b>
<b>Course Code : USIT6P3</b>	<b>Subject Name : Business Intelligence</b>

Date	<b>INDEX</b>	Pg. No.	Sign.
	<p><b>Theory-1 : Loading</b></p> <p><b>Practical-1:</b> Import the legacy data from different sources such as (Excel , - SqlServer, Oracle etc.) and load in the target system. (You can download sample database such as Adventure works, North wind, food mart etc.) <b>(IT Lab)</b></p> <p><b>Example-1:</b> Import the legacy data from different sources such as Excel. <b>(IT Lab)</b></p> <p><b>Example-2:</b> Show Implementation of Classification algorithm in R -(<b>Homework</b>)</p> <p><b>Example -3:</b> Import the legacy data from different sources such as Sql Server. <b>(Homework)</b></p> <p><b>Practical-1:</b> Familiarizing Quantum GIS: Installation of QGIS, datasets for both Vector and Raster data, Maps. <b>(IT Lab)</b></p>		
	<p><b>Theory-2: Extraction</b></p> <p><b>Practical-2:</b> Perform the Extraction Transformation and Loading (ETL) -process to construct the database in the Sql server. <b>(IT Lab)</b></p> <p><b>Example-1:</b> Perform the Extraction. <b>(IT Lab)</b></p> <p><b>Example-2:</b> (A) Create the cube with suitable dimension and fact tables based on ROLAP, MOLAP and HOLAP model. <b>(Homework)</b></p> <p><b>Example -3:</b> Perform Transformation. <b>(Homework)</b></p>		
	<p><b>Theory-3: Data staging</b></p> <p><b>Practical-3:</b> a. Create the Data staging area for the selected database. b. Create the cube with suitable dimension and fact tables based on ROLAP, MOLAP and HOLAP model. <b>(IT Lab)</b></p> <p><b>Example-1:</b> Create the cube with suitable dimension and fact tables based on ROLAP. <b>(IT Lab)</b></p> <p><b>Example-2:</b> Perform the data clustering using clustering algorithm in R Programming. <b>(Homework)</b></p> <p><b>Example -3:</b> Create the cube with suitable dimension and fact tables based on MOLAP. <b>(Homework)</b></p>		
	<p><b>Theory-4: ETL</b></p> <p><b>Practical-4:</b> a. Create the ETL map and setup the schedule for execution. – b. Execute the MDX queries to extract the data from data ware house. <b>(IT Lab)</b></p> <p><b>Example-1:</b> Execute the MDX queries to extract the data from the Excel. <b>(IT Lab)</b></p> <p><b>Example-2:</b> Perform the Linear regression on the given data warehouse data. <b>(Homework)</b></p> <p><b>Example-3:</b> Execute the MDX queries to extract the data from the SQL server. <b>(Homework)</b></p>		
	<p><b>Theory-5: Data ware house</b></p> <p><b>Practical-5:</b> a. Import the data ware house data in Microsoft Excel and create the Pivot table and PivotChart. b. Import the cube in Microsoft Excel and create the Pivot table and Pivot Chart to perform data analysis. <b>(IT Lab)</b></p> <p><b>Example-1:</b> Import the data ware house data in Microsoft Excel and create the Pivot table. <b>(IT Lab)</b></p> <p><b>Example-2:</b> Show prediction Using Linear Regression. <b>(Homework)</b></p> <p><b>Example-3:</b> Import the data ware house data in Microsoft Excel and create the PivotChart. <b>(Homework)</b></p>		
	<p><b>Theory-6: Data ware house data</b></p> <p><b>Practical-6:</b> Apply the what – if Analysis for data visualization. Design and generate necessary reports based on the data ware house data. <b>(IT Lab)</b></p> <p><b>Example-1:</b> Show waterfall graph on data in power bi. <b>(IT Lab)</b></p> <p><b>Example-2:</b> perform the logistic regression on the given data warehouse data <b>(Homework)</b></p> <p><b>Example-3:</b> Show use of table and matrix. <b>(Homework)</b></p>		

	<p><b>Theory-7: Classification</b></p> <p><b>Practical-7:</b> Perform the data classification using classification algorithm (<b>IT Lab</b>)</p> <p><b>Example-1:</b> Show use of slicer on data. (<b>IT Lab</b>)</p> <p><b>Example-2:</b> Perform the data clustering using clustering algorithm in R – Programming. (<b>Homework</b>)</p> <p><b>Example-3:</b> Use filters on data. (<b>Homework</b>)</p>		
	<p><b>Theory-7: Classification</b></p> <p><b>Practical-7:</b> Perform the data classification using classification algorithm (<b>IT Lab</b>)</p> <p><b>Example-1:</b> Show use of slicer on data. (<b>IT Lab</b>)</p> <p><b>Example-2:</b> Perform the data clustering using clustering algorithm in R -Programming. (<b>Homework</b>)</p> <p><b>Example-3:</b> Use filters on data. (<b>Homework</b>)</p>		
	<p><b>Theory-8: Clustering</b></p> <p><b>Practical-8:</b> Perform the data clustering using clustering algorithm. (<b>IT Lab</b>)</p> <p><b>Example-1:</b> Transform less structured data in power bi. (<b>IT Lab</b>)</p> <p><b>Example-2:</b> Use merge query in power bi. (<b>Homework</b>)</p>		
	<p><b>Theory-9: Linear regression</b></p> <p><b>Practical-9:</b> Perform the Linear regression on the given data ware house data.- (<b>IT Lab</b>)</p> <p><b>Example-1:</b> Optimize models for reporting. (<b>IT Lab</b>)</p> <p><b>Example-2:</b> Show map visualization. (<b>Homework</b>)</p>		
	<p><b>Theory-10: logistic regression</b></p> <p><b>Practical-10:</b> Perform the logistic regression on the given data ware house -data. (<b>IT Lab</b>)</p> <p><b>Example-1:</b> Perform ETL transformation on the above data by converting the attribute Name from lowercase to uppercase. (<b>IT Lab</b>)</p> <p><b>Example-2:</b> What is pinning on data set? (<b>IT Lab</b>)</p> <p><b>Example-3:</b> publish a report to the web from power bi. (<b>Homework</b>)</p>		



## SHRI GPM DEGREE COLLEGE OF SCIENCE & COMMERCE

Department of Computer

Affiliated to University of Mumbai

Vision.. Innovation.. Solution.. Presentation..

**Practical-1: Import the legacy data from different sources such as (Excel , - Sql Server, Oracle etc.) and load in the target system. (You can download sample database such as Adventure works, North wind, food mart etc.)**

### **Aims:**

1. To understand and implement the process of loading legacy data from various sources such as Excel, SQL Server, and Oracle into a target system efficiently.

### **Learning Objectives:**

1. Understand the importance of data migration and loading techniques.
2. Gain hands-on experience in importing data from different data sources.
3. Learn how to transform and load data into a target database.
4. Identify common challenges in data migration and methods to overcome them.

### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

### **Theory-1: Loading**

Loading is the process of importing legacy data from different sources such as Excel, SQL Server, Oracle, and other database systems into a target system. This is a crucial step in data migration and ETL (Extract, Transform, Load) processes, ensuring seamless data integration for further processing and analysis.

### **Process of Loading**

The loading phase consists of several key steps:

1. **Data Extraction:** Extract data from different legacy sources while maintaining data integrity.
2. **Data Transformation:** Perform necessary transformations such as data cleaning, validation, and mapping to match the target system's schema.
3. **Data Loading:** Transfer the transformed data into the target system, ensuring minimal downtime and data consistency.

### **Types of Loading**

- **Full Load:** A one-time transfer of all data from the source system to the target system.
- **Incremental Load:** Only new or updated records are loaded periodically to optimize performance.
- **Batch Loading:** Data is loaded in predefined chunks to manage system resources efficiently.
- **Real-time Loading:** Continuous streaming of data to support real-time analytics and reporting.

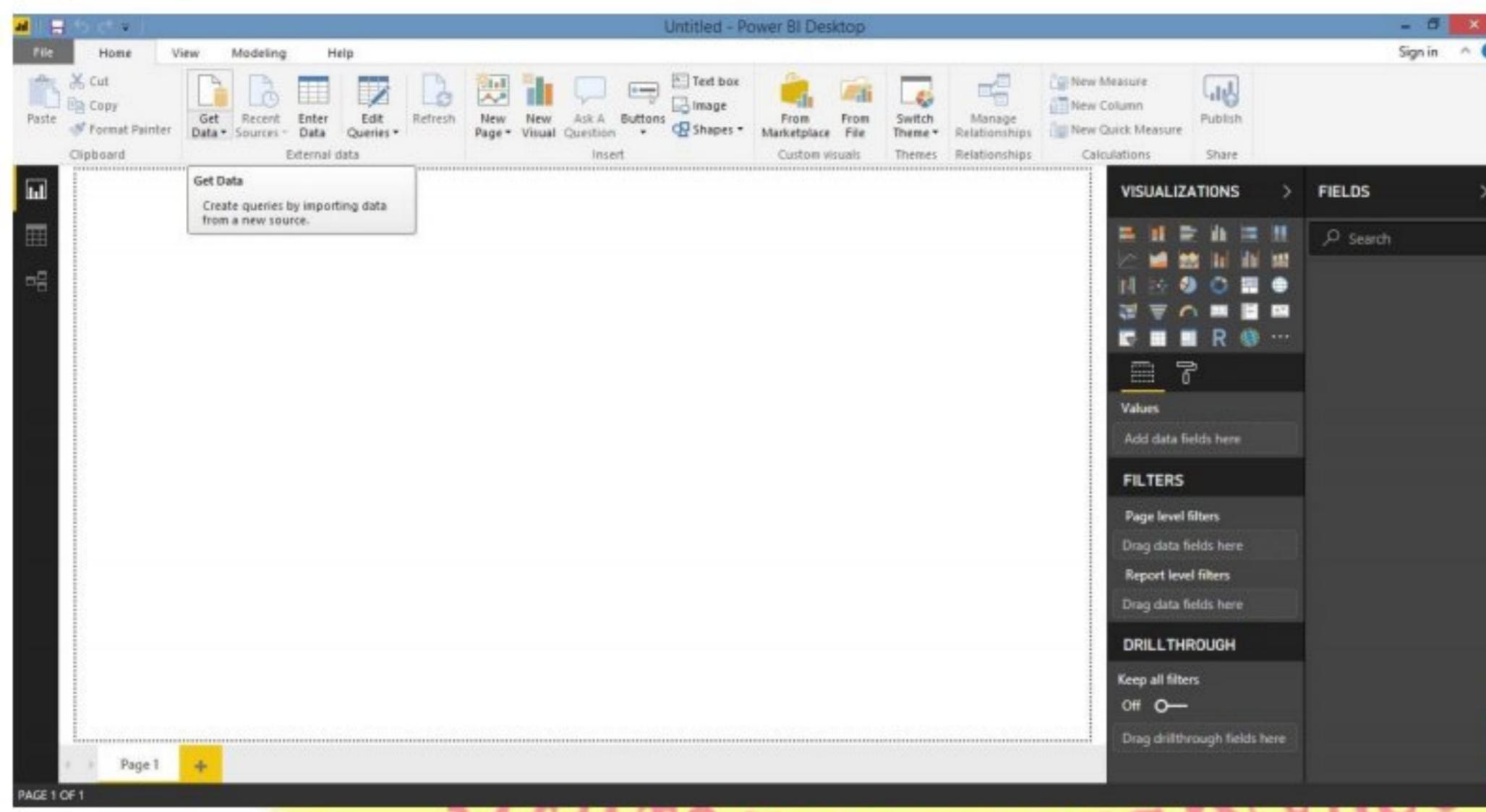
## Challenges and Considerations

- **Data Compatibility:** Ensure the legacy data format aligns with the target system's structure.
- **Performance Optimization:** Efficient indexing and batch processing help improve load speed.
- **Error Handling:** Implement logging and rollback mechanisms to handle failures and ensure data consistency.
- **Security Compliance:** Maintain data confidentiality by implementing encryption and access control.

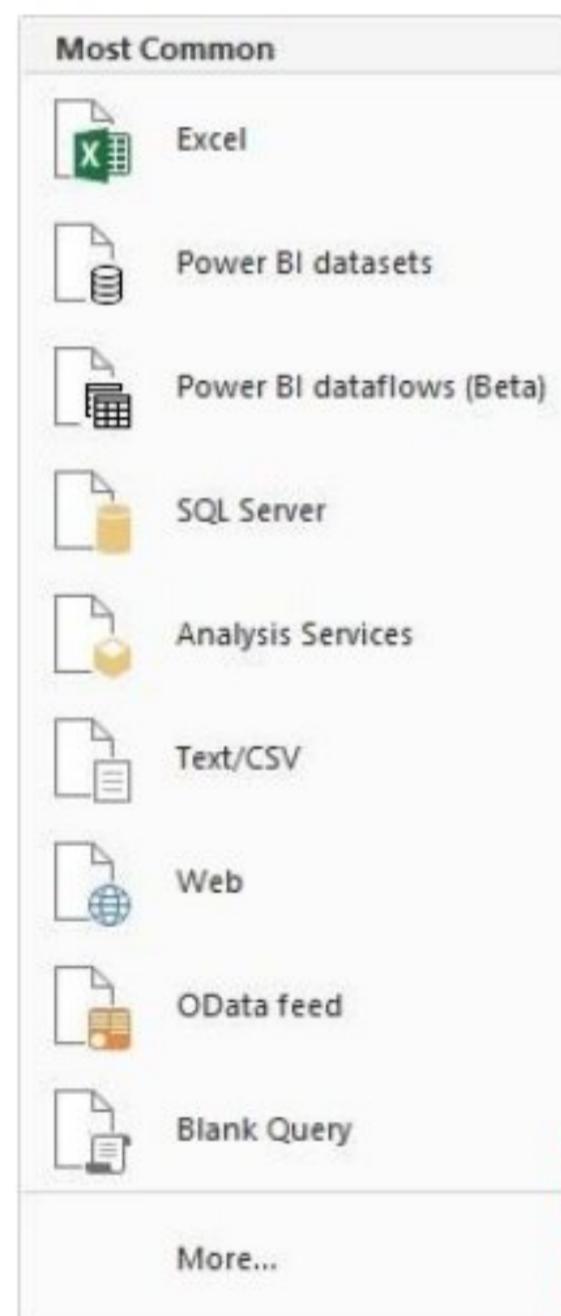


## PRACTICAL 1

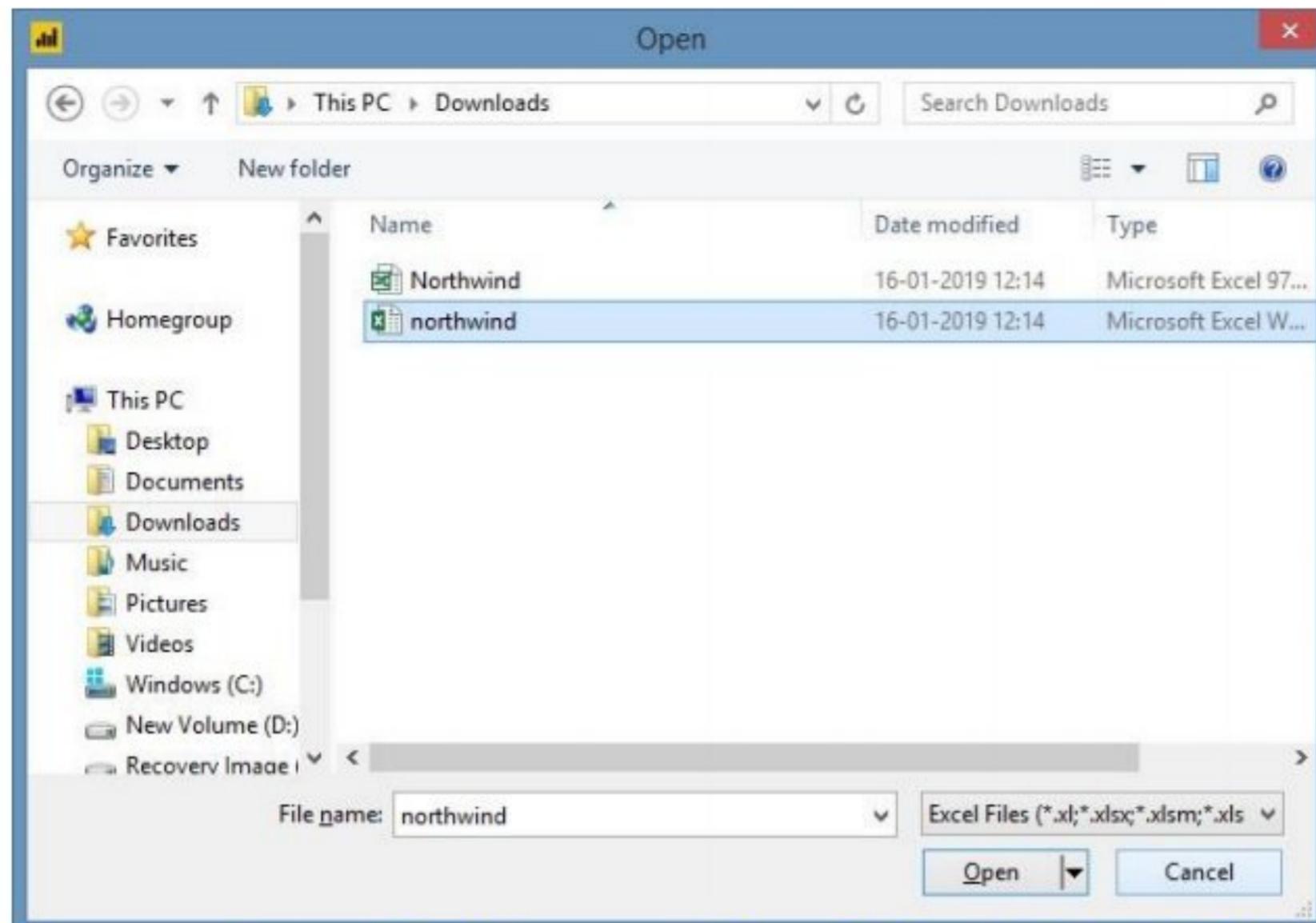
- a. Import the legacy data from different sources such as ( Excel , SqlServer, Oracle etc.) and load in the target system. ( You can download sample database such as Adventureworks, Northwind, foodmart etc.)  
Step 1: Open Power BI



Step 2: Click on Get data following list will be displayed → select Excel



Step 3: Select required file and click on Open, Navigator screen appears

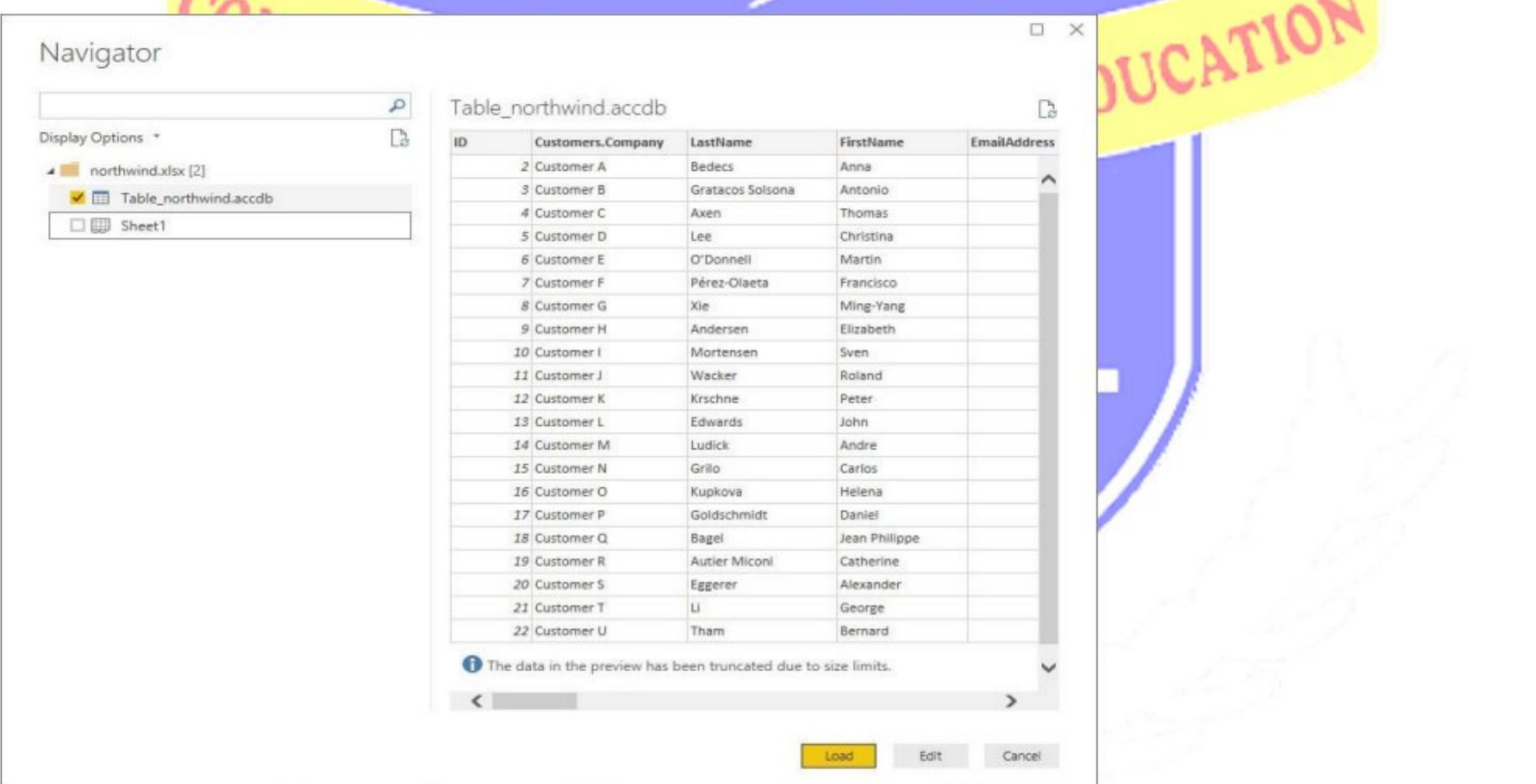
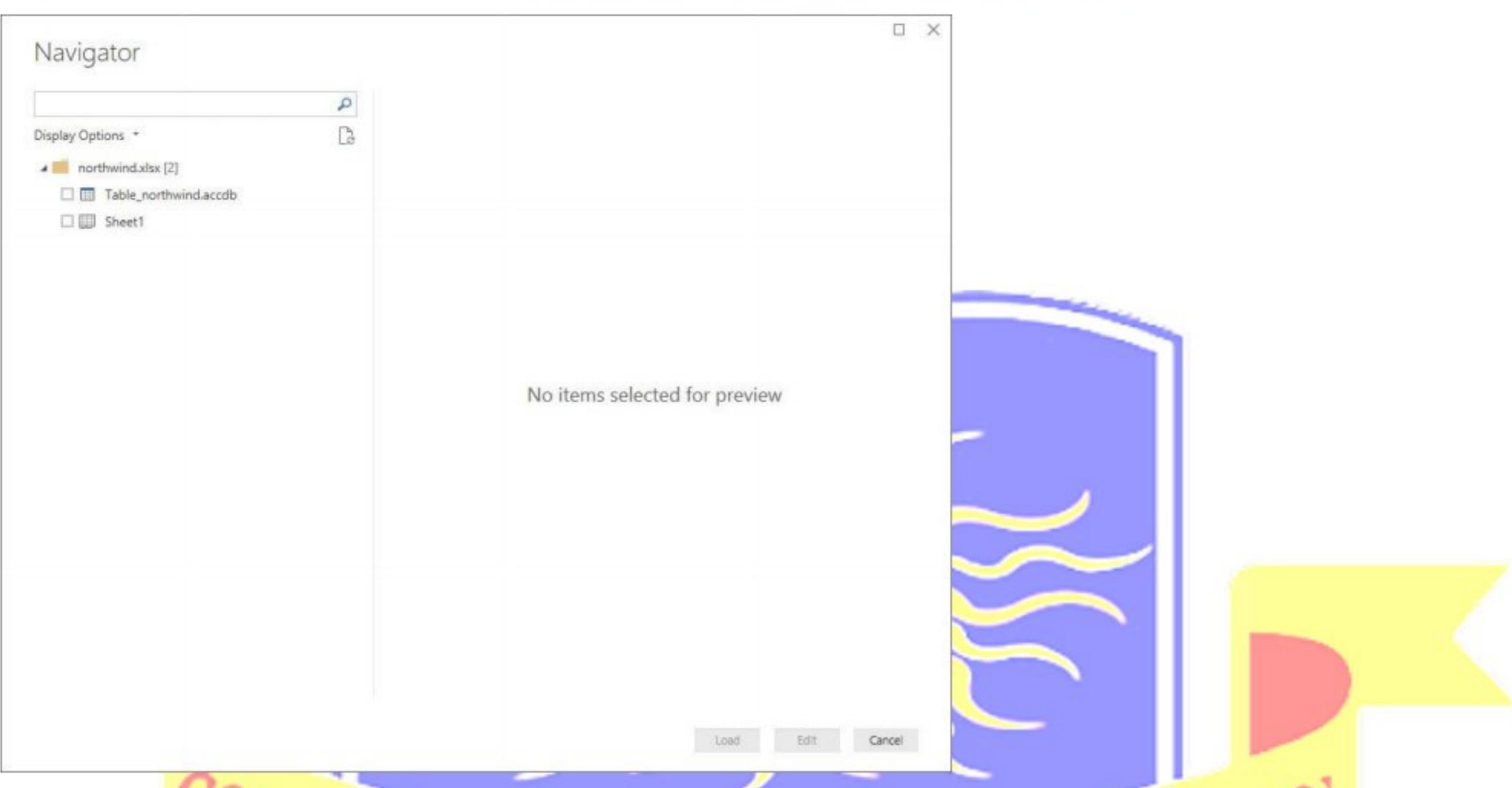


Step 4: Select file and click on edit

**COMMITTED TO EXCELLENCE IN EDUCATION**



**निर्मलसेह उत्तम सेवाधर्म**



Step 5: Power query editor appears



Untitled - Power Query Editor

File Home Transform Add Column View Help

New Source Recent Enter Data Data source settings Manage Parameters Refresh Preview Manage Choose Columns Remove Columns Keep Rows Remove Rows Sort Data Type: Whole Number Use First Row as Headers Split Column Group By Replace Values Combine Transform

Queries [1] Table\_northwind accdb

ID	Customer Company	Last Name	First Name	Email Address	Job Title
1	Customer A	Bedecs	Anna	null	Owner
2	Customer B	Gratacos Solsona	Antonio	null	Owner
3	Customer C	Axen	Thomas	null	Purchasing
4	Customer D	Lee	Christina	null	Purchasing
5	Customer E	O'Donnell	Martin	null	Owner
6	Customer F	Pérez-Olaeta	Francisco	null	Purchasing
7	Customer G	Xie	Ming-Yang	null	Owner
8	Customer H	Andersen	Elizabeth	null	Purchasing
9	Customer I	Mortensen	Sven	null	Purchasing
10	Customer J	Wacker	Roland	null	Purchasing
11	Customer K	Krschne	Peter	null	Purchasing
12	Customer L	Edwards	John	null	Purchasing
13	Customer M	Ludick	Andre	null	Purchasing
14	Customer N	Grilo	Carlos	null	Purchasing
15	Customer O	Kupkova	Helena	null	Purchasing
16					

23 COLUMNS, 26 ROWS PREVIEW DOWNLOADED AT 12:22

QUERY SETTINGS

Properties Name: Table\_northwind accdb All Properties

Applied Steps Source: Navigation: Changed Type

Step 6: Again, go to Get Data and select OData feed

Most Common

- Excel
- Power BI datasets
- Power BI dataflows (Beta)
- SQL Server
- Analysis Services
- Text/CSV
- Web
- OData feed**
- Blank Query
- More...

Step 7:

Paste url as <http://services.odata.org/V3/Northwind/Northwind.svc/> Click on ok



OData feed

Basic    Advanced

URL

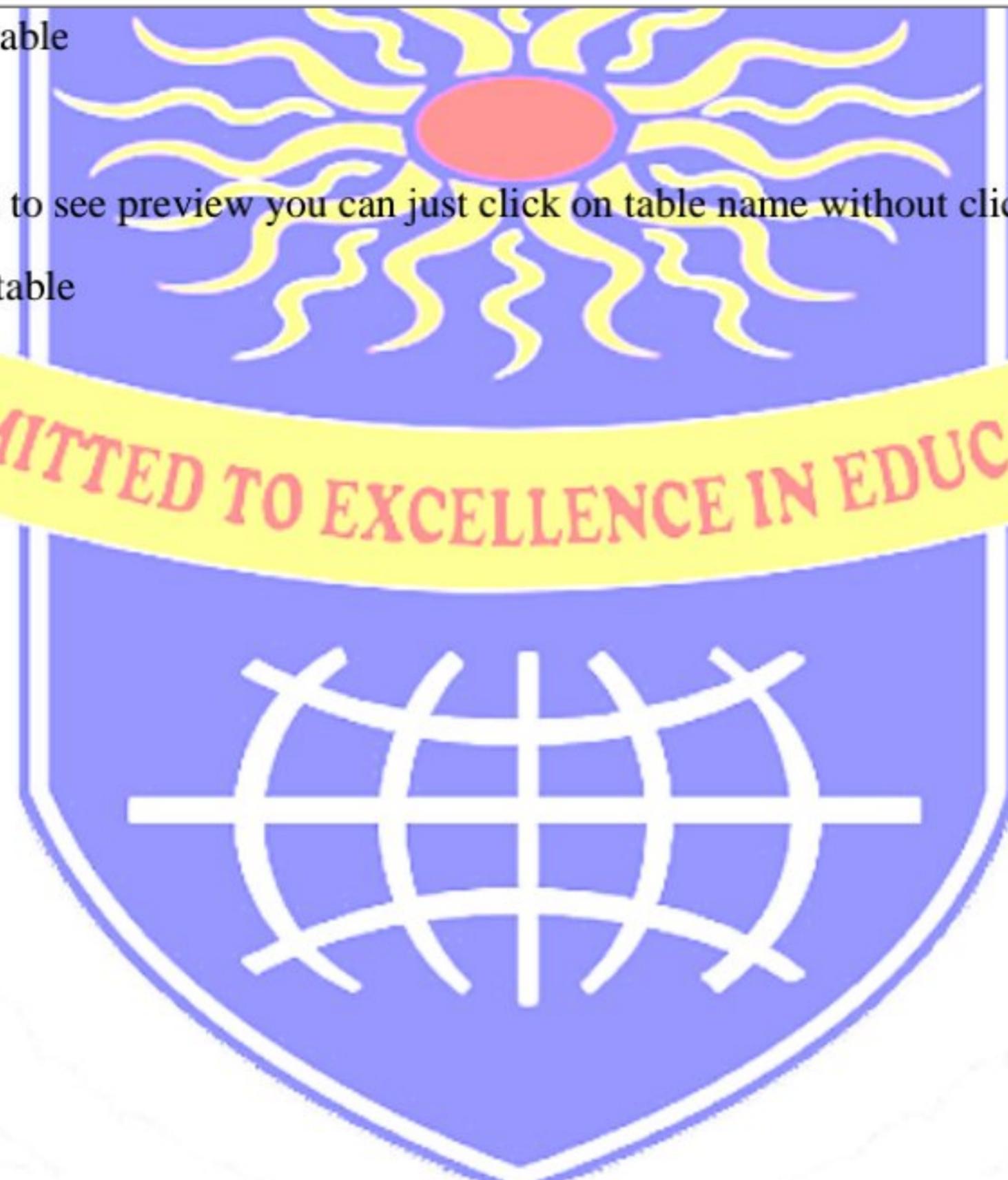
OK   :   Cancel

Step 8: Select orders table

And click on edit

Note: If you just want to see preview you can just click on table name without clicking on checkbox

Click on edit to view table



निर्मलस्नेह उत्तम सेवाधर्म



Navigator

Display Options ▾

- http://services.odata.org/V3/Northwind/No...
  - Alphabetical\_list\_of\_products
  - Categories
  - Category\_Sales\_for\_1997
  - Current\_Product\_Lists
  - Customer\_and\_Suppliers\_by\_Cities
  - CustomerDemographics
  - Customers
  - Employees
  - Invoices
  - Order\_Details
  - Order\_Details\_Extendeds
  - Order\_Subtotals
  - Orders**
  - Orders\_Qries
  - Product\_Sales\_for\_1997
  - Products
  - Products\_Above\_Average\_Prices
  - Products\_by\_Categories
  - Regions

Orders

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate
10248	VINET	5	04-07-1996 00:00:00	01-08-199
10249	TOMSP	6	05-07-1996 00:00:00	16-08-199
10250	HANAR	4	08-07-1996 00:00:00	05-08-199
10251	VICTE	3	08-07-1996 00:00:00	05-08-199
10252	SUPRD	4	09-07-1996 00:00:00	06-08-199
10253	HANAR	3	10-07-1996 00:00:00	24-07-199
10254	CHOPS	5	11-07-1996 00:00:00	08-08-199
10255	RICSU	9	12-07-1996 00:00:00	09-08-199
10256	WELLI	3	15-07-1996 00:00:00	12-08-199
10257	HILAA	4	16-07-1996 00:00:00	13-08-199
10258	ERNSH	1	17-07-1996 00:00:00	14-08-199
10259	CENTC	4	18-07-1996 00:00:00	15-08-199
10260	OTTIK	4	19-07-1996 00:00:00	16-08-199
10261	QUED	4	19-07-1996 00:00:00	16-08-199
10262	RATTC	8	22-07-1996 00:00:00	19-08-199
10263	ERNSH	9	23-07-1996 00:00:00	20-08-199
10264	FOLKO	6	24-07-1996 00:00:00	21-08-199
10265	BLONP	2	25-07-1996 00:00:00	22-08-199
10266	WARTH	3	26-07-1996 00:00:00	06-09-199
10267	FRANK	4	29-07-1996 00:00:00	26-08-199
10268	GROSR	8	30-07-1996 00:00:00	27-08-199
10269	WHITC	5	31-07-1996 00:00:00	14-08-199
10270	WARTH	1	01-08-1996 00:00:00	29-08-199

Select Related Tables      Load      Edit      Cancel

Untitled - Power Query Editor

File Home Transform Add Column View Help

New & Apply Close New Source Recent Sources Enter Data Data source settings Manage Parameters Refresh Preview Advanced Editor Properties Manage Choose Columns Remove Columns Keep Rows Remove Rows Split Column Group By Data Type: Whole Number Use First Row as Headers Replace Values Transform

Queries [2]

- Table\_northwind accdb
- Orders**

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	Shipped
10248	VINET	5	04-07-1996 00:00:00	01-08-1996 00:00:00	16-0
10249	TOMSP	6	05-07-1996 00:00:00	16-08-1996 00:00:00	10-0
10250	HANAR	4	08-07-1996 00:00:00	05-08-1996 00:00:00	12-0
10251	VICTE	3	08-07-1996 00:00:00	05-08-1996 00:00:00	15-0
10252	SUPRD	4	09-07-1996 00:00:00	06-08-1996 00:00:00	11-0
10253	HANAR	3	10-07-1996 00:00:00	24-07-1996 00:00:00	16-0
10254	CHOPS	5	11-07-1996 00:00:00	08-08-1996 00:00:00	23-0
10255	RICSU	9	12-07-1996 00:00:00	09-08-1996 00:00:00	15-0
10256	WELLI	3	15-07-1996 00:00:00	12-08-1996 00:00:00	17-0
10257	HILAA	4	16-07-1996 00:00:00	13-08-1996 00:00:00	22-0
10258	ERNSH	1	17-07-1996 00:00:00	14-08-1996 00:00:00	23-0
10259	CENTC	4	18-07-1996 00:00:00	15-08-1996 00:00:00	25-0
10260	OTTIK	4	19-07-1996 00:00:00	16-08-1996 00:00:00	29-0
10261	QUED	4	19-07-1996 00:00:00	16-08-1996 00:00:00	30-0
10262	RATTC	8	22-07-1996 00:00:00	19-08-1996 00:00:00	25-0

18 COLUMNS, 830 ROWS      PREVIEW DOWNLOADED AT 12:25

QUERY SETTINGS

Properties Name: Orders All Properties

Applied Steps Source: Navigation

**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

- 1. Define data loading.**
- 2. What is incremental loading?**
- 3. Why validate data during loading?**
- 4. Mention one data loading challenge.**

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



## **Practical-2: Perform the Extraction Transformation and Loading (ETL) -process to construct the database in the Sql server.**

### **Aims:**

1. To implement a complete ETL process by extracting data from multiple sources, transforming it as necessary, and loading it into a SQL Server database.
2. To build a reliable and optimized database using ETL techniques for improved data management and reporting.

### **Learning Objectives:**

1. Understand the ETL methodology and its role in data warehousing.
2. Gain proficiency in extracting data from diverse sources such as Excel, SQL Server, and Oracle.
3. Learn to perform data transformations including cleansing, formatting, and aggregation.
4. Master the process of loading transformed data into SQL Server and validating its integrity.

### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

### **Theory-2: Extraction**

Extraction is the first phase of the ETL (Extract, Transform, Load) process, which is essential for constructing a database in SQL Server. This step involves retrieving data from multiple sources such as relational databases, flat files, APIs, and cloud storage systems to ensure that accurate and relevant data is available for further processing.

### **Process of Extraction**

The extraction process includes the following steps:

1. **Identifying Data Sources:** Determine and analyze the structure and format of legacy data sources.
2. **Data Retrieval:** Extract data using various methods such as SQL queries, API calls, or file parsing.
3. **Data Staging:** Store the extracted data temporarily in a staging area to maintain integrity before transformation.
4. **Data Validation:** Check data completeness, consistency, and correctness to ensure reliability.

### **Types of Extraction**

- **Full Extraction:** Extracts all data from the source system at once, typically used for initial loads.
- **Incremental Extraction:** Only new or modified data is extracted periodically, reducing system load.
- **Real-time Extraction:** Continuous extraction to support real-time analytics and processing.

## Challenges and Considerations

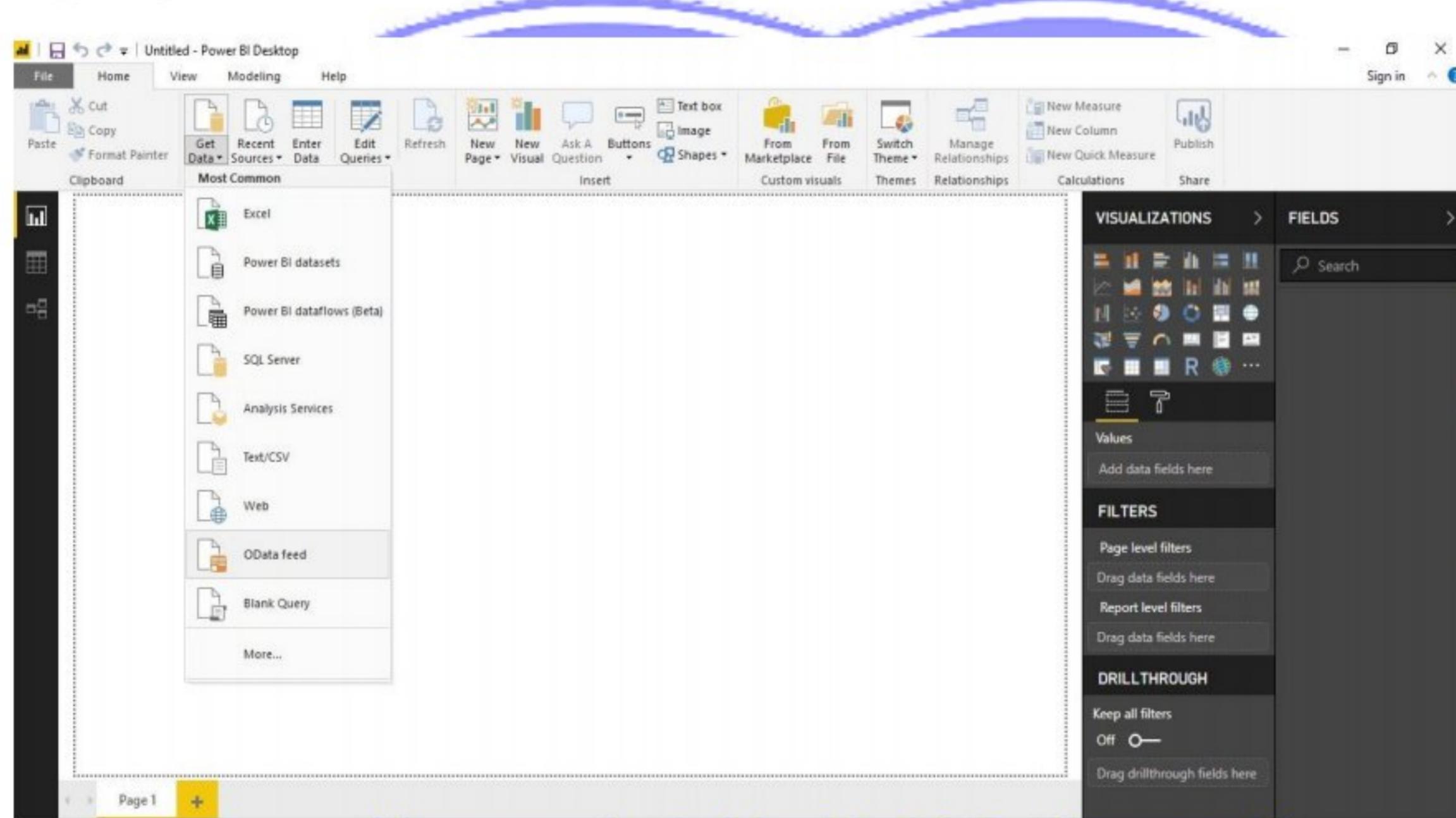
- **Performance Issues:** Large data volumes can impact system performance; optimizing queries and indexing helps mitigate this.
- **Data Integrity:** Ensure extracted data remains consistent and unaltered during transfer.
- **Security Concerns:** Implement encryption and secure connections to protect sensitive data.
- **Handling Data Anomalies:** Implement error detection and correction mechanisms to avoid loading faulty data.



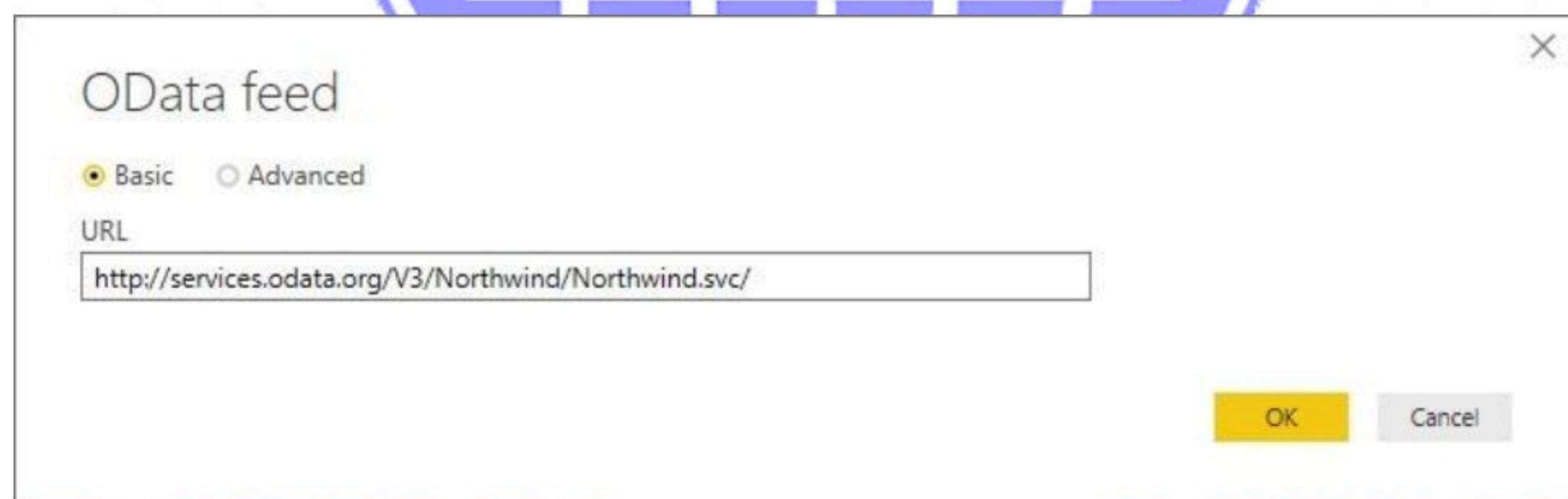
## PRACTICAL 2

Perform the Extraction Transformation and Loading (ETL) process to construct the database in the Power BI.

Step 1: Open Power BI, Click on Get Data → OData Feed



Paste Url : <http://services.odata.org/V3/Northwind/Northwind.svc/> And Click OK



Step 2: Click on Check Box of Products table and then click on Edit



Navigator

Display Options ▾

- http://services.odata.org/V3/Northwind/Northwind.svc/Products
- Alphabetical\_list\_of\_products
- Categories
- Category\_Sales\_for\_1997
- Current\_Product\_Lists
- Customer\_and\_Suppliers\_by\_Cities
- CustomerDemographics
- Customers
- Employees
- Invoices
- Order\_Details
- Order\_Details\_Extendeds
- Order\_Subtotals
- Orders
- Orders\_Qries
- Product\_Sales\_for\_1997
- Products**
- Products\_Above\_Average\_Prices
- Products\_by\_Categories
- Regions

Products

ProductID	ProductName	SupplierID	CategoryID	Quan
1	Chai	1	1	10 boxes x 20 bags
2	Chang	1	1	24 - 12 oz bottles
3	Aniseed Syrup	1	2	12 - 550 ml bottles
4	Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars
5	Chef Anton's Gumbo Mix	2	2	36 boxes
6	Grandma's Boysenberry Spread	3	2	12 - 8 oz jars
7	Uncle Bob's Organic Dried Pears	3	7	12 - 1 lb pkgs.
8	Northwoods Cranberry Sauce	3	2	12 - 12 oz jars
9	Mishi Kobe Niku	4	6	18 - 500 g pkgs.
10	Ikura	4	8	8 - 200 ml jars
11	Queso Cabrales	5	4	1 kg pkg.
12	Queso Manchego La Pastora	5	4	10 - 500 g pkgs.
13	Konbu	6	8	2 kg box
14	Tofu	6	7	40 - 100 g pkgs.
15	Genen Shouyu	6	2	24 - 250 ml bottles
16	Pavlova	7	3	32 - 16 oz jars
17	Alice Mutton	7	6	20 - 3.5 oz jars
18	Carmarvon Tigers	7	8	14 - 12 oz jars
19	Teatime Chocolate Biscuits	8	3	10 - 12 oz jars
20	Sir Rodney's Marmalade	8	3	30 - 3.5 oz jars
21	Sir Rodney's Scones	8	3	24 - 12 oz jars
22	Gustaf's Knäckebroð	9	5	24 - 1.5 kg bags
23	Tunnbröd	9	5	12 - 500 g pkgs.

Select Related Tables Load Edit Cancel



### 1) Remove other columns to only display columns of interest

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).

Untitled - Power Query Editor

File Home Transform Add Column View Help

Close & Apply New Source Recent Enter Data Data source settings Manage Parameters Refresh Preview Advanced Editor Choose Columns Remove Rows Keep Rows Remove Columns Manage Columns Sort Data Type: Any Use First Row as Headers Close New Query Data Sources Parameters Query Manage Rows Reduce Rows Group By Split Column Replace Values Transform

Queries [1]

Products

ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit
1	Chai	1	1	10 boxes x 20 bags
2	Chang	1	1	24 - 12 oz bottles
3	Aniseed Syrup	1	2	12 - 550 ml bottles
4	Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars
5	Chef Anton's Gumbo Mix	2	2	36 boxes
6	Grandma's Boysenberry Spread	3	2	12 - 8 oz jars
7	Uncle Bob's Organic Dried Pears	3	7	12 - 1 lb pkgs.
8	Northwoods Cranberry Sauce	3	2	12 - 12 oz jars
9	Mishi Kobe Niku	4	6	18 - 500 g pkgs.
10	Ikura	4	8	8 - 200 ml jars
11	Queso Cabrales	5	4	1 kg pkg.
12	Queso Manchego La Pastora	5	4	10 - 500 g pkgs.
13	Konbu	6	8	2 kg box
14	Tofu	6	7	40 - 100 g pkgs.
15	Genen Shouyu	6	2	24 - 250 ml bottles

13 COLUMNS, 77 ROWS PREVIEW DOWNLOADED AT 00:53

Select Remove Columns > Remove Other Columns from the ribbon, or rightclick on a column header and click Remove Other Columns



Untitled - Power Query Editor

File Home Transform Add Column View Help

Close & Apply New Source Recent Sources Enter Data Data source settings Manage Parameters Refresh Preview Properties Advanced Editor Choose Columns Remove Columns Keep Rows Remove Rows Sort Split Column Group By Data Type: Any Use First Row as Headers Replace Values Combine

Close New Query Data Sources Parameters Query Manage Columns Reduce Rows Sort Transform

Queries [1] Products

	ProductID	ProductName	CategoryID	QuantityPerUnit
1	1	Chai	1	10 boxes x 20 bags
2	2	Chang	1	24 - 12 oz bottles
3	3	Aniseed Syrup	2	12 - 550 ml bottles
4	4	Chef Anton's Cajun Seasoning	2	48 - 6 oz jars
5	5	Chef Anton's Gumbo Mix	2	36 boxes
6	6	Grandma's Boysenberry Spread	2	12 - 8 oz jars
7	7	Uncle Bob's Organic Dried Pears	7	12 - 1 lb pkgs.
8	8	Northwoods Cranberry Sauce	2	12 - 12 oz jars
9	9	Mishi Kobe Niku	6	18 - 500 g pkgs.
10	10	Ikura	8	12 - 200 ml jars
11	11	Queso Cabrales	4	1 kg pkg.
12	12	Queso Manchego La Pastora	4	10 - 500 g pkgs.
13	13	Konbu	8	2 kg box
14	14	Tofu	7	40 - 100 g pkgs.
15	15	Genen Shouyu	2	24 - 250 ml bottles
16				

13 COLUMNS, 77 ROWS PREVIEW DOWNLOADED AT 00:53

QUERY SETTINGS

PROPERTIES Name: Products All Properties

APPLIED STEPS Source: Navigation

COMMITTED TO EXCELLENCE IN EDUCATION

After selecting Remove Other Columns only selected four columns are displayed other columns are discarded.

Untitled - Power Query Editor

File Home Transform Add Column View Help

Close & Apply New Source Recent Sources Enter Data Data source settings Manage Parameters Refresh Preview Properties Advanced Editor Choose Columns Remove Columns Keep Rows Remove Rows Sort Split Column Group By Data Type: Whole Number Use First Row as Headers Replace Values Combine

Close New Query Data Sources Parameters Query Manage Columns Reduce Rows Sort Transform

Queries [1] Products

	ProductID	QuantityPerUnit	ProductName	UnitsInStock
1	1	10 boxes x 20 bags	Chai	39
2	2	24 - 12 oz bottles	Chang	17
3	3	12 - 550 ml bottles	Aniseed Syrup	13
4	4	48 - 6 oz jars	Chef Anton's Cajun Seasoning	53
5	5	36 boxes	Chef Anton's Gumbo Mix	0
6	6	12 - 8 oz jars	Grandma's Boysenberry Spread	120
7	7	12 - 1 lb pkgs.	Uncle Bob's Organic Dried Pears	15
8	8	12 - 12 oz jars	Northwoods Cranberry Sauce	6
9	9	18 - 500 g pkgs.	Mishi Kobe Niku	29
10	10	12 - 200 ml jars	Ikura	31
11	11	1 kg pkg.	Queso Cabrales	22
12	12	10 - 500 g pkgs.	Queso Manchego La Pastora	86
13	13	2 kg box	Konbu	24
14	14	40 - 100 g pkgs.	Tofu	35
15	15	24 - 250 ml bottles	Genen Shouyu	39
16	16	32 - 500 g boxes	Pavlova	29

4 COLUMNS, 77 ROWS PREVIEW DOWNLOADED AT 00:53

QUERY SETTINGS

PROPERTIES Name: Products All Properties

APPLIED STEPS Source: Navigation Removed Other Columns



## 2. Change the data type of the UnitsInStock column

- a) Select the UnitsInStock column.

Check if the data type of selected column is a Whole number

Untitled - Power Query Editor

File Home Transform Add Column View Help

New Recent Enter Data Source Sources Data Sources Manage Parameters Refresh Preview Advanced Editor Choose Columns Remove Rows Keep Rows Remove Rows Sort Split Column Group By Replace Values Use First Row as Headers Combine

Queries [1] Products

	ProductID	QuantityPerUnit	ProductName	UnitsInStock
1	1	10 boxes x 20 bags	Chai	39
2	2	24 - 12 oz bottles	Chang	17
3	3	12 - 550 ml bottles	Aniseed Syrup	13
4	4	48 - 6 oz jars	Chef Anton's Cajun Seasoning	53
5	5	36 boxes	Chef Anton's Gumbo Mix	0
6	6	12 - 8 oz jars	Grandma's Boysenberry Spread	120
7	7	12 - 1 lb pkgs.	Uncle Bob's Organic Dried Pears	15
8	8	12 - 12 oz jars	Northwoods Cranberry Sauce	6
9	9	18 - 500 g pkgs.	Mishi Kobe Niku	29
10	10	12 - 200 ml jars	Ikura	31
11	11	1 kg pkg.	Queso Cabrales	22
12	12	10 - 500 g pkgs.	Queso Manchego La Pastora	86
13	13	2 kg box	Konbu	24
14	14	40 - 100 g pkgs.	Tofu	35
15	15	24 - 250 ml bottles	Genen Shouyu	39
16	16	32 - 500 g boxes	Pavlova	29

4 COLUMNS, 77 ROWS PREVIEW DOWNLOADED AT 00:53

QUERY SETTINGS

PROPERTIES Name: Products All Properties

APPLIED STEPS Source Navigation Removed Other Columns

- b) Select the Data Type drop-down button in the Home ribbon.  
c) 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).



Queries [1] Products

	ProductID	QuantityPerUnit	ProductName	UnitsInStock
1	1	10 boxes x 20 bags	Chai	39
2	2	24 - 12 oz bottles	Chang	17
3	3	12 - 550 ml bottles	Aniseed Syrup	13
4	4	48 - 6 oz jars	Chef Anton's Cajun Seasoning	53
5	5	36 boxes	Chef Anton's Gumbo Mix	0
6	6	12 - 8 oz jars	Grandma's Boysenberry Spread	120
7	7	12 - 1 lb pkgs.	Uncle Bob's Organic Dried Pears	15
8	8	12 - 12 oz jars	Northwoods Cranberry Sauce	6
9	9	18 - 500 g pkgs.	Mishi Kobe Niku	29
10	10	12 - 200 ml jars	Ikura	31
11	11	1 kg pkg.	Queso Cabrales	22
12	12	10 - 500 g pkgs.	Queso Manchego La Pastora	86
13	13	2 kg box	Konbu	24
14	14	40 - 100 g pkgs.	Tofu	35
15	15	24 - 250 ml bottles	Genen Shouyu	39
16	16	32 - 500 g boxes	Pavlova	29

After clicking on Whole number, you can see the changed Datatype in column header of UnitsInStock.

Queries [1] Products

	ProductID	QuantityPerUnit	ProductName	UnitsInStock
1	1	10 boxes x 20 bags	Chai	39
2	2	24 - 12 oz bottles	Chang	17
3	3	12 - 550 ml bottles	Aniseed Syrup	13
4	4	48 - 6 oz jars	Chef Anton's Cajun Seasoning	53
5	5	36 boxes	Chef Anton's Gumbo Mix	0
6	6	12 - 8 oz jars	Grandma's Boysenberry Spread	120
7	7	12 - 1 lb pkgs.	Uncle Bob's Organic Dried Pears	15
8	8	12 - 12 oz jars	Northwoods Cranberry Sauce	6
9	9	18 - 500 g pkgs.	Mishi Kobe Niku	29
10	10	12 - 200 ml jars	Ikura	31
11	11	1 kg pkg.	Queso Cabrales	22
12	12	10 - 500 g pkgs.	Queso Manchego La Pastora	86
13	13	2 kg box	Konbu	24
14	14	40 - 100 g pkgs.	Tofu	35
15	15	24 - 250 ml bottles	Genen Shouyu	39
16	16	32 - 500 g boxes	Pavlova	29

After above step, close query editor and click on Yes to save changes.

Now you can view fields of Products table on right side, check all the fields of table to get representation in charts form.

The screenshot shows the Power BI Desktop interface. The ribbon menu includes File, Home, View, Modeling, and Help. The Home tab is selected, showing various data entry and visualization tools like Get Data, Refresh, New Page, and Insert. A scatter plot visual is displayed on the canvas, showing data points for ProductID and UnitsInStock across different ProductNames and QuantityPerUnits. The Fields pane on the right lists the Products table with fields: ProductID, ProductName, QuantityPerUnit, and UnitsInStock.

### 3. Expand the Orders table

Once You have loaded a data source, you can click on Recent Sources to select desired table (Orders).

The screenshot shows the Power BI Desktop interface with the Recent Sources pane expanded. The pane lists 'Most Recent' sources, including 'http://services.odata.org/V3/Northwind/Northwind.svc' and 'List of students 2018-19.xlsx'. The main canvas area displays a scatter plot similar to the one above, using the same data source. The Fields pane on the right remains the same, listing the Products table with its four fields.



After selecting the URL, Navigator window will appear from which you can select Orders table.

Click on Edit.



निर्मलस्नेह उत्तम सेवाधृष्ट



Navigator

Display Options ▾

- http://services.odata.org/V3/Northwind/No...
  - Alphabetical\_list\_of\_products
  - Categories
  - Category\_Sales\_for\_1997
  - Current\_Product\_Lists
  - Customer\_and\_Suppliers\_by\_Cities
  - CustomerDemographics
  - Customers
  - Employees
  - Invoices
  - Order\_Details
  - Order\_Details\_Extendeds
  - Order\_Subtotals
  - Orders**
  - Orders\_Qries
  - Product\_Sales\_for\_1997
  - Products
  - Products\_Above\_Average\_Prices
  - Products\_by\_Categories
  - Regions

Orders

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate
10248	VINET	5	04-07-1996 00:00:00	01-08-199
10249	TOMSP	6	05-07-1996 00:00:00	16-08-199
10250	HANAR	4	08-07-1996 00:00:00	05-08-199
10251	VICTE	3	08-07-1996 00:00:00	05-08-199
10252	SUPRD	4	09-07-1996 00:00:00	06-08-199
10253	HANAR	3	10-07-1996 00:00:00	24-07-199
10254	CHOPS	5	11-07-1996 00:00:00	08-08-199
10255	RICSU	9	12-07-1996 00:00:00	09-08-199
10256	WELLI	3	15-07-1996 00:00:00	12-08-199
10257	HILAA	4	16-07-1996 00:00:00	13-08-199
10258	ERNSH	1	17-07-1996 00:00:00	14-08-199
10259	CENTC	4	18-07-1996 00:00:00	15-08-199
10260	OTTIK	4	19-07-1996 00:00:00	16-08-199
10261	QUEDE	4	19-07-1996 00:00:00	16-08-199
10262	RATTC	8	22-07-1996 00:00:00	19-08-199
10263	ERNSH	9	23-07-1996 00:00:00	20-08-199
10264	FOLKO	6	24-07-1996 00:00:00	21-08-199
10265	BLONP	2	25-07-1996 00:00:00	22-08-199
10266	WARTH	3	26-07-1996 00:00:00	06-09-199
10267	FRANK	4	29-07-1996 00:00:00	26-08-199
10268	GROSR	8	30-07-1996 00:00:00	27-08-199
10269	WHITC	5	31-07-1996 00:00:00	14-08-199
10270	WARTH	1	01-08-1996 00:00:00	29-08-199

Select Related Tables      Load      Edit      Cancel

Query Editor Window will appear

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.



The screenshot shows the Power Query Editor interface. The 'Transform' tab is active in the ribbon. A context menu is open over the 'Order\_Details' column, with the 'Expand' option selected. The 'QUERY SETTINGS' pane on the right shows the query name is set to 'Orders'. The preview area at the bottom right indicates it was downloaded at 00:53.

After clicking on OK following screen appears with combined columns

The screenshot shows the Power BI Desktop interface after expanding the 'Order\_Details' column. The 'Transform' tab is active. The 'QUERY SETTINGS' pane now lists 'Navigation' and 'Expanded Order\_Details' as applied steps. The preview area at the bottom right indicates it was downloaded at 00:53.

#### 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:

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



Screenshot of the Power Query Editor showing the 'Orders' query. The 'Applied Steps' pane shows the 'Expanded Order\_Details' step. The 'Properties' pane shows the 'Name' as 'Orders'. The main grid displays columns: ProductID, UnitPrice, Quantity, and Shipper.

- b) In the Custom Column dialog box, in the Custom Column Formula textbox, enter [Order\_Details.UnitPrice] \* [Order\_Details.Quantity] by selecting from available columns and click on insert for each column.
- c) In the New column name textbox, enter LineTotal.
- d) Click OK.

Screenshot of the 'Custom Column' dialog box. The 'New column name' field contains 'LineTotal'. The 'Custom column formula:' field contains '= [Order\_Details.UnitPrice]\*[Order\_Details.Quantity]'. The 'Available columns:' list includes Order\_Details.ProductID, Order\_Details.UnitPrice, Order\_Details.Quantity, and Shipper. The 'OK' button is visible at the bottom right.



Untitled - Power Query Editor

File Home Transform Add Column View Help

Queries [2] Products Orders

uctID	1.2 Order_Details.UnitPrice	1.2 Order_Details.Quantity	Shipper	ABC 123 LineTotal
1	11	14	12 Record	168
2	42	9.8	10 Record	98
3	72	34.8	5 Record	174
4	14	18.6	9 Record	167.4
5	51	42.4	40 Record	1696
6	41	7.7	10 Record	77
7	51	42.4	35 Record	1484
8	65	16.8	15 Record	252
9	22	16.8	6 Record	100.8
10	57	15.6	15 Record	234
11	65	16.8	20 Record	336
12	20	64.8	40 Record	2592
13	33	2	25 Record	50
14	60	27.2	40 Record	1088
15	31	10	20 Record	200
16				

21 COLUMNS, 999+ ROWS PREVIEW DOWNLOADED AT 00:53

QUERY SETTINGS

- PROPERTIES Name: Orders All Properties
- APPLIED STEPS Source, Navigation, Expanded Order\_Details, **Added Custom**

## 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.

- a) In Query Editor, drag the LineTotal column to the left, after ShipCountry.

Untitled - Power Query Editor

File Home Transform Add Column View Help

Queries [2] Products Orders

	A <sup>B</sup> C ShipPostalCode	A <sup>B</sup> C ShipCountry	ABC 123 LineTotal	Customer	Employee
1	null	51100	168	Record	Record
2	null	51100	98	Record	Record
3	null	51100	174	Record	Record
4	null	44087	167.4	Record	Record
5	null	44087	1696	Record	Record
6	05454-876	Brazil	77	Record	Record
7	05454-876	Brazil	1484	Record	Record
8	05454-876	Brazil	252	Record	Record
9	null	69004	100.8	Record	Record
10	null	69004	234	Record	Record
11	null	69004	336	Record	Record
12	null	B-6000	2592	Record	Record
13	null	B-6000	50	Record	Record
14	null	B-6000	1088	Record	Record
15	05454-876	Brazil	200	Record	Record
16					

21 COLUMNS, 999+ ROWS PREVIEW DOWNLOADED AT 00:53

QUERY SETTINGS

- PROPERTIES Name: Orders All Properties
- APPLIED STEPS Source, Navigation, Expanded Order\_Details, **Added Custom**, **Reordered Columns**



- b)** 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.

The screenshot shows the Power Query Editor interface with the 'Orders' query selected. The 'Queries [2]' pane on the left lists 'Products' and 'Orders'. The main area displays a table with three columns: 'ProductID', 'UnitPrice', and 'Quantity'. The 'UnitPrice' and 'Quantity' columns have their names partially deleted, showing '1.2.' before the original names. The 'QUERY SETTINGS' pane on the right shows the 'Name' as 'Orders' and the 'Applied Steps' list, which includes 'Renamed Columns'.

## 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 & Apply.



The screenshot shows the Power Query Editor interface. On the left, there's a navigation pane with 'Close & Apply' buttons, 'Recent Sources', 'Data Sources', 'Parameters', 'Preview', and 'Manage'. The main area displays two queries: 'Products' and 'Orders'. The 'Orders' query is currently selected. The preview pane below shows 15 records from the Orders table, with columns including ProductID, UnitPrice, Quantity, and Shipper. The transform ribbon at the top has various tools like 'Close & Apply', 'New Source', 'Data Settings', 'Manage Parameters', 'Refresh', 'Properties', 'Advanced Editor', 'Manage Columns', 'Choose Columns', 'Remove Columns', 'Keep Rows', 'Remove Rows', 'Reduce Rows', 'Sort', 'Split Column', 'Group By', 'Replace Values', and 'Transform'. A 'QUERY SETTINGS' pane on the right shows the query name 'Orders' and applied steps like 'Source', 'Navigation', 'Expanded Order\_Details', 'Added Custom', 'Reordered Columns', and 'Renamed Columns'.

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

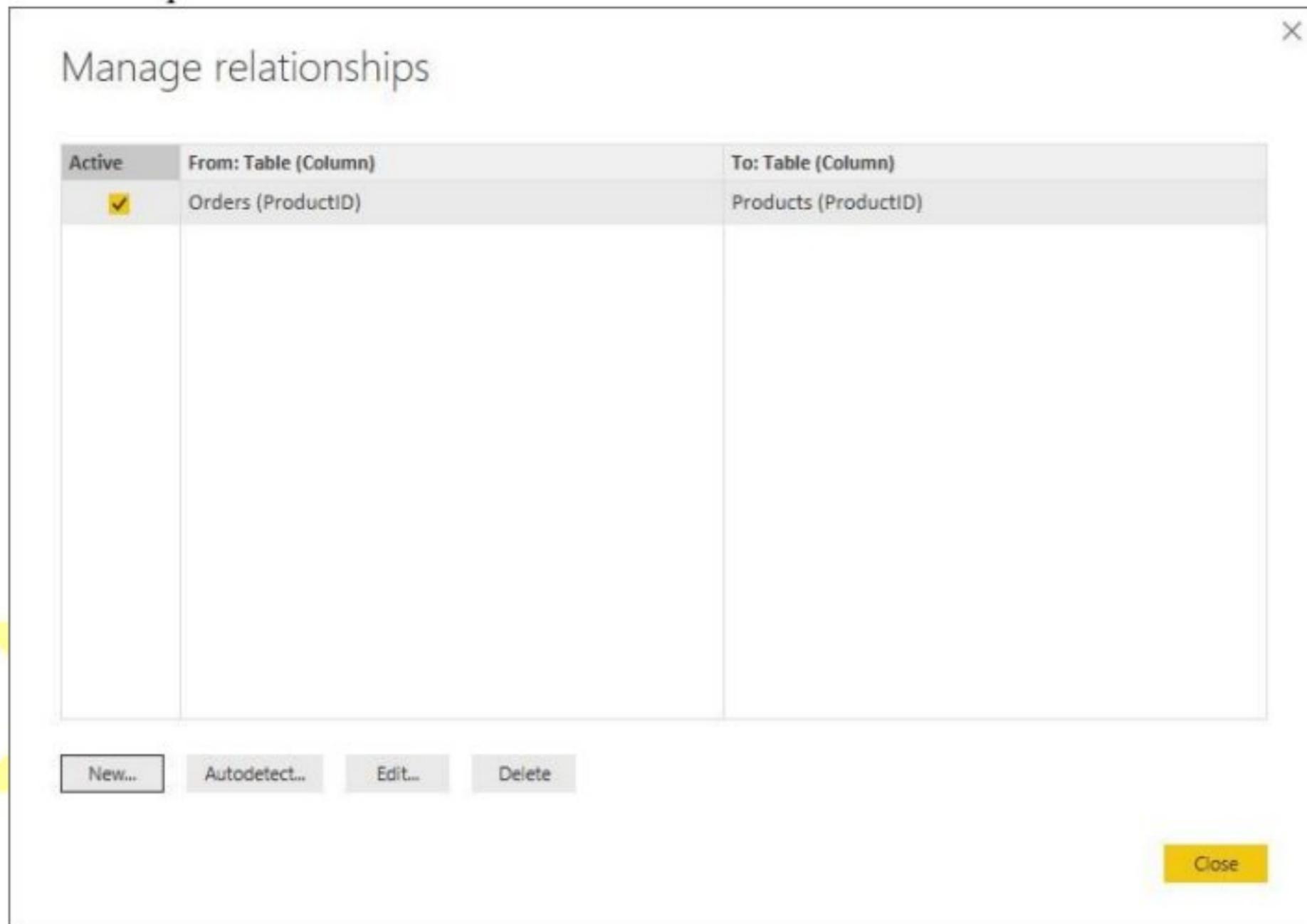
The screenshot shows a 'Apply query changes' dialog box. It says 'Orders' and 'Loading data to model...'. There is a 'Cancel' button at the bottom right. The background shows a decorative banner with the text 'CE IN EDUCATION'.

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

The screenshot shows the Power BI Desktop ribbon. The 'Home' tab is selected. The ribbon includes 'Clipboard' (Cut, Copy, Paste), 'External data' (Get Data, Refresh), 'Insert' (Text box, Buttons, Shapes, From Marketplace, From File, Switch Theme, Themes), 'Relationships' (Manage Relationships), 'Calculations' (New Measure, New Column, New Quick Measure), and 'Share'. The 'Manage Relationships' button is highlighted. The 'FIELDS' pane on the right lists tables 'Orders' and 'Products' with their respective fields: ProductID, ProductName, QuantityPerUnit, and UnitsInStock. The 'VALUES' section has a placeholder 'Add data fields here'. The 'FILTERS' section has placeholders for 'Page level filters' and 'Report level filters'. The 'DRILLTHROUGH' section has a placeholder 'Drag drillthrough fields here'.



Step 4. Select the New... button



Step 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.

निर्मलेन्ह उत्तम सेवाधर्म



Create relationship

Select tables and columns that are related.

Products

ProductID	QuantityPerUnit	ProductName	UnitsInStock
1	10 boxes x 20 bags	Chai	39
2	24 - 12 oz bottles	Chang	17
3	12 - 550 ml bottles	Aniseed Syrup	13

Orders

Name	ShipAddress	ShipCity	ShipRegion	ShipPostalCode	ShipCountry	LineTotal	ProductID	Ur
K-Stop	Taucherstraße 10	Cunewalde	null	01307	Germany	595.2	10	
K-Stop	Taucherstraße 10	Cunewalde	null	01307	Germany	150	31	
K-Stop	Taucherstraße 10	Cunewalde	null	01307	Germany	40	33	

Cardinality

One to many (1:\*)

Cross filter direction

Single

Make this relationship active

Assume referential integrity

Apply security filter in both directions

There's already a relationship between these two columns.

OK Cancel

Step 6: Select Cancel, and then select Relationship view in Power BI Desktop.

निर्मलेन्ह उत्तम सेवाधर्म



The screenshot shows the Power BI Desktop interface. The ribbon at the top includes File, Home, View, Modeling, and Help tabs. The Home tab is selected. The ribbon bar contains various icons for data management (Get Data, Sources, Enter Data, Edit Queries, Refresh), visualization creation (New Page, New Visual, Ask A Question, Buttons, Text box, Image, Shapes), and publishing (From Marketplace, From File, Switch Theme, Manage Relationships, New Measure, New Column, New Quick Measure, Publish). The main workspace is currently empty. The right side features the 'FIELDS' pane, which displays a search bar and a tree view of data fields. Under the 'Products' node, fields like ProductID, ProductName, QuantityPerUnit, and UnitsInStock are listed. Below the Fields pane are sections for Visualizations, Filters, and Drillthrough.

Step 7: We see the following, which visualizes the relationship between the queries.

The screenshot shows the Power BI Desktop interface with the ribbon tabs File, Home, Modeling, and Help. The Home tab is selected. The ribbon bar includes icons for Get Data, Sources, Enter Data, Edit Queries, Refresh, New Page, New Visual, Ask A Question, Buttons, Text box, Image, Shapes, From Marketplace, From File, Switch Theme, Manage Relationships, New Measure, New Column, New Quick Measure, Publish, Calculations, and Share. The main workspace is empty. The right side features the 'Relationships' pane, which shows a relationship between the 'Products' and 'Orders' tables. A line connects the 'ProductID' field in the 'Products' table to the 'OrderID' field in the 'Orders' table. The 'Products' table has a '1' and the 'Orders' table has an asterisk (\*) at the connection point.

Step 8: When you double-click the arrow on the line that connects the two queries, an Edit Relationship dialog appears.



Edit relationship

Select tables and columns that are related.

Orders

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	ShippedDate	ShipVia	F
10273	QUICK	3	05-08-1996 00:00:00	02-09-1996 00:00:00	12-08-1996 00:00:00	3	
10273	QUICK	3	05-08-1996 00:00:00	02-09-1996 00:00:00	12-08-1996 00:00:00	3	
10273	QUICK	3	05-08-1996 00:00:00	02-09-1996 00:00:00	12-08-1996 00:00:00	3	

Products

ProductID	QuantityPerUnit	ProductName	UnitsInStock
1	10 boxes x 20 bags	Chai	39
2	24 - 12 oz bottles	Chang	17
3	12 - 550 ml bottles	Aniseed Syrup	13

Cardinality

Many to one (\*;1)

Cross filter direction

Single

Make this relationship active

Assume referential integrity

Apply security filter in both directions

OK Cancel

Step 9: No need to make any changes, so we'll just select Cancel to close the Edit Relationship dialog.

निर्मलेन्ह उत्तम सेवाधर्म

**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

1. **What is extraction?**
2. **Name an extraction tool.**
3. **What is incremental extraction?**
4. **Why is extraction important?**

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



**Practical-3: a. Create the Data staging area for the selected database.  
b. Create the cube with suitable dimension and fact tables based on ROLAP, MOLAP and HOLAP model.**

**Aims:**

1. To create a dedicated data staging area for the selected database, enabling efficient data cleansing, transformation, and preparation for analysis.
2. To design and construct a multidimensional cube with appropriate dimension and fact tables using ROLAP, MOLAP, and HOLAP models for enhanced data analysis.

**Learning Objectives:**

1. Understand the role of data staging in the ETL and data warehousing process.
2. Gain hands-on experience in setting up a staging area to extract, clean, and transform raw data.
3. Learn the fundamentals of cube design, including the creation of dimension and fact tables.
4. Explore the differences and use cases for ROLAP, MOLAP, and HOLAP models in multidimensional analysis.

**Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

**Theory-3: Data staging**

Data staging is a crucial phase in data warehousing where extracted data is temporarily stored, cleansed, and transformed before being loaded into the target database. This ensures data quality, integrity, and optimized performance for analytical processing.

**Creating a Data Staging Area**

1. **Database Selection:** Choose a suitable database system (e.g., SQL Server, Oracle) to store the staged data.
2. **Schema Design:** Define staging tables to hold raw, intermediate, and transformed data.
3. **Data Loading:** Extracted data from multiple sources is loaded into the staging area for preprocessing.
4. **Data Cleansing:** Perform data validation, deduplication, and format standardization.
5. **Transformation:** Apply necessary business rules, data aggregation, and normalization before transferring data to the data warehouse.

**Creating a Cube with Suitable Dimensions and Fact Tables**

## **1. Choosing the OLAP Model:**

- **ROLAP (Relational OLAP):** Stores data in relational databases and processes queries dynamically.
- **MOLAP (Multidimensional OLAP):** Stores pre-aggregated data in multidimensional cubes for faster access.

## **2. Defining Fact Tables:**

- Contains measurable business data (e.g., sales, revenue, transaction count).
- Linked with dimension tables using foreign keys.

## **3. Defining Dimension Tables:**

- Stores descriptive attributes (e.g., time, product, customer, location).
- Supports data slicing, dicing, and drill-down operations.

## **4. Building the Cube:**

- Organize fact and dimension tables within the chosen OLAP model.
- Precompute aggregations to enhance query performance.



### PRACTICAL 3 b

Create the cube with suitable dimension and fact tables based on

OLAP

Step 1: 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.

Downloading "Data\_WareHouse\_SQLScript.zip" from the article  
<https://www.codeproject.com/Articles/652108/Create-First-Data-Warehouse>



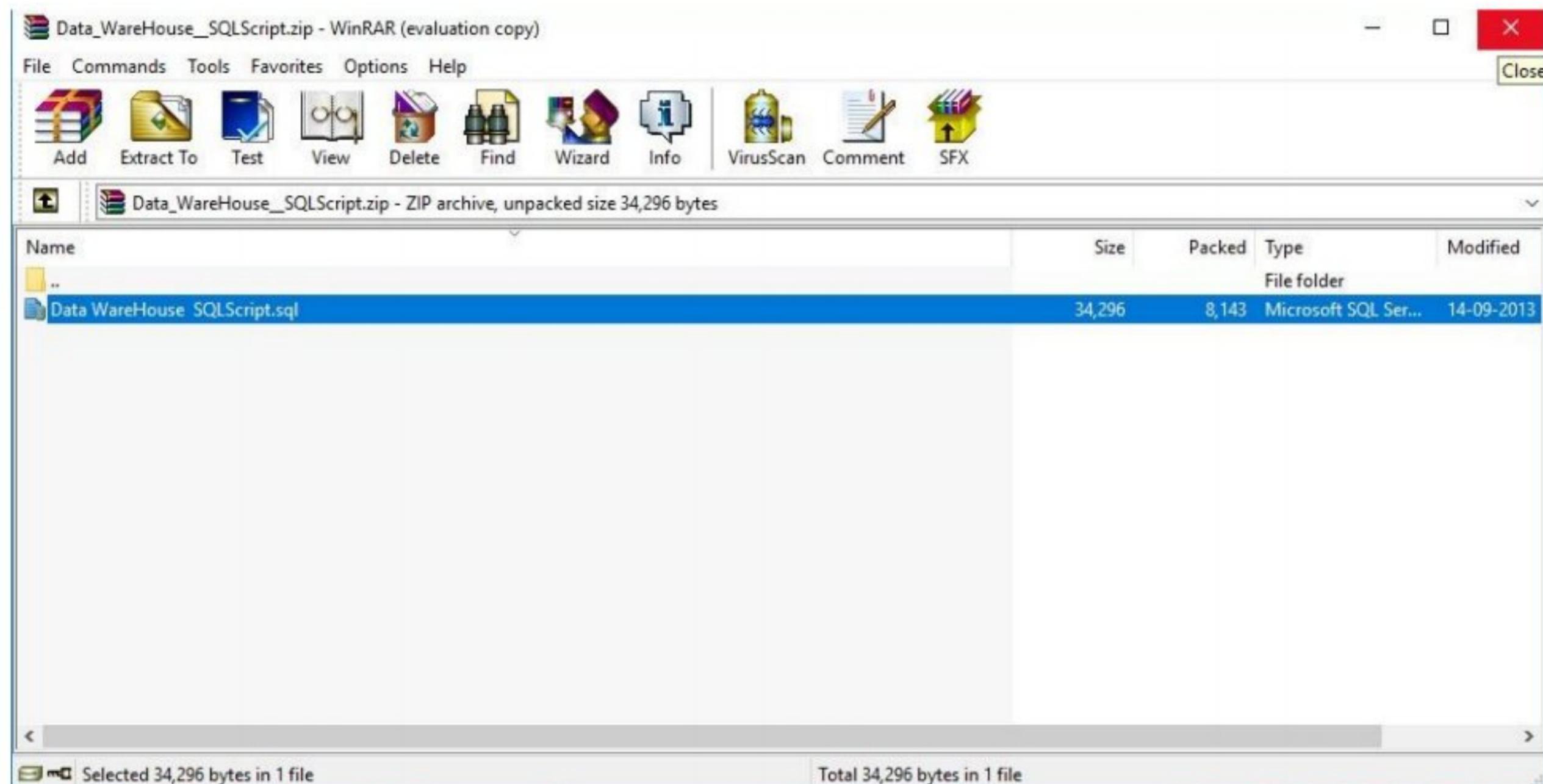
निर्मलेन्ह उत्तम सेवाधर्म

You must logon to download zip files. To ensure that our limited resources are not abused we must ask that all readers wishing to download zip files log in first. Please check the FAQ if you encounter problems.

**Existing Members**  
Sign in to your account  
Your Email   
Password   
  
Forgot your password?

**New Membership**  
Download, Vote, Comment, Publish.  
Your Email  ansurkar.gauri@gmail.com  
Optional Password   
  
 I have read and agree to the [Terms of Service](#) and [Privacy Policy](#).  
 Please subscribe me to the CodeProject newsletters.

Or sign in using



After downloading extract file in folder.

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

- 1.** Open SQL Server Management Studio 2012
- 2.** Connect Database Engine



Password for sa : admin123 (as given during installation)

Click Connect.

**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

OR

**1.** Go to the extracted sql file and double click on it.

**2.** New Sql Query Editor will be opened containing Sales\_DW Database.

निर्माण नोहे उत्तम सेवाधर्म



Data WareHouse SQLScript.sql - LAPTOP-Q7CMJ9K1\HP.Sales\_DW (LAPTOP-Q7CMJ9K1\Gauri (55)) - Microsoft SQL Server Management Studio

File Edit View Query Project Debug Tools Window Help

Sales DW Data WareHouse S...CMJ9K1\Gauri (55) X

Object Explorer Connect ▾

LAPTOP-Q7CMJ9K1\HP (SQL Server 11)

Databases System Databases master Views Synonyms Programmability Service Broker Storage Security model msdb tempdb Security Server Objects Replication Management

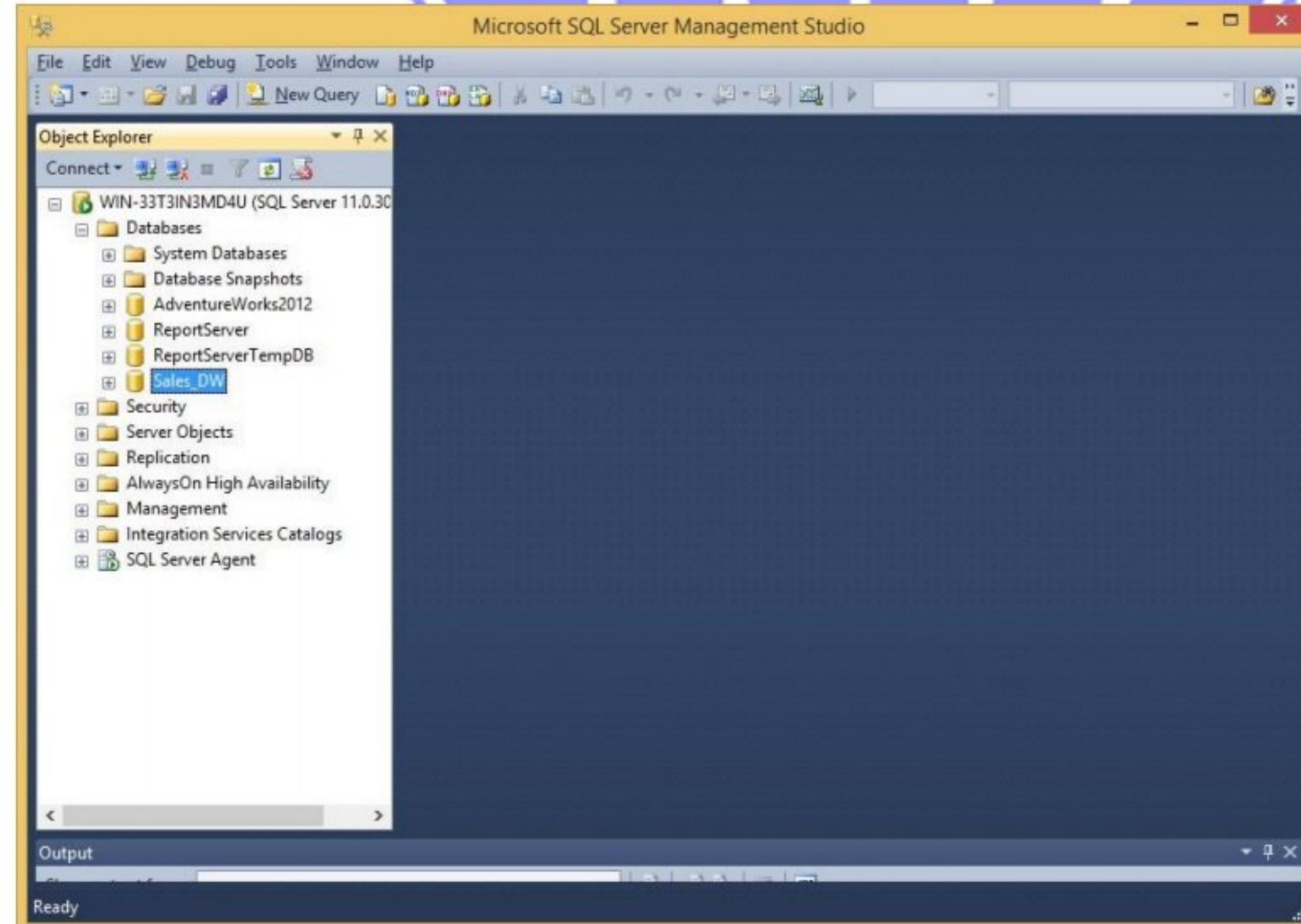
--DROP DATABASE Sales\_DM  
GO  
Create database Sales\_DM  
Go  
  
Use Sales\_DM  
Go  
  
--Create Customer dimension table in Data Warehouse which will hold customer personal details.  
  
Create table DimCustomer  
(  
CustomerID int primary key identity,  
CustomerAltID varchar(10) not null,  
CustomerName varchar(50),  
Gender varchar(20)  
)  
go  
  
--Fill the Customer dimension with sample Values  
  
Insert into DimCustomer(CustomerAltID,CustomerName,Gender)values  
('IMI-001','Henry Ford','M'),  
('IMI-002','Bill Gates','M'),  
('IMI-003','Muskan Shaikh','F'),  
('IMI-004','Richard Thrubin','M'),  
('IMI-005','Emma Wattson','F');  
Go  
  
--Create basic level of Product Dimension table without considering any Category or Subcategory.  
  
Create table DimProduct  
(  
CustomerID int primary key identity,  
CustomerAltID varchar(10) not null,  
CustomerName varchar(50),  
Gender varchar(20)  
)  
go  
  
100 %

Connected. (1/1) LAPTOP-Q7CMJ9K1\HP (11.0 RTM) | LAPTOP-Q7CMJ9K1\Gauri ... | Sales DW | 00:00:00 | 0 rows

Ln 309 Col 1 Ch 1 INS

Matches: 0

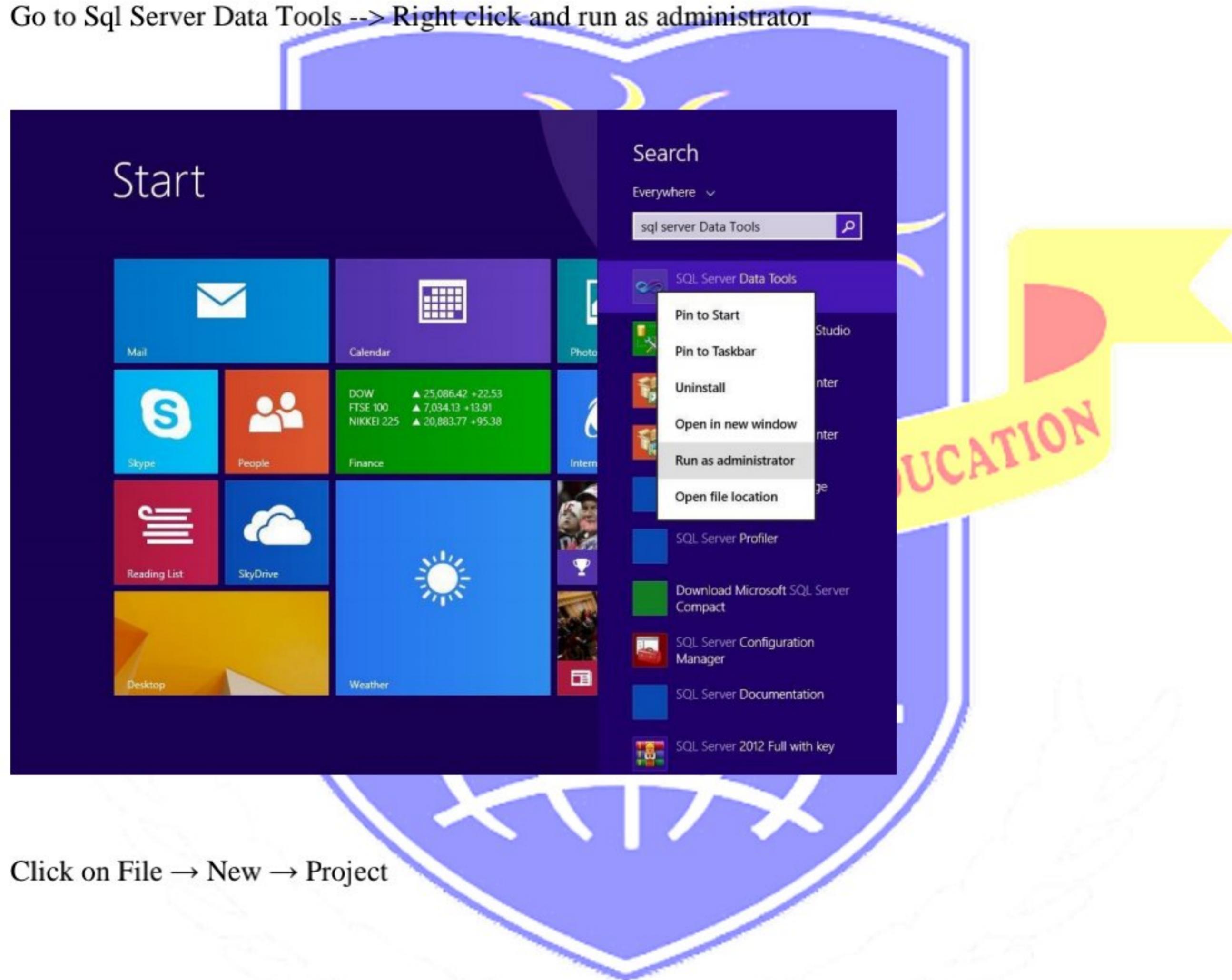
- "DEDICATED TO EXCELLENCE IN EDUCATION"**
3. Click on execute or press F5 by selecting query one by one or directly click on Execute.
  4. After completing execution save and close SQL Server Management studio & Reopen to see Sales\_DW in Databases Tab.





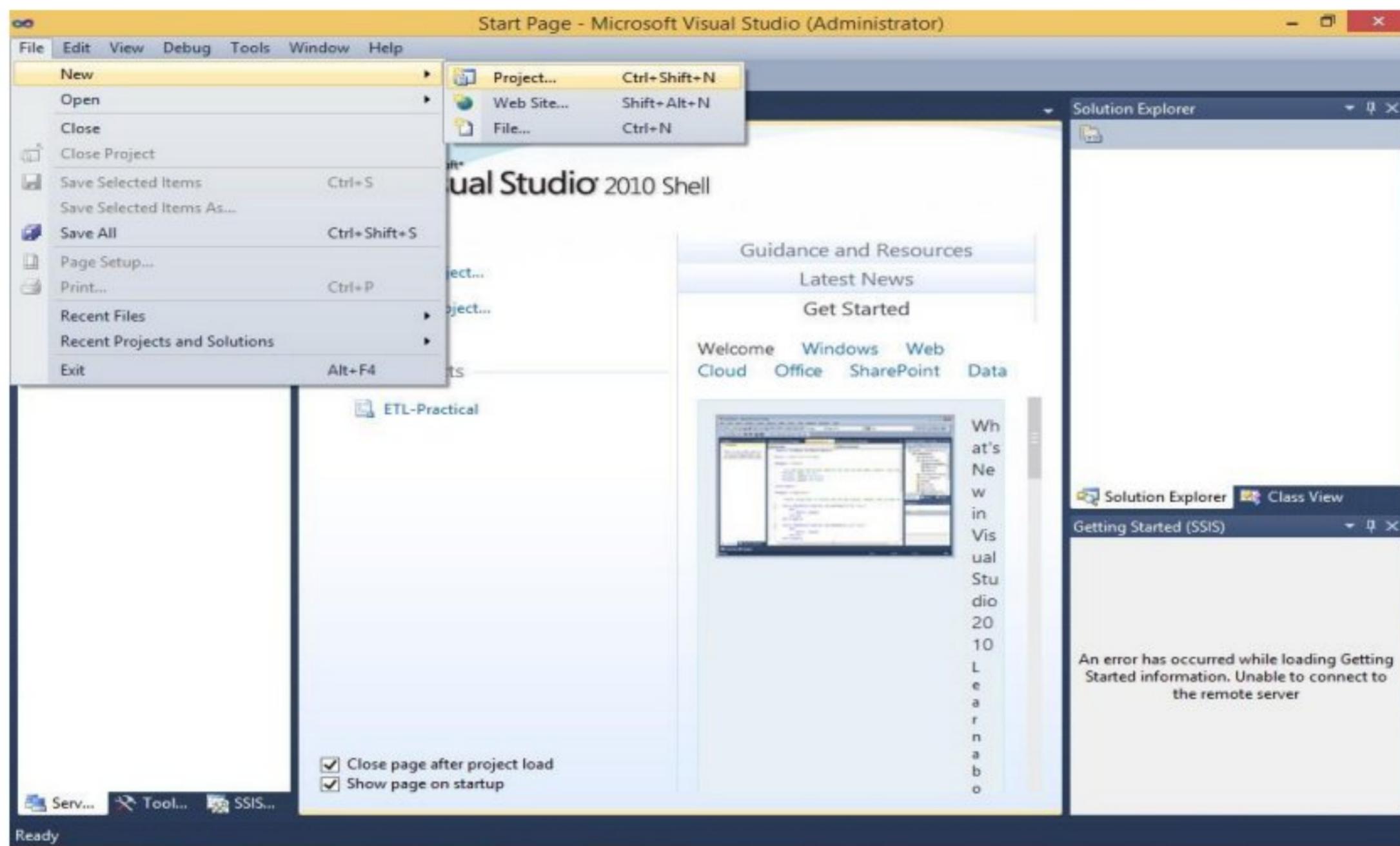
Step 2: Start SSDT environment and create New Data Source

Go to Sql Server Data Tools --> Right click and run as administrator



Click on File → New → Project

निर्माण नोह उत्तम सेवाधर्म

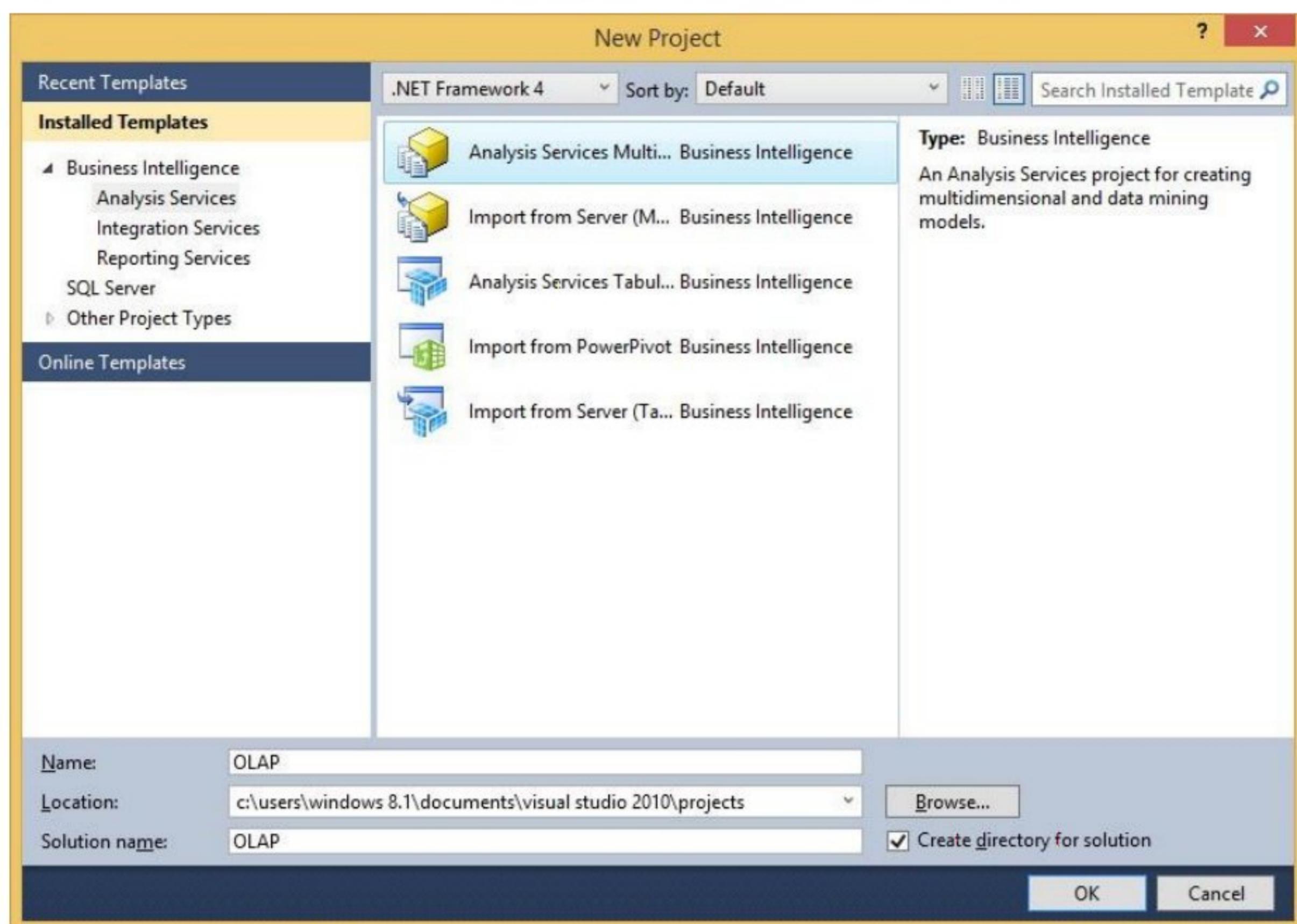


**COMMITTED TO EXCELLENCE IN EDUCATION**

In Business Intelligence → Analysis Services Multidimensional and Data Mining models → appropriate project name → click OK

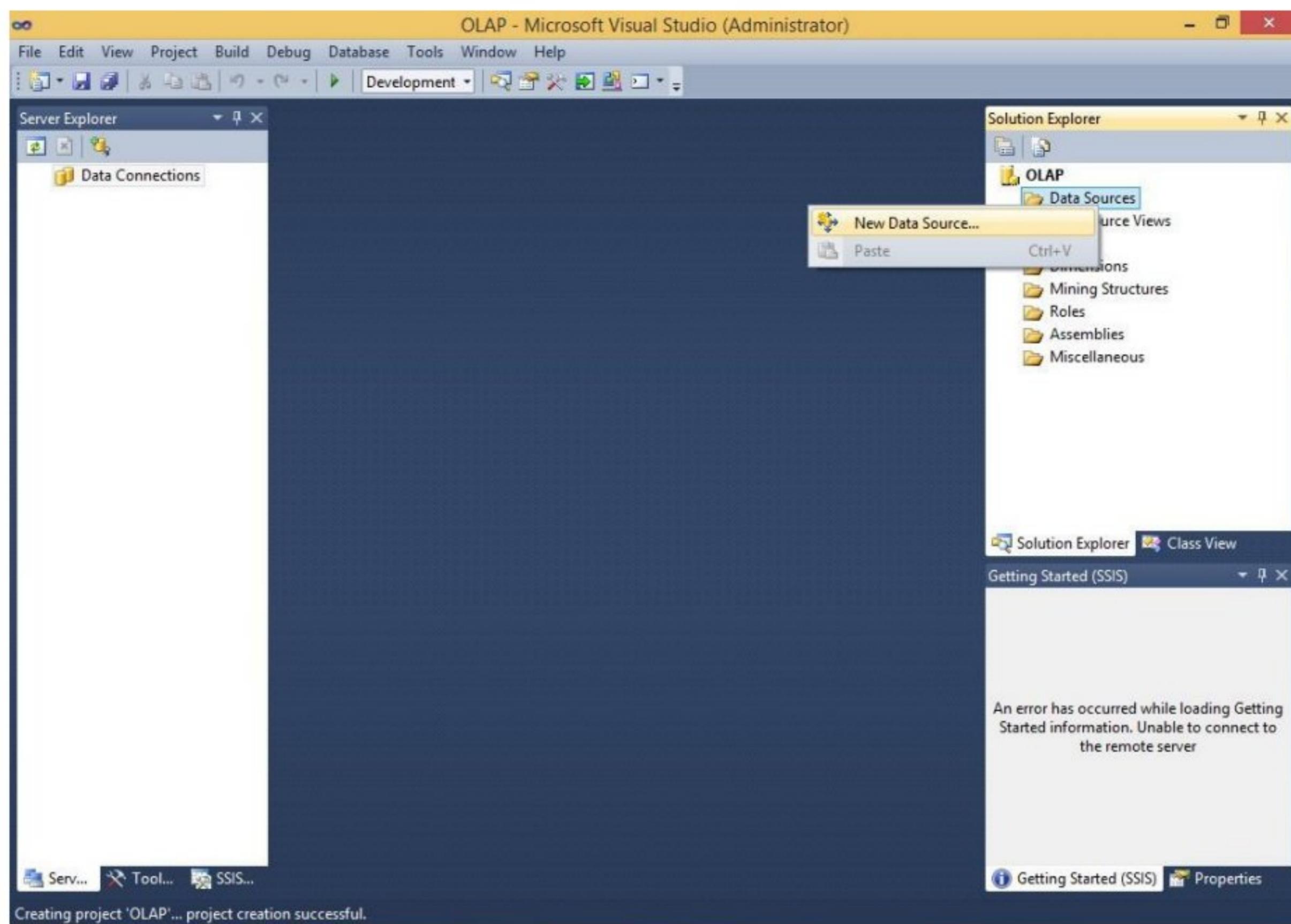


**निर्मलेन्ह उत्तम सेवाधर्म**

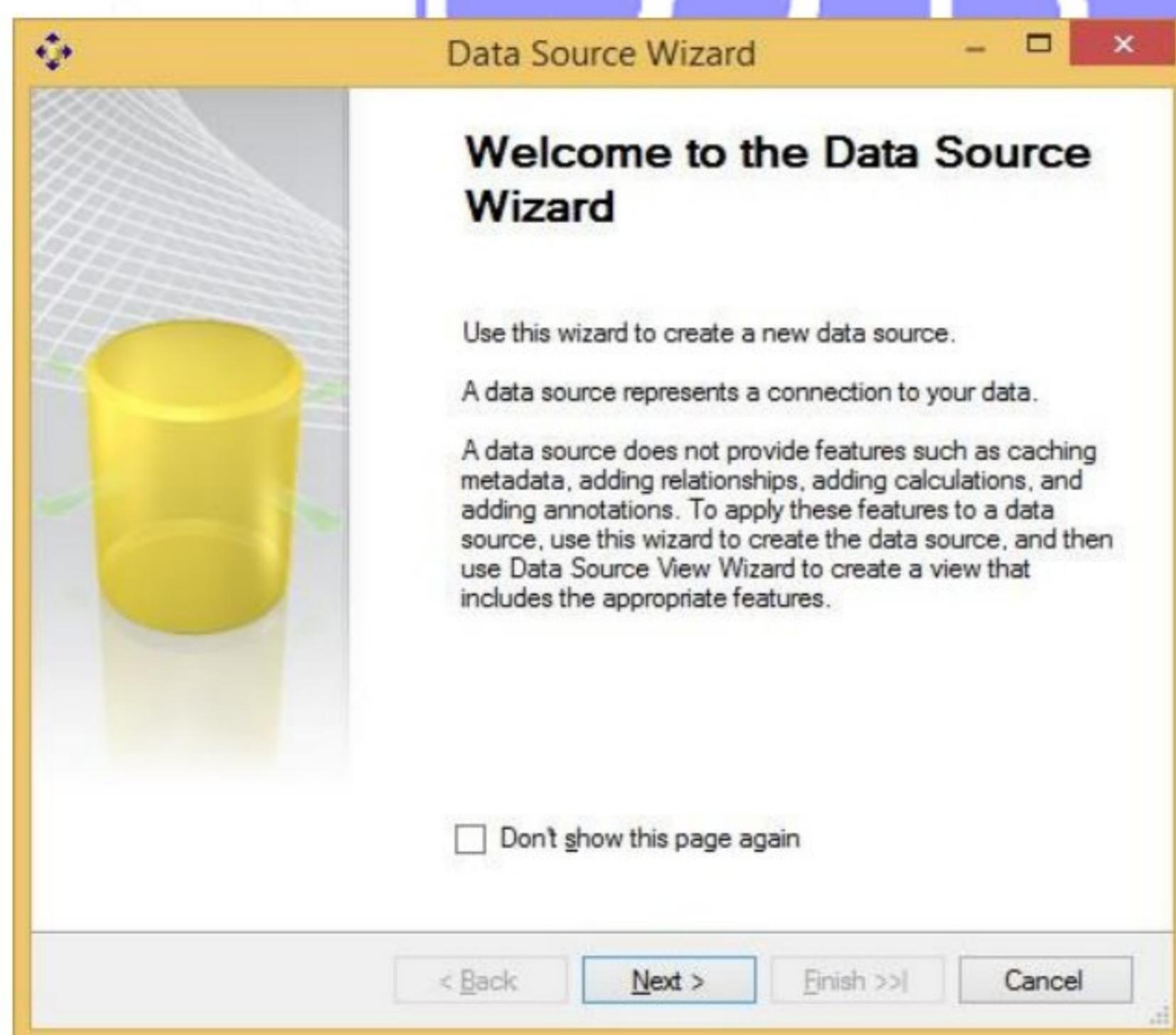


Right click on Data Sources in solution explorer → New Data Source

निर्माण नोहे उत्तम सेवाधर्म

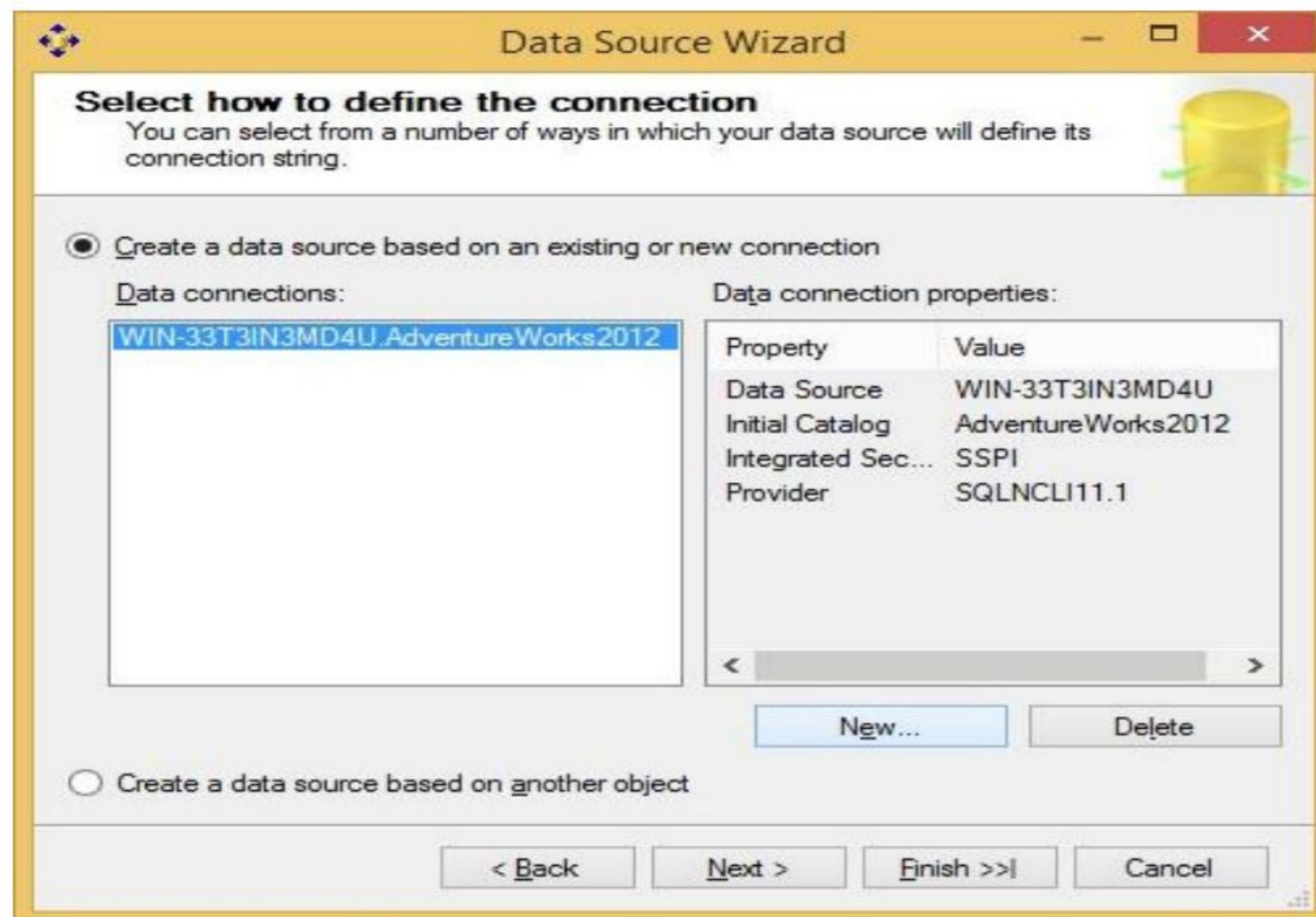


Data Source Wizard appears





Click on New



Select Server Name → select Use SQL Server Authentication → Select or enter a database name (Sales\_DW)

Note : Password for sa : admin123 (as given during installation of SQL 2012 full version)

निर्माण नोह उत्तम सेवाधर्म



Connection Manager

Provider: Native OLE DB\SQL Server Native Client 11.0

**Connection**

Server name: WIN-33T3IN3MD4U Refresh

Log on to the server

Use Windows Authentication  
 Use SQL Server Authentication

User name: sa  
Password: \*\*\*\*\*  
 Save my password

Connect to a database

Select or enter a database name:  
Sales\_DW

Attach a database file:  
  Browse...  
Logical name:  

Test Connection OK Cancel Help

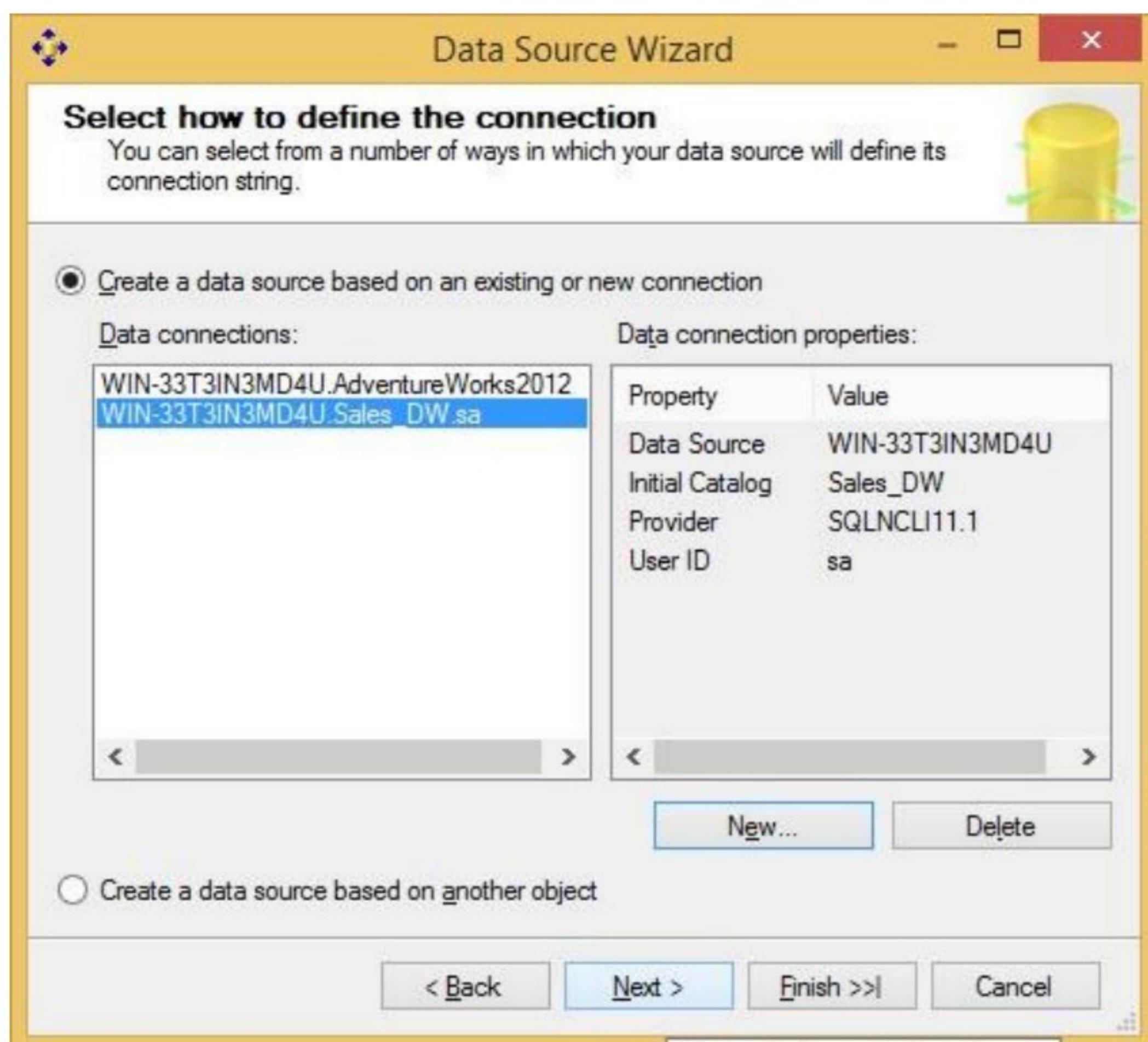
Connection Manager

Test connection succeeded.

OK

Click Next

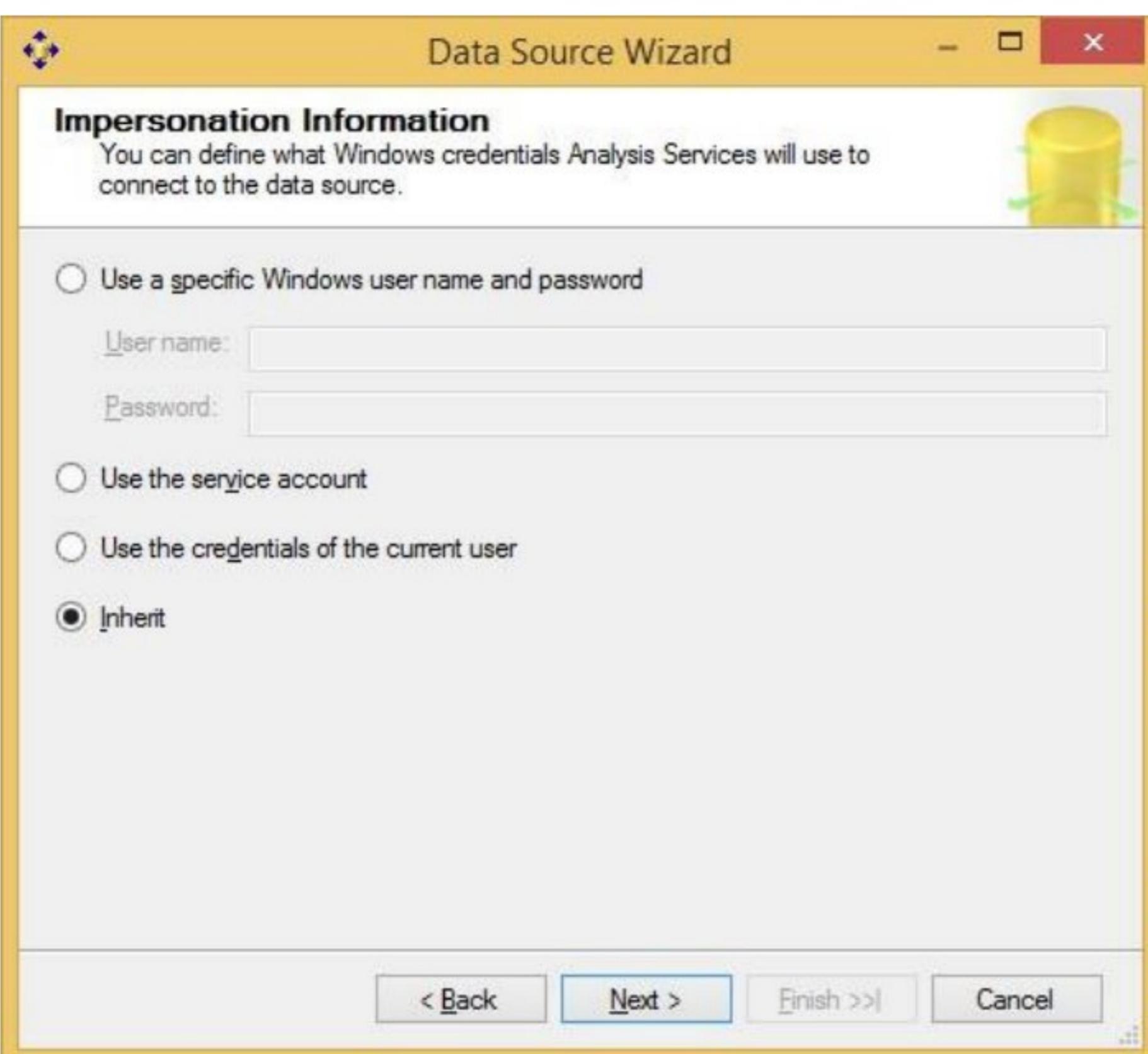
निर्मलेन्ह उत्तम सेवाधर्म



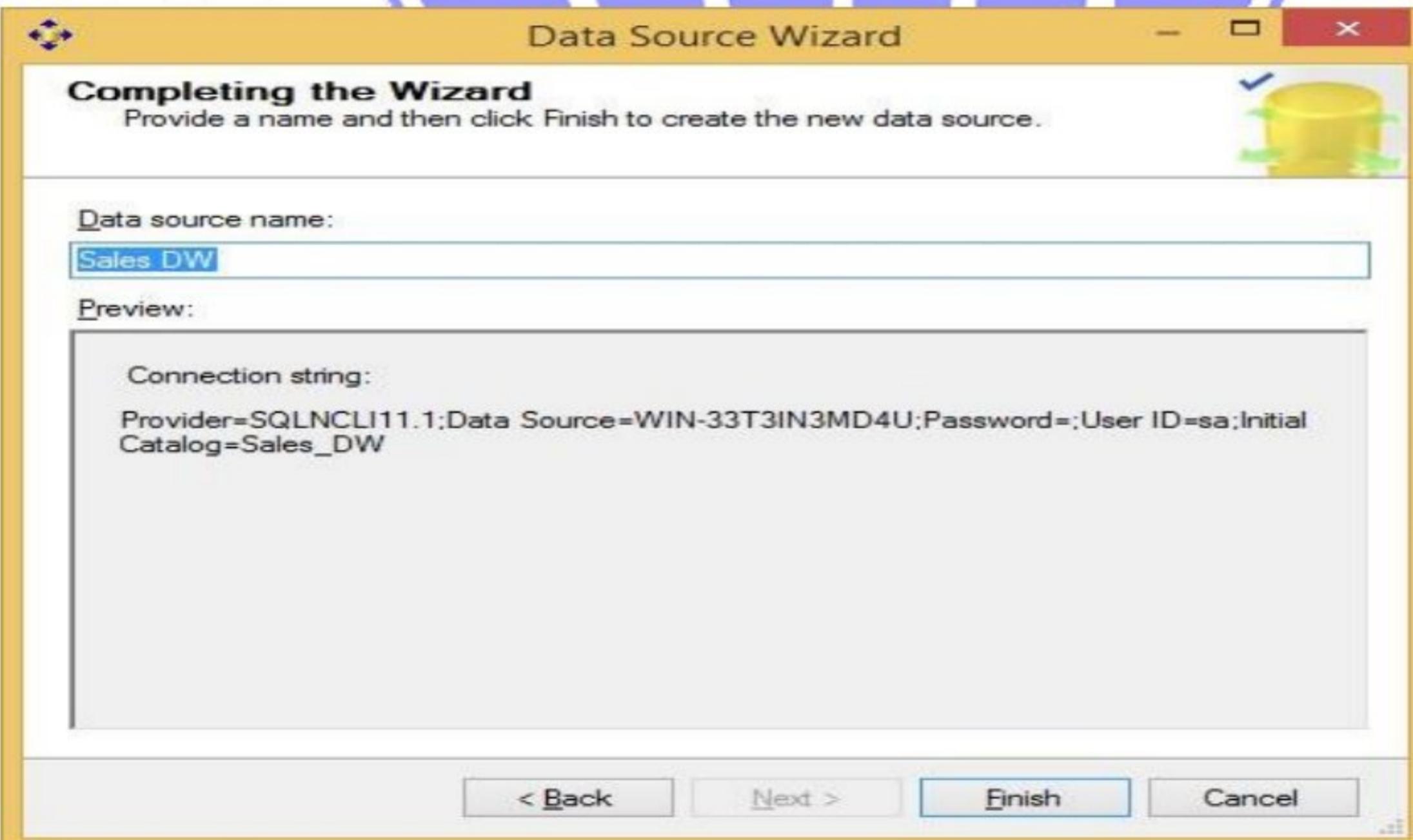
Select Inherit → Next



निर्मलेन्ह उत्तम सेवाधर्म

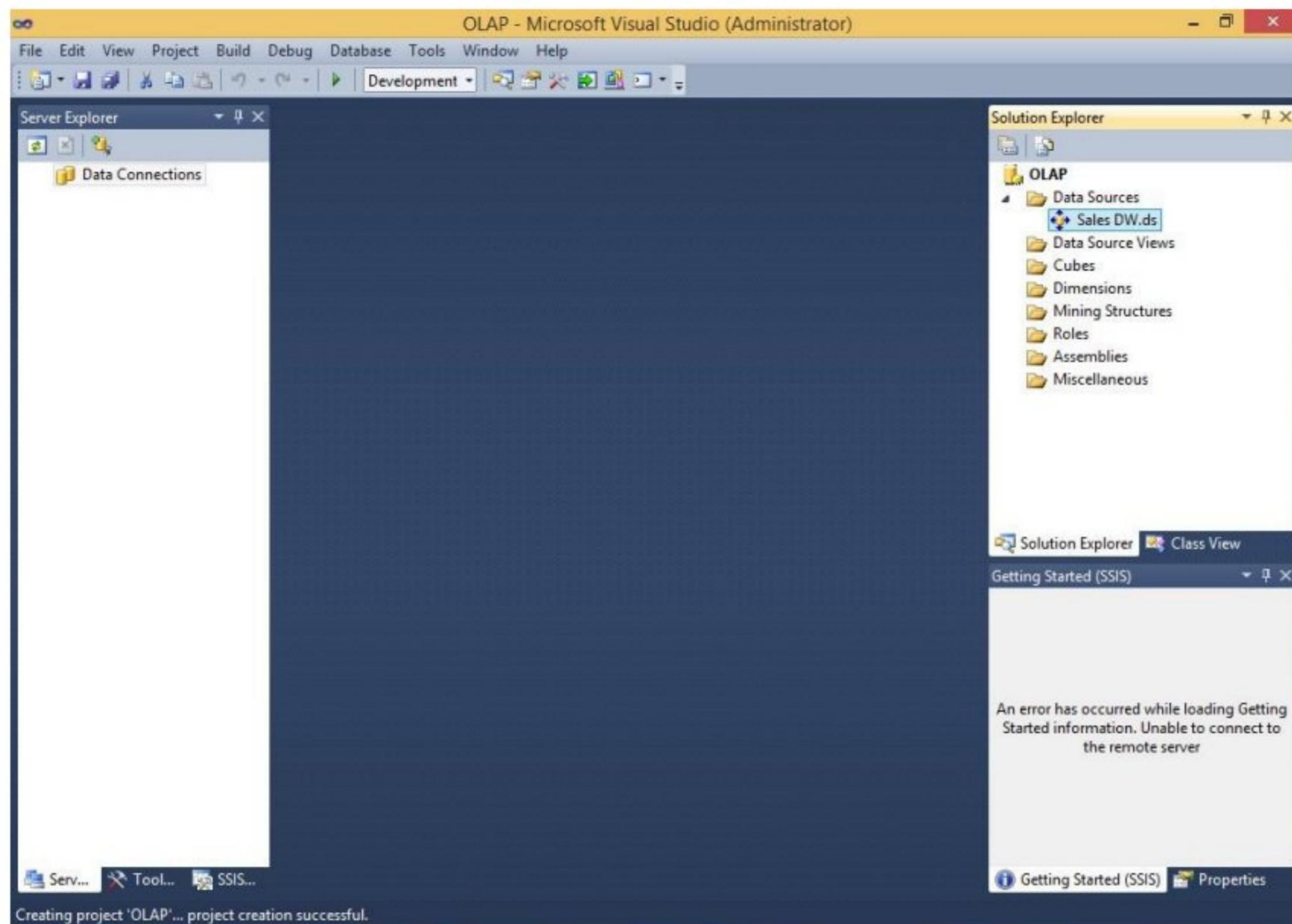


Click Finish





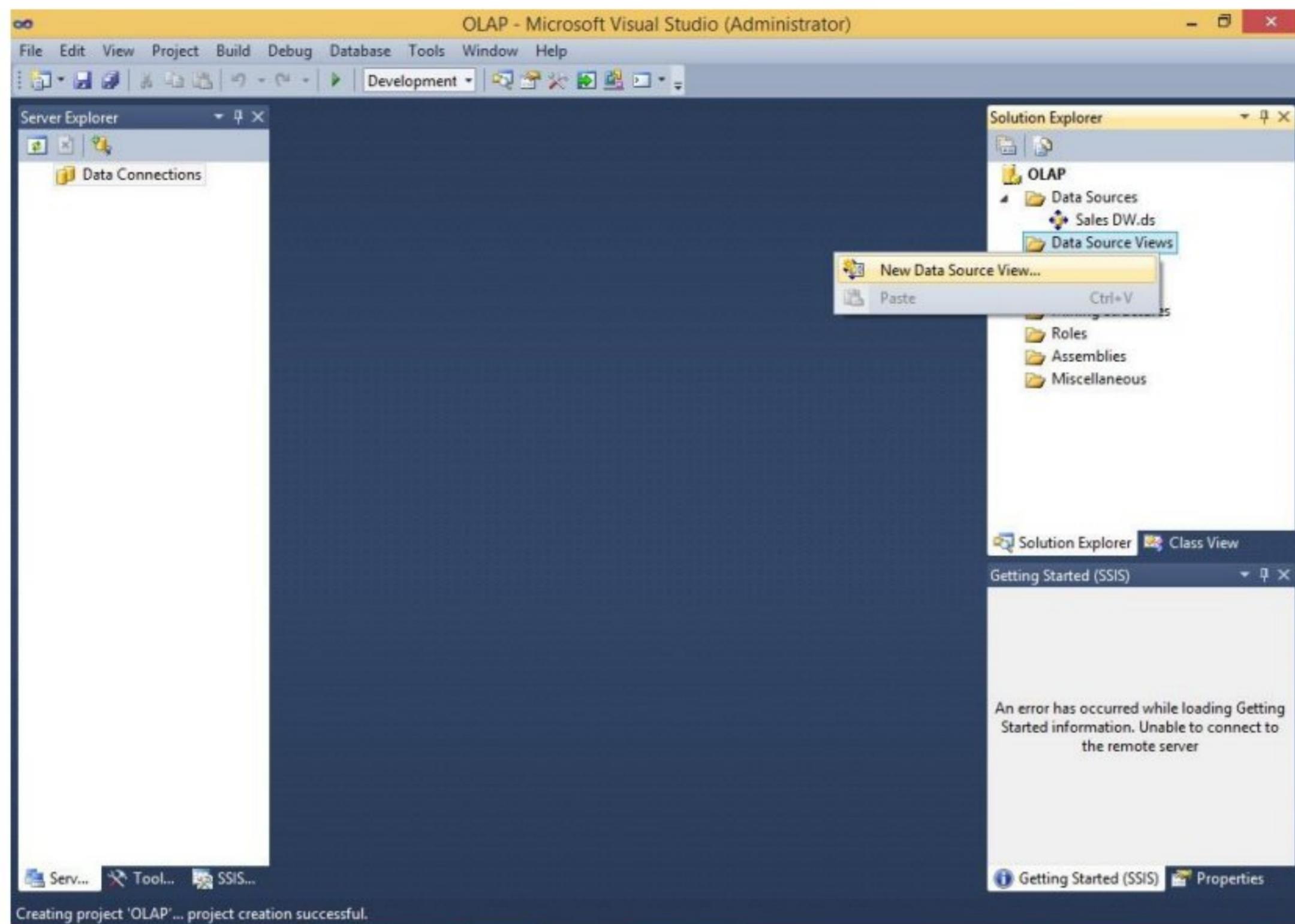
Sales\_DW.ds gets created under Data Sources in Solution Explorer



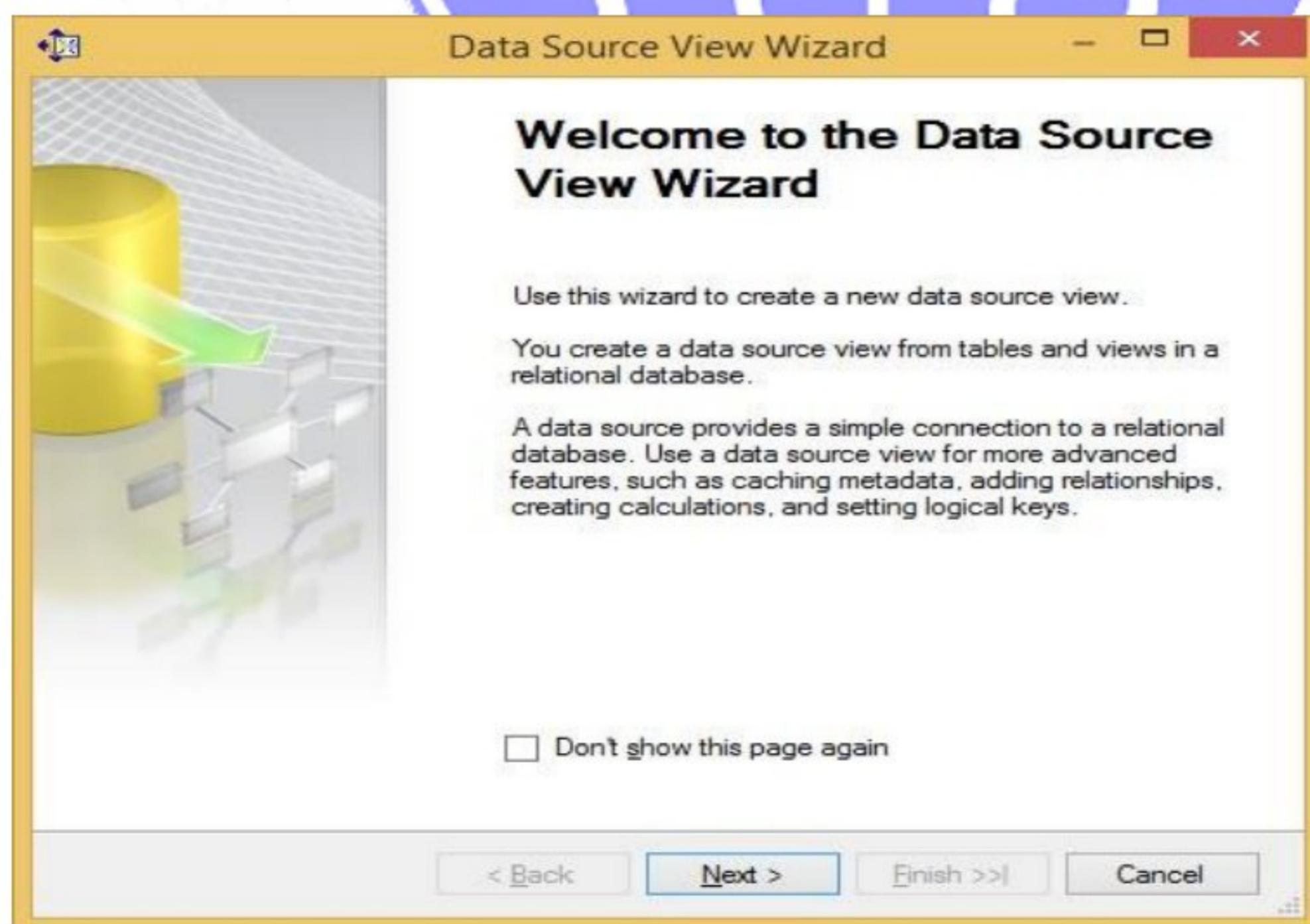
Step 3: Creating New Data Source View

In Solution explorer right click on Data Source View → Select New Data Source View

निष्ठानोह उत्तम सेवाधर्म

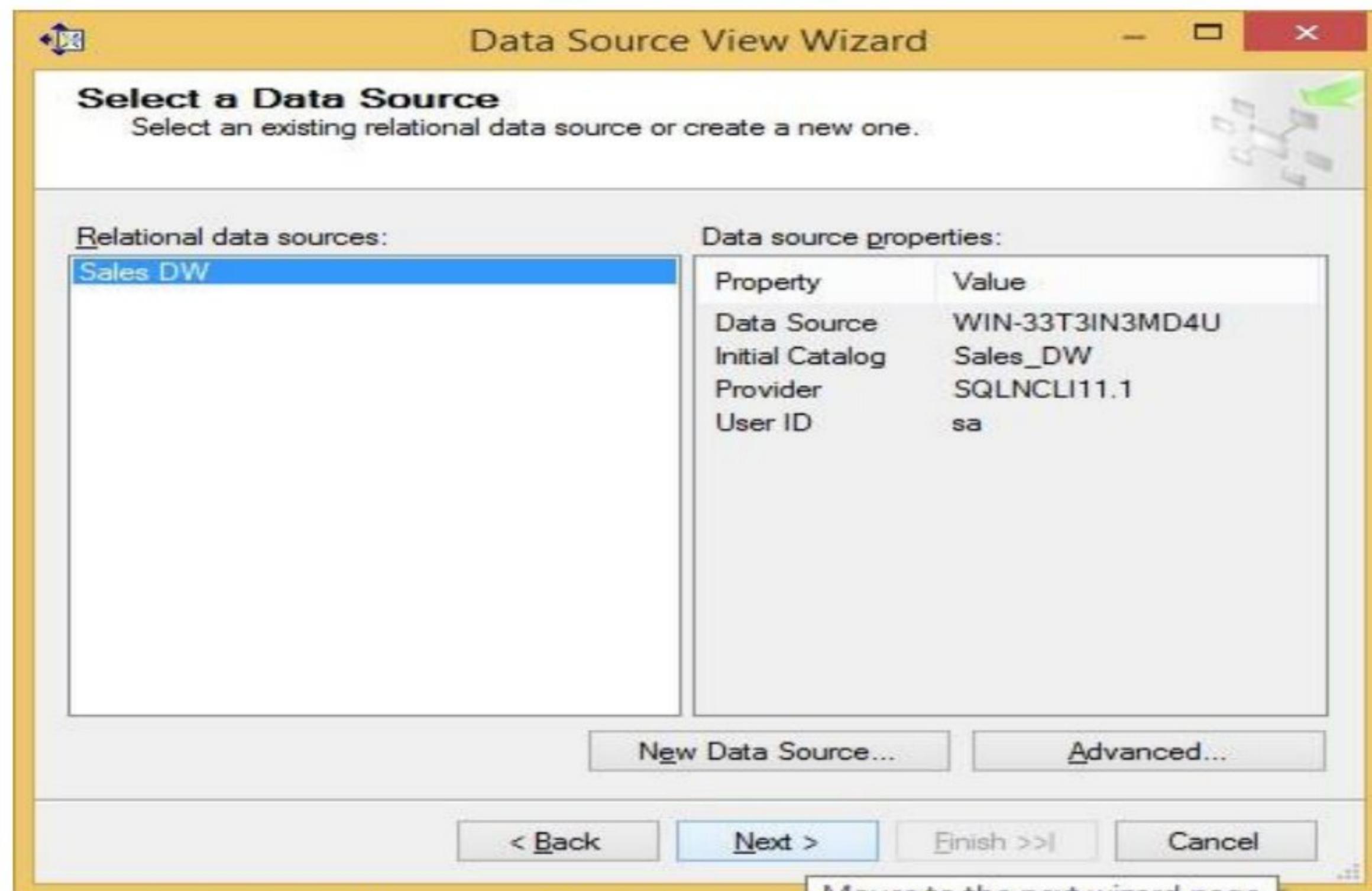


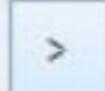
Click Next





Click Next



Select FactProductSales(dbo) from Available objects and put in Includes Objects by clicking on 

निर्मलेन्ह उत्तम सेवाधर्म



Data Source View Wizard

### Select Tables and Views

Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type
DimCustomer (dbo)	Table
DimDate (dbo)	Table
DimProduct (dbo)	Table
DimSalesPerson (dbo)	Table
DimStores (dbo)	Table
DimTime (dbo)	Table
FactProductSales (dbo)	Table

Included objects:

Name	Type

> < >> <<

Filter:

Show system objects

Add Related Tables

< Back Next > Finish >> Cancel



निर्मलेन्ह उत्तम सेवाधर्म



Data Source View Wizard

### Select Tables and Views

Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type
DimCustomer (dbo)	Table
DimDate (dbo)	Table
DimProduct (dbo)	Table
DimSalesPerson (dbo)	Table
DimStores (dbo)	Table
DimTime (dbo)	Table

Included objects:

Name	Type
FactProductSales (dbo)	Table

> < >> <<

Filter:

Show system objects

Add Related Tables

< Back Next > Finish >> Cancel

Click on Add Related Tables



निर्मलेन्ह उत्तम सेवाधर्म



Data Source View Wizard

Select Tables and Views  
Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type
DimCustomer (dbo)	Table
DimDate (dbo)	Table
DimProduct (dbo)	Table
DimSalesPerson (dbo)	Table
DimStores (dbo)	Table
DimTime (dbo)	Table

Included objects:

Name	Type
FactProductSales (dbo)	Table

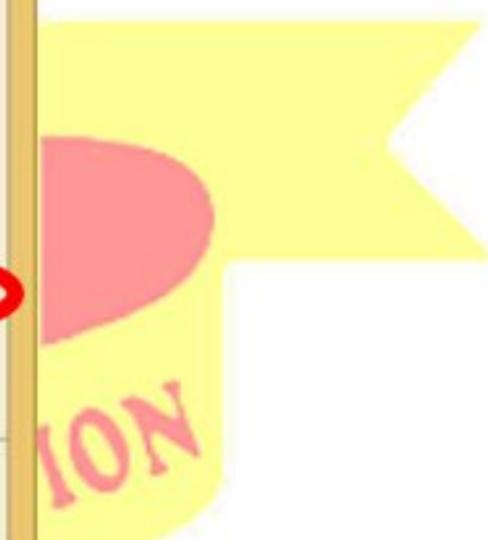
> < >> <<

Filter:

Show system objects

Add Related Tables

< Back Next > Finish >> Cancel



Click Next



निर्मल नेह उत्तम सेवाधर्म



Data Source View Wizard

**Select Tables and Views**  
Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type

Included objects:

Name	Type
FactProductSales (dbo)	Table
DimStores (dbo)	Table
DimProduct (dbo)	Table
DimTime (dbo)	Table
DimDate (dbo)	Table
DimCustomer (dbo)	Table
DimSalesPerson (dbo)	Table

> < >> <<

Filter:

Show system objects

Add Related Tables

< Back Next > Finish >> Cancel

Click Finish

Data Source View Wizard

**Completing the Wizard**  
Provide a name, and then click Finish to create the new data source view.

Name:

Preview:

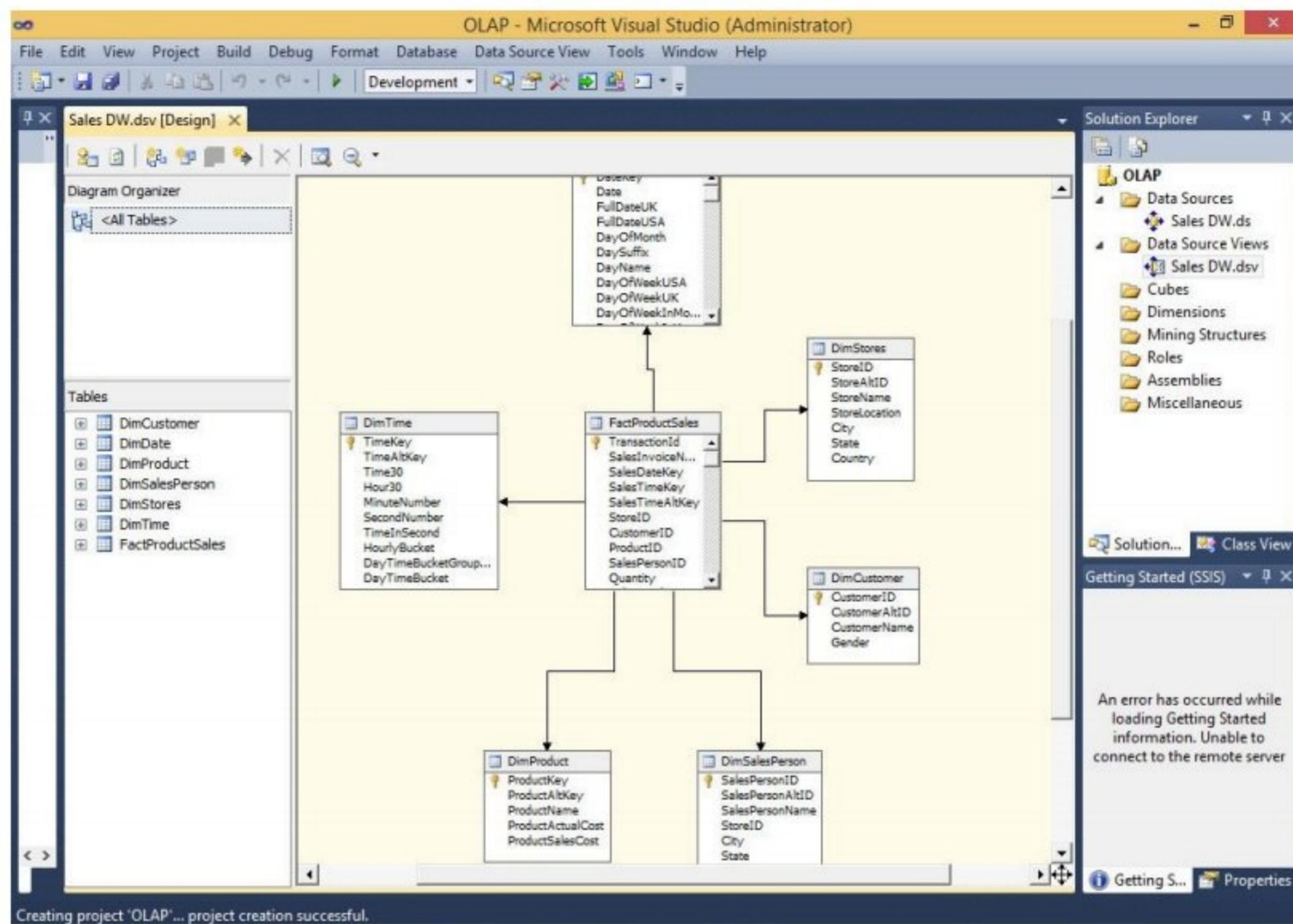
- Sales DW
  - FactProductSales (dbo)
  - DimStores (dbo)
  - DimProduct (dbo)
  - DimTime (dbo)
  - DimDate (dbo)
  - DimCustomer (dbo)
  - DimSalesPerson (dbo)

< Back Next > Finish Cancel

Completed the wizard



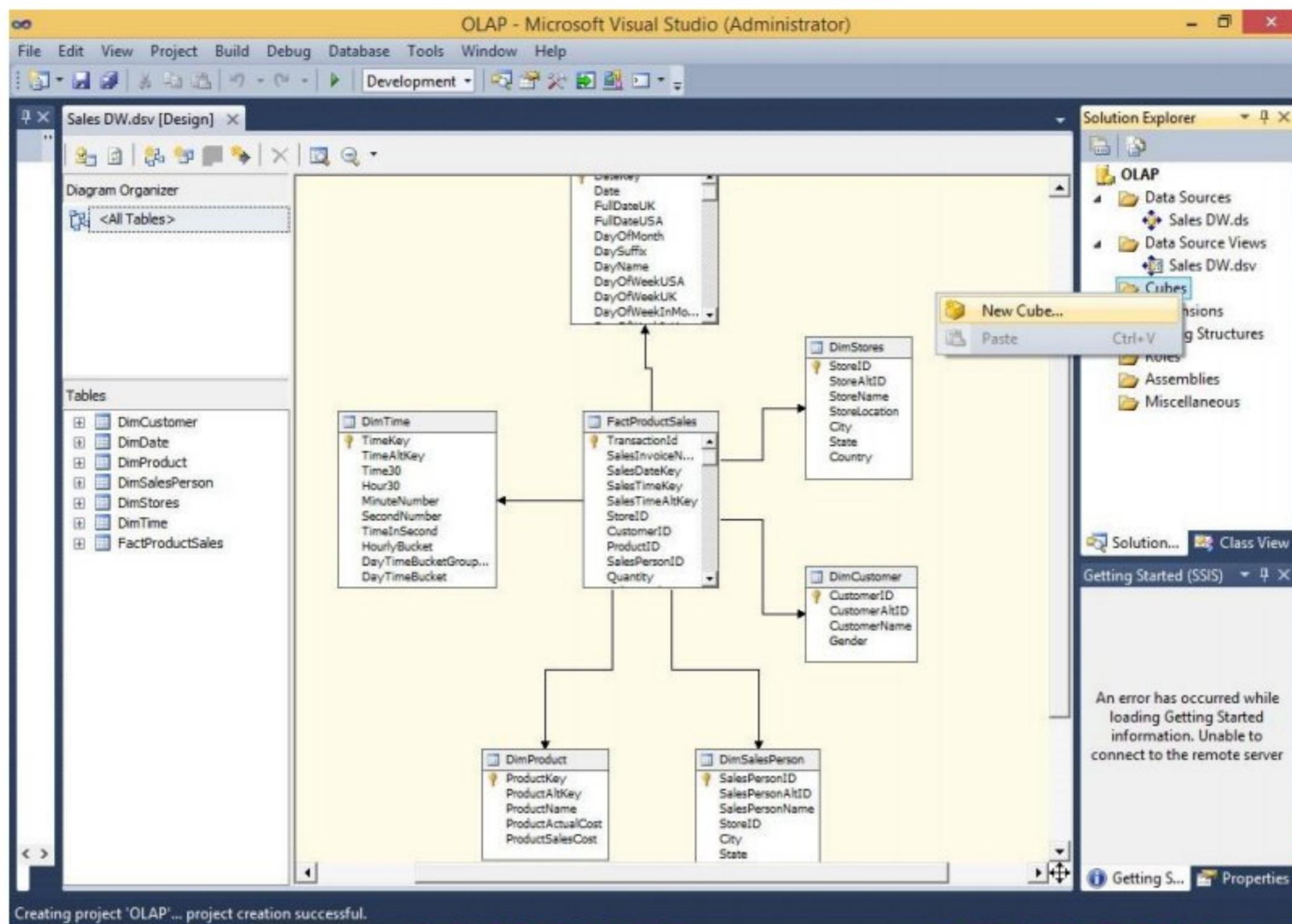
Sales DW.dsv appears in Data Source Views in Solution Explorer.



Step 4: Creating new cube

Right click on Cubes → New Cube

निर्माण नोहे उत्तम सेवाधर्म



Cube Wizard

## Welcome to the Cube Wizard

Use this wizard to create a new cube. First, you select the data source view and tables for the cube, and then you set its properties. You can also opt to create a cube without using a data source.

Don't show this page again

< Back Next > Finish >> Cancel

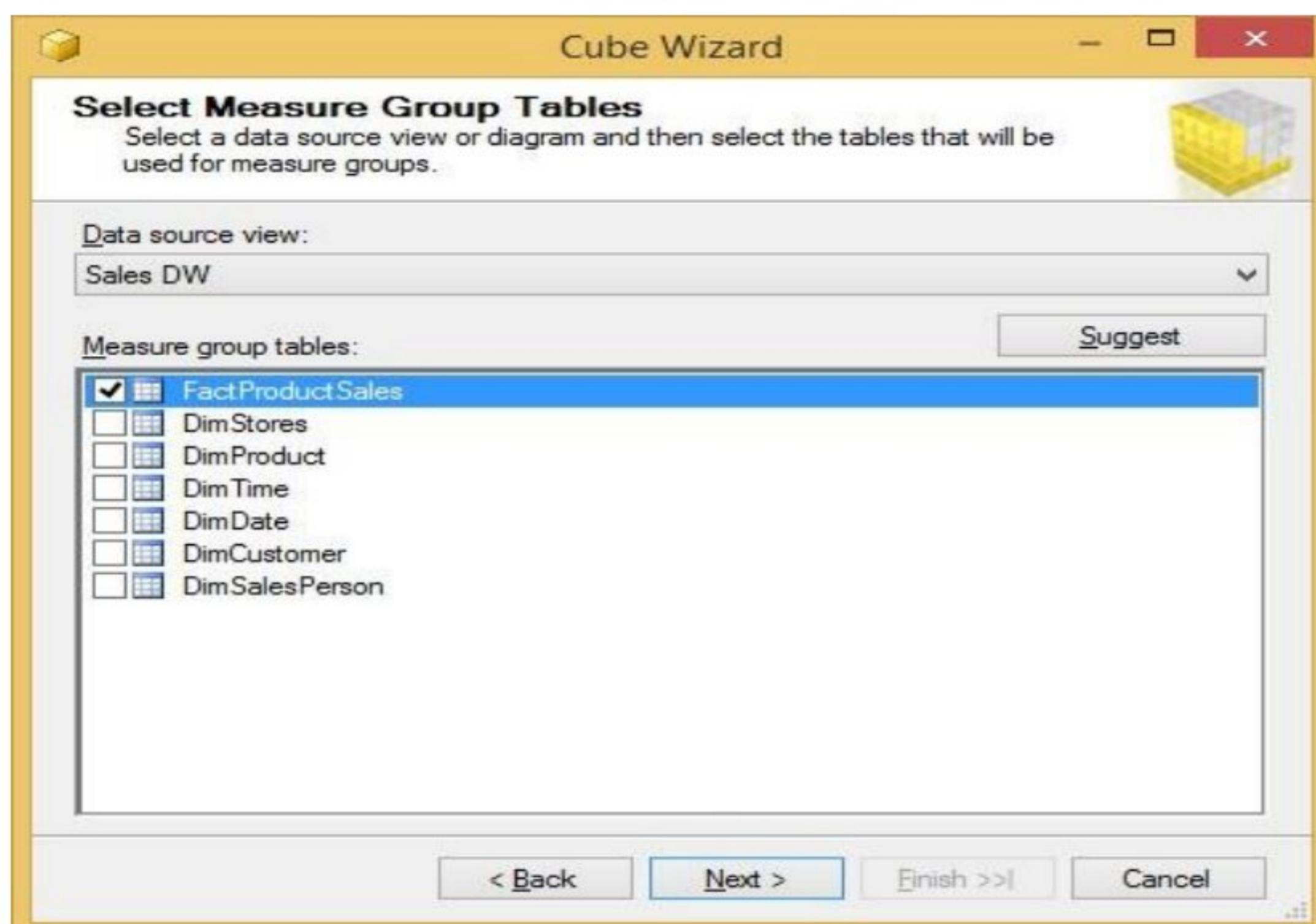
Select Use existing tables in Select Creation Method → Next



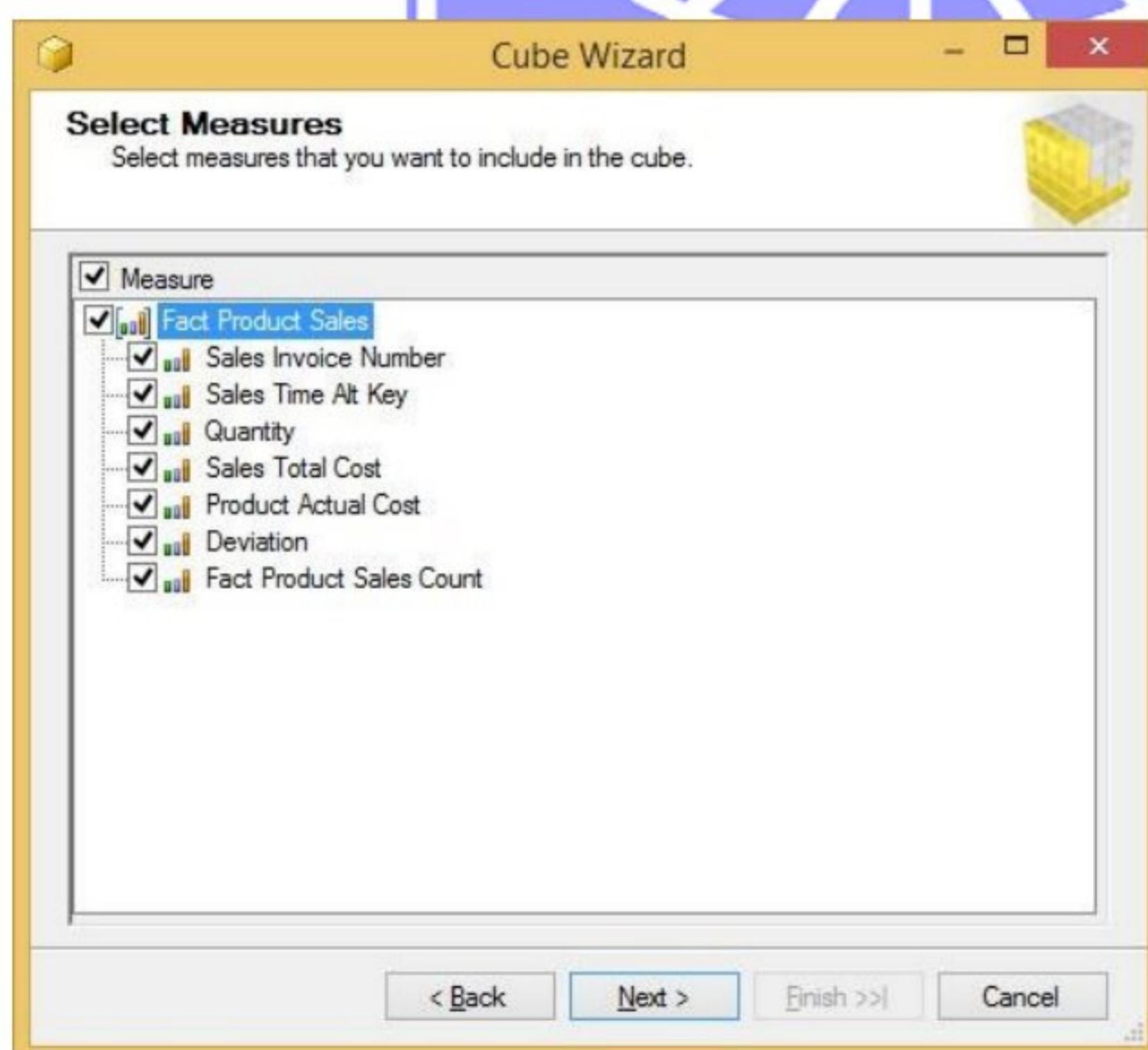
In Select Measure Group Tables → Select FactProductSales → Click Next



निर्माण नोहे उत्तम सेवाधर्म

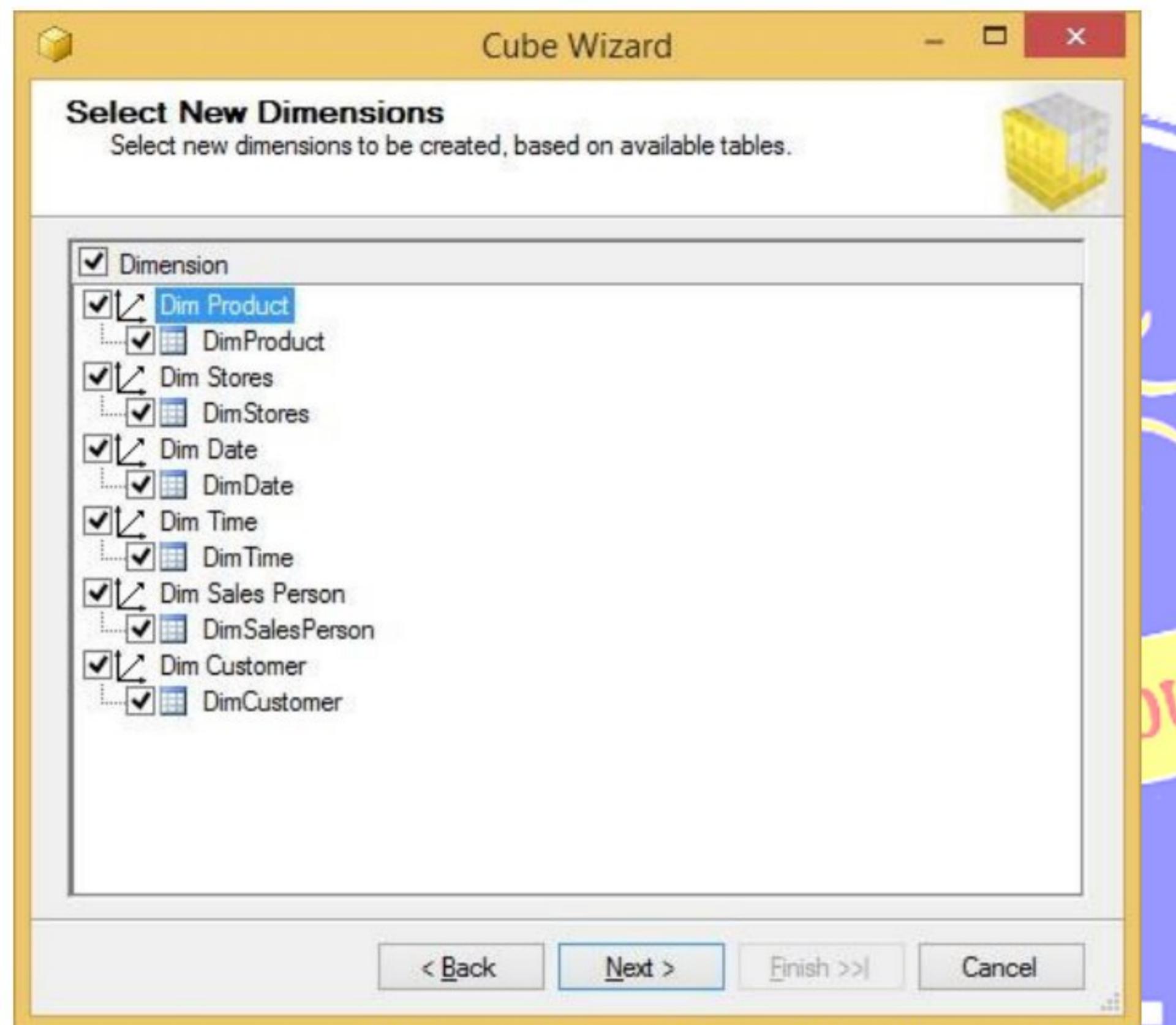


In Select Measures → check all measures → Next





In Select New Dimensions → Check all Dimensions → Next

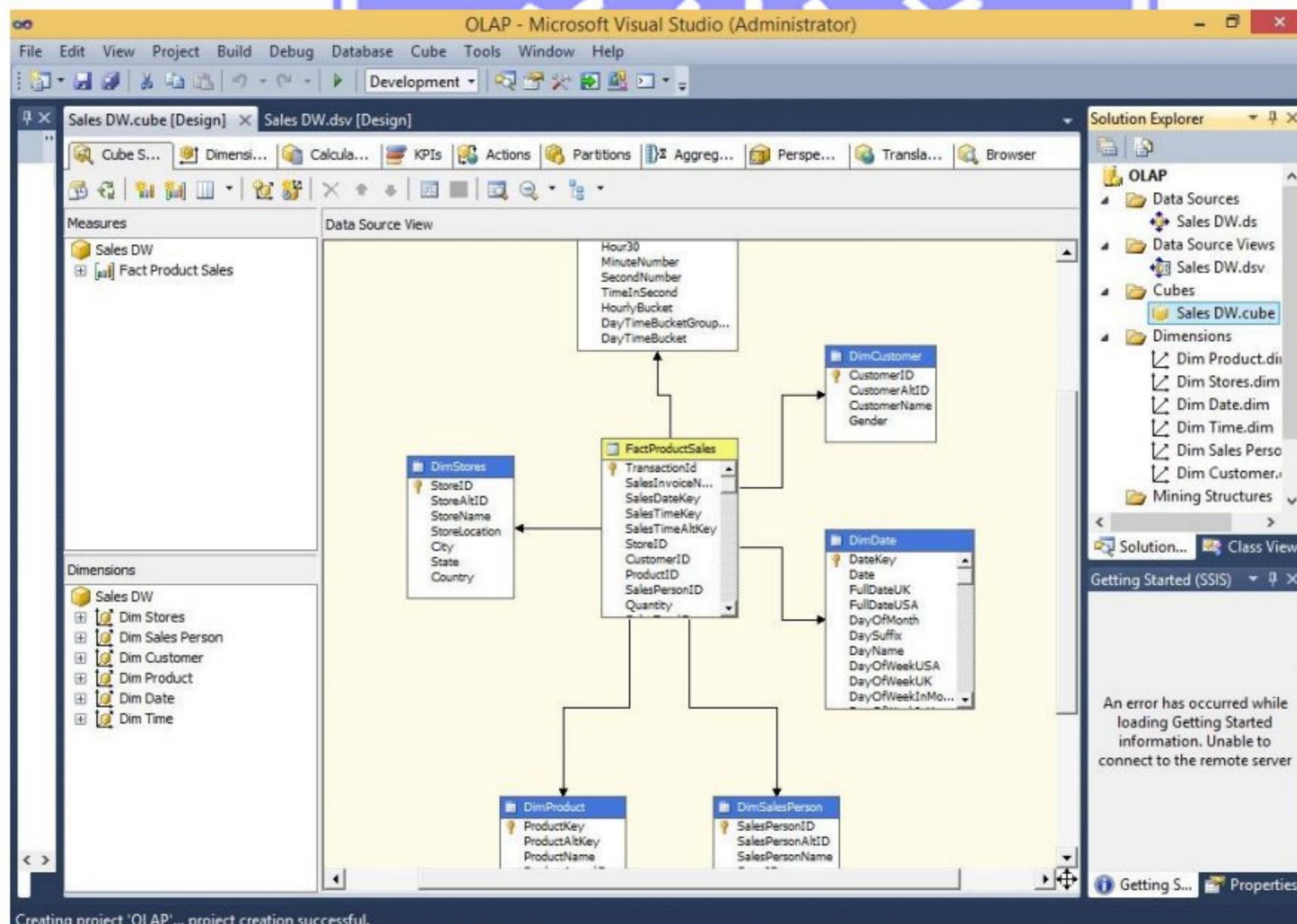


Click on Finish

निर्मलेन्ह उत्तम सेवाधर्म



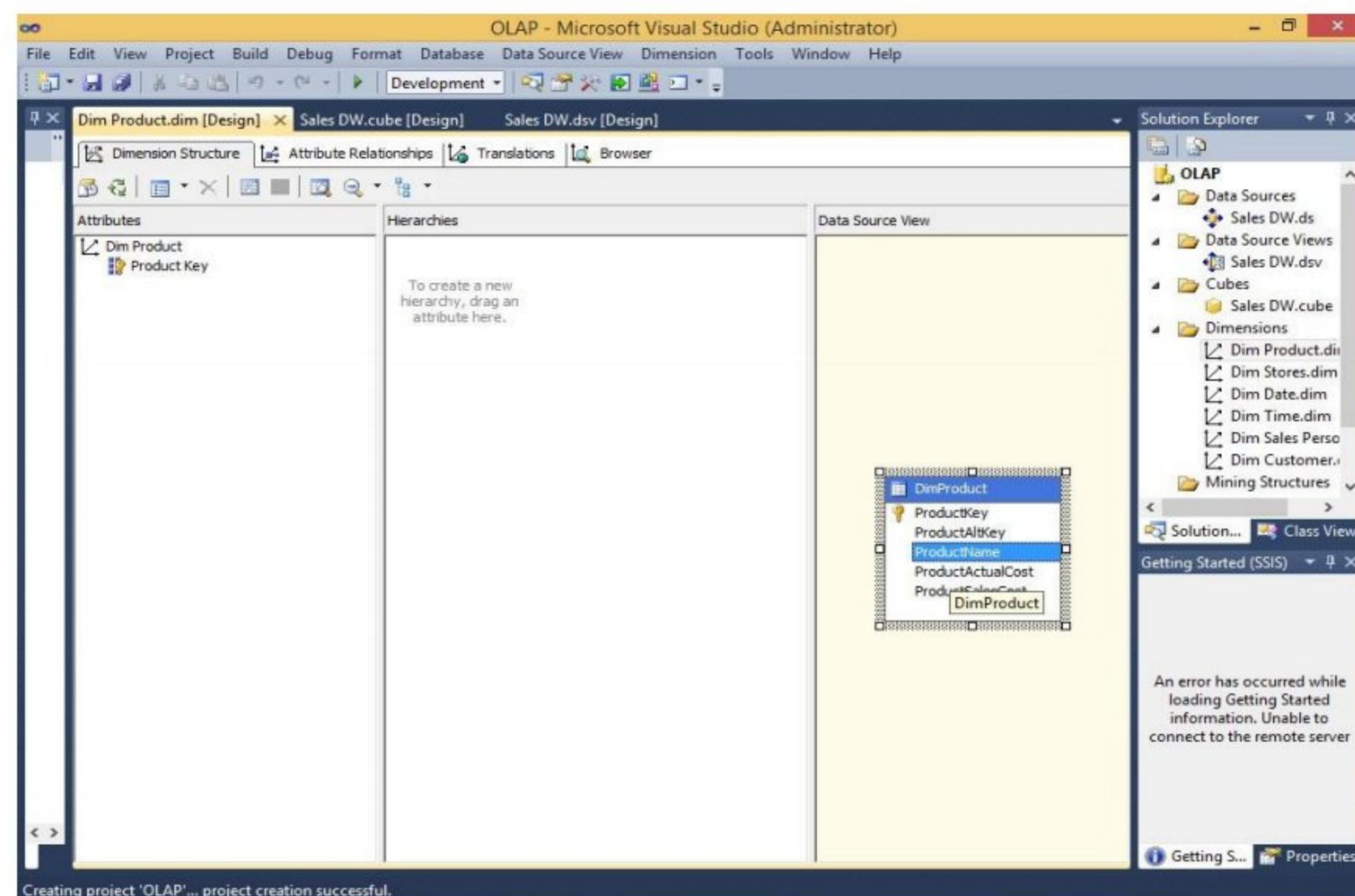
Sales\_DW.cube is created





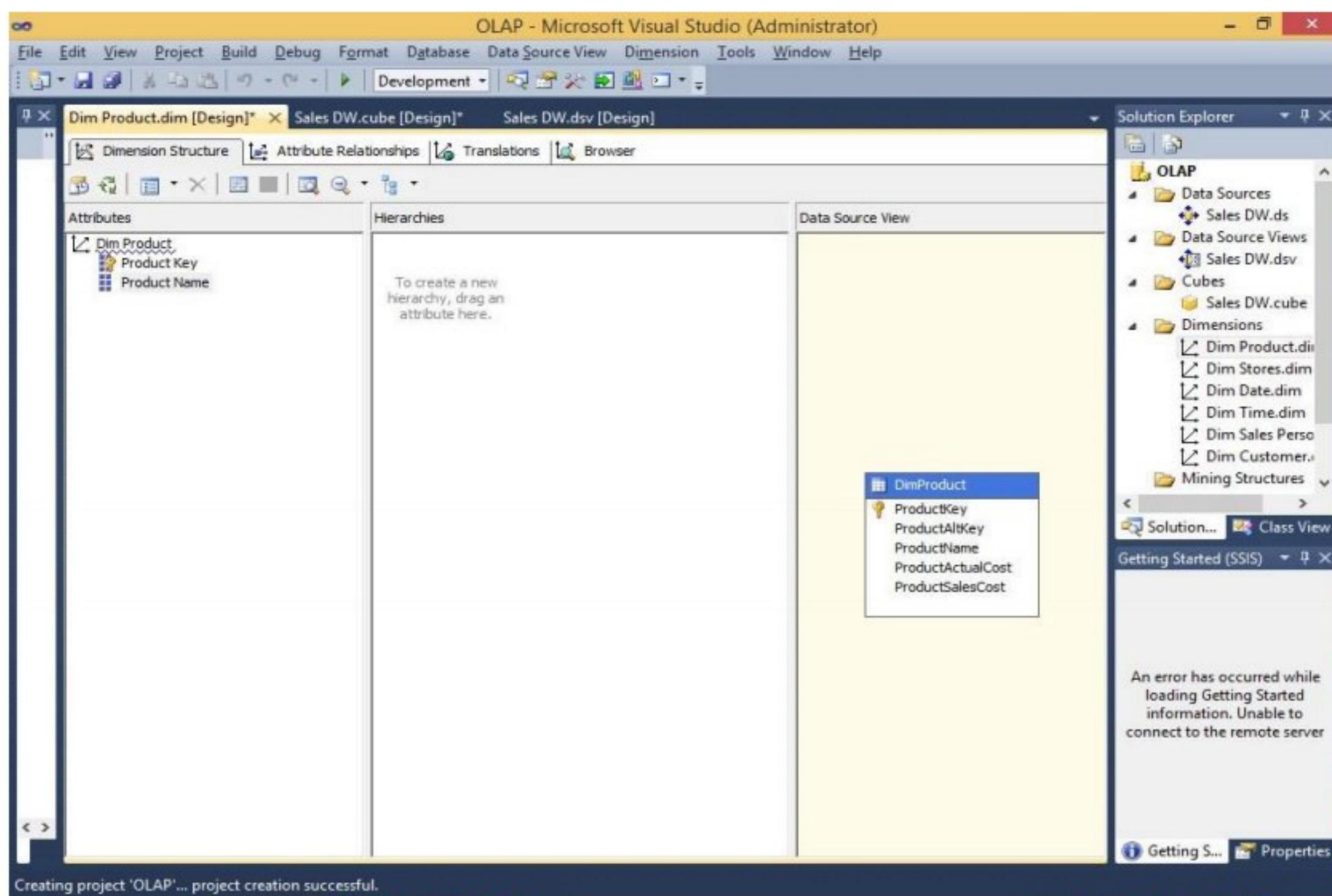
## Step 5: Dimension Modification

In dimension tab → Double Click Dim Product.dim



Drag and Drop Product Name from Table in Data Source View and Add in Attribute Pane at left side

निर्मलेन्ह उत्तम सेवाधर्म

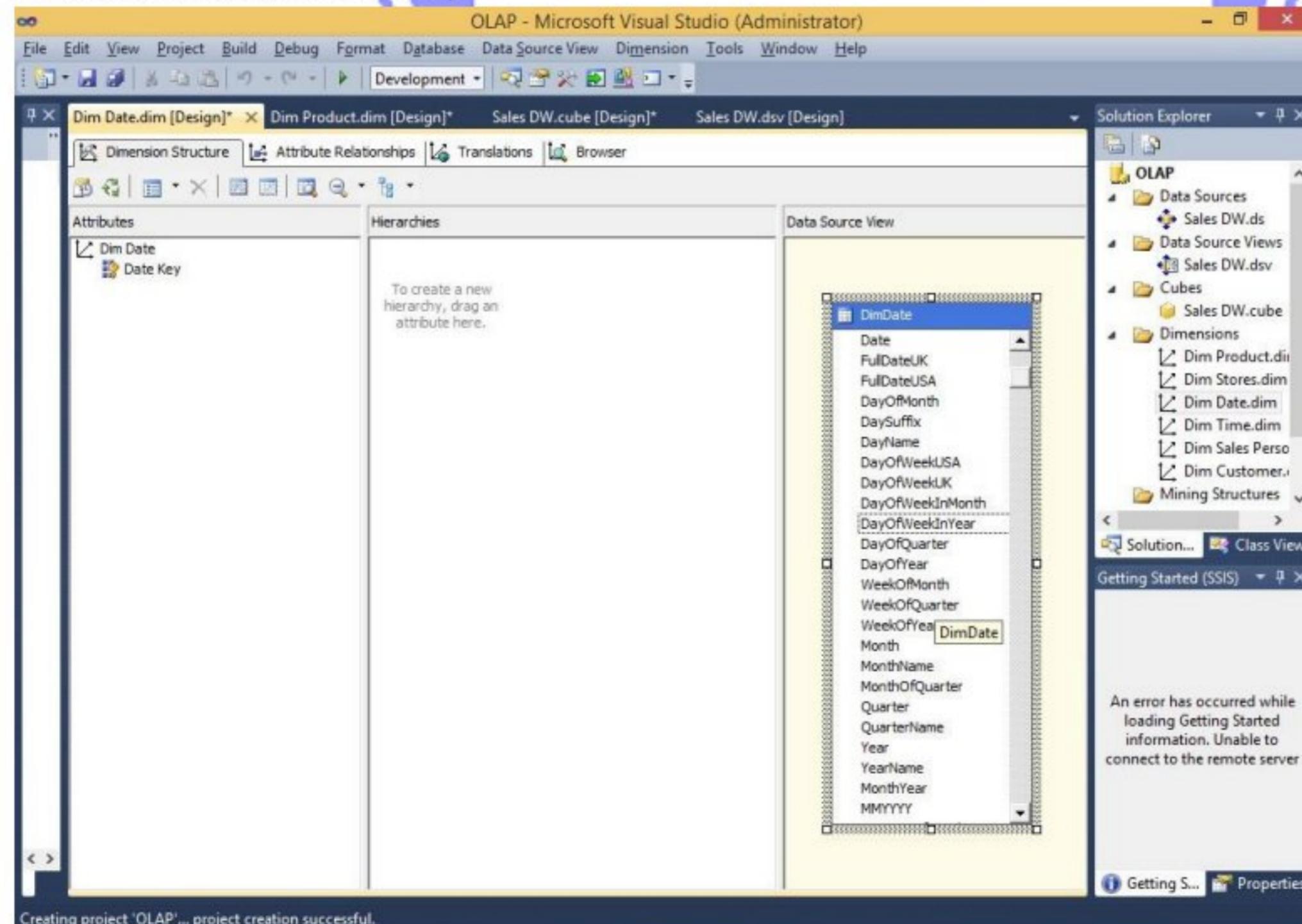


Creating project 'OLAP'... project creation successful.

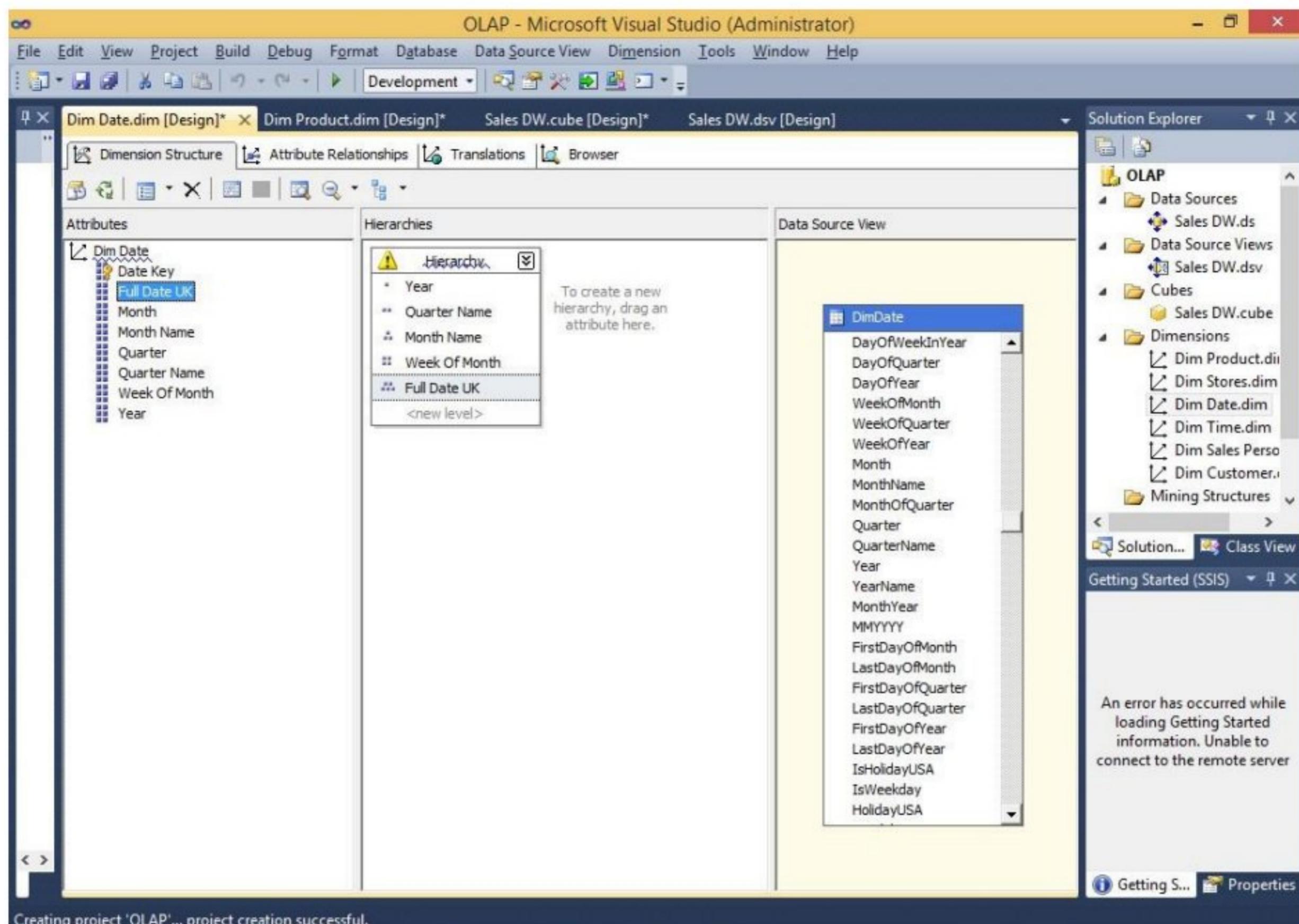
## WELCOME TO EXCELLENCE IN EDUCATION

Step 6: 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)



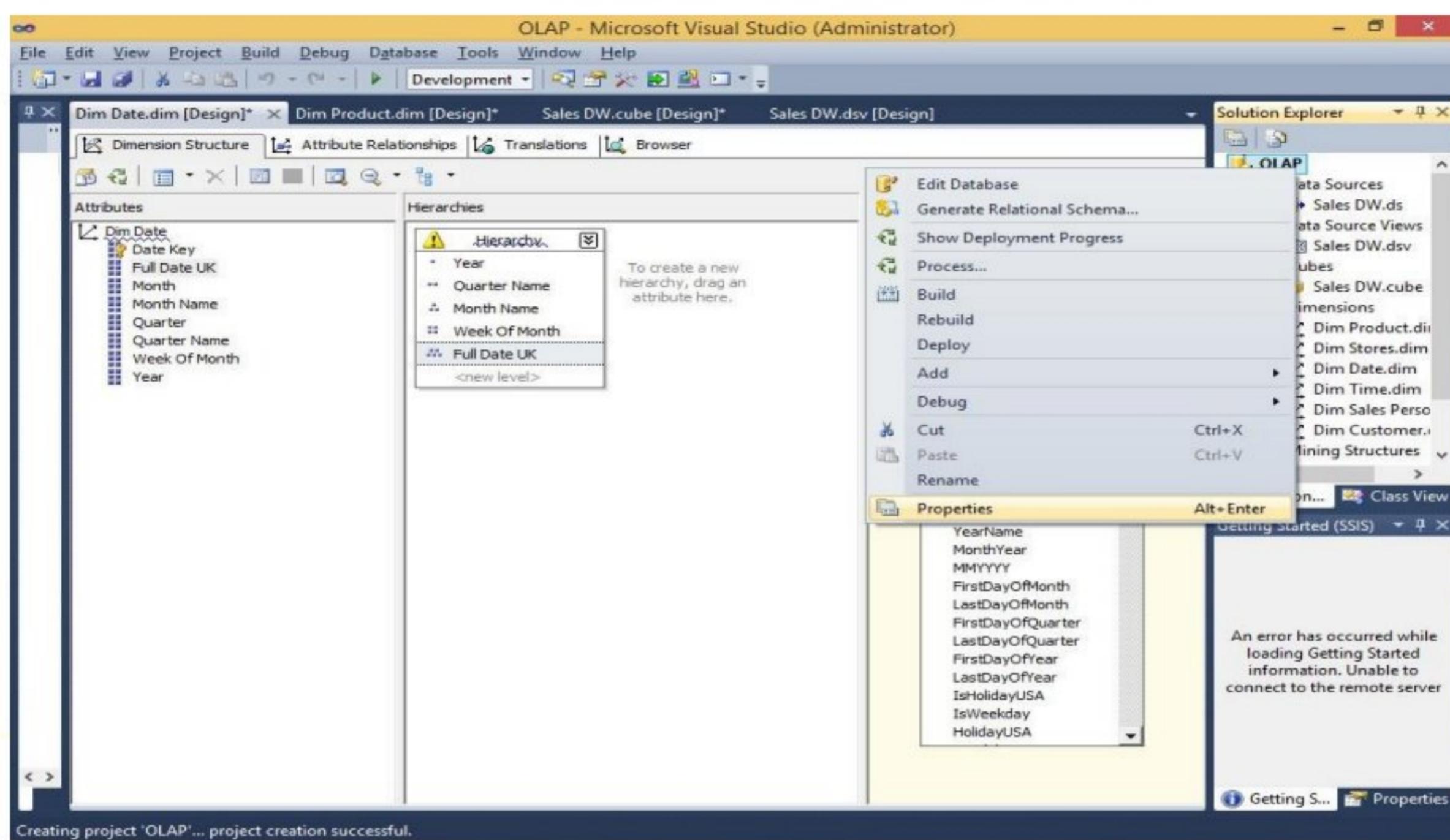
Creating project 'OLAP'... project creation successful.



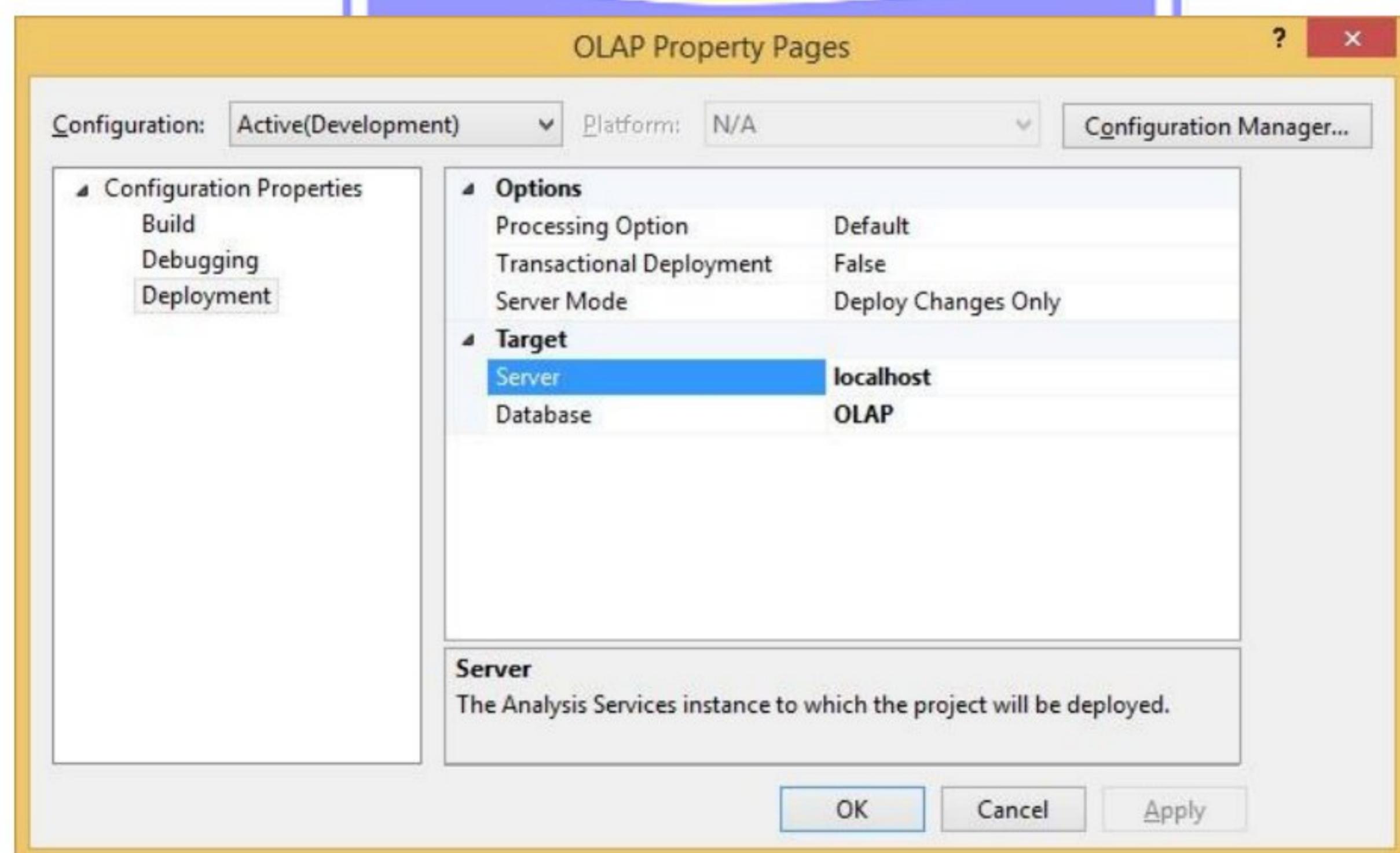
Step 7: Deploy Cube

Right click on Project name → Properties

निर्माण नोहे उत्तम सेवाधर्म



This window appears



Do following changes and click on Apply & ok



OLAP Property Pages

Configuration: Active(Development) Platform: N/A Configuration Manager...

▲ Configuration Properties  
Build  
Debugging  
Deployment

▲ Options  
Processing Option: Do Not Process  
Transactional Deployment: False  
Server Mode: Deploy All

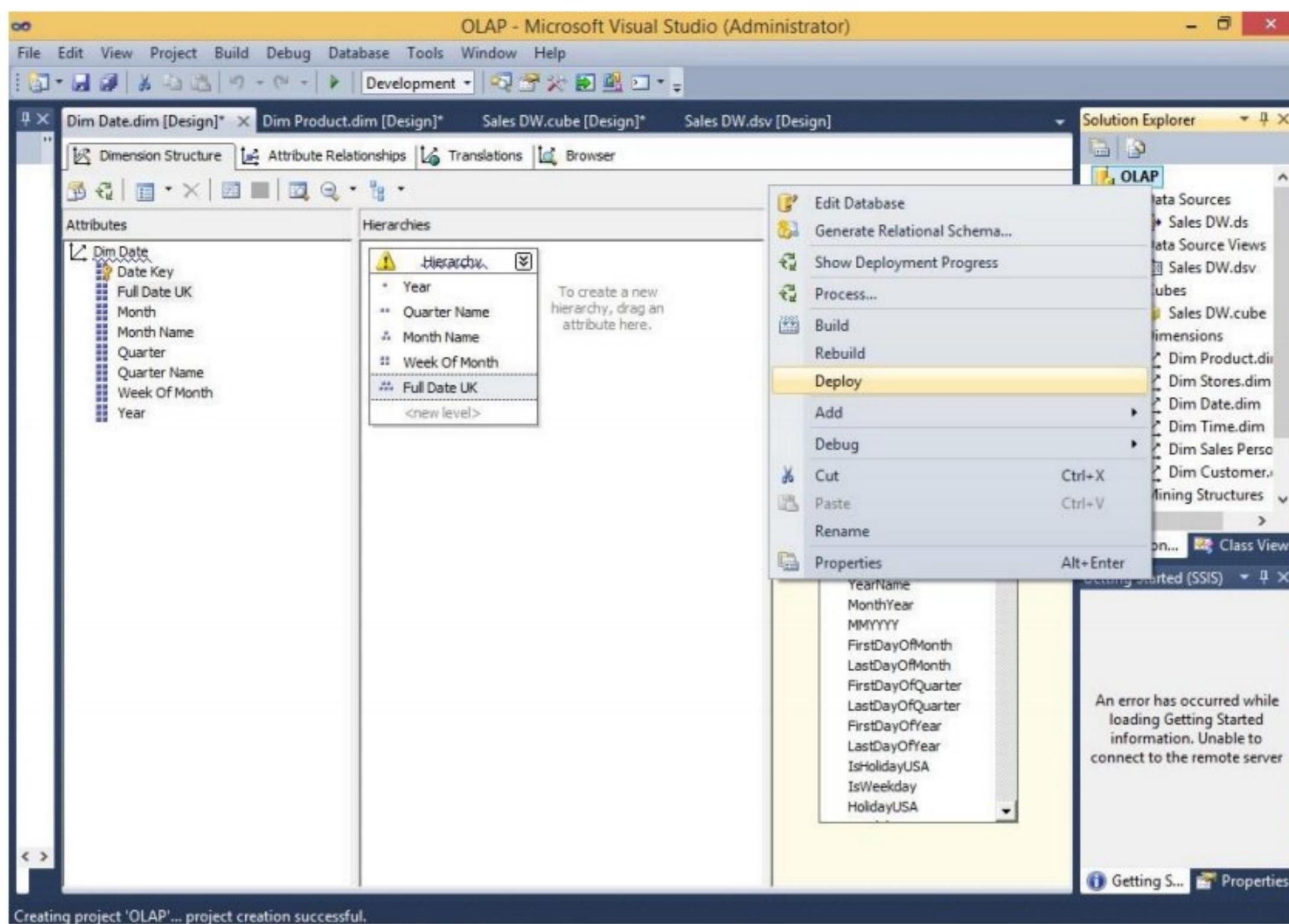
▲ Target  
Server: localhost  
Database: OLAP

**Server Mode**  
Specifies whether only changed objects or all objects should be deployed.

OK Cancel Apply

Right click on project name → Deploy

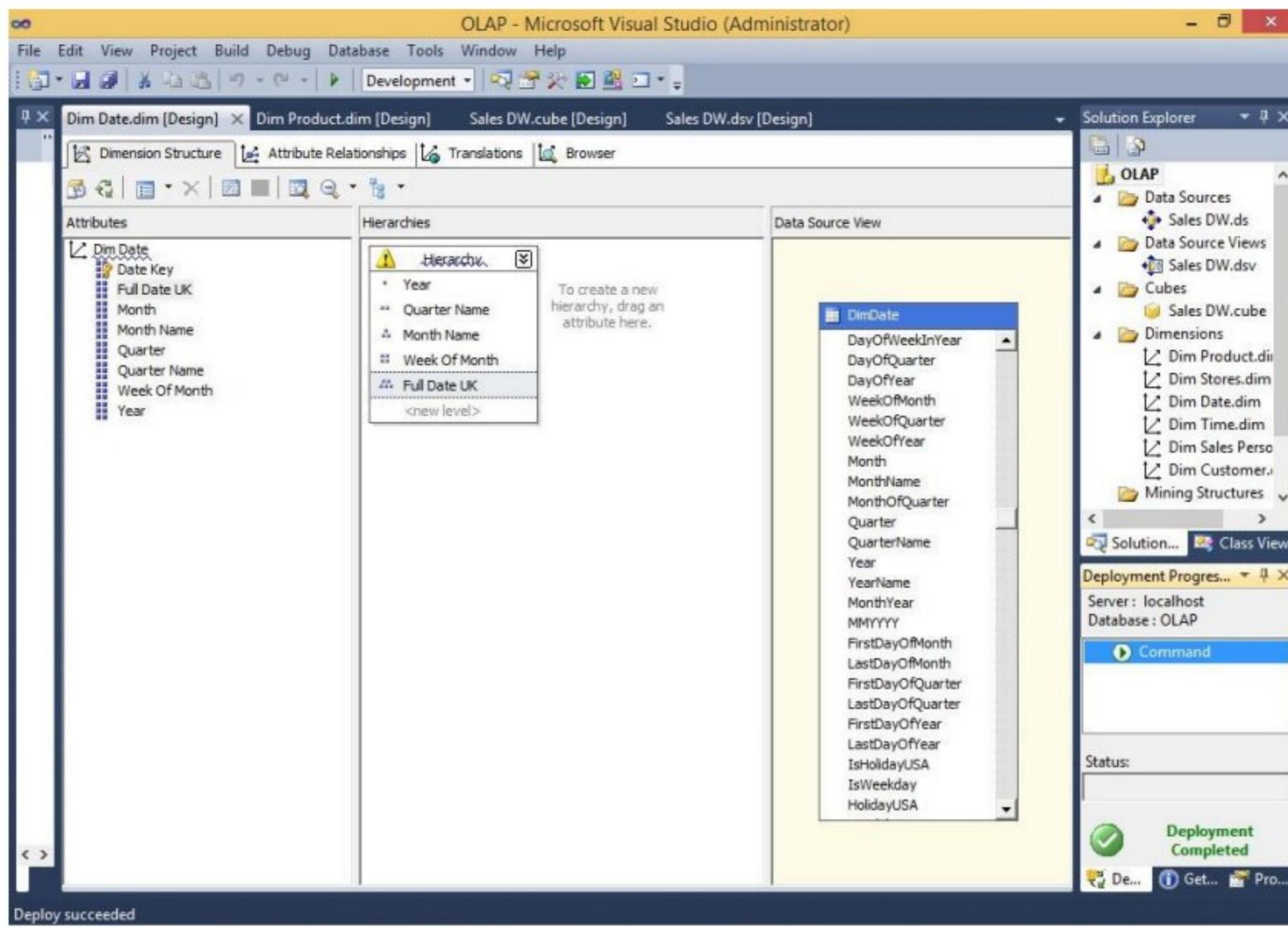
निर्माण नोहे उत्तम सेवाधर्म



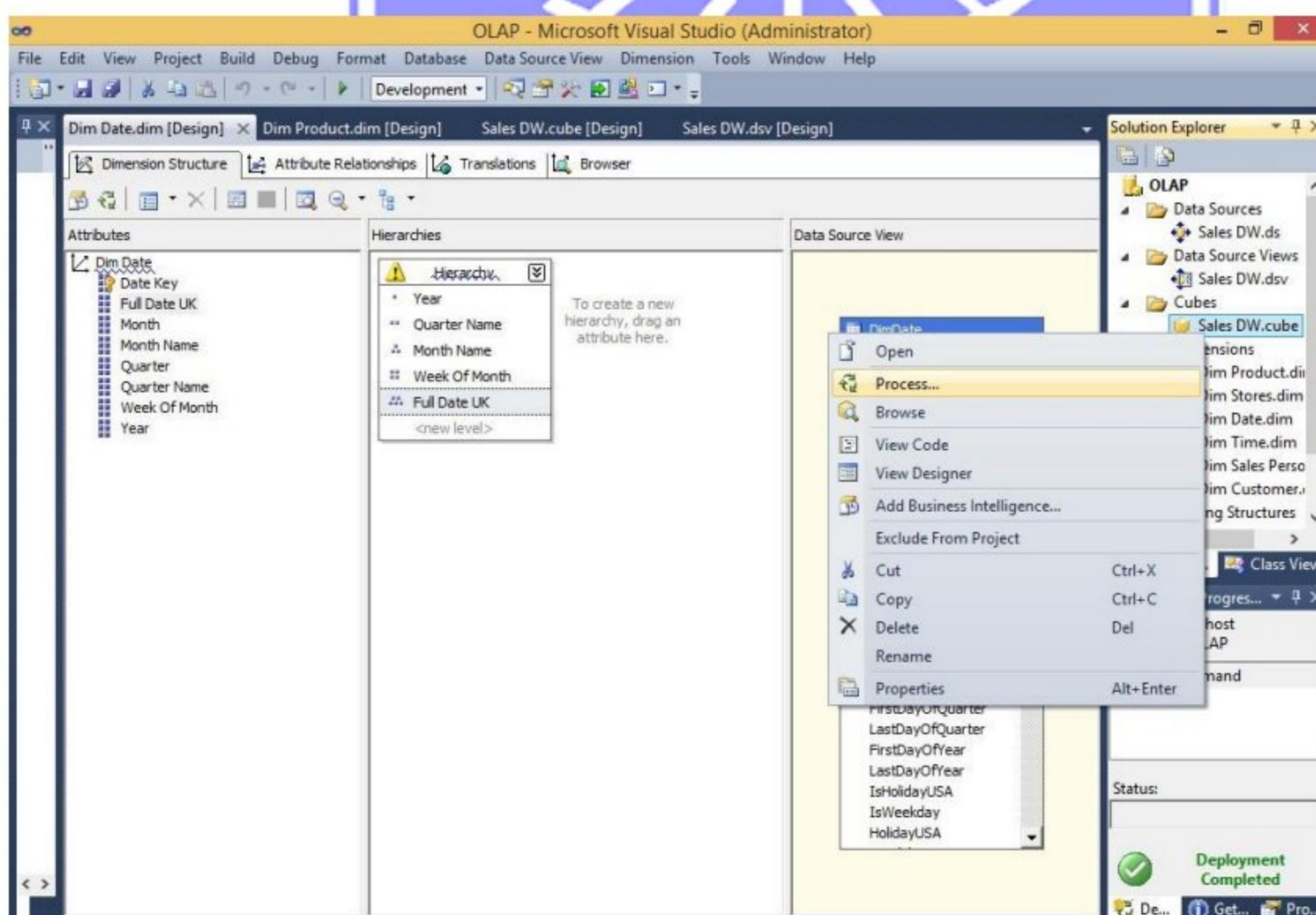
Deployment successful



निर्मलेन्ह उत्तम सेवाधर्म

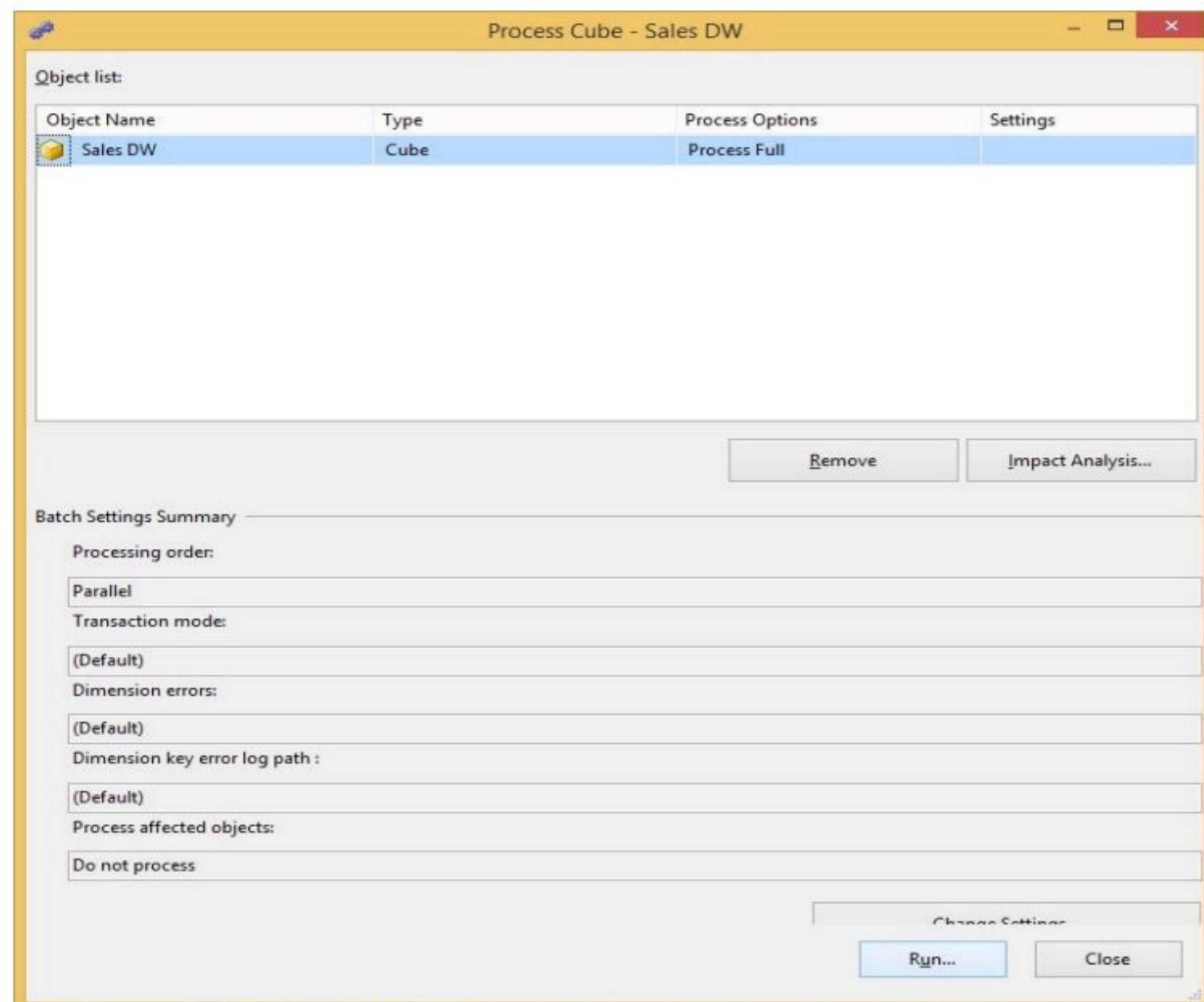


To process cube right click on Sales\_DW(cube) → Process

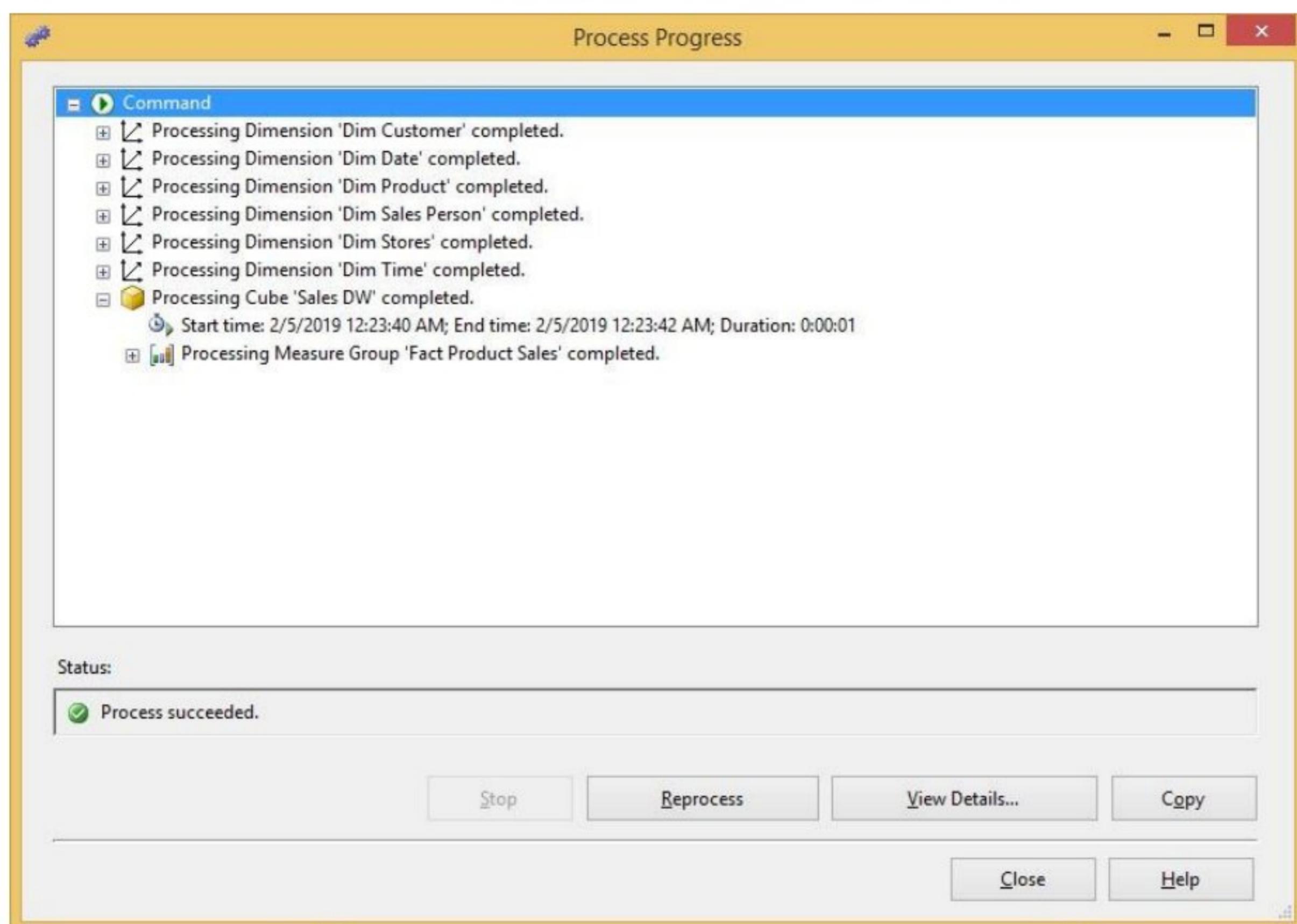




Click run



निर्मलेन्ह उत्तम सेवाधर्म



Browse the cube for analysis in solution explorer



निर्मलेन्ह उत्तम सेवाधर्म



OLAP - Microsoft Visual Studio (Administrator)

File Edit View Project Build Debug Database Cube Tools Window Help

Solution Explorer

OLAP

- Data Sources
- Data Source Views
- Sales DW.ds
- Sales DW.dsv
- Cubes
- Sales DW(cube)

Dim Date.dim [Design] Dim Product.dim [Design] Sales DW(cube) [Design] Sales DW.dsv [Design]

Attributes

Hierarchies

To create a new hierarchy, drag an attribute here.

Year

Quarter Name

Month Name

Week Of Month

Full Date UK

<new level>

Data Source View

Open

Process...

Browse

View Code

View Designer

Add Business Intelligence...

Exclude From Project

Cut Ctrl+X

Copy Ctrl+C

Delete Del

Rename

Properties Alt+Enter

FirstDayOfQuarter

LastDayOfQuarter

FirstDayOfYear

LastDayOfYear

IsHolidayUSA

IsWeekday

HolidayUSA

Status:

Deployment Completed

Deploy Get... Properties

Deploy succeeded

OLAP - Microsoft Visual Studio (Administrator)

File Edit View Project Build Debug Database Cube Tools Window Help

Solution Explorer

OLAP

- Data Sources
- Data Source Views
- Sales DW.ds
- Sales DW.dsv
- Cubes
- Sales DW(cube)
- Dimensions
- Dim Product.dim
- Dim Stores.dim
- Dim Date.dim
- Dim Time.dim
- Dim Sales Person
- Dim Customer
- Mining Structures
- Roles
- Assemblies

Cube S... Dimensions Calculations KPIs Actions Partitions Aggregates Perspectives Translations Browser

Dim Date.dim [Design] Dim Product.dim [Design] Sales DW(cube) [Design] Sales DW.dsv [Design]

Language: Default

Edit as Text Import...

Sales DW

Measure Group: <All>

Sales DW

- Measures
- Fact Product Sales
  - Deviation
  - Fact Product Sales Count
  - Product Actual Cost
  - Quantity
  - Sales Invoice Number
  - Sales Time Alt Key
  - Sales Total Cost
- KPIs
- Dim Customer
- Dim Date
  - Date Key
  - Full Date UK
  - Month
  - Month Name
  - Quarter
  - Quarter Name

Calculated Members

Drag levels or measures here to add to the query.

Status:

Deployment Completed

Deploy Get... Properties

Deploy succeeded

**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

1. Define staging.
2. Why is staging important?
3. Name one staging component.
4. How to validate data in staging?

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



### **Practical-4: a. Create the ETL map and setup the schedule for execution.**

**– b. Execute the MDX queries to extract the data from data ware house.**

#### **Aims:**

1. To design an ETL map outlining the flow of data from source systems to the data warehouse.
2. To set up and schedule the ETL process for automated execution.
3. To execute MDX queries for extracting and analyzing data from the data warehouse.

#### **Learning Objectives:**

1. Understand the principles of ETL mapping and how it drives data integration.
2. Gain hands-on experience in scheduling ETL jobs to ensure timely data processing.
3. Learn to write and execute MDX queries to extract multidimensional data for reporting and analysis.
4. Evaluate the effectiveness and performance of the ETL process and query results.

#### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

#### **Theory-4: ETL**

ETL (Extract, Transform, Load) is a core process in data warehousing that integrates data from multiple sources into a centralized repository. This theory focuses on two key aspects: designing an ETL map with an execution schedule and using MDX queries to extract data from the data warehouse.

#### **Creating the ETL Map and Scheduling Execution**

##### **1. Designing the ETL Map**

- **Source Identification:** Identify all relevant data sources (databases, files, APIs) and understand their formats and structures.
- **Mapping Data Flow:** Create a visual map that outlines the flow of data from source systems to the staging area, detailing transformation steps and the final load into the data warehouse.
- **Defining Transformation Rules:** Document the business rules needed for data cleansing, conversion, aggregation, and enrichment.
- **Error Handling:** Establish procedures for logging errors, issuing alerts, and rolling back transactions if issues occur during the ETL process.

##### **2. Scheduling the ETL Process**

- **Determining Frequency:** Decide whether the ETL jobs should run in batch mode (e.g., nightly, hourly) or in real time, based on data freshness requirements and system load.
- **Automation:** Utilize ETL automation tools (such as SQL Server Integration Services, Informatica, or Apache NiFi) to schedule and execute jobs reliably.

- **Monitoring and Alerts:** Set up monitoring dashboards and notification systems to track job performance and address any failures promptly.

## Executing MDX Queries to Extract Data from the Data Warehouse

### 1. Understanding MDX:

- MDX (Multidimensional Expressions) is a query language designed for OLAP (Online Analytical Processing) systems. It enables the retrieval and analysis of multidimensional data stored in cubes.
- MDX queries help extract aggregated, detailed, or time-series data across various dimensions like product, geography, and time.

### 2. Running MDX Queries:

- **Connection:** Establish a secure connection to the OLAP data warehouse.
- **Query Construction:** Build MDX queries to retrieve relevant data. For example, a query to extract total sales by product category for the year 2024 might look like this:
- **Result Analysis:** Analyze the query results to generate actionable insights, support reporting, and enable dynamic data visualizations.



### PRACTICAL 4 b

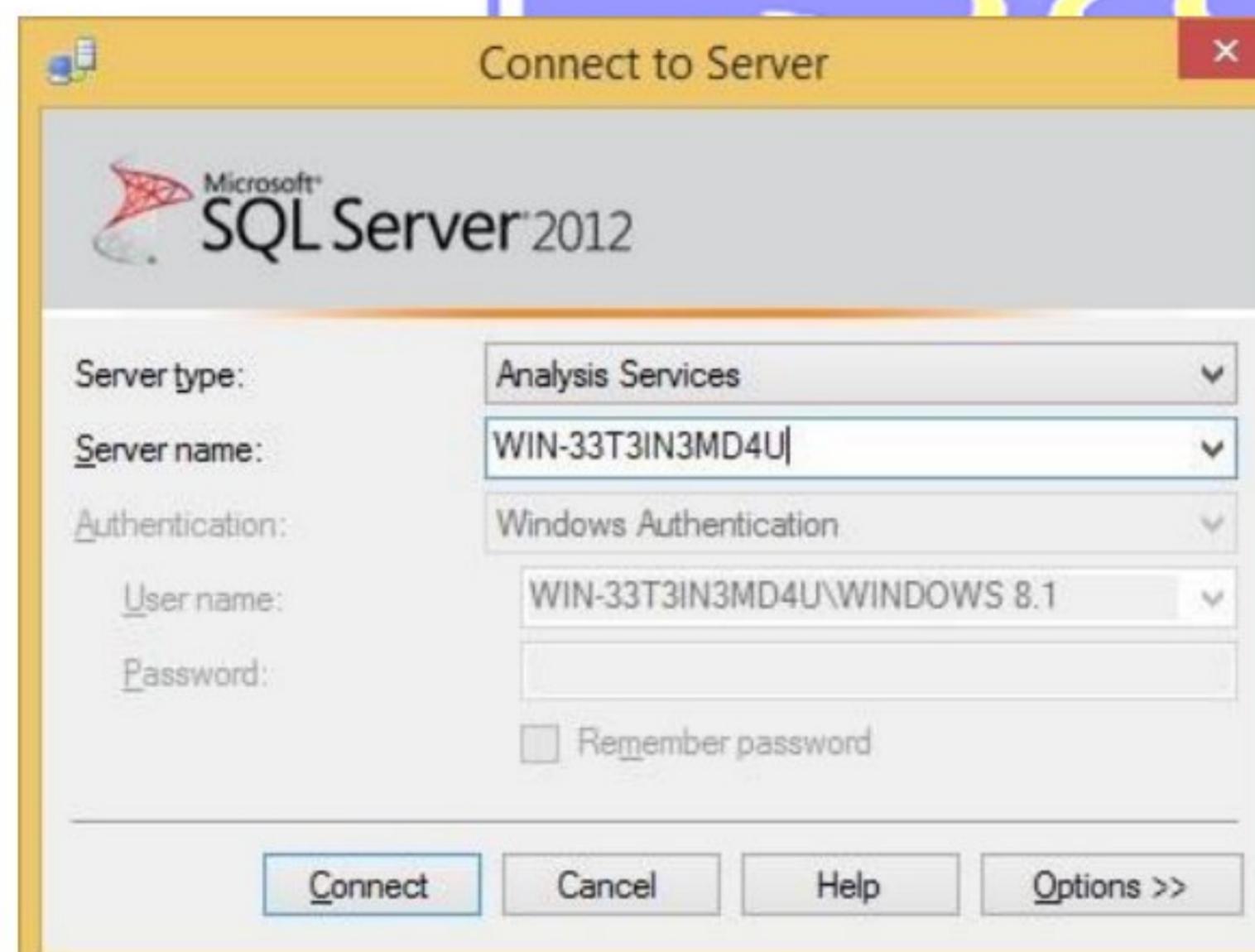
Execute the MDX queries to extract the data from the datawarehouse.

Step 1: Open SQL Server Management Studio and connect to Analysis Services.

Server type: Analysis Services

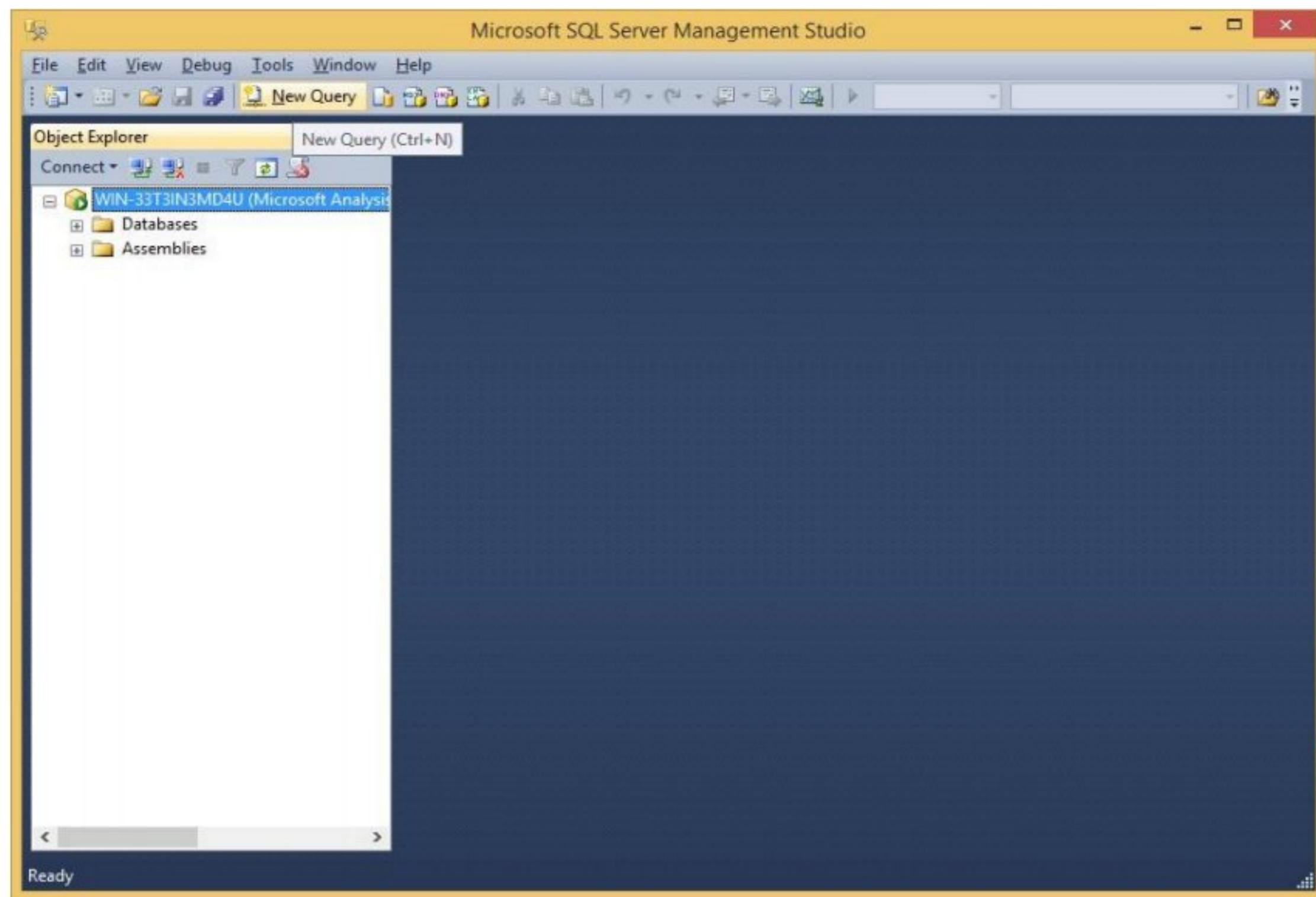
Server Name: (according to base machine)

Click on connect



Step 2: Click on New Query & type following query based on Sales\_DW

निर्माण नोह उत्तम सेवाधर्म



select [Measures].[Sales Time Alt Key] on columns

from [Sales DW] Click on  
execute

निर्मलेन्ह उत्तम सेवाधर्म



MDXQuery1.mdx - WIN-33T3IN3MD4U.OLAP (WIN-33T3IN3MD4U\WINDOWS 8.1)\* - Microsoft SQL Server Management Studio

File Edit View Query Project Debug Tools Window Help

OLAP

MDXQuery1.mdx - D4U\WINDOWS 8.1)\*

Cube: Sales DW

Measure Group: <All>

Sales DW

Fact Product Sales

Deviation

Fact Product Sales Count

Product Actual Cost

Quantity

Sales Invoice Number

Sales Time Alt Key

Sales Total Cost

KPIs

Dim Customer

Dim Date

Dim Product

Dim Sales Person

Dim Stores

Dim Time

select [Measures].[Sales Time Alt Key] on columns  
[from [Sales DW]]

Messages Results

Sales Time Alt Key  
3631639

Query executed successfully.

Ready

LN 1 Col 8 Ch 8 INS

select [Measures].[Quantity] on columns from [Sales DW]

निर्मलेन्ह उत्तम सेवाधर्म



MDXQuery1.mdx - WIN-33T3IN3MD4U.OLAP (WIN-33T3IN3MD4U\WINDOWS 8.1)\* - Microsoft SQL Server Management Studio

File Edit View Query Project Debug Tools Window Help

OLAP Execute

Object Expl... MDXQuery1.mdx - D4U\WINDOWS 8.1)\*

Cube: Sales DW

Measure Group: <All>

Sales DW

Measures

Fact Product Sales

Deviation

Fact Product Sales Count

Product Actual Cost

Quantity

Sales Invoice Number

Sales Time Alt Key

Sales Total Cost

KPIs

Dim Customer

Dim Date

Dim Product

Dim Sales Person

Dim Stores

Dim Time

select [Measures].[Quantity] on columns  
from [Sales DW]

Messages Results

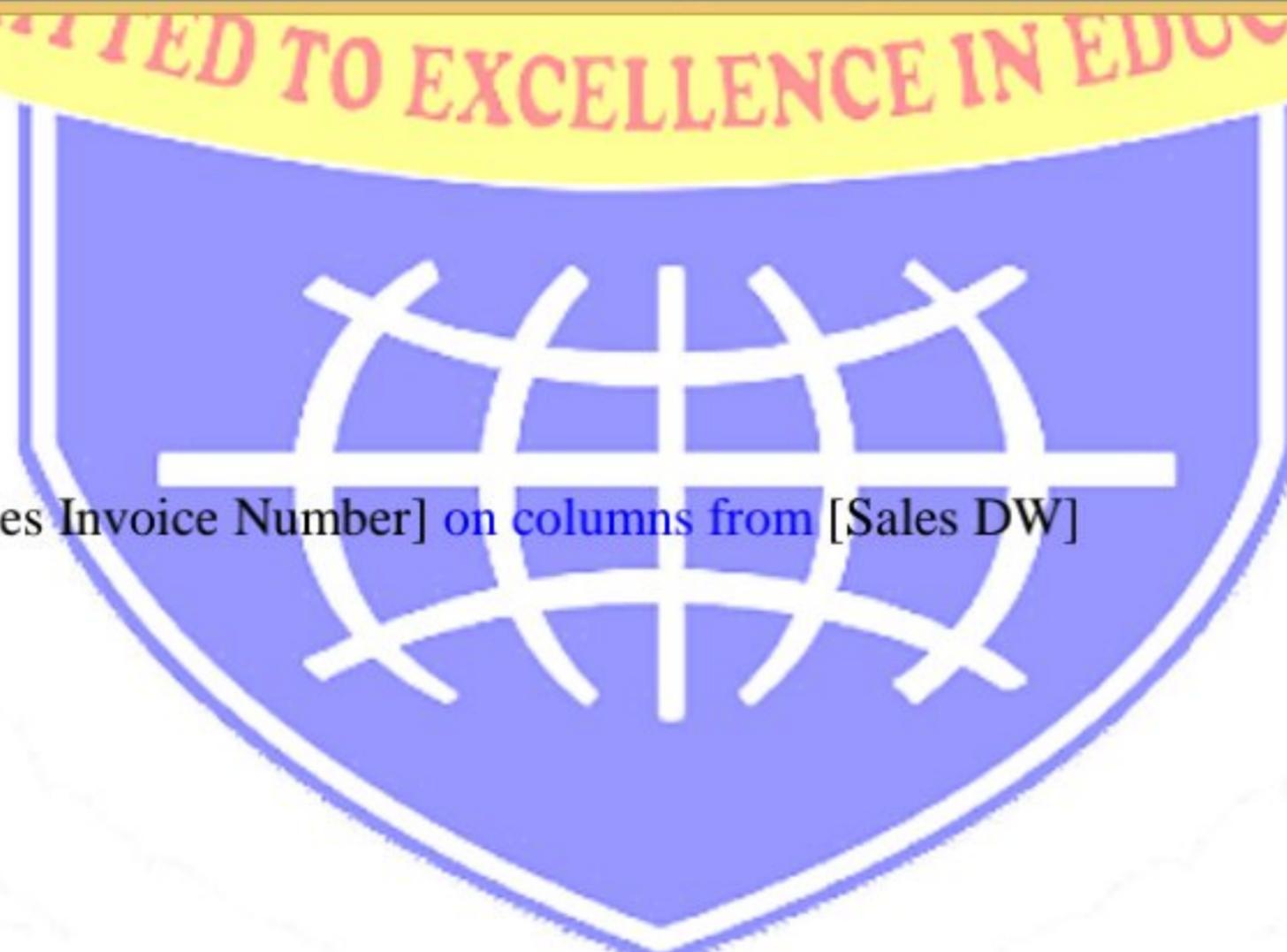
Quantity 43

Query executed successfully.

WIN-33T3IN3MD4U | WIN-33T3IN3MD4U\WINDOW... | OLAP | 00:00:01

Ready

Ln 1 Col 8 Ch 8 INS



select [Measures].[Sales Invoice Number] on columns from [Sales DW]

निर्मलेन्ह उत्तम सेवाधर्म



MDXQuery1.mdx - WIN-33T3IN3MD4U.OLAP (WIN-33T3IN3MD4U\WINDOWS 8.1)\* - Microsoft SQL Server Management Studio

File Edit View Query Project Debug Tools Window Help

OLAP

New Query Execute

MDXQuery1.mdx - ...D4U\WINDOWS 8.1)\*

Cube: Sales DW

Measure Group: <All>

Sales DW

Measures

- Fact Product Sales
  - Deviation
  - Fact Product Sales Count
  - Product Actual Cost
  - Quantity
  - Sales Invoice Number
  - Sales Time Alt Key
  - Sales Total Cost
- KPIs
- Dim Customer
- Dim Date
- Dim Product
- Dim Sales Person
- Dim Stores
- Dim Time

select [Measures].[Sales Invoice Number] on columns from [Sales DW]

Messages Results

Sales Invoice Number

139

Query executed successfully.

Ready

LN 1 Col 42 Ch 42 INS

select [Measures].[Sales Total Cost] on columns from [Sales DW]

MDXQuery1.mdx - WIN-33T3IN3MD4U.OLAP (WIN-33T3IN3MD4U\WINDOWS 8.1)\* - Microsoft SQL Server Management Studio

File Edit View Query Project Debug Tools Window Help

OLAP

New Query Execute

MDXQuery1.mdx - ...D4U\WINDOWS 8.1)\*

Cube: Sales DW

Measure Group: <All>

Sales DW

Measures

- Fact Product Sales
  - Deviation
  - Fact Product Sales Count
  - Product Actual Cost
  - Quantity
  - Sales Invoice Number
  - Sales Time Alt Key
  - Sales Total Cost
- KPIs
- Dim Customer
- Dim Date
- Dim Product
- Dim Sales Person
- Dim Stores
- Dim Time

select [Measures].[Sales Total Cost] on columns from [Sales DW]

Messages Results

Sales Total Cost

1231.5

Query executed successfully.

Ready

LN 1 Col 37 Ch 37 INS



select [Measures].[Sales Total Cost] on columns  
, [Dim Date].[Year].[Year] on rows from [Sales DW]

```
MDXQuery1.mdx - WIN-33T3IN3MD4U.OLAP (WIN-33T3IN3MD4U\WINDOWS 8.1)* - Microsoft SQL Server Management Studio - 100 %
```

File Edit View Query Project Debug Tools Window Help

Object Expl... Connect OLAP New Query Execute

Cube: Sales DW

Measure Group: <All>

- Quantity
- Sales Invoice Number
- Sales Time Alt Key
- Sales Total Cost

KPIs

Dim Customer

Dim Date

- Date Key
- Full Date UK
- Month
- Month Name
- Quarter
- Quarter Name
- Week Of Month
- Year
- Members
- Year
- Hierarchy

Dim Product

Dim Sales Person

MDXQuery1.mdx - D4U\WINDOWS 8.1)\*

```
select [Measures].[Sales Total Cost] on columns
, [Dim Date].[Year].[Year] on rows
from [Sales DW]
```

Messages Results

	Sales Total Cost
2013	1231.5
2014	(null)
Unknown	(null)

Ready

Query executed successfully.

WIN-33T3IN3MD4U | WIN-33T3IN3MD4U\WINDOW... | OLAP | 00:00:01

Ln 3 Col 16 Ch 16 INS

select [Measures].[Sales Total Cost] on columns , NONEMPTY({[Dim Date].[Year].[Year]}) on rows from [Sales DW]

निर्मलेश उत्तम सेवाधर्म



MDXQuery1.mdx - WIN-33T3IN3MD4U.OLAP (WIN-33T3IN3MD4U\WINDOWS 8.1)\* - Microsoft SQL Server Management Studio

File Edit View Query Project Debug Tools Window Help

OLAP Execute

MDXQuery1.mdx -...D4U\WINDOWS 8.1)\*

Cube: Sales DW

Measure Group: <All>

Sales DW

Measures

- Fact Product Sales
  - Deviation
  - Fact Product Sales Curr
  - Product Actual Cost
  - Quantity
  - Sales Invoice Number
  - Sales Time Alt Key
  - Sales Total Cost
- KPIs
- Dim Customer
- Dim Date
  - Date Key
  - Full Date UK
  - Month
  - Month Name
  - Quarter
  - Quarter Name
  - Week Of Month

100 %

Messages Results

Sales Total Cost

2013	1231.5
------	--------

Query executed successfully.

WIN-33T3IN3MD4U | WIN-33T3IN3MD4U\WINDOW... | OLAP | 00:00:01

Ready Ln 3 Col 16 Ch 16 INS

select [Measures].[Sales Total Cost] on columns from [Sales DW]  
Where [Dim Date].[Year].[Year].&[2013]

निर्मलेन्ह उत्तम सेवाधर्म



MDXQuery1.mdx - WIN-33T3IN3MD4U.OLAP (WIN-33T3IN3MD4U\WINDOWS 8.1)\* - Microsoft SQL Server Management Studio -

File Edit View Query Project Debug Tools Window Help

Object Expl... OLAP MDXQuery1.mdx ~D4U\WINDOWS 8.1)\*

Cube: Sales DW

Measure Group: <All>

Sales Time At Key  
Sales Total Cost  
KPIs  
Dim Customer  
Dim Date  
Date Key  
Full Date UK  
Month  
Month Name  
Quarter  
Quarter Name  
Week Of Month  
Year  
Members  
Year  
Member Properties  
2013  
2014  
Unknown  
Hierarchy

```
select [Measures].[Sales Total Cost] on columns
from [Sales DW]
Where [Dim Date].[Year].[Year].&[2013]
```

100 % Messages Results

Sales Total Cost  
1231.5

Ready

Query executed successfully.

WIN-33T3IN3MD4U | WIN-33T3IN3MD4U\WINDOW... | OLAP | 00:00:01

Ln 2 Col 1 Ch 1 INS



निर्मल सेवा द्वाधम

**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

1. **What is ETL?**
2. **Name the three steps in ETL.**
3. **Why is ETL important?**
4. **Name an ETL tool.**

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



### **Practical-5:**

- a. Import the data ware house data in Micros Excel and create the Pivot table and PivotChart.**
- b. Import the cube in Microsoft Excel and create- the Pivot table and Pivot Chart to perform data analysis.**

### **Aims:**

1. To import data from a data warehouse into Microsoft Excel and create PivotTables and PivotCharts for effective data analysis.
2. To import a multidimensional cube into Excel and use its interactive features to analyze data through PivotTables and Pivot Charts.

### **Learning Objectives:**

1. Understand how to connect Excel to a data warehouse and cube.
2. Learn to create and customize PivotTables and Pivot Charts to summarize and visualize data.
3. Develop skills in analyzing large datasets and deriving actionable insights using Excel's data analysis tools.

### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

### **Theory-5: Data ware house**

Data warehousing consolidates data from various sources into a centralized repository, enabling efficient analysis and reporting. Microsoft Excel serves as a versatile tool for importing data, creating PivotTables, and generating PivotCharts, which empower users to explore and visualize complex datasets.

#### **A. Importing Data Warehouse Data into Microsoft Excel**

##### **1. Data Importation:**

- o Utilize Excel's Data tab to establish a connection to the data warehouse using options like ODBC, SQL Server, or other database connectors.
- o Configure the connection to securely access the required tables or views from the warehouse.

##### **2. Creating a PivotTable:**

- o Once the data is imported, go to the Insert tab and select PivotTable to initiate a new PivotTable based on the imported data.

- Organize the data by dragging and dropping fields into Rows, Columns, Values, and Filters, enabling a multidimensional view of the data.

### 3. Generating a PivotChart:

- With the PivotTable in place, use the PivotChart feature to create a visual representation of your summarized data.
- Customize the chart type (bar, line, pie, etc.) to effectively highlight key trends and insights.

## B. Importing the OLAP Cube into Microsoft Excel

### 1. Connecting to the Cube:

- In Excel, select Data > Get Data > From Other Sources > From Analysis Services to connect directly to an OLAP cube.
- Input the necessary server and database credentials to access the multidimensional data.

### 2. Building a PivotTable from the Cube:

- Excel will automatically present the cube's dimensions and measures.
- Create a PivotTable by dragging cube fields into the respective areas, leveraging the inherent hierarchical structure for detailed analysis.

### 3. Creating a PivotChart for Cube Data:

- After setting up the PivotTable, insert a PivotChart to visualize the aggregated cube data.
- Utilize interactive features like drill-downs to explore underlying data and uncover hidden insights.



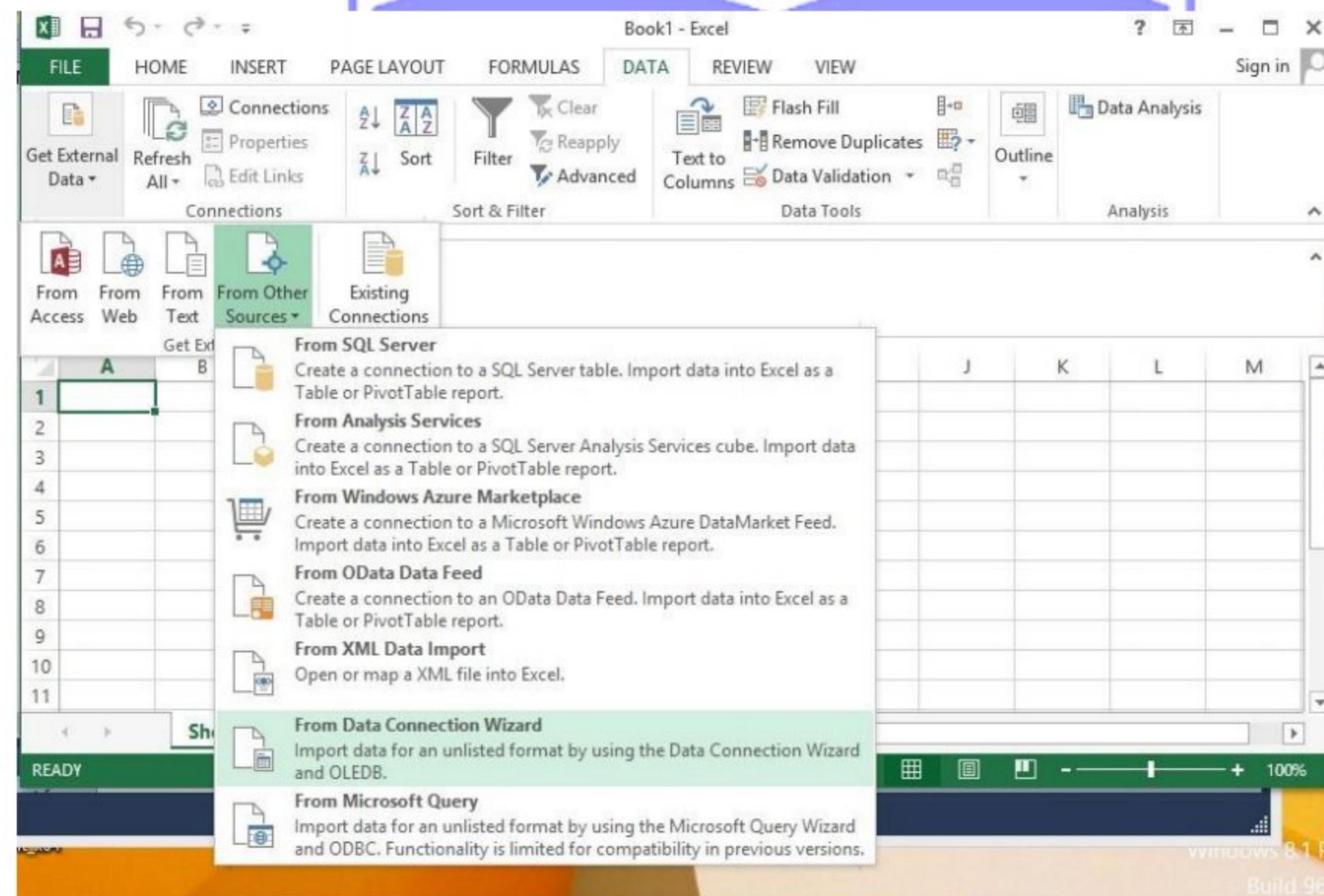
## PRACTICAL 5 a

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

(Ms Office Professional is used to make sure Power View is enabled for visualization.)

Step 1: Open Excel 2013 (Professional)

Go to Data tab → Get External Data → From Other Sources → From Data Connection Wizard



Step 2: In Data Connection Wizard → Select Microsoft SQL Server → Click on Next

निर्मलेश उत्तम सेवाधर्म



Step 3: In connect to Database Server provide Server name( Microsoft SQL Server Name)

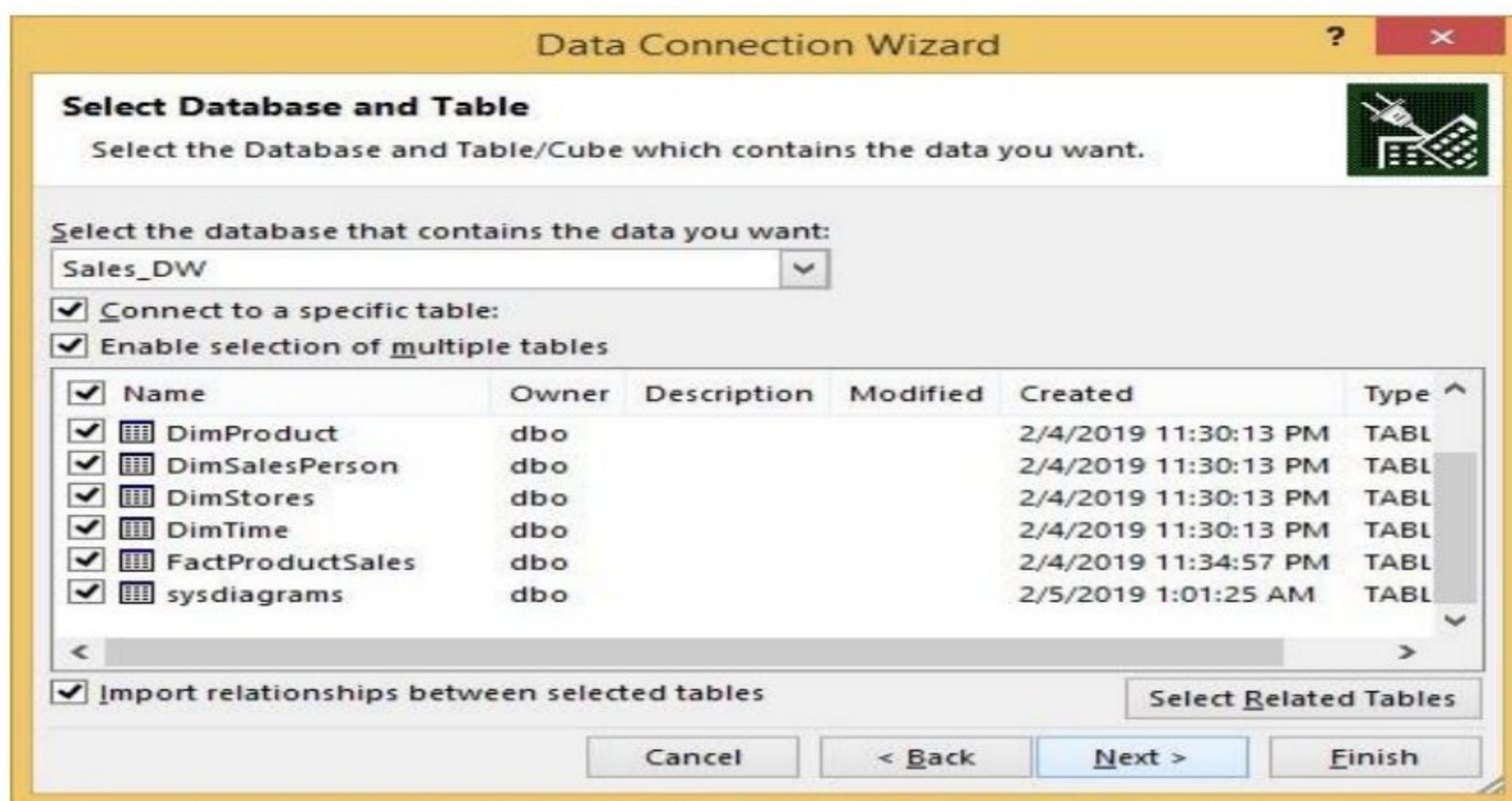
Provide password for sa account as given during installation of SQL Server 2012 full version)

Password: admin123

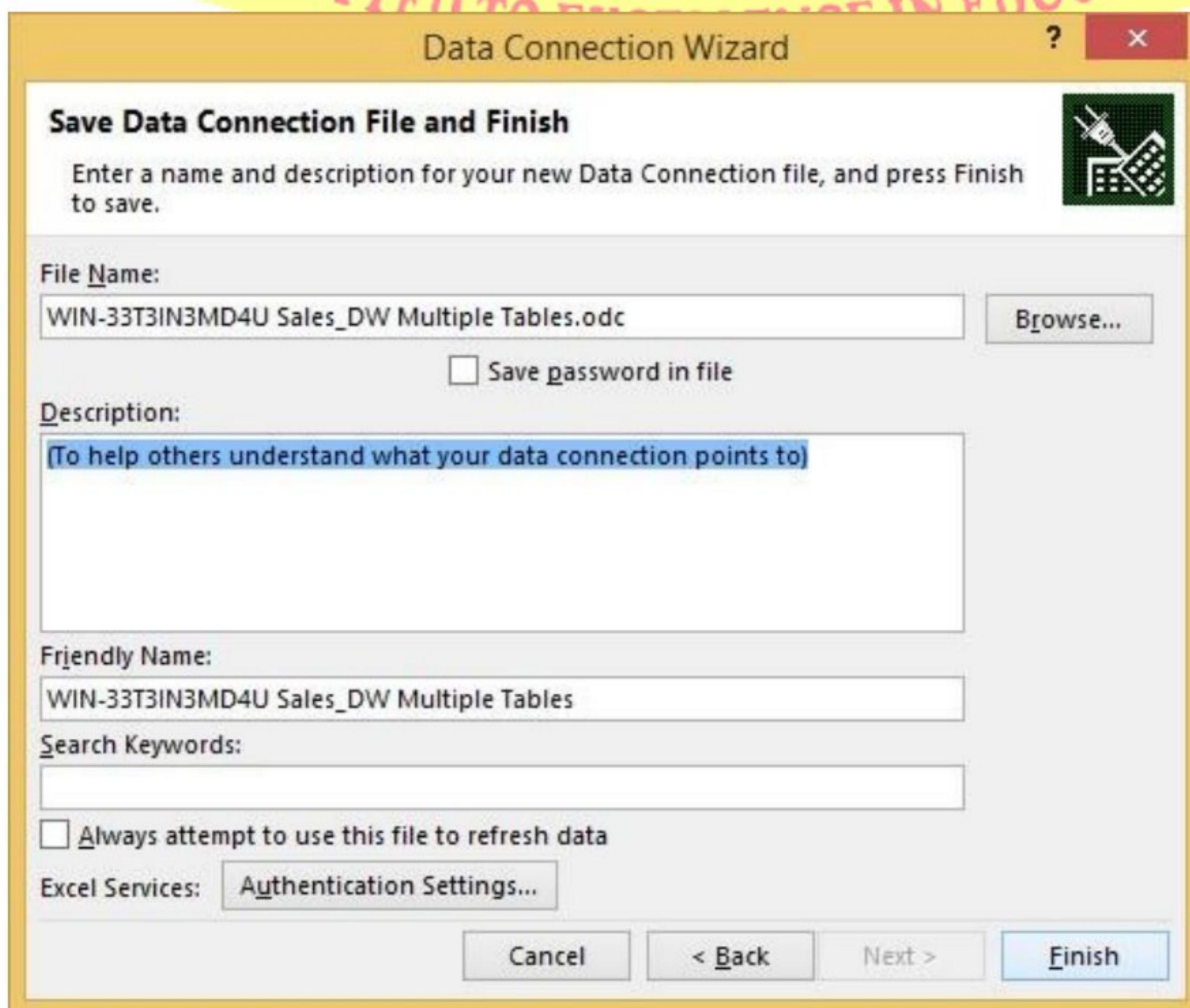
Click on Next



Step 4: In Select Database and Table → Select Sales\_DW (already created in SQL) → check all dimensions and import relationships between selected tables

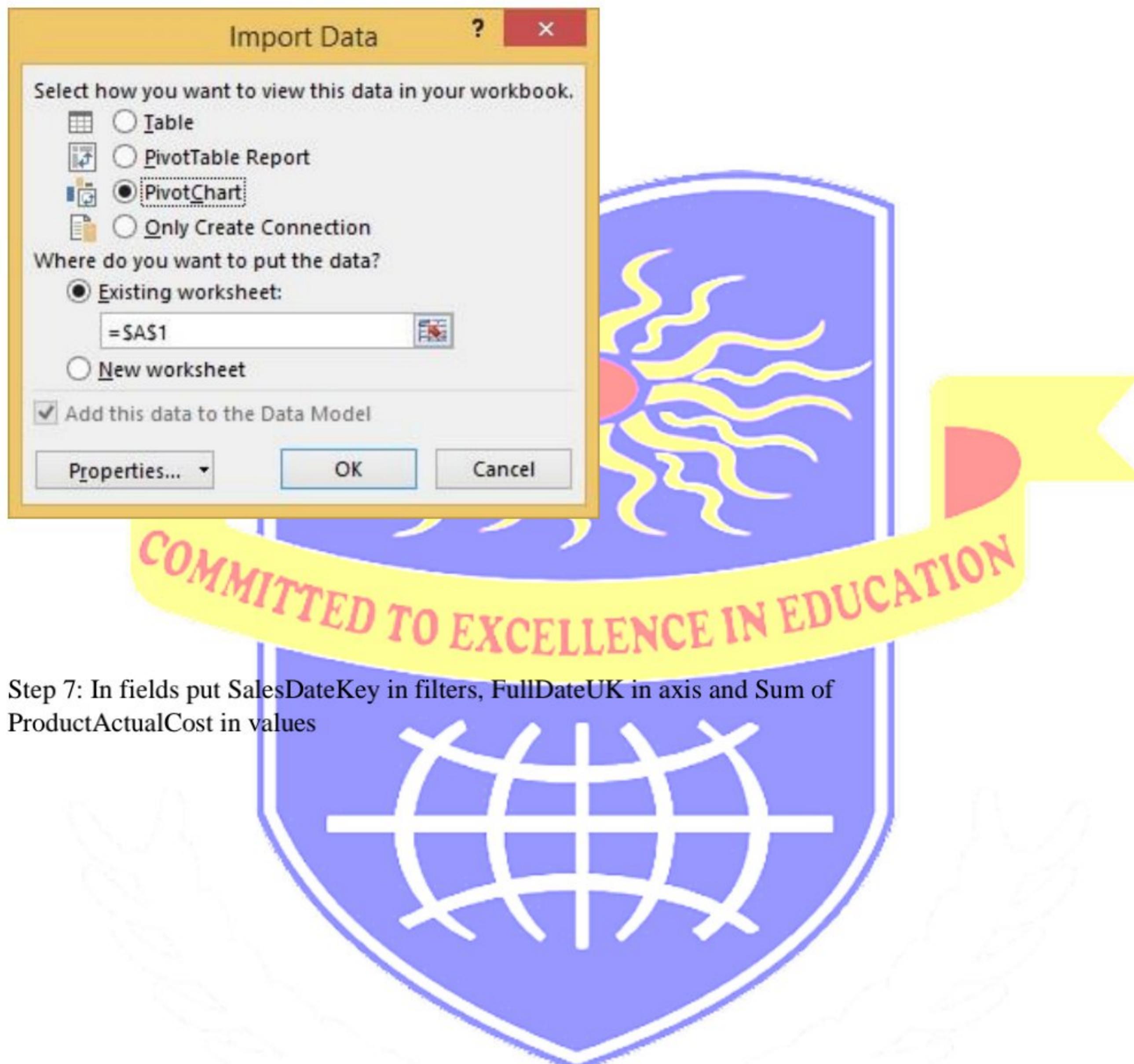


Step 5: In save data connection files browse path and click on Finish



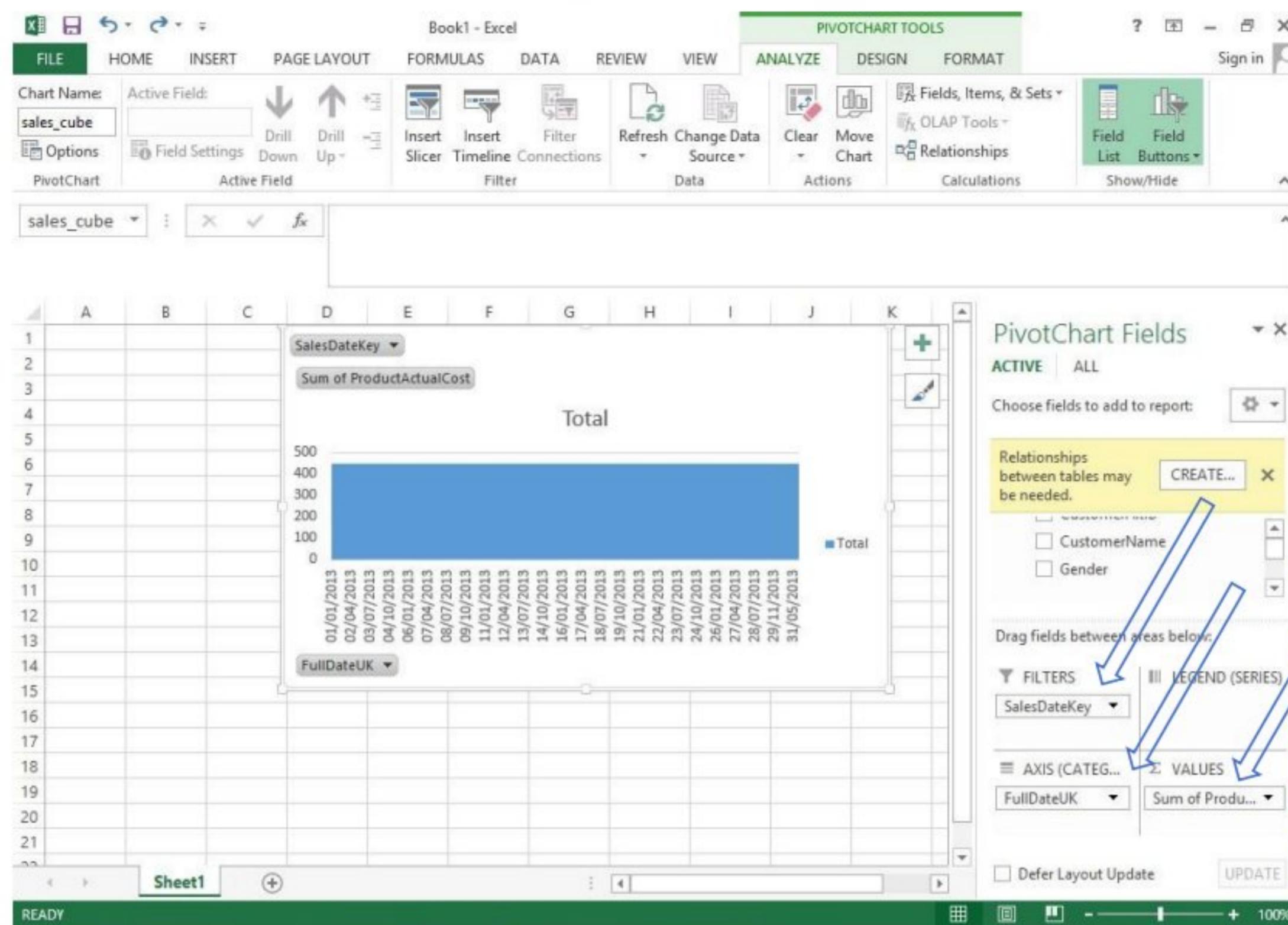


Step 6: In import data select Pivot Chart and click on OK



Step 7: In fields put SalesDateKey in filters, FullDateUK in axis and Sum of ProductActualCost in values

निर्मलेन्ह उत्तम सेवाधर्म



Step 8: In Insert Tab → go to Pivot Table

निर्मलेन्ह उत्तम सेवाधर्म



Book1 - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

PivotTable Recommended Table PivotTables Tables Illustrations Apps Recommended Charts PivotChart Charts Sparklines Filters Links Text Symbols Power View Reports

PivotTable

Easily arrange and summarize complex data in a PivotTable.

FYI: You can double-click a value to see which detailed values make up the summarized total.

Tell me more

SalesDateKey Sum of ProductActualCost

Total

500  
400  
300  
200  
100  
0

01/01/2013 02/04/2013 03/07/2013 04/10/2013 06/01/2013 07/04/2013 08/07/2013 09/10/2013 11/01/2013 12/04/2013 13/07/2013 14/10/2013 16/01/2013 17/04/2013 18/07/2013 19/10/2013 21/01/2013 22/04/2013 23/07/2013 24/10/2013 26/01/2013 27/04/2013 28/07/2013 29/11/2013 31/05/2013

FullDateUK

Sheet1

READY

100%

A yellow ribbon banner with the letters 'DON' written in red, positioned to the right of the Excel window.

Step 9: Click on Choose Connection to select existing connection with Sales\_DW and click on open



निर्मलेन्ह उत्तम सेवाधर्म



Create PivotTable

Choose the data that you want to analyze

Select a table or range  
 Use an external data source  
Table/Range:   
Choose Connection...  
Connection name: WIN-33T3IN3MD4U Sales\_DW Multiple Tabl

Choose where you want the PivotTable report to be placed

New Worksheet  
 Existing Worksheet  
Location: Sheet1!\$A\$16

Choose whether you want to analyze multiple tables

Add this data to the Data Model

OK Cancel

Existing Connections

Select a Connection or Table

Connections Tables

Show: All Connections

Connections in this Workbook

WIN-33T3IN3MD4U Sales\_DW Multiple Tables  
[Blank]

Connection files on the Network

<No connections found>

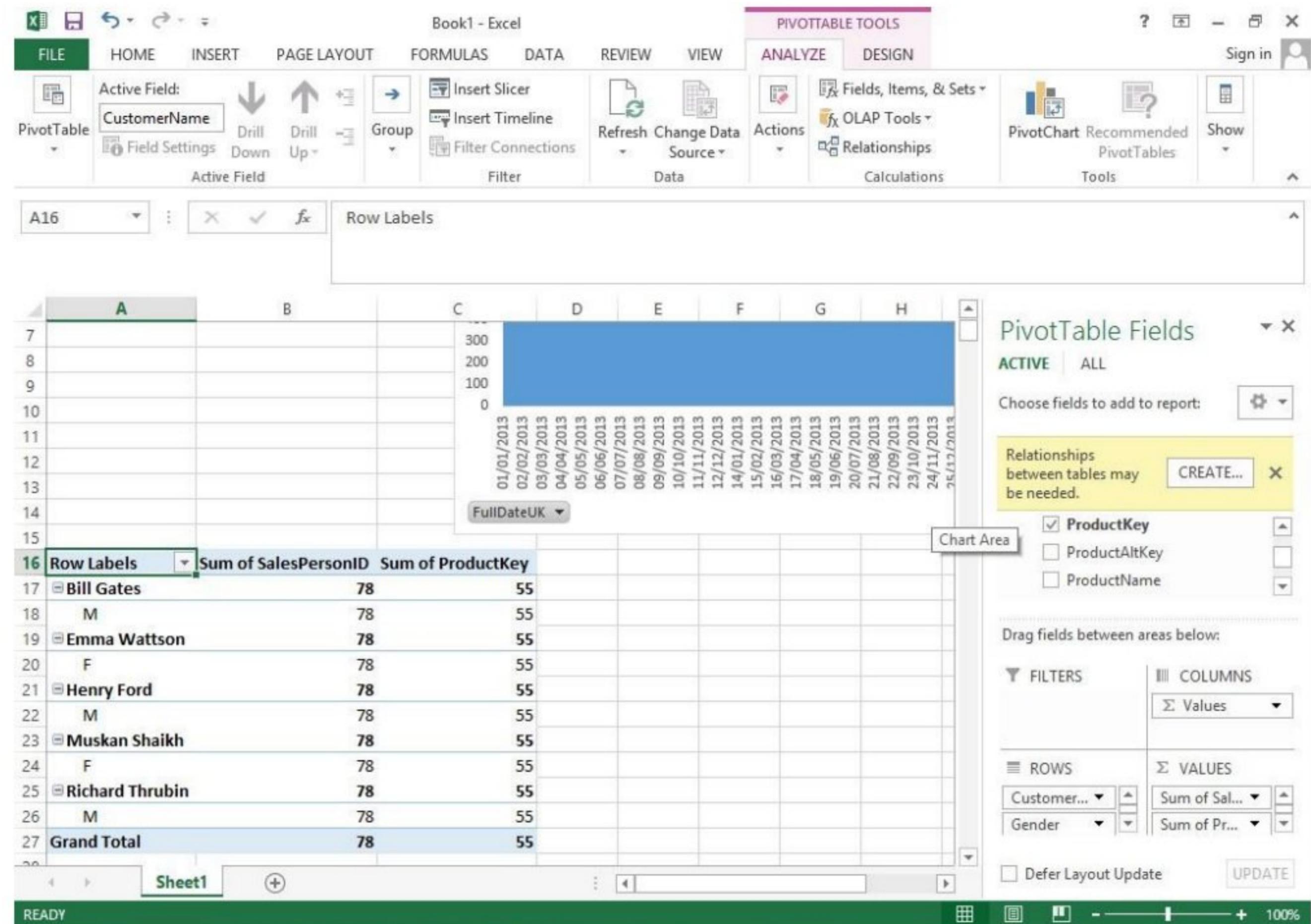
Connection files on this computer

WIN-33T3IN3MD4U Sales\_DW Multiple Tables  
[Blank]

Browse for More... Open Cancel



Pivot table and Pivot chart is created



**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

- 1. What is a data warehouse?**
- 2. Name a key feature of a data warehouse.**
- 3. Why are data warehouses used?**
- 4. What is the difference between OLTP and OLAP?**

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



---

## **Practical-6: Apply the what – if Analysis for data visualization. Design and generate necessary reports based on the data ware house data.**

---

### **Aims:**

1. To apply what-if analysis techniques to simulate various business scenarios using data warehouse data.
2. To design and generate dynamic reports that help in understanding potential outcomes and supporting decision-making.

### **Learning Objectives:**

1. Understand the concept and benefits of what-if analysis in a data warehousing context.
2. Learn to configure and manipulate what-if parameters to simulate changes in key business metrics.
3. Gain hands-on experience in designing interactive dashboards and reports that reflect hypothetical scenarios.
4. Develop skills in interpreting data-driven insights to inform business strategies.

### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

### **Theory-6: Data ware house data**

Data warehouses consolidate information from diverse sources, creating a centralized repository for strategic decision-making. Applying what-if analysis enables analysts to simulate different scenarios, assess potential impacts, and visualize outcomes. This approach transforms raw data into actionable insights, guiding strategic planning and operational improvements.

### **What-If Analysis for Data Visualization**

- **Definition:**

What-if analysis involves modifying key input variables to explore alternative outcomes. It helps identify risks, forecast trends, and understand the sensitivity of various business metrics.

- **Techniques and Tools:**

Utilize tools like Microsoft Excel, Power BI, or Tableau to perform scenario analysis. These tools support dynamic simulations using data models, adjustable parameters, and interactive dashboards.

- **Application Areas:**

- **Financial Forecasting:** Model revenue or expense changes under different economic conditions.
- **Sales Projections:** Analyze the impact of pricing strategies or market fluctuations on sales volumes.
- **Resource Allocation:** Predict outcomes of different staffing or inventory levels.

## Designing and Generating Reports

### 1. Data Preparation and Integration:

- Import data warehouse records into Excel or a BI tool using secure connections (e.g., ODBC, SQL Server connectors).
- Cleanse and aggregate the data to ensure accuracy for analysis.

### 2. Report Layout and Structure:

- **Interactive Dashboards:** Create dashboards that integrate PivotTables, PivotCharts, and slicers to allow dynamic filtering.
- **Custom Metrics:** Incorporate key performance indicators (KPIs) tailored to the business context, such as profit margins, growth rates, or cost variances.



## PRACTICAL 6

Apply the what – if Analysis for data visualization. Design and generate necessary reports based on the data warehouse data.

A book store and have 100 books in storage. You sell a certain % for the highest price of \$50 and a certain % for the lower price of \$20.

The screenshot shows an Excel spreadsheet titled "Book1 - Excel". The Data tab is selected. In the formula bar, the formula =E10\*F10+E11\*F11 is entered in cell F12. A black arrow points from the formula bar to cell F12. Another black arrow points from cell F12 to cell D10, which contains the value 3800. Cell D10 is highlighted with an orange oval. The spreadsheet contains the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2																	
3																	
4	total books	% sold for highest price															
5	100	60															
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	

If you sell 60% for the highest price, cell D10 calculates a total profit of  $60 * 50 + 40 * 20 = 3800$ .

Create Different Scenarios But what if you sell 70% for the highest price? And what if you sell 80% for the highest price? Or 90%, or even 100%? Each different percentage is a different scenario. You can use the Scenario Manager to create these scenarios.

Note: To type different percentage into cell C4 to see the corresponding result of a scenario in cell D10 we use what if analysis.

What-if analysis enables you to easily compare the results of different scenarios.

Step 1: In Excel, On the Data tab, in the Data tools group, click What-If Analysis



The screenshot shows a Microsoft Excel spreadsheet titled "Book1 - Excel". The formula bar at the top displays the formula  $=E10*F10+E11*F11$ . The worksheet contains the following data:

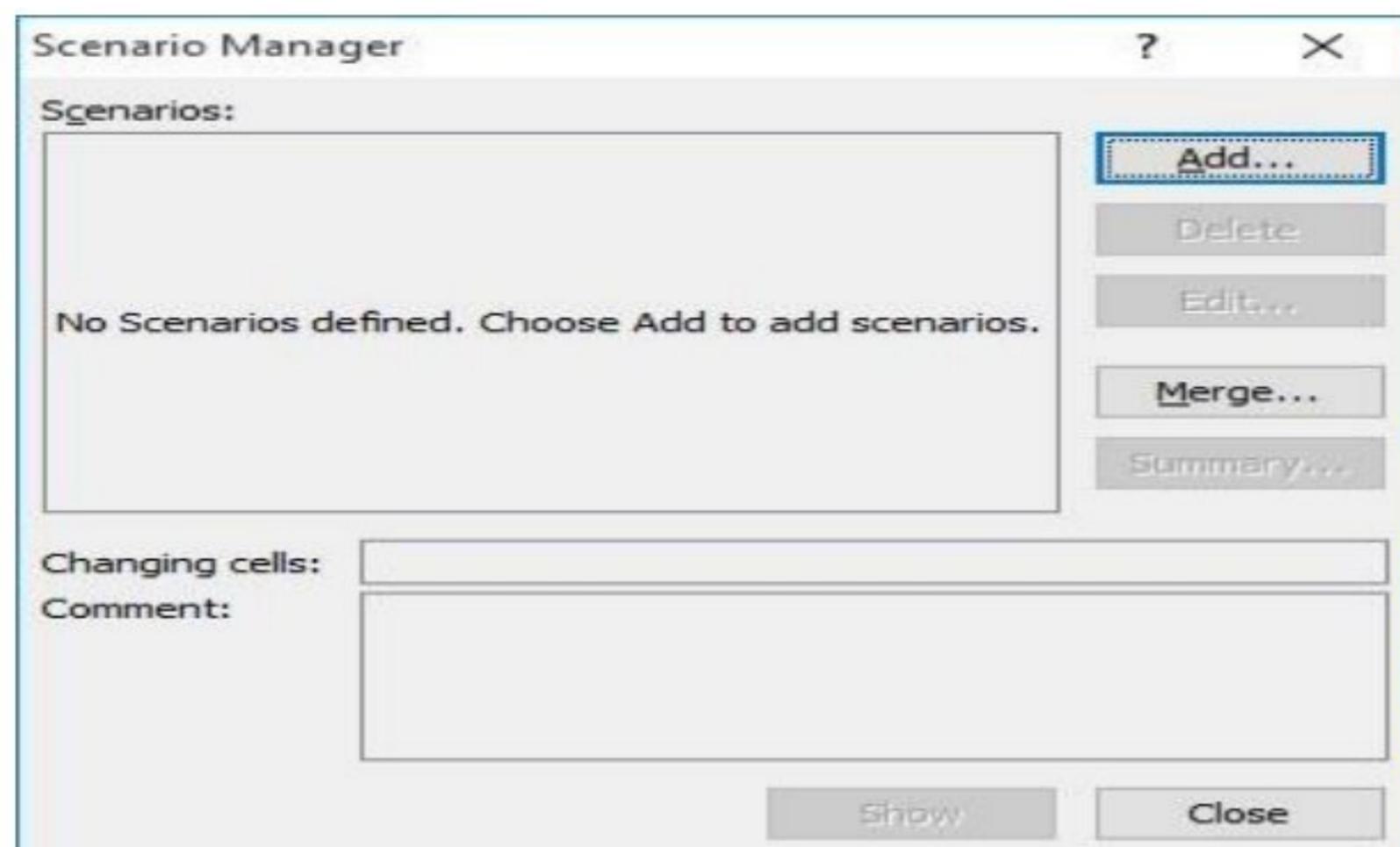
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2																	
3																	
4		total books	% sold for highest price														
5		100	60														
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	

Step 2: Click on What-if-Analysis and select scenario manager.

The screenshot shows the same Microsoft Excel spreadsheet as above. The "Data" tab is selected, and the "What-If Analysis" group is active. The "Scenario Manager..." button is highlighted. A tooltip for the "Scenario Manager" button is displayed, stating: "Create different groups of values or scenarios, and switch between them." The formula bar still shows the formula  $=E10*F10+E11*F11$ .

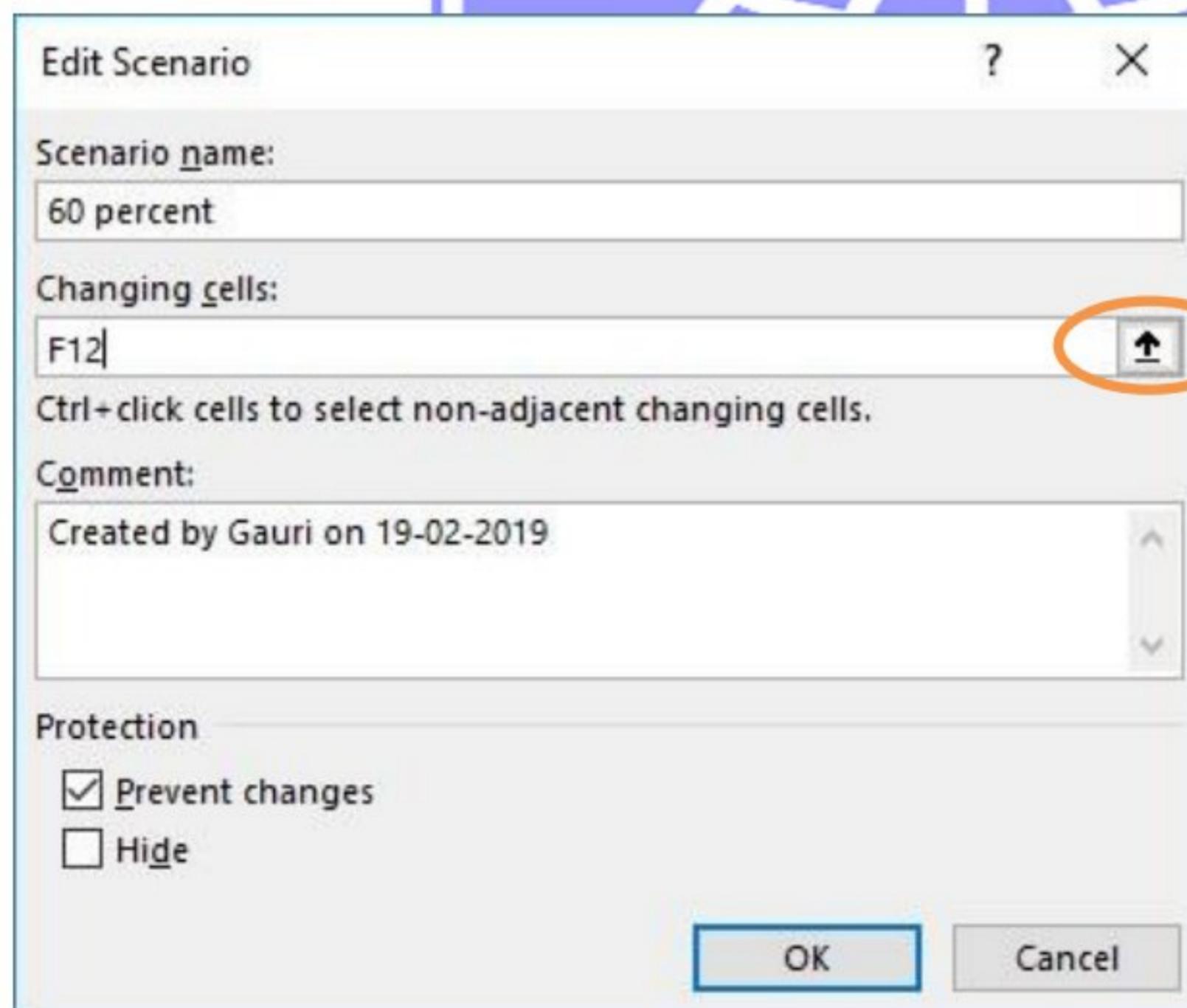


The Scenario Manager Dialog box appears. Step 3: Add a scenario by clicking on Add.



Step 4: Type a name (60percent), select cell F10 (% sold for the highest price) for the Changing cells and click on OK.

Click on icon which is circled.



Select F10 cell.



Add Scenario - Changing cells:

SFS10

Click back on the icon again and then click OK

Edit Scenario

Scenario name:  
60 percent

Changing cells:  
SFS10

Ctrl+click cells to select non-adjacent changing cells.

Comment:  
Created by Gauri on 19-02-2019

Protection

Prevent changes  
 Hide

OK Cancel

Step 5: Enter the corresponding value 0.6 and click on OK again.

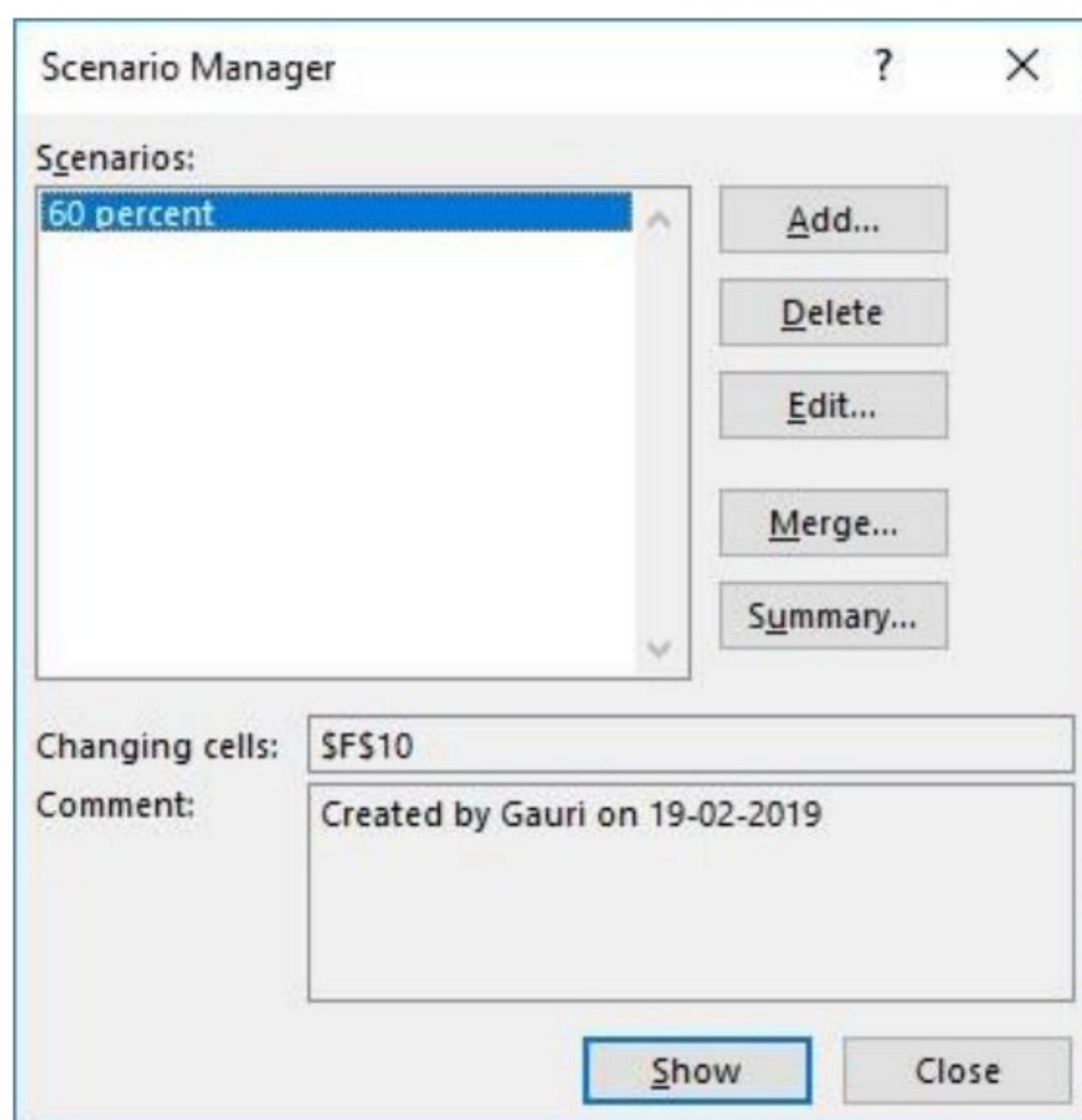
Scenario Values

Enter values for each of the changing cells.

1: SFS10 0.6

Add OK Cancel

निर्मलेह उत्तम सेवाधर्म



Step 6: To apply scenarios click on Show

निर्मलेन्ह उत्तम सेवाधर्म



Book1 - Excel

Gauri Ansarkar

F12 : =E10\*F10+E11\*F11

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2																	
3																	
4	total books	% sold for highest price															
5	100	60															
6																	
7																	
8																	
9							no of books	unit profit									
10							highest	60	0.6								
11							lowest	40	20								
12							total price		836								
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	

Scenario Manager

Scenarios:

- 60 percent
- 70 percent
- 80 percent
- 90 percent
- 100 percent

Changing cells: SF\$10

Comment: Created by Gauri on 19-02-2019

Show Close

Step 7: Next, add 4 other scenarios (70%, 80%, 90% and 100%)

Finally, your Scenario Manager should be consistent with the picture below:

Scenario Manager

Scenarios:

- 60% highest
- 70% highest
- 80% highest
- 90% highest
- 100% highest

Changing cells: SCS4

Comment: Created by excel-easy.com on 2/21/2017

Show Close

ह उत्तम सेवाधर्म

**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

1. **What is data warehousing?**
2. **Define data warehouse data.**
3. **Why is data warehousing important?**
4. **Data warehouse vs. database?**

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



## Practical-7: Perform the data classification using classification algorithm

### **Aims:**

1. To apply classification algorithms for categorizing data into predefined classes.
2. To understand how data classification supports predictive analytics and decision-making.

### **Learning Objectives:**

1. Comprehend the fundamental concepts and techniques in classification within data mining and machine learning.
2. Gain hands-on experience in implementing classification algorithms using programming libraries.
3. Evaluate model performance through metrics such as accuracy, precision, recall, and F1 score.
4. Develop skills in data preprocessing, model training, and interpretation of classification results.

### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

## **Theory-7: Classification**

Data classification is a supervised machine learning technique aimed at categorizing data into predefined classes or labels. By analyzing historical, labeled data, classification algorithms learn patterns and relationships, enabling them to accurately predict the class of new, unseen instances.

### **Process of Data Classification**

1. **Data Preparation**
  - o **Collection & Cleaning:** Gather a representative dataset and clean it by handling missing values, removing outliers, and correcting errors.
  - o **Feature Selection & Engineering:** Identify key features that influence the classification outcome and transform raw data into a suitable format for analysis.
2. **Dataset Partitioning**
  - o **Training Set:** Allocate a portion of the data to train the classification model, ensuring it learns the underlying patterns.
  - o **Test Set:** Reserve another portion to validate and evaluate the model's performance, thereby preventing overfitting.
3. **Algorithm Selection**

- Choose an appropriate classification algorithm based on the nature of the data and problem requirements. Common choices include:
  - **Decision Trees:** Offer intuitive, rule-based classification.
  - **Naive Bayes:** Uses probabilistic reasoning for efficient classification.
  - **Support Vector Machines (SVM):** Effective in high-dimensional feature spaces.
  - **K-Nearest Neighbors (KNN):** A simple, instance-based approach.

#### 4. Model Training and Evaluation

- **Training:** Input the training data into the selected algorithm to develop a predictive model.
- **Evaluation:** Use performance metrics—such as accuracy, precision, recall, and F1 score—to assess how well the model classifies new data using the test set.

#### 5. Optimization and Deployment

- **Parameter Tuning:** Employ techniques like cross-validation and grid search to optimize model parameters and enhance performance.
- **Deployment:** Once validated, deploy the model to classify incoming data in real-time or batch processes, ensuring it continuously supports decision-making with up-to-date predictions.



## PRACTICAL 7

Perform the data classification using classification algorithm.

OR

Data Analysis using Time Series Analysis

Software required: R 3.5.1

Time series is a series of data points in which each data point is associated with a timestamp. A simple example is the price of a stock in the stock market at different points of time on a given day. Another example is the amount of rainfall in a region at different months of the year. R language uses many functions to create, manipulate and plot the time series data. The data for the time series is stored in an R object called time-series object. It is also a R data object like a vector or data frame.

The time series object is created by using the ts() function.

Syntax

The basic syntax for ts() function in time series analysis is – timeseries.object.name <- ts(data, start, end, frequency)

Following is the description of the parameters used –

- data is a vector or matrix containing the values used in the time series.
- start specifies the start time for the first observation in time series.
- end specifies the end time for the last observation in time series.
- frequency specifies the number of observations per unit time. Except the parameter "data" all other parameters are optional

Consider the annual rainfall details at a place starting from January 2012. We create an R time series object for a period of 12 months and plot it.

Code to run in R

```
# Get the data points in form of a R vector.
```

```
rainfall <- c(799,1174.8,865.1,1334.6,635.4,918.5,685.5,998.6,784.2,985,882.8,1071)
```

```
# Convert it to a time series object.
```

```
rainfall.timeseries <- ts(rainfall,start = c(2012,1),frequency = 12)
```



```
# Print the timeseries data. print(rainfall.timeseries)
```

```
# Give the chart file a name. png(file =  
"rainfall.png")
```

```
# Plot a graph of the time series.  
plot(rainfall.timeseries)
```

```
# Save the file.  
dev.off()
```

After this again plot to get chart plot(rainfall.timeseries)

Output:

When we execute the above code, it produces the following result and chart –

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	2012	799.0	1174.8		
865.1	1334.6	635.4	918.5	685.5	998.6	784.2		Oct	Nov	Dec	985.0	882.8	1071.0



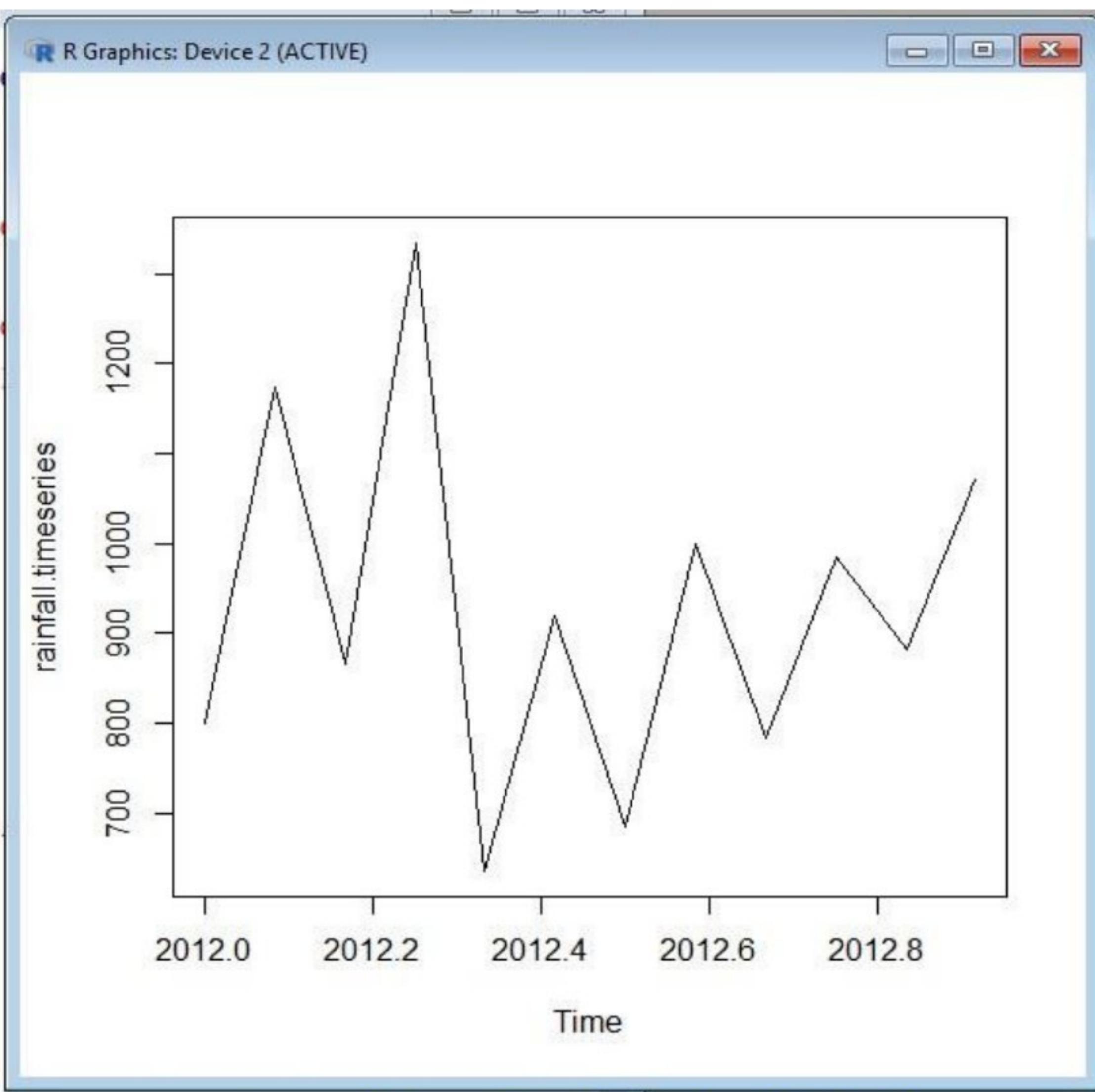
RGui (64-bit)

File Edit View Misc Packages Windows Help

Type 'q()' to quit R.

```
> # Get the data points in form of a R vector.  
> rainfall <- c(799,1174.8,865.1,1334.6,635.4,918.5,685.5,998.6,784.2,985,882.8,1071)  
> # Convert it to a time series object.  
> rainfall.timeseries <- ts(rainfall,start = c(2012,1),frequency = 12)  
> # Print the timeseries data.  
> print(rainfall.timeseries)  
    Jan     Feb     Mar     Apr     May     Jun     Jul     Aug     Sep     Oct  
2012  799.0  1174.8  865.1  1334.6  635.4  918.5  685.5  998.6  784.2  985.0  
      Nov     Dec  
2012  882.8  1071.0  
> # Give the chart file a name.  
> png(file = "rainfall.png")  
> # Plot a graph of the time series.  
> plot(rainfall.timeseries)  
> # Save the file.  
> dev.off()  
null device  
1  
> plot(rainfall.timeseries)  
> |
```





**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

- 1. What is classification?**
- 2. What is supervised classification?**
- 3. Why is classification important?**
- 4. Example of a classification task?**

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



## **Practical-8: Perform the data clustering using clustering algorithm.**

### **Aims:**

1. To understand the fundamentals of clustering and its application in grouping similar data points.
2. To discover inherent patterns and segments within datasets using unsupervised learning techniques.

### **Learning Objectives:**

1. Comprehend the basic concepts of unsupervised learning and clustering methods.
2. Gain hands-on experience with popular clustering algorithms such as K-means and hierarchical clustering.
3. Learn how to preprocess, analyze, and visualize data for effective clustering.
4. Evaluate the quality of clusters using metrics like silhouette scores.

### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

### **Theory-8: Clustering**

Clustering is an unsupervised learning technique that groups similar data points into clusters, uncovering inherent structures in the dataset without predefined labels. This approach is useful for segmenting data, discovering patterns, and reducing data complexity in areas such as market segmentation, anomaly detection, and image analysis.

### **Process of Data Clustering**

#### **1. Data Preparation**

- o **Data Collection & Cleaning:** Gather a comprehensive dataset and clean it by removing noise, handling missing values, and eliminating outliers.
- o **Normalization & Scaling:** Standardize or normalize features to ensure that no single feature dominates the clustering due to scale differences.

#### **2. Feature Selection and Extraction**

- o **Feature Selection:** Identify key variables that capture the essence of the data, ensuring that the chosen features enhance the distinction between clusters.
- o **Dimensionality Reduction:** Apply techniques such as Principal Component Analysis (PCA) if necessary to reduce dimensionality while preserving important information.

#### **3. Algorithm Selection**

- **K-Means:** A widely used algorithm that partitions data into k clusters by minimizing the variance within each cluster.
- **Hierarchical Clustering:** Builds a tree of clusters (dendrogram) and does not require specifying the number of clusters upfront.
- **DBSCAN:** Groups data points based on density, which is effective for discovering clusters of arbitrary shapes and handling noise.

#### 4. Model Training and Evaluation

- **Clustering Execution:** Run the selected algorithm on the prepared dataset. For example, with K-Means, choose an initial value for k and iterate until the cluster centroids stabilize.
- **Cluster Validation:** Evaluate the quality of the clusters using metrics such as silhouette scores, Davies-Bouldin index, or within-cluster sum of squares (WCSS).
- **Parameter Tuning:** Adjust parameters (like the number of clusters in K-Means or the neighborhood radius in DBSCAN) to refine cluster quality.

PRACTICAL 8

Perform the data 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 21, 96, 33

Cluster means:

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

Clustering vector:

Within cluster sum of squares by cluster:

```
[1] 17.669524 118.651875 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	17	0	33
versicolor	4	46	0
virginica	0	50	0

```
#Plot the clusters and their centers
```

```
plot(newiris[c("Sepal.Length","Sepal.Width")],col=kc$cluster)
```

```
points(kc$centers[,c("Sepal.Length","Sepal.Width")],col=1:3,pch=8,cex=2) dev.off()
```

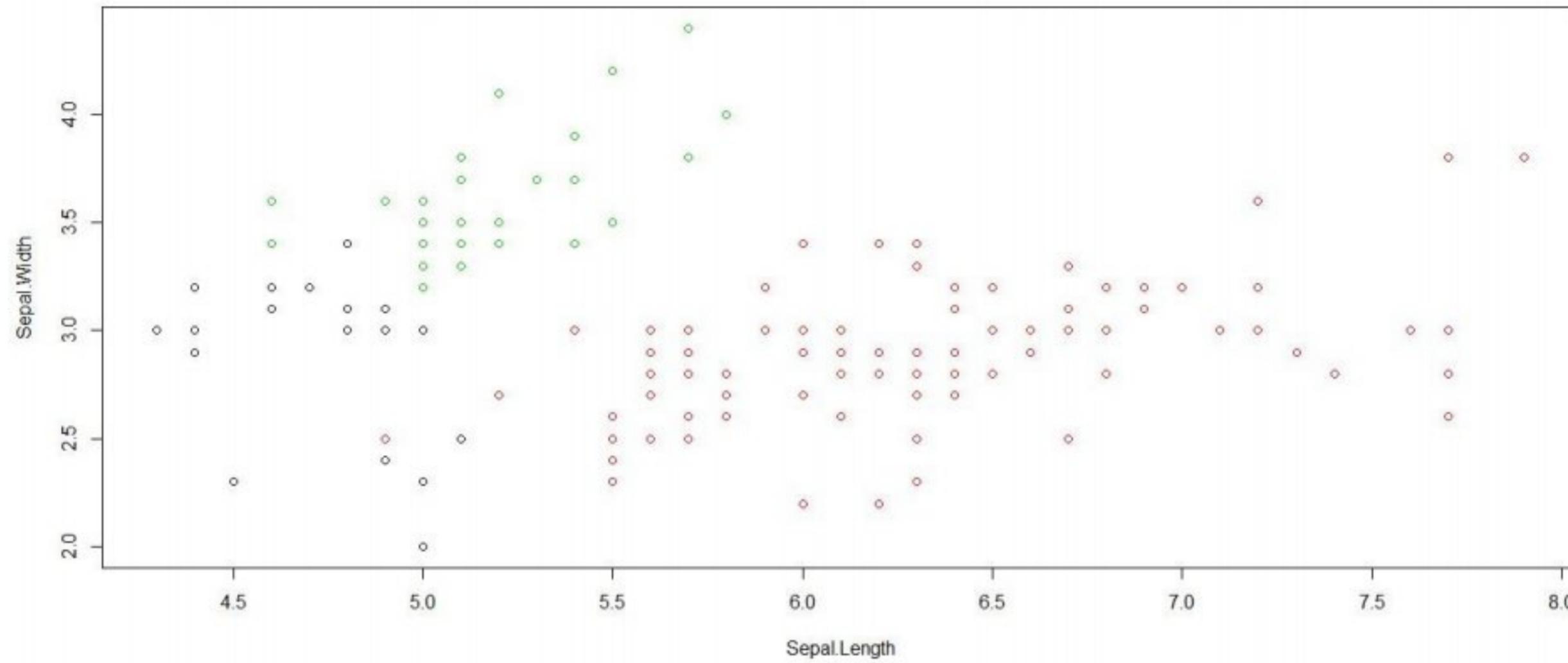
```
#Plot the clusters and their centre
```

```
plot(newiris[c("Sepal.Length","Sepal.Width")],col=kc$cluster)
```



**SHRI G.P.M. DEGREE COLLEGE**  
Department of Computer  
Vision.. Innovation.. Solution.. Presentation

RGui (64-bit) - [R Graphics: Device 2 (ACTIVE)]  
File History Resize Windows  
Return focus to Console



**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

- 1. What is clustering?**
- 2. Name a clustering algorithm.**
- 3. What is the purpose of clustering?**
- 4. Difference between clustering and classification?**

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



## **Practical-9: Perform the Linear regression on the given data ware house data.**

### **Aims:**

1. To apply linear regression techniques on data extracted from a data warehouse.
2. To forecast relationships between variables and derive predictive insights for informed decision-making.

### **Learning Objectives:**

1. Understand the fundamental concepts and assumptions of linear regression.
2. Gain practical experience in preprocessing and cleaning data for regression analysis.
3. Learn to implement linear regression models using Python and evaluate model performance using metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).
4. Interpret the regression coefficients to assess the impact of independent variables on the target variable.

### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

### **Theory-9: Linear regression**

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It helps identify trends, make predictions, and derive insights from data warehouse information. This technique is widely used for forecasting, performance analysis, and decision-making in business intelligence.

### **Steps to Perform Linear Regression**

#### **1. Data Preparation and Extraction**

- Extract structured data from the data warehouse using SQL queries or data connectors.
- Identify the target variable (dependent) and relevant predictors (independent variables).
- Perform data cleaning by handling missing values, removing duplicates, and standardizing formats.

#### **2. Feature Selection and Preprocessing**

- Analyze feature correlations to select the most relevant independent variables.
- Normalize or scale variables if they have different units to improve model accuracy.
- Split the dataset into training and testing subsets (e.g., 80% for training, 20% for testing).

#### **3. Model Evaluation and Interpretation**

- Assess the model's performance using metrics such as:
  - **R-squared ( $R^2$ ):** Measures how well the model explains the variability in data.
  - **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):** Evaluate prediction accuracy.
  - **Residual Analysis:** Check for normal distribution and homoscedasticity of residuals.

#### 4. Prediction and Business Application

- Apply the trained model to new warehouse data for forecasting and decision-making.
- Use insights from the model to optimize business strategies, such as sales predictions, resource allocation, and trend analysis.



## PRACTICAL 9

Perform the Linear regression on the given data warehouse data.

Input Data

Below is the sample data representing the observations –

# Values of height

151, 174, 138, 186, 128, 136, 179, 163, 152, 131

# Values of weight

63, 81, 56, 91, 47, 57, 76, 72, 62, 48

lm() Function :

This function creates the relationship model between the predictor and the response variable.

Syntax :

The basic syntax for lm() function in linear regression is – lm(formula,data)

Following is the description of the parameters used :-

- formula is a symbol presenting the relation between x and y.
- data is the vector on which the formula will be applied.

**A. Create Relationship Model & get the Coefficients # Values**  
of height x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)

# Values of width y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62,

48)



# Apply the lm() function.

```
relation <- lm(y~x) print(relation)
```

OUTPUT:

```
Call:  
lm(formula = y ~ x)
```

Coefficients:

(Intercept)	x
-38.4551	0.6746

**B.** Get the Summary of the Relationship # Values of height x <-  
c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)

```
# Values of width y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62,
```

48)

# Apply the lm() function. relation <-

```
lm(y~x) print(summary(relation))
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.3002	-1.6629	0.0412	1.8944	3.9775

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-38.45509	8.04901	-4.778	0.00139 **
x	0.67461	0.05191	12.997	1.16e-06 ***

---

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*\*' 0.01 '\*\*' 0.05 '\*' 0.1 '.' 1

Residual standard error: 3.253 on 8 degrees of freedom

Multiple R-squared: 0.9548, Adjusted R-squared: 0.9491

F-statistic: 168.9 on 1 and 8 DF, p-value: 1.164e-06

**predict()** Function



## Syntax

The basic syntax for predict() in linear regression is – predict(object, newdata)

Following is the description of the parameters used –

- object is the formula which is already created using the lm() function.
- newdata is the vector containing the new value for predictor variable.

C. Predict the weight of new persons  
# The predictor vector.  
  
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)

*COMMITTED TO EXCELLENCE IN EDUCATION*

# The response vector. y <- c(63, 81, 56, 91, 47, 57, 76,

72, 62, 48)

# Apply the lm() function.  
relation <- lm(y~x)

# Find weight of a person with height 170.

a <- data.frame(x = 170) result <-  
predict(relation,a) print(result)

OUTPUT:

1  
76.22869

D. Visualize the Regression Graphically # Create the predictor and  
response variable. x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)  
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48) relation <- lm(y~x)



```
# Give the chart file a name.
```

```
png(file = "linearregression.png")
```

```
# Plot the chart.
```

```
plot(y,x,col = "blue",main = "Height & Weight Regression", abline(lm(x~y)),cex = 1.3,pch = 16,xlab =  
"Weight in Kg",ylab = "Height in cm")
```

```
# Save the file.
```

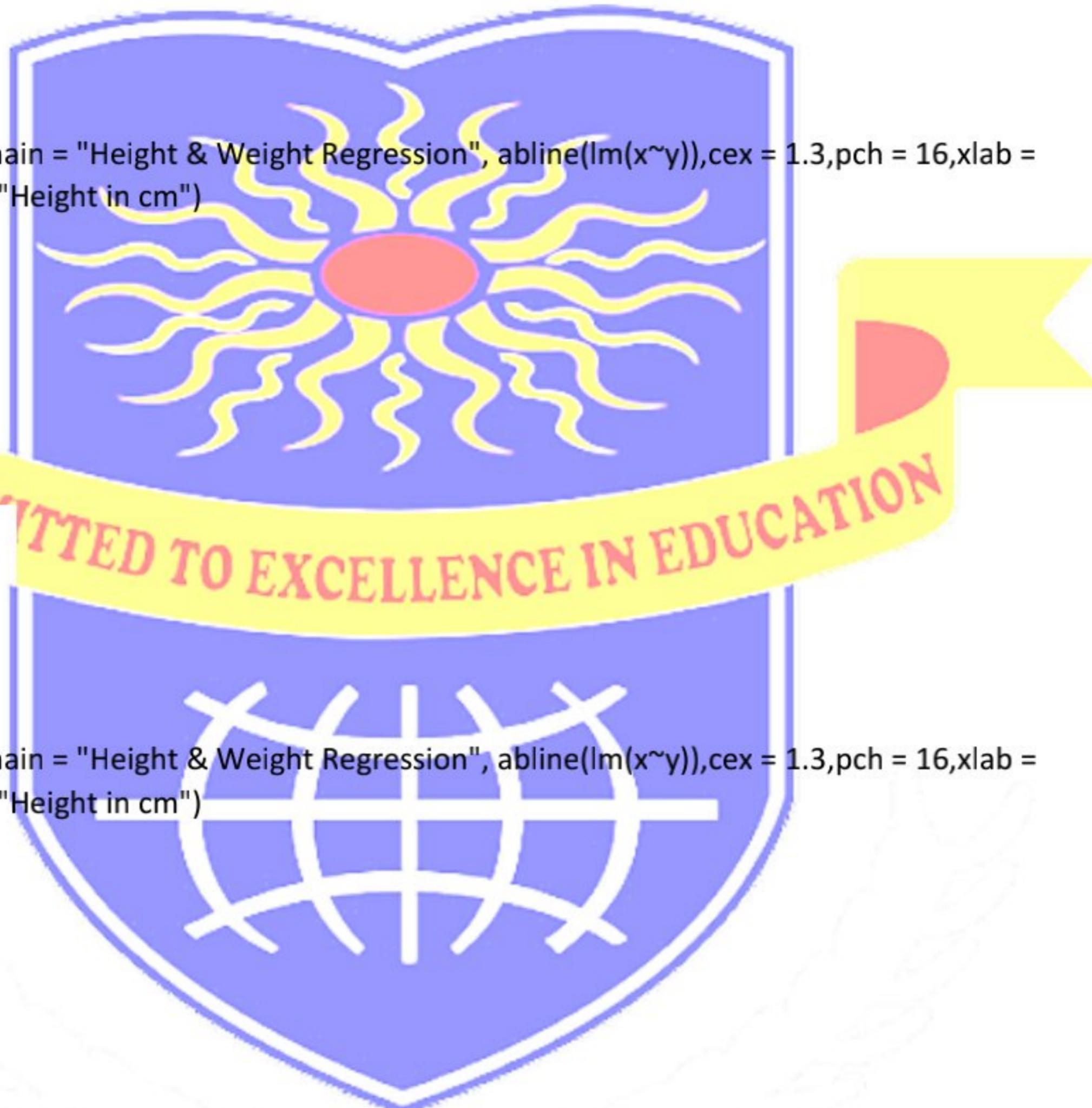
```
dev.off()
```

```
null device  
1
```

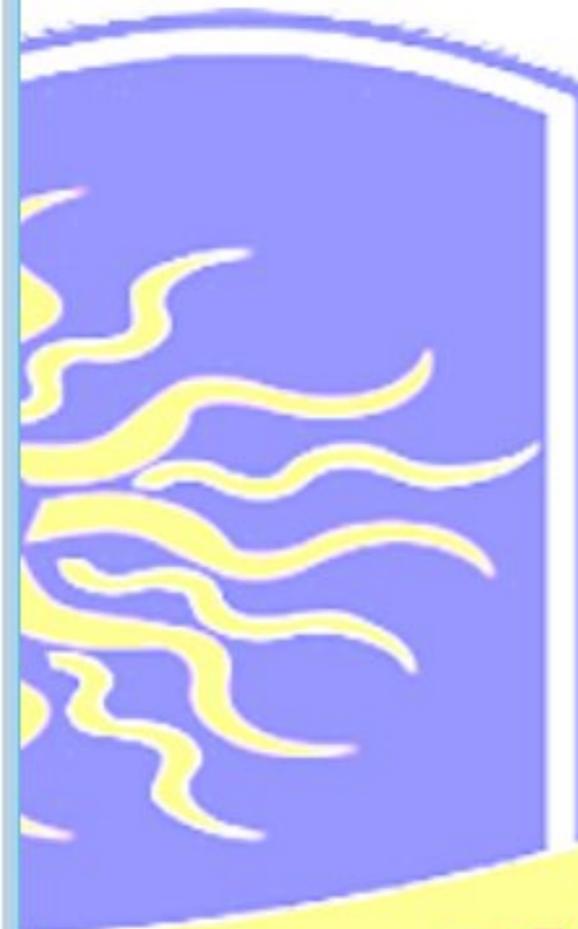
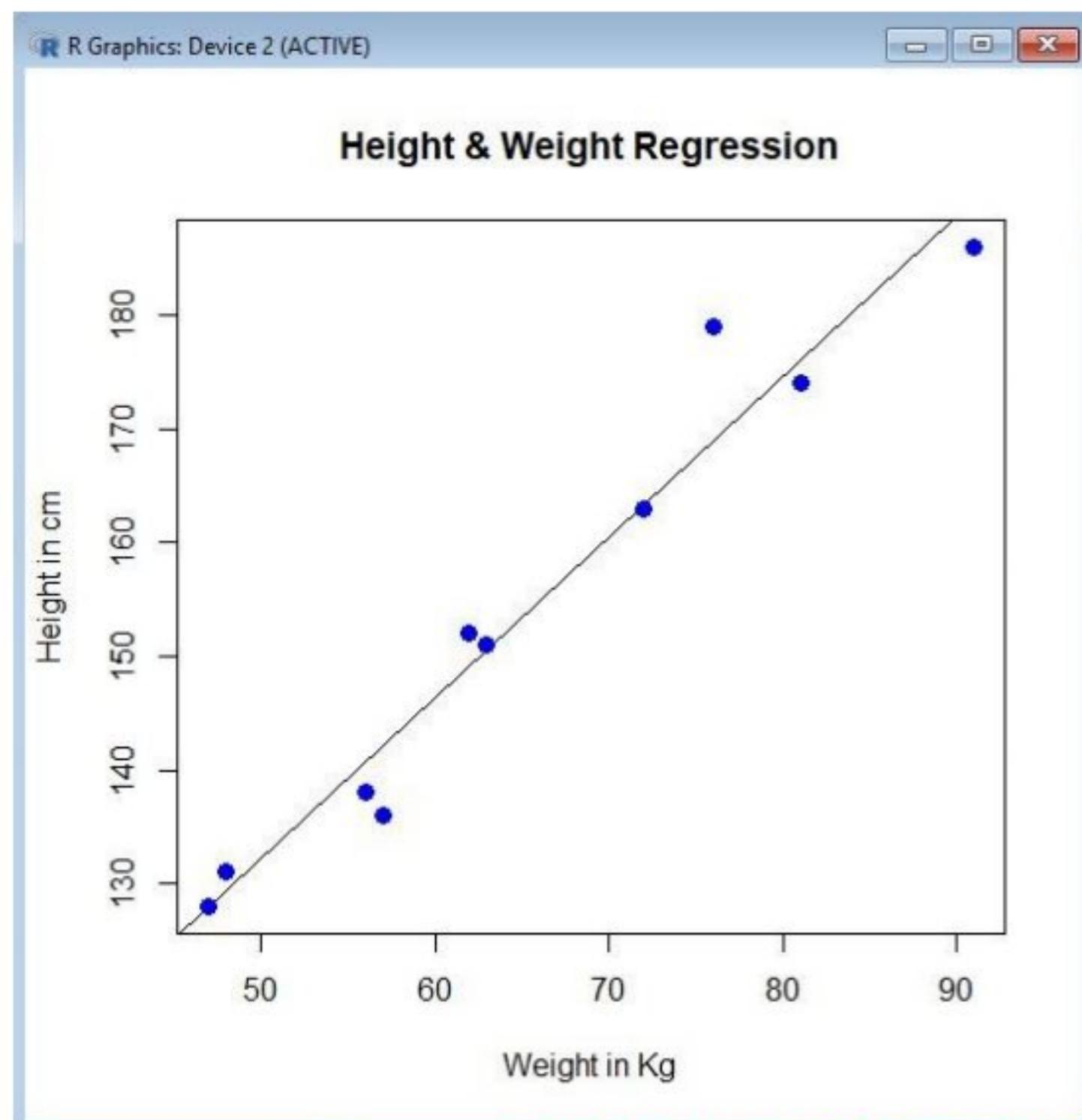
```
# Plot the chart.
```

```
plot(y,x,col = "blue",main = "Height & Weight Regression", abline(lm(x~y)),cex = 1.3,pch = 16,xlab =  
"Weight in Kg",ylab = "Height in cm")
```

OUTPUT:



निर्मल लोह उत्तम सेवाधर्म



**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

1. **What is linear regression?**
2. **Name one assumption of linear regression.**
3. **What is the purpose of the slope coefficient?**
4. **How is the R-squared value interpreted?**

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]



## **Practical-10: Perform the logistic regression on the given data ware house -data.**

### **Aims:**

1. To apply logistic regression techniques on data extracted from a data warehouse for binary classification tasks.
2. To predict the probability of a binary outcome based on multiple predictor variables.

### **Learning Objectives:**

1. Understand the fundamentals and assumptions of logistic regression.
2. Gain hands-on experience in preparing and preprocessing data for logistic regression analysis.
3. Evaluate model performance using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
4. Interpret model coefficients to derive insights for decision-making.

### **Tool & Technologies used:**

1. Power BI is a powerful business intelligence tool used for data visualization and analysis.

### **Theory-10: logistic regression**

Logistic regression is a statistical and machine learning technique used for classification problems, where the target variable is categorical (e.g., binary classification: yes/no, success/failure). Unlike linear regression, logistic regression models the probability that an instance belongs to a particular category using the logistic (sigmoid) function.

### **Steps to Perform Logistic Regression**

#### **1. Data Preparation and Extraction**

- Extract structured data from the data warehouse using SQL queries or data connectors.
- Identify the dependent variable (categorical outcome) and independent variables (predictors).
- Perform data cleaning by handling missing values, standardizing formats, and removing duplicates.

#### **2. Feature Selection and Preprocessing**

- Select the most relevant independent variables using correlation analysis.
- Convert categorical variables into numerical format using encoding techniques (e.g., one-hot encoding).
- Normalize or scale numerical variables if required.
- Split the dataset into training and testing subsets (e.g., 80% training, 20% testing).

### **3. Model Evaluation and Performance Metrics**

- **Accuracy Score:** Measures overall correctness of predictions.
- **Confusion Matrix:** Displays true positives, false positives, true negatives, and false negatives.
- **Precision, Recall, and F1-Score:** Evaluate classification performance.
- **ROC Curve & AUC Score:** Analyze the model's ability to distinguish between classes.

### **4. Prediction and Business Application**

- Apply the trained model to new warehouse data for predictive analysis.
- Use logistic regression to classify outcomes such as customer churn, fraud detection, risk assessment, and marketing segmentation.



## PRACTICAL 10

Perform the logistic regression on the given data warehouse data.

To perform this you need to download quality.csv file from following link:

<https://github.com/TarekDib03/Analytics/tree/master/Week3%20%20Logistic%20Regression/Data>

```
#provide path of file where it is saved on your machine quality <-
read.csv('C:/Users/Gauri/Downloads/quality.csv')

> #analysing the quality dataset
> str(quality)
'data.frame': 131 obs. of 14 variables:
 $ MemberID    : int 1 2 3 4 5 6 7 8 9 10 ...
 $ InpatientDays : int 0 1 0 8 2 16 2 24 ...
 $ ERVisits     : int 0 1 0 1 2 0 1 0 1 2 ...
 $ OfficeVisits  : int 18 6 5 19 19 9 8 8 4 0 ...
 $ Narcotics     : int 1 1 3 0 3 2 1 0 3 2 ...
 $ DaysSinceLastERVisit: num 731 411 731 158 449 ...
 $ Pain          : int 10 0 10 34 10 6 4 5 5 2 ...
 $ TotalVisits   : int 18 8 5 20 29 11 25 10 7 6 ...
 $ ProviderCount : int 21 27 16 14 24 40 19 11 28 21 ...
 $ MedicalClaims  : int 93 19 27 59 51 53 40 28 20 17 ...
 $ ClaimLines     : int 222 115 148 242 204 156 261 87 98 66 ...
 $ StartedOnCombination: logi FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ AcuteDrugGapSmall : int 0 1 5 0 0 4 0 0 0 0 ...
 $ PoorCare        : int 0 0 0 0 1 0 0 1 0 ...

> table(quality$PoorCare)
```



98 33

> 98/131

[1] 0.7480916

> install.packages("caTools")

Installing package into 'C:/Users/Gauri/Documents/R/win-library/3.5'

**(as 'lib' is unspecified)**

--- Please select a CRAN mirror for use in this session --- also installing the dependency 'bitops'

trying URL

'[http://mirror.its.dal.ca/cran/bin/windows/contrib/3.5/bitops\\_1.0-6.zip](http://mirror.its.dal.ca/cran/bin/windows/contrib/3.5/bitops_1.0-6.zip)' Content type

'application/zip' length 38894 bytes (37 KB) downloaded 37 KB

trying URL

'[http://mirror.its.dal.ca/cran/bin/windows/contrib/3.5/caTools\\_1.17.1.1.zip](http://mirror.its.dal.ca/cran/bin/windows/contrib/3.5/caTools_1.17.1.1.zip)'

Content type 'application/zip' length 329665 bytes (321 KB) downloaded 321 KB

**package 'bitops' successfully unpacked and MD5 sums checked**

**package 'caTools' successfully unpacked and MD5 sums**

**checked**

The downloaded binary packages are in

C:\Users\Gauri\AppData\Local\Temp\RtmpmUN9oK\downloaded\_packages

> library(caTools) Warning

message:

**package 'caTools' was built under R version 3.5.2**

> set.seed(88)



```
> split = sample.split(quality$PoorCare, SplitRatio = 0.75) >  
  
> split  
  
[1] TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE TRUE  
FALSE TRUE  
[28] TRUE TRUE FALSE FALSE FALSE FALSE TRUE TRUE TRUE FALSE  
TRUE TRUE TRUE FALSE FALSE TRUE TRUE FALSE TRUE FALSE TRUE  
FALSE TRUE TRUE FALSE FALSE TRUE  
  
[55] TRUE  
FALSE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE  
TRUE TRUE TRUE TRUE TRUE  
  
[82] TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE  
  
[109] TRUE FALSE FALSE TRUE TRUE FALSE TRUE TRUE TRUE FALSE  
TRUE TRUE FALSE TRUE TRUE FALSE TRUE TRUE FALSE TRUE TRUE TRUE FALSE  
  
> qualityTrain = subset(quality, split == TRUE) > qualityTest =  
subset(quality, split == FALSE)  
  
> nrow(qualityTrain)  
  
[1] 99  
  
> nrow(qualityTest)  
  
[1] 32  
  
> QualityLog = glm(PoorCare ~ OfficeVisits + Narcotics, data=qualityTrain, family=binomial)  
  
> summary(QualityLog)
```

Call:

```
glm(formula = PoorCare ~ OfficeVisits + Narcotics, family = binomial, data = qualityTrain)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.06303	-0.63155	-0.50503	-0.09689	2.16686

Coefficients:



Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.64613 0.52357 -5.054 4.33e-07 \*\*\*

OfficeVisits 0.08212 0.03055 2.688 0.00718 \*\*

Narcotics 0.07630 0.03205 2.381 0.01728 \*

---

**Signif. codes:** 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 111.888 on 98 degrees of freedom

Residual deviance: 89.127 on 96 degrees of freedom

AIC: 95.127

Number of Fisher Scoring iterations: 4

```
> predictTrain = predict(QualityLog, type="response")
```

```
> summary(predictTrain)
```

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.06623 0.11912 0.15967 0.25253 0.26765 0.98456

```
> tapply(predictTrain, qualityTrain$PoorCare, mean)
```

**0** 1

0.1894512 0.4392246

```
> table(qualityTrain$PoorCare, predictTrain > 0.5)
```

FALSE TRUE

**0** 70 4

**1** 15 10

> 10/25



[1] 0.4

> 70/74

[1] 0.9459459

> table(qualityTrain\$PoorCare, predictTrain > 0.7)

FALSE TRUE

**0** 73 1

**1** 17 8

> 8/25

[1] 0.32

> 73/74

[1] 0.9864865

> table(qualityTrain\$PoorCare, predictTrain > 0.2)

FALSE TRUE

**0** 54 20

**1** 9 16

> 16/25

[1] 0.64

> 54/74

[1] 0.7297297

> install.packages("ROCR")

Installing package into 'C:/Users/Gauri/Documents/R/win-library/3.5'

(as 'lib' is unspecified) also installing the

dependencies 'gtools', 'gdata', 'gplots'

trying URL



'http://mirror.its.dal.ca/cran/bin/windows/contrib/3.5/gtools\_3.8.1.zip' Content type

'application/zip' length 325812 bytes (318 KB) downloaded 318 KB

trying URL

'http://mirror.its.dal.ca/cran/bin/windows/contrib/3.5/gdata\_2.18.0.zip' Content type

'application/zip' length 1260728 bytes (1.2 MB) downloaded 1.2 MB

trying URL

'http://mirror.its.dal.ca/cran/bin/windows/contrib/3.5/gplots\_3.0.1.1.zip' Content type 'application/zip'

length 656764 bytes (641 KB) downloaded 641 KB

trying URL

'http://mirror.its.dal.ca/cran/bin/windows/contrib/3.5/ROCR\_1.0-7.zip' Content type

'application/zip' length 201823 bytes (197 KB) downloaded 197 KB

**package 'gtools' successfully unpacked and MD5 sums checked**

**package 'gdata' successfully unpacked and MD5 sums checked**

**package 'gplots' successfully unpacked and MD5 sums checked**

**package 'ROCR' successfully unpacked and MD5 sums checked**

The downloaded binary packages are in

C:\Users\Gauri\AppData\Local\Temp\RtmpmUN9oK\downloaded\_packages

> library(ROCR)

Loading required package: gplots

**Attaching package: 'gplots'**

The following object is masked from 'package:stats':



lowess

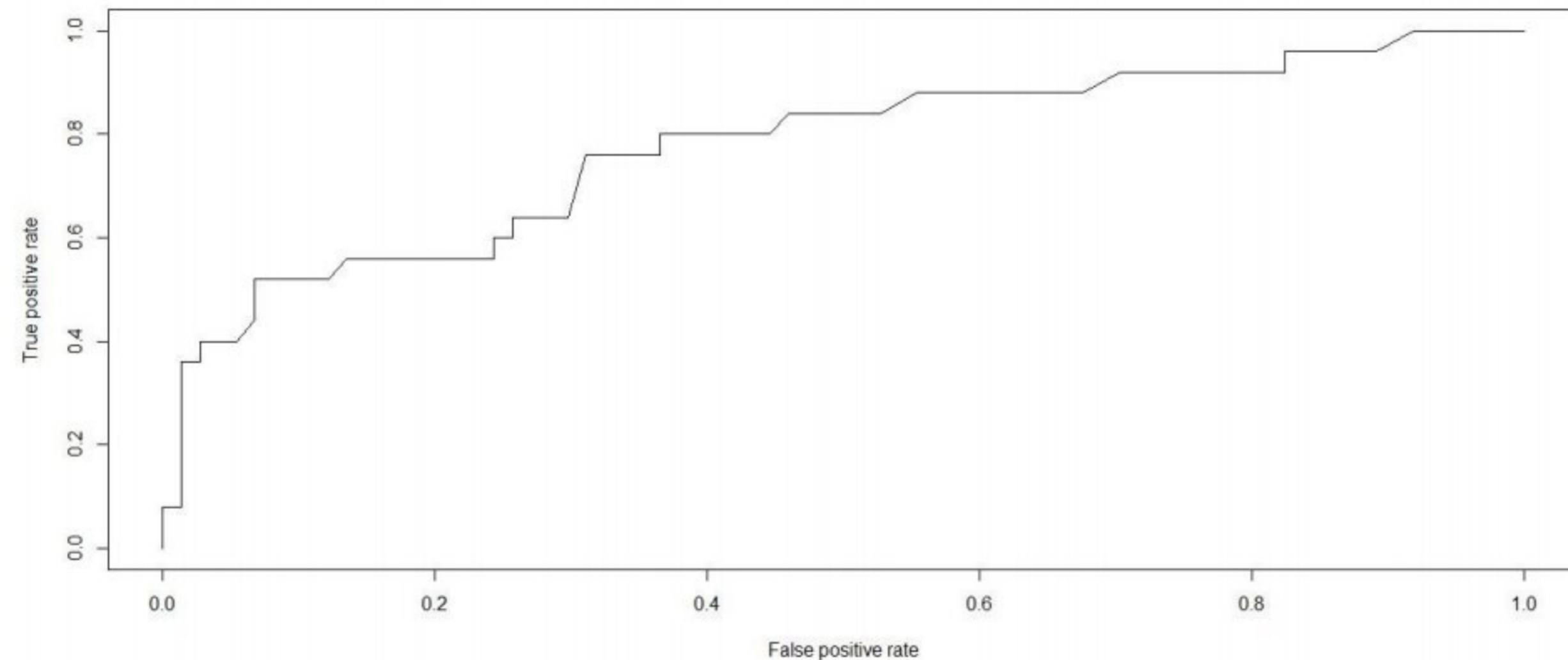
Warning messages:

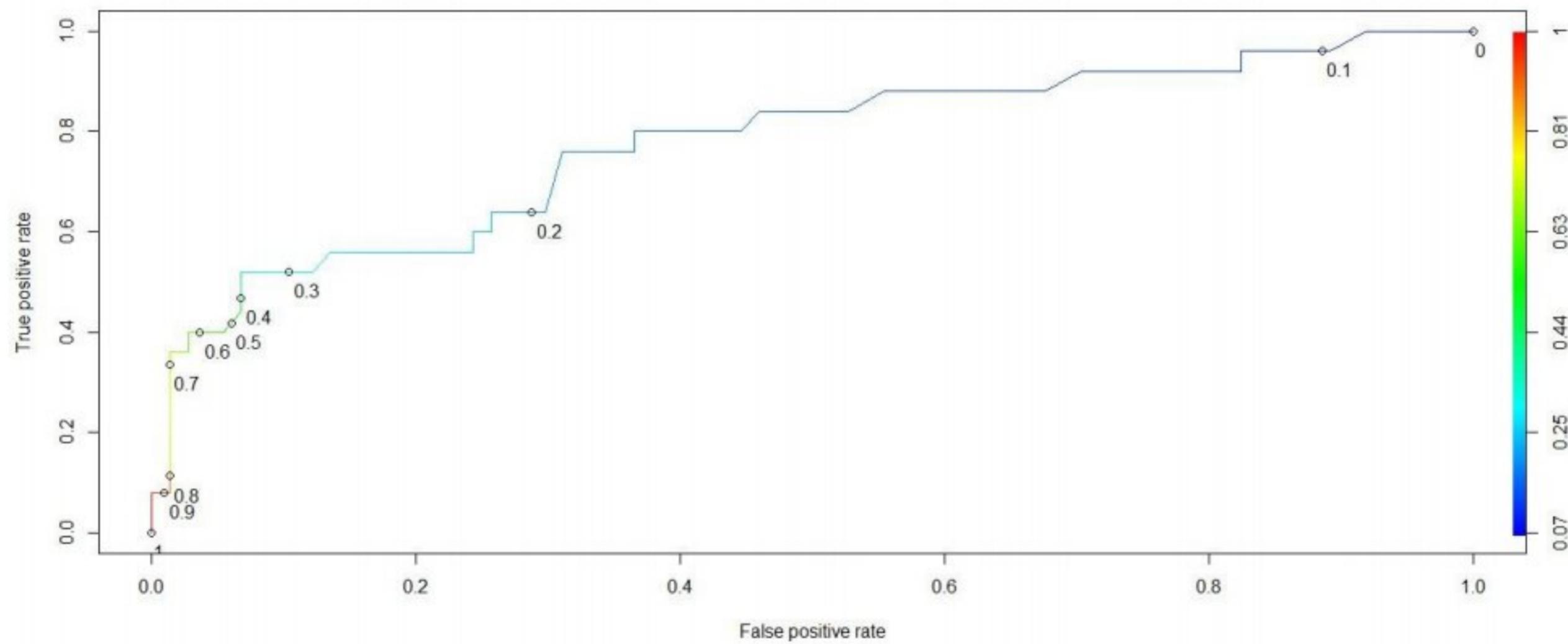
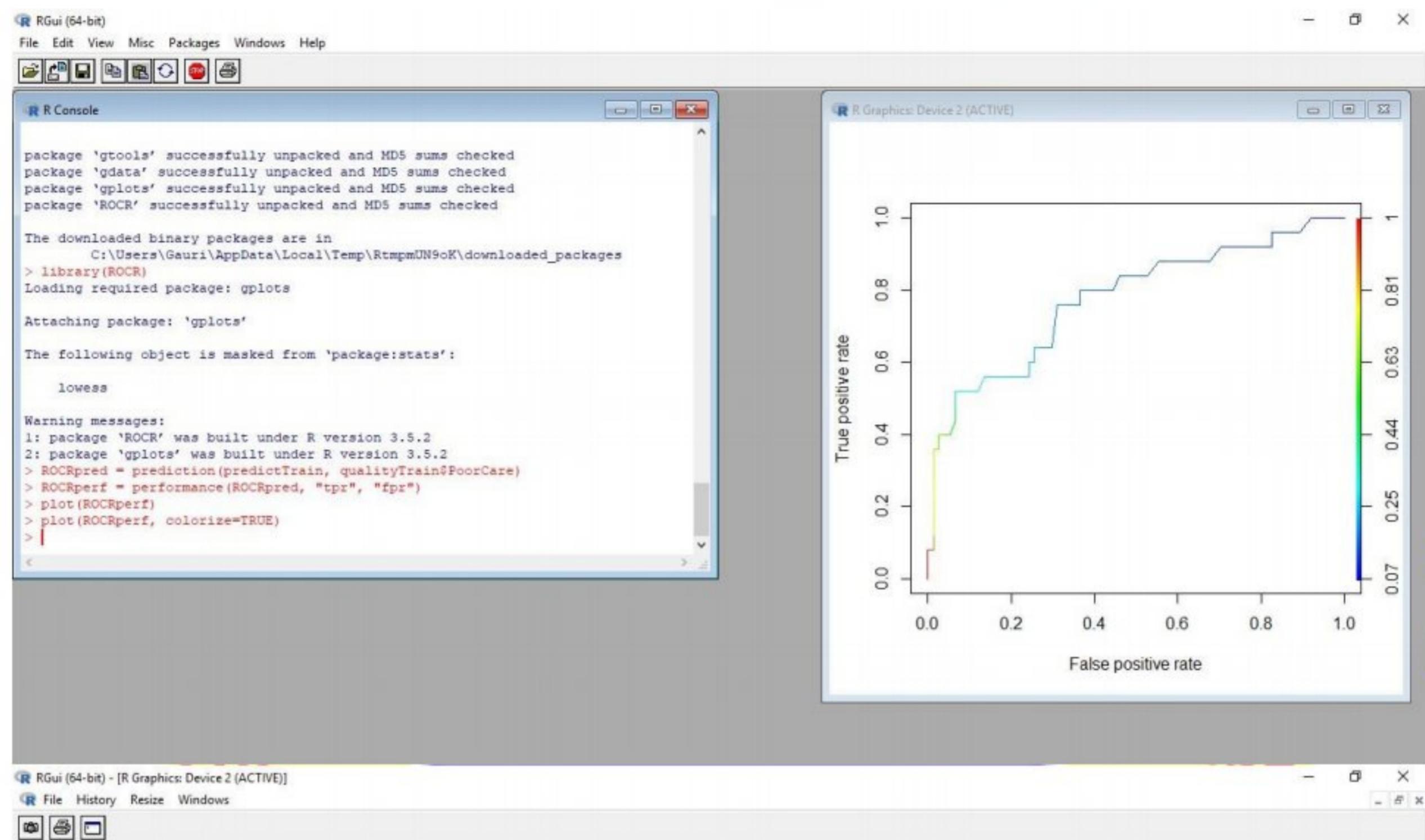
1: package 'ROCR' was built under R version 3.5.2

**2: package 'gplots' was built under R version 3.5.2**

```
> ROCRpred = prediction(predictTrain, qualityTrain$PoorCare)  
> ROCRperf = performance(ROCRpred, "tpr", "fpr")  
> plot(ROCRperf)  
> plot(ROCRperf, colorize=TRUE)  
> plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))  
>
```

RGui (64-bit) - [R Graphics: Device 2 (ACTIVE)]  
File History Resize Windows





**Learning Outcomes:**

**Course Outcomes:**

**Conclusion:**

**Viva Questions:**

1. Define logistic regression.
2. Purpose of the logit function?
3. Outcome variable type?
4. Example use case?

**For Faculty use:**

Correction Parameters	Formative Assessment[40%]	Timely Completion of Practical[40%]	Attendance Learning Attitude[20%]
-----------------------	---------------------------	-------------------------------------	-----------------------------------