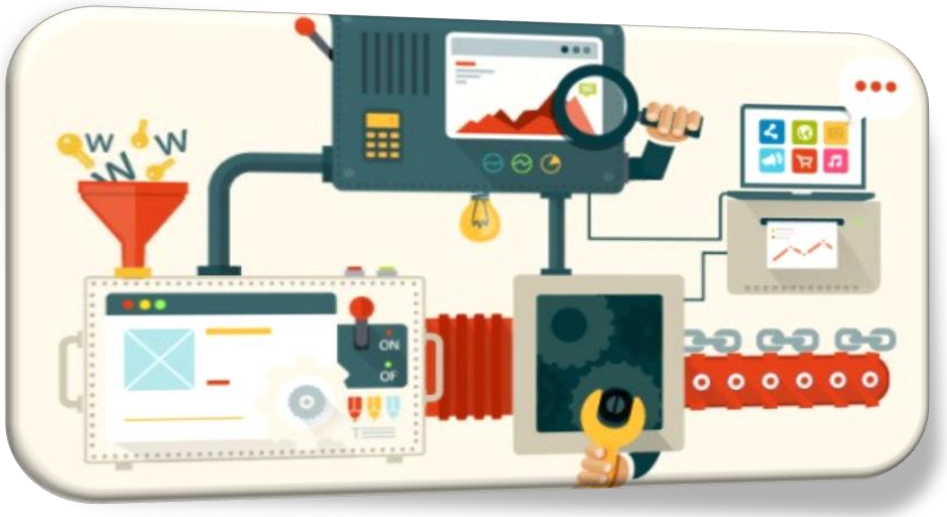
**Intelligent science and Technology Series**

# 《Data Warehousing and Data Mining》 Course Project Tutorial

**(2019 revised Edition)**

**Authors**: Muhammad Hassaan Farooq Butt



**2019.03**

# Content


Experiment 1. Building Data Warehouse and Analyzing Data with OLAP ..... - 1 -
Section One: Several Important Concepts ..... - 1 -
Section Two: The Brief Introduction of ETL Tools ..... - 3 -
Section Three: The Brief Introduction of OLAP ..... - 4 -
Section Four: Preparation ..... - 4 -
Experiment 1.1 Using SSIS for ETL on Data ..... - 7 -
Section One: The Goal of the Experiment ..... - 7 -
Section Two: The Content of Experiment ..... - 7 -
Section Three: The Procedure of Experiment ..... - 7 -
Experiment 1.2 ..... - 35 -
Building a Data Cube ..... - 35 -
Section One: The Goal of the Experiment ..... - 35 -
Section Two: The Content of Experiment ..... - 35 -
Section Three: The Procedure of Experiment ..... - 35 -
Experiment 1.3 ..... - 43 -
The OLAP for Data Analysis ..... - 43 -
Section One: The Goal of the Experiment ..... - 43 -
Section Two: The Content of Experiment ..... - 43 -
Section Three: The Procedure of Experiment ..... - 43 -
Experiment 1.4 ..... - 49 -
Build a data warehouse to manage 'Frequent-Flyer Flight Segment' information which is your homework ..... - 49 -
in Chapter 4. ..... - 49 -
Experiment 1.5 ..... - 49 -
Do OLAP based on the data warehouse of Exp 1.4 ..... - 49 -
Experiment 2. Data Mining ..... - 50 -
Section One: Several Important Concepts ..... - 50 -
Section two: Techniques of Data Mining ..... - 50 -
Experiment 2.1 Mining Association Rules ..... - 51 -
Section One: The Goals of Experiment ..... - 51 -
Section Two: The Content of Experiment ..... - 51 -
Section Three: The Procedure of Experiment ..... - 51 -
Experiment 2.2 K-Means Clustering Algorithm ..... - 61 -
Section One: The Goal of Experiment ..... - 61 -
Section Two: The Content of Experiment ..... - 61 -
Section Three: The Procedures of Experiment ..... - 61 -
Experiment 2.3 Classification Algorithm ..... - 67 -
The Goal of Experiment ..... - 67 -
Requirements of Your Experiment Report ..... - 68 -

# Experiment 1. Building Data Warehouse and Analyzing Data with OLAP

## Section One: Several Important Concepts

### 1. Data Warehouse

Data warehouse is designed for providing a tool or an environment which is a need for the whole enterprise to transform an operation system into a decision support system. It tries to solve problems including what is the topic data and how to extract those data from traditional operational processing system and how to convert the dispersed, inconsistent operational data into an integrated, low noise data.

A data warehouse is a subject-oriented, integrated, non-volatile and time-variant collection of data in support of management decisions. The data warehouse is not a product that you can buy but a kind of data storage scheme.

The concept of data warehouse can be understood in the following steps: First, a data warehouse is used to support decisions and process analysis-oriented data, and it is different from the operational database, which is used to improve transaction efficiency. Second, the data warehouse is used to operate several heterogeneous data sources which is distributed in the enterprise and to save the data model in a new way, according to the topic of decisions. Finally, the data saved in the data warehouse usually won't be modified.

### 2. Data Cube

Data cube is a multidimensional data set that consists of all tables or part of tables extracted from the data warehouse. It is a multidimensional matrix which let users analyze data warehouse from multiple perspectives.

### 3. ROLAP and MOLAP

ROLAP is based on the relationship model of online analytical processing system. In this model, data is saved in the forms of rows and columns. However, MOLAP is based on multidimensional model of online analysis processing system. The table below is a comparison between these two models.

Table 1.1 ROLAP vs. MOLAP

| | Storage | Basic Technology | Function and Performance |
|---|---|---|---|
| ROLAP | 1. Data is considered as relationship tables saved in the data warehouse.<br>2. You can obtain the sum of the data in details.<br>3. The storage of the data is very huge.<br>4. You can access all data from the storage of the data | 1. Obtaining data from the data warehouse is through complex SQL.<br>2. Using ROLAP engine to build the data cube.<br>3. The presentation layer can represent the multidimensional view. | 1. It is a famous environment and can obtain more tools.<br>2. It also has some limits in the complex function.<br>3. It is difficult to across the dimension down. |

| | | | |
|---|---|---|---|
| | warehouse. | | |
| MOLAP | 1. Data is considered as relationship tables that saved in the data warehouse.<br>2. Different kinds of the data are collected in the (multidimensional) database.<br>3. The storage of the data is moderate.<br>4. Access the total data from the multidimensional database and access the detail data from the data warehouse. | 1. Using MOLAP engine to build data cube in advance.<br>2. Saving the multi-dimensional view in the array through special technology instead of tables.<br>3. You can get the matrix data in the retrieve operation.<br>4. You can manage the summary of the sparse data though the technology of the sparse matrix. | 1. Much faster access.<br>2. It has a large library of functions to perform complex calculations.<br>3. You can easily analyze the data no matter how many dimensions it has. |

## 4. Star Schema, Snowflake Schema and Fact Constellations

Star schema is a relational data warehouse schema that contains a fact table and many dimensional tables. The fact table is placed at the center and joined to the dimensional tables that are not connected with each other. One of the star schemata is shown in Figure 1.

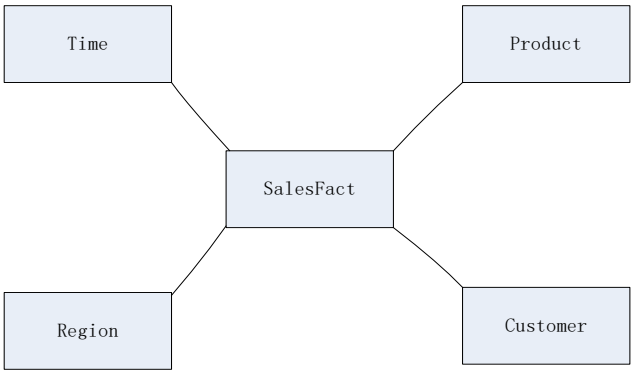



Figure 1. Star Schema

Snowflake schema is one of star schema. Some dimensional hierarchy is normalized into a set of smaller dimension tables, which forms a shape like snowflake. The snowflake schema is shown in Figure 2.

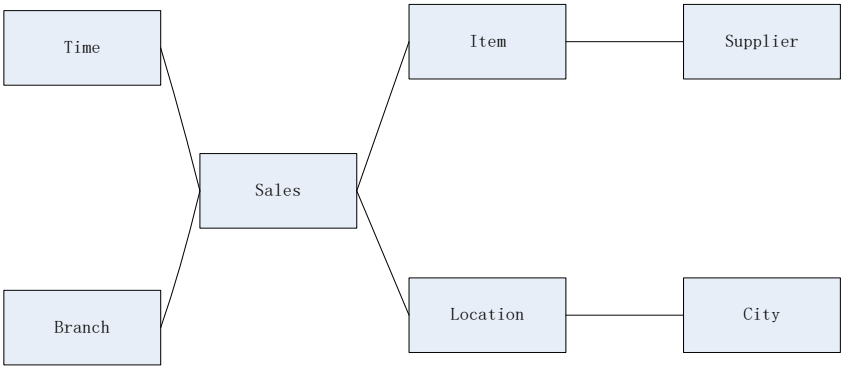



Figure 2. Snowflake Schema

Fact constellations schema is viewed as a collection of stars, in which multiple fact tables share dimension tables. That's why it's called galaxy schema or fact constellations. The fact constellations are shown in Figure 3.

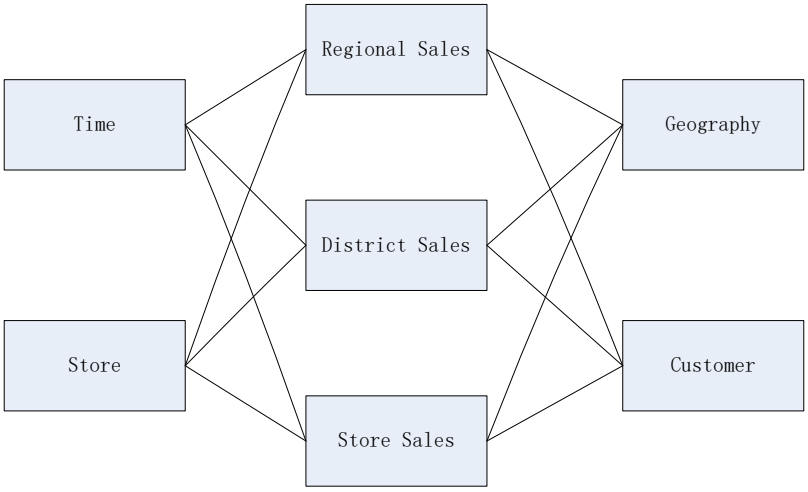



Figure 3. Fact Constellations Schema

## 5. The Process of ETL

You need to operate the data before you build a data warehouse. Here, you use ETL techniques which contain extraction, translation and loading.

## 6. The Analysis and Operations of OLAP

There are four steps for the analysis and operations of OLAP:

(1) **Slice:** Slice refers to creating a specific projection of the multidimensional data space.

(2) **Dice:** Dice refers to creating an extraction of the multidimensional data space without necessarily reducing the dimensionality of data.

(3) **Drill:** Drill contains roll-up and drill down. Roll-up removes detail typically by dropping a report header and provides an aggregation of the more detailed data, and Drill down is the counter-operation of Roll-up. Here, you do not give more details about Drill-up.

(4) **Pivot:** Pivot reorganizes the result space by reorganizing the dimensional attributes and the operation typically performed by the front-end reporting tool.

# Section Two: The Brief Introduction of ETL Tools

The tool that you use in this experiment is SQL 2008, and the ETL tool is SQL Server Integration Services (SSIS) accordingly. It is often used to extract, transform and load data. ETL tool can complete some tasks as follows:

(1) Extract data from lots of relationship databases that belong to some advanced companies.

(2) Extract data from the old databases, index files and flat files.

(3) Transform source field to target field in different format.

(4) Implement the process of normalized transformation and of redefining the keys and the change of structure.

(5) Provide the route to check data from source to target.

(6) Apply the rules of extraction and transformation.

(7) Add serval records into a whole target record in source system.

(8) Record and manage the metadata.

# Section Three: The Brief Introduction of OLAP

SQL Server 2008 Analysis Services, the abbreviation of which is SSAS, is used for analyzing data in SQL Server 2008. It is very complicated to use, so it requires the users must have a lot of background knowledge to analyze data warehouse and design multidimensional data.

The data cube is a collection of multidimensional data and it consists of all or parts of tables which are extracted from the data warehouse. OLAP is a process of accessing data source by multidimensional structures and organizing or gathering the data. Analysis Services is a tool used for building and managing the multidimensional data and analyzing the cube. You can use the traditional bottom-up design methods in this tool; besides it also supports the top-down design methods.

In addition, OLAP also has the following functions:

(1) It can achieve the multidimensional display of the data.

(2) It can achieve some operations such as accumulation, summary, calculation in advance and deduction.

(3) It has some formulae and complicated computation in the library.

(4) It can achieve the computation across different dimensions.

(5) It can achieve some operations including pivoting, crossing tables, drilling down and generalizing according to one dimension or multiple dimensions.

# Section Four: Preparation

1. Install the SQL server 2014, and open Microsoft SQL Server Management Studio, then you will see the following interface. Set the server's name, then click the button. It will connect to your local server.

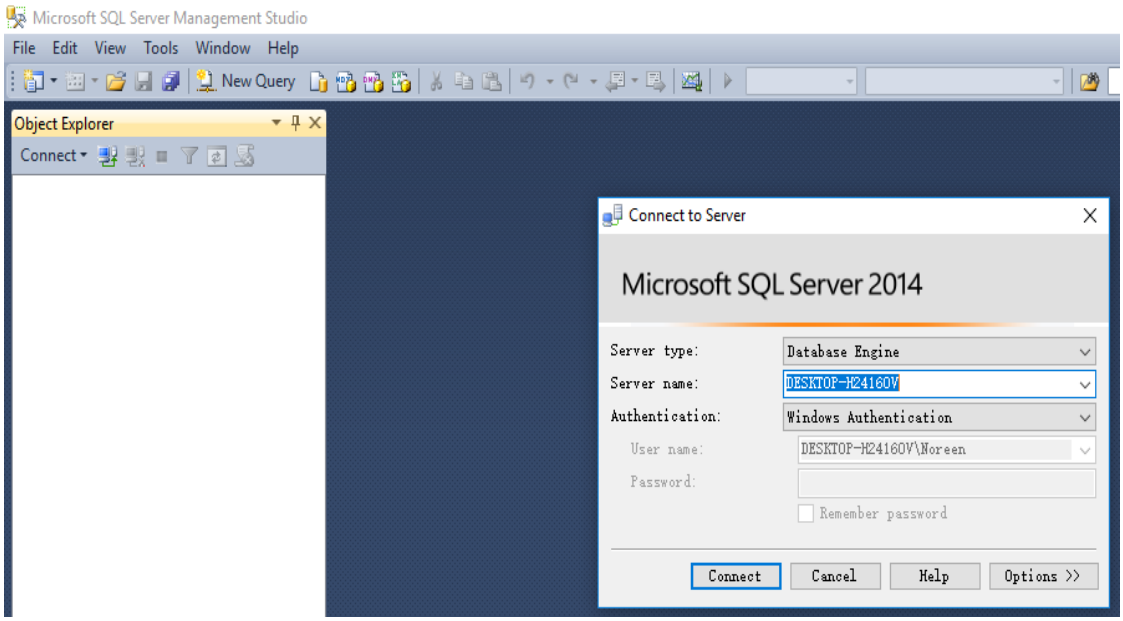



Figure 4. Connection to Server

2. Right-click Database to fold the panel below. Click “Restore Database…” item. Restore “AdventureWorks” database by “AdventureWorksDW.bak” file. You can do it as follows.

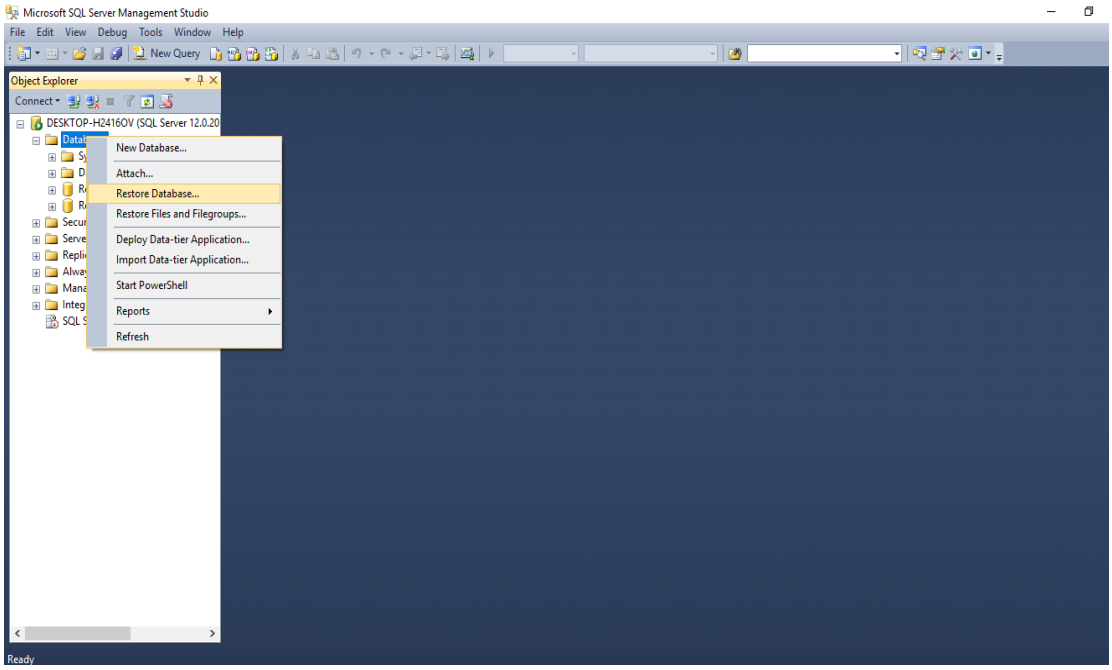



Figure 5. Database Restoration Menu

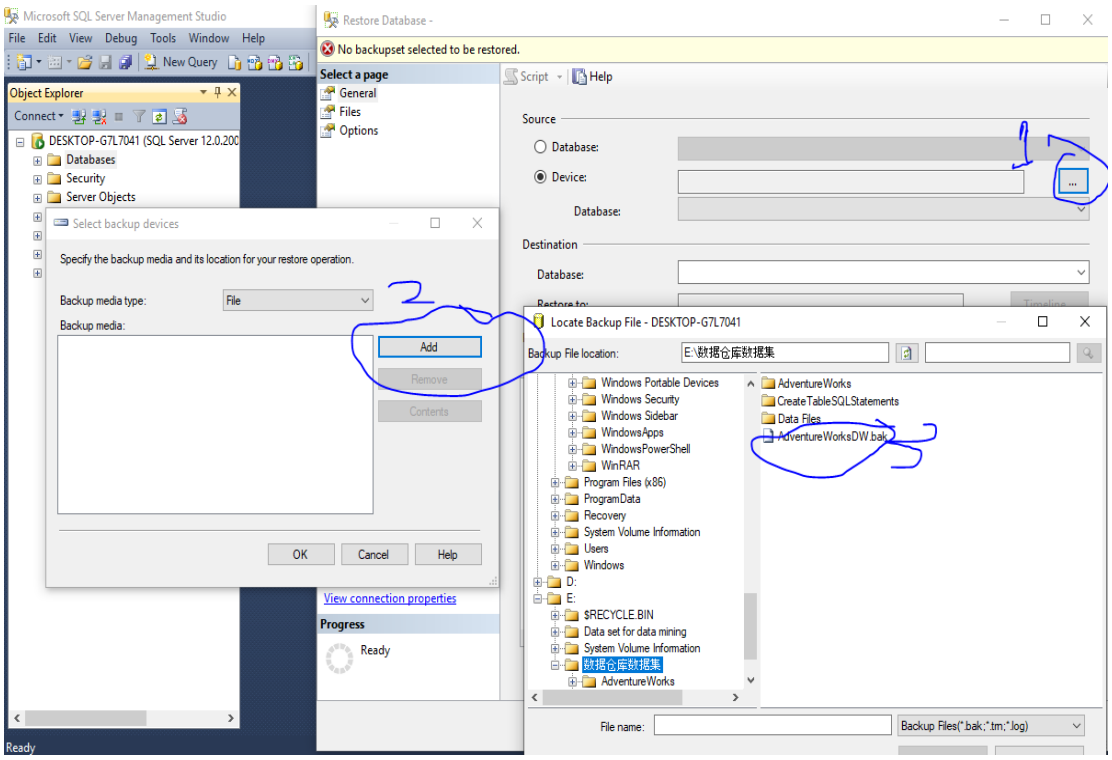



Figure 6. Database Restoration

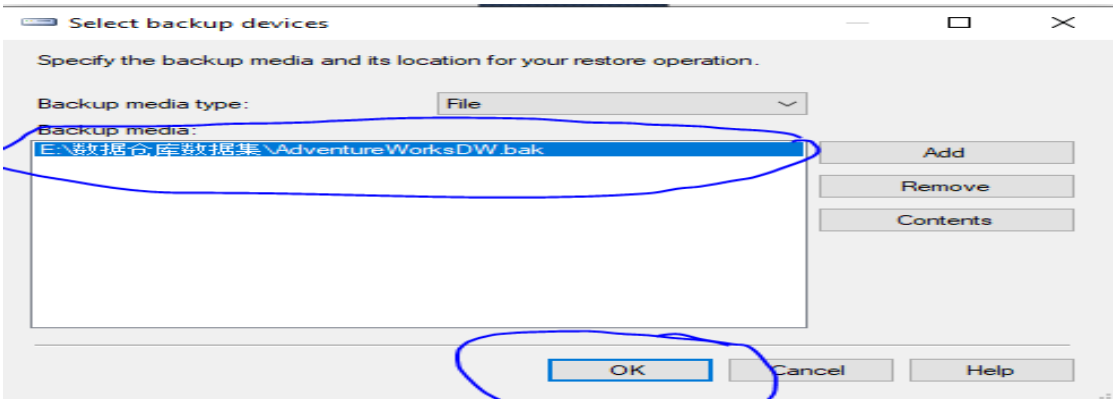



Figure 7. Restoration Settings

3. If you succeeded to restore the database specified, you will get the following figure.

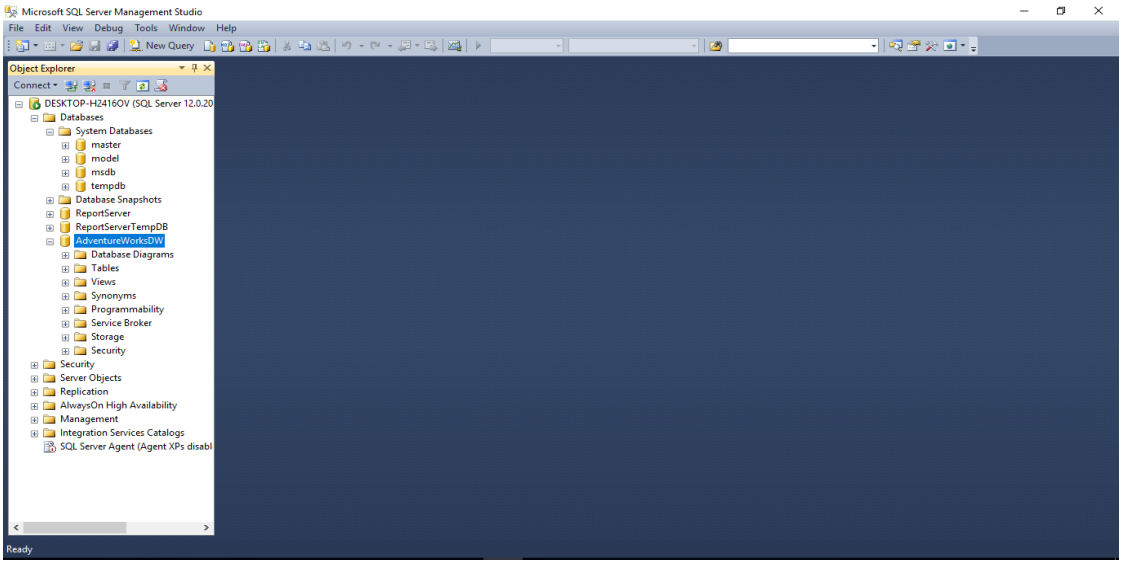



Figure 7. “AdventureWorks” Database

# Experiment 1.1 Using SSIS for ETL on Data

## Section One: The Goal of the Experiment

(1) Be familiar with the use of SSIS.

(2) Deepen the understanding of ETL (Extraction, Transform and Loading).

## Section Two: The Content of Experiment

The experiment is based on the company of Adventure Works Cycle. The company wanted to add five new sales regions, but the sales data of these regions were not collected to a database before. Now the company needed these data, so it requires the supervisors of the five regions to import all the information of customers into a text file named *customers.txt*. The data of the five regions mixes together completely, even some of them are invalid. Now the main task is to use the SSIS tool to import the data of *customers.txt* into database by regions. You should do your process of Extraction, Transform and Loading, and save the error data in a file at the same time.

## Section Three: The Procedure of Experiment

### 1. Generate the Solutions to SSIS

In order to generate the solutions for SSIS, you need to follow these steps:

(1) Open SQL Server Business Intelligence Development Studio choose 【File】 →【New】→【Project】, choose "Integration Services Project". Here you name the project as "alldemo" (you can name it by yourself) ， and then【OK】.

(2) In Solution Explorer, below the "SSIS packages" is "Package.dtsx". You should rename it as "alldemo.dtsx" which is easy to remember.

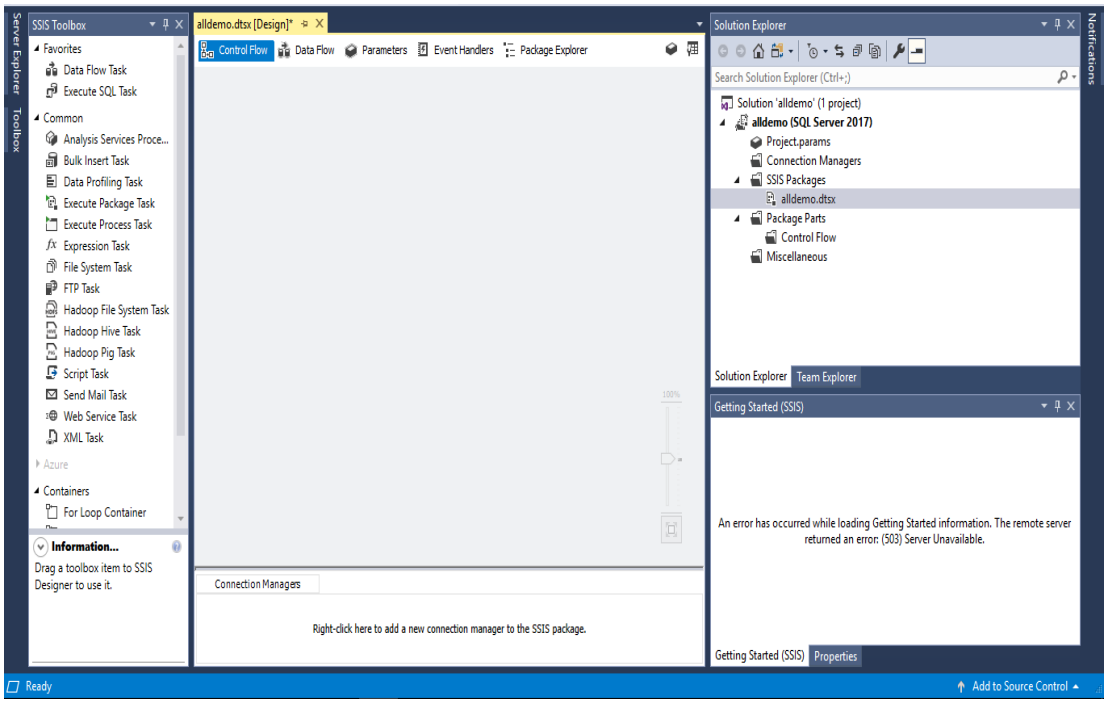



Figure 1.1.1 "alldemo" Project Panel

(3) Right-click in the area of “Connection Managers” arbitrarily. Choose “New OLE DB Connection”.

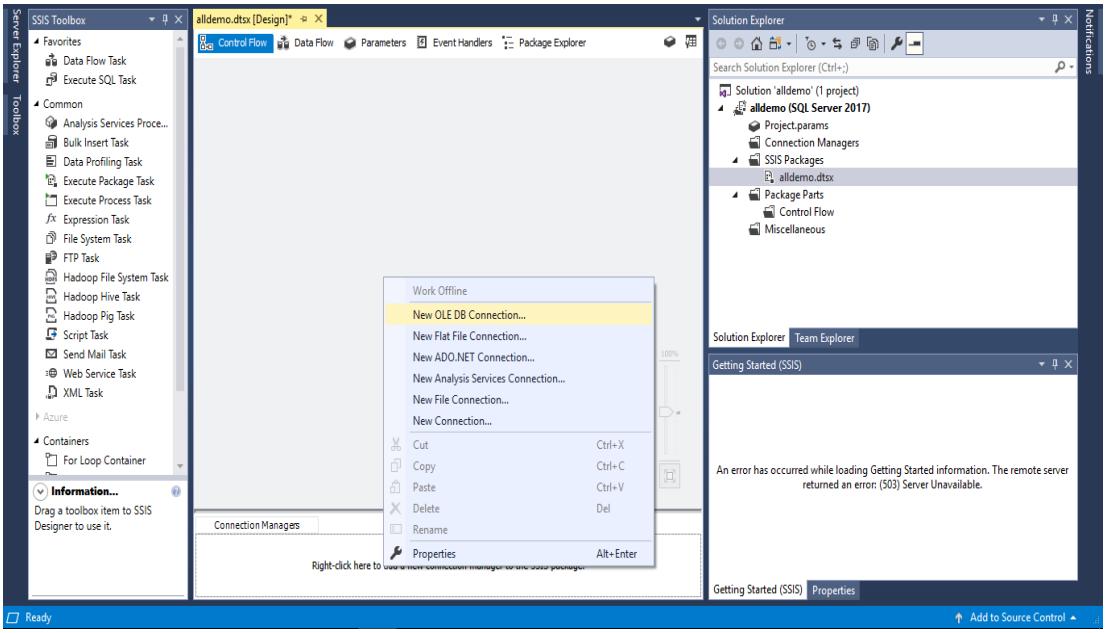



Figure 1.1.2 Creation of OLE DB Connection

Click【New】.

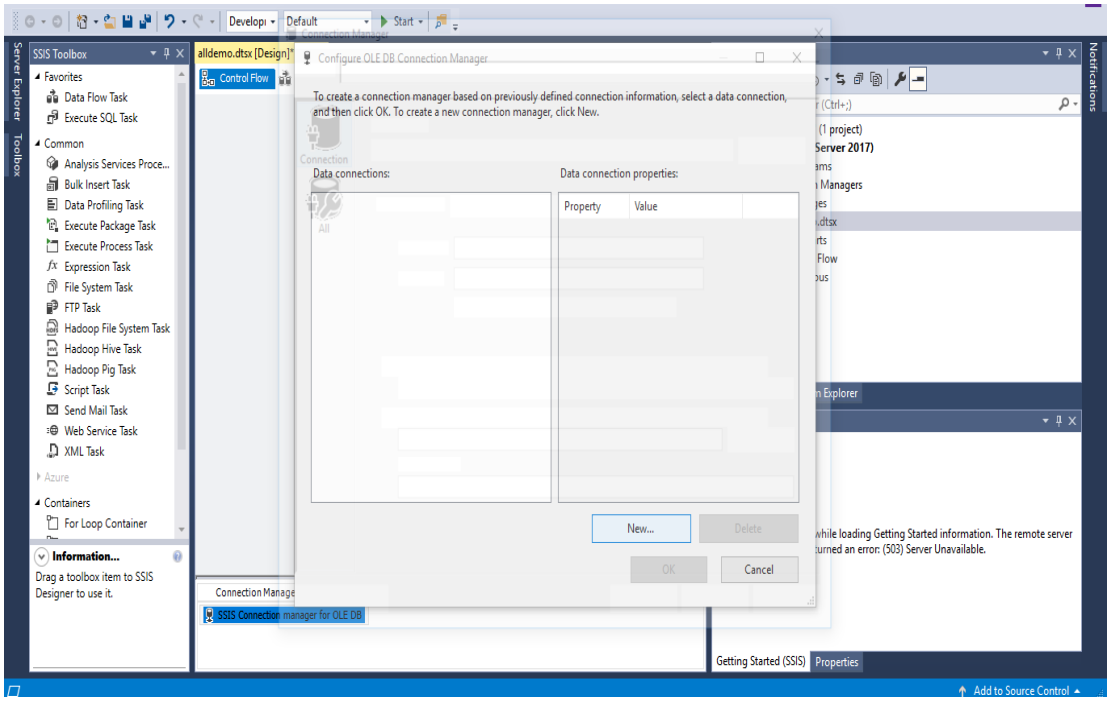



Figure 1.1.3 Creation of a Data Connection

Once in the Dialog box of “Connection Manager for Configuring OLE DB”. Then choose AdventureWorks database of the corresponding server (the name of computer).

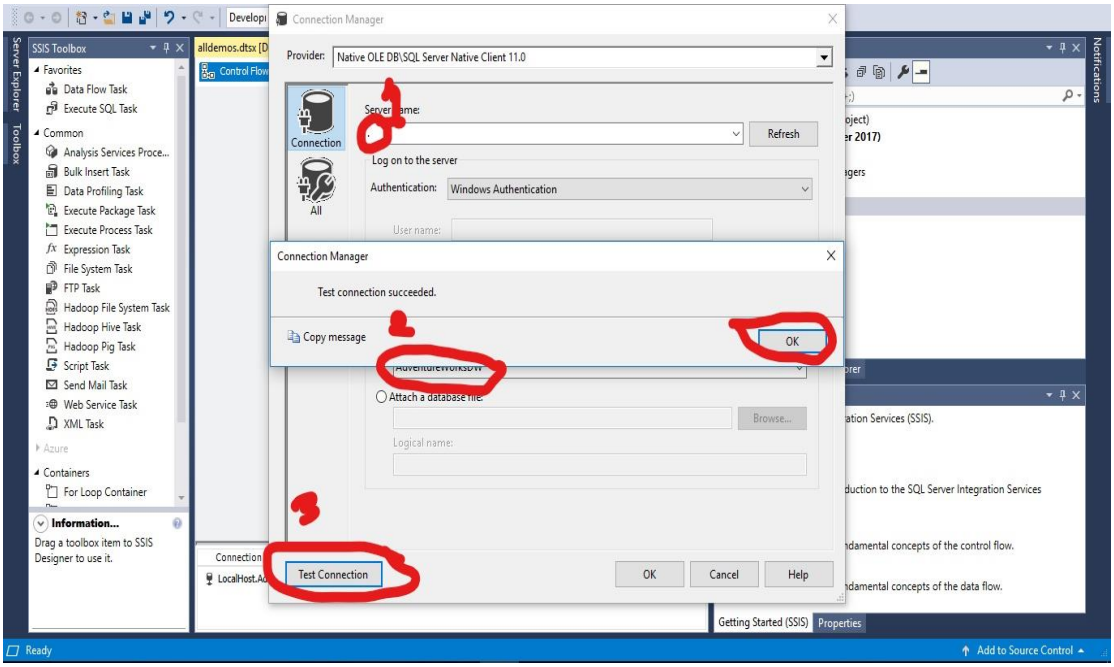
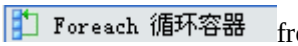



Figure 1.1.4 Connection Settings

After setting well, you can test the connection and successfully connection established shows as in picture.

## 2. The Designment of Control Flow

(1) Drag the Component 

from Toolbox to Control Flow, and name it as "*Foreach Cycle Operation SQL Sentence*".

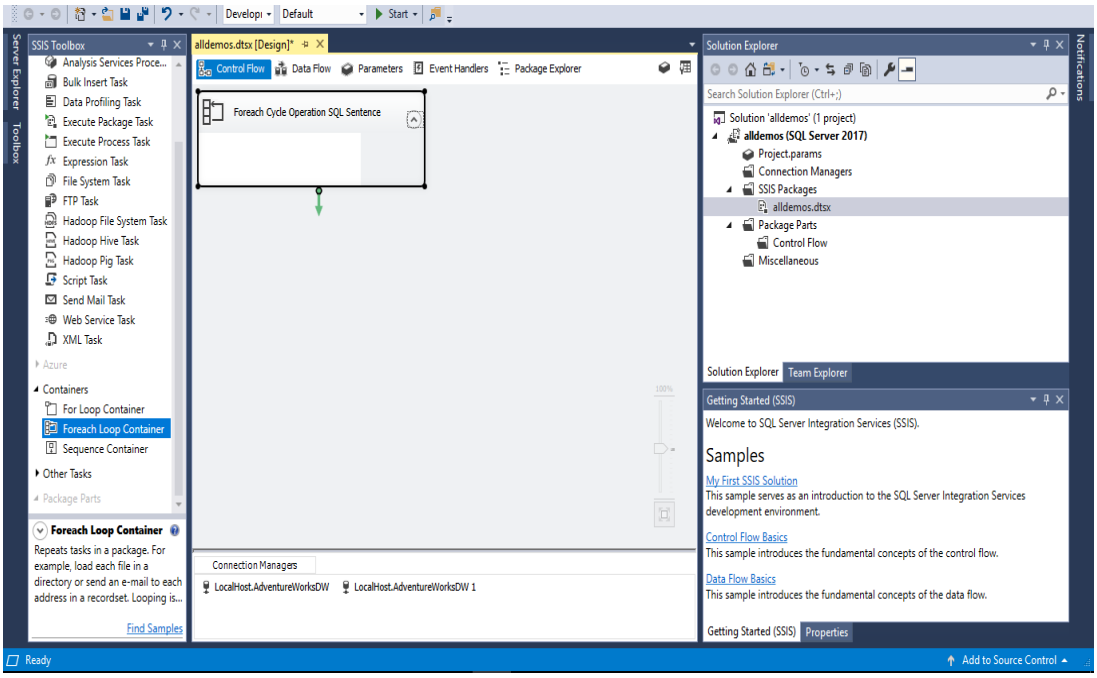



Figure 1.1.6 Drag the Foreach Loop Container

Right-click the Foreach Loop Container choose “Edit” and click on “Set”. In Foreach Loop Editor, you can find “Collection” item. Do the same settings as follows.

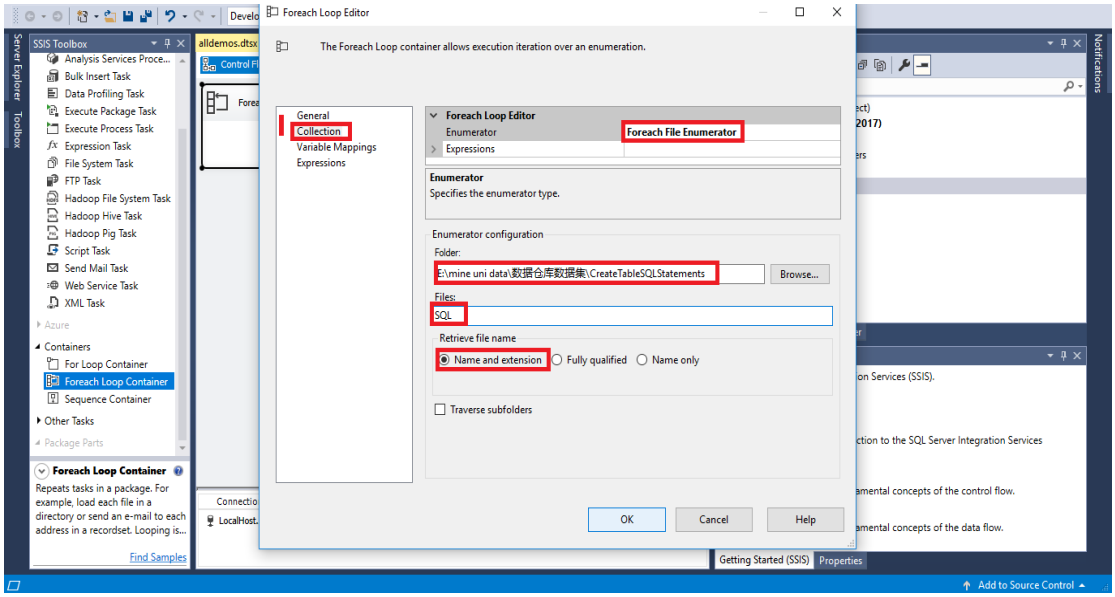



Figure 1.1.7 Collection Setting

(2) Click “Variable Mappings” item. Choose "<New variable...>" in the drop-down list of the “Variable” and set in a pop-up window shown in the following figure. In this step, you will create a user variable named “*vfileName*”. Click 【OK】to return to “Control Flow” window.

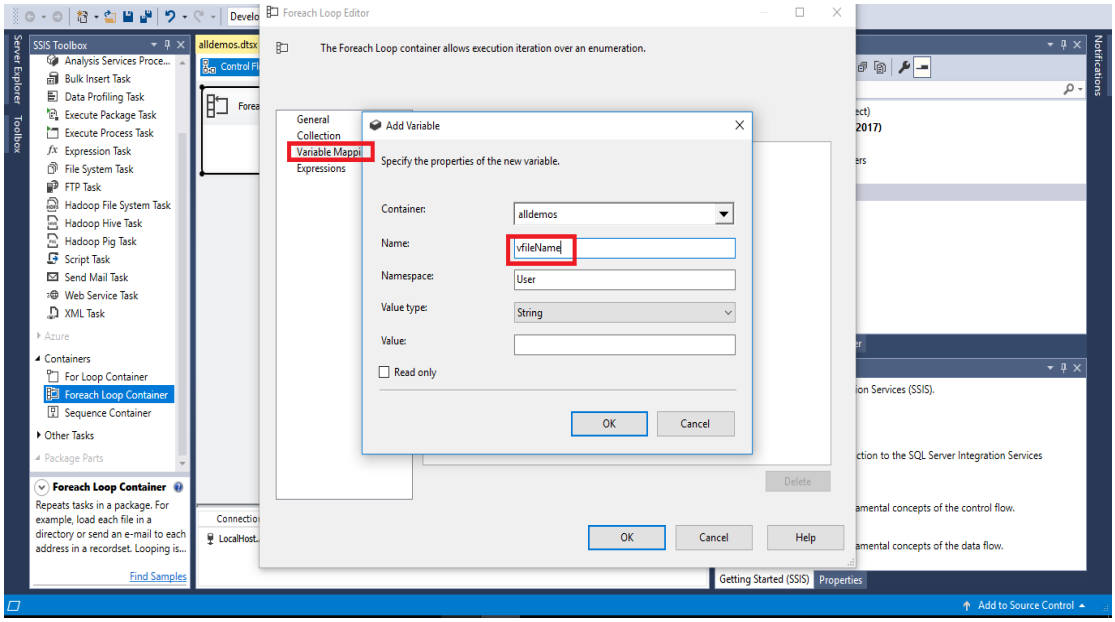



Figure 1.1.8 Variable Mappings Setting

(3) With the intention of making the Foreach Loop operate successfully, you need to create a connection file of SQL. Right-click in the Connection Managers select 【New File Connection】. Then choose “Existing file” for “Usage type”. Click “Browse…” to find SQL files which are the data set of the project.

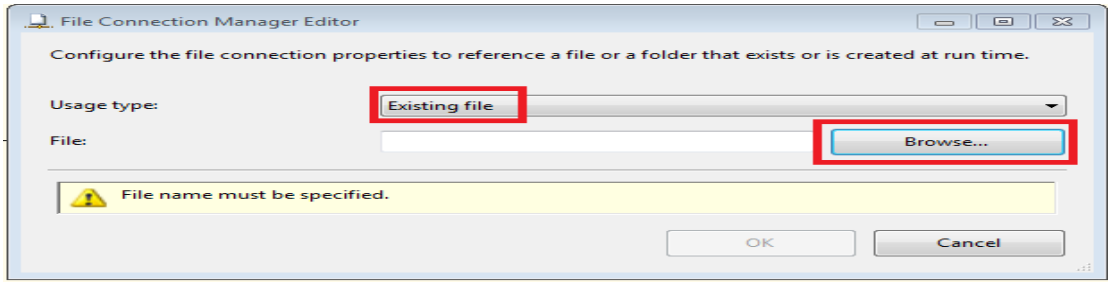



Figure 1.1.9 Connection to File

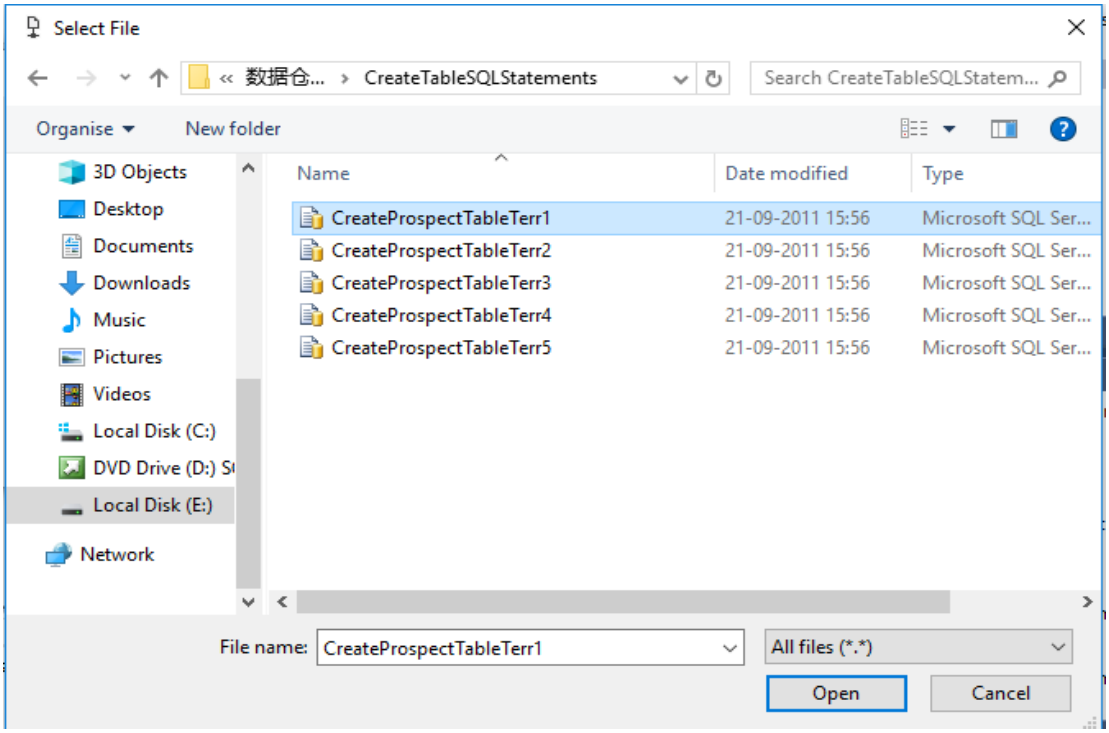



Figure 1.1.10 SQL File Exploration

(4) Modify the attribute of the file connection named "*CreateProspectTableTerr1.sql*". Choose "Expression".

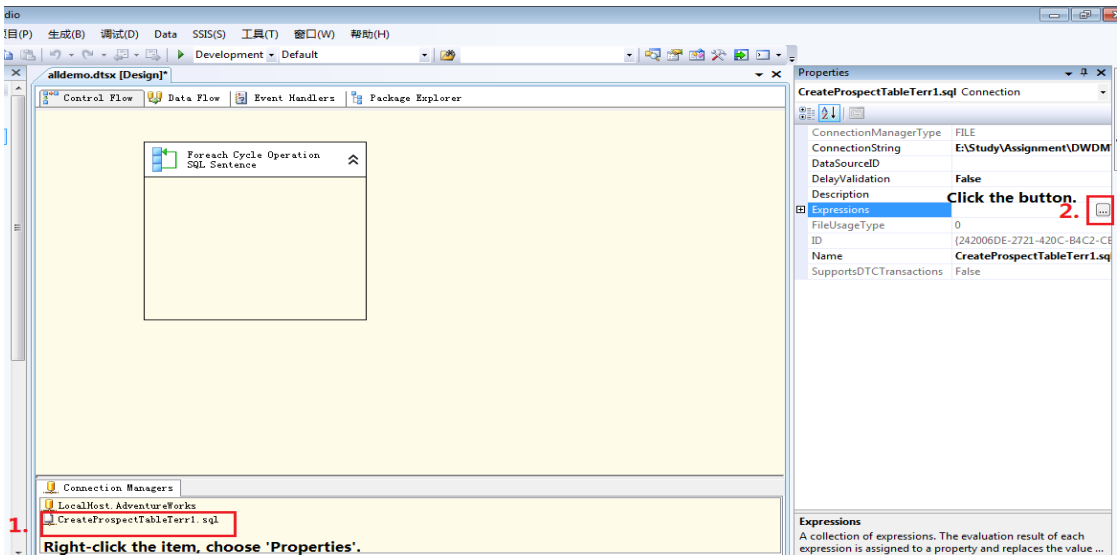



Figure 1.1.11 SQL File Connection Setting

In Property Expressions Editor, choose “ConnectionString”, and click “…”.

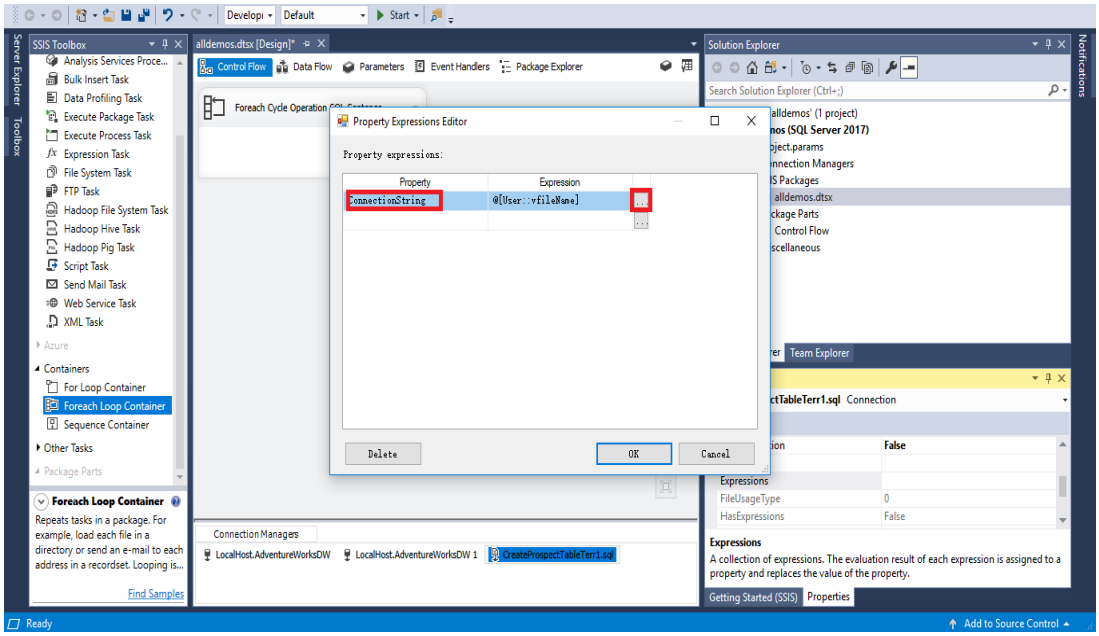



Figure 1.1.12 Property Expression Setting

Then, you will open Expression Builder window, drag the variable named “User::vfileName” to the blank frame of “Expression”.

Attention: if you can’t run your experiment because of losing sql file, you can try to change the ‘expression’ to the sql path with vfileName. Eg, change

@[User::vfileName to

"D:\\SQL\\CreateTableSQLStatements\\"+@[User::vfileName]

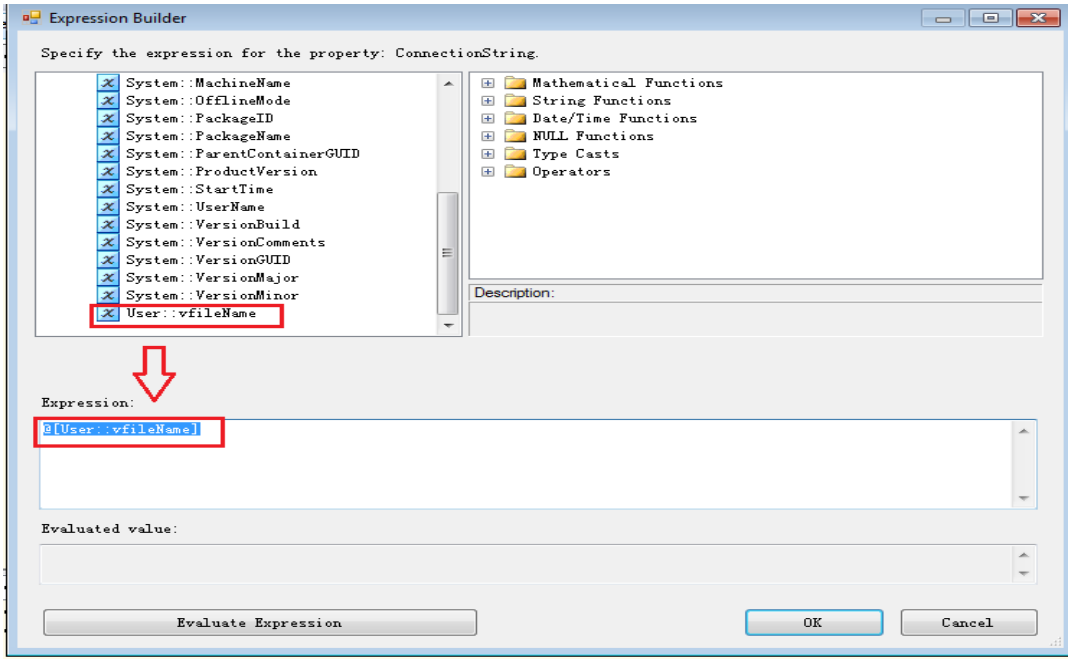



Figure 1.1.13 Specify the Expression for the Property

(5) Choose 

in the pane of "Toolbox", drag it into the foreach loop container which named “Foreach Cycle Operation SQL Sentence”.

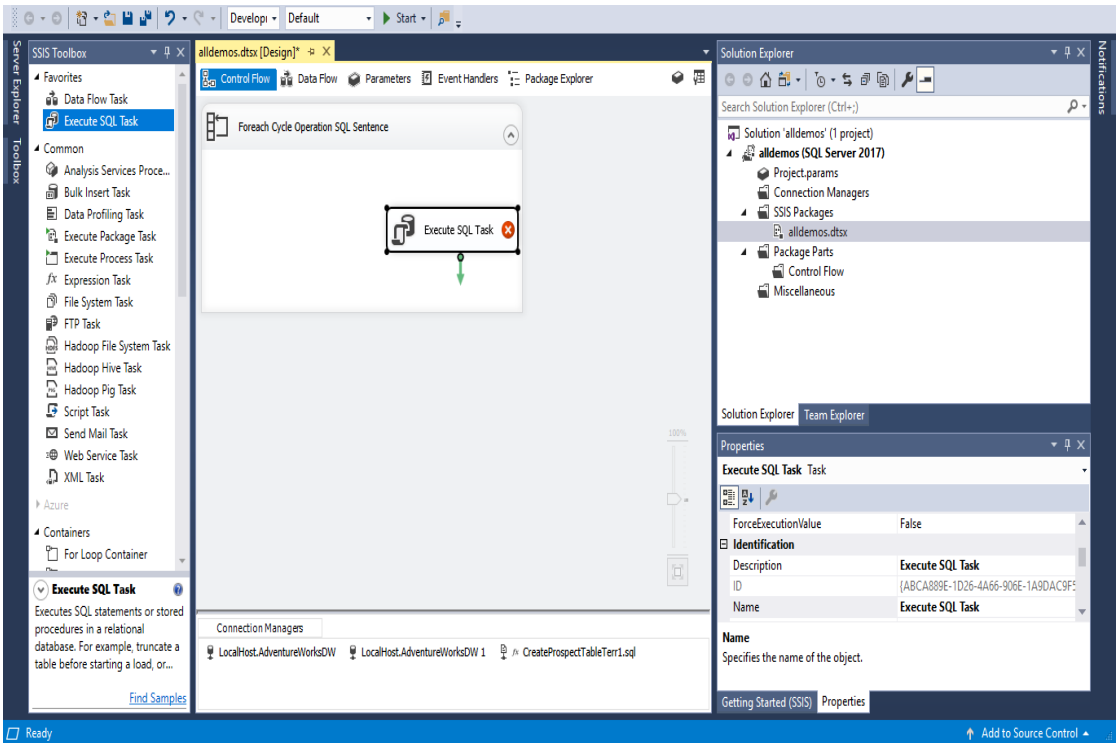



Figure 1.1.14 Drag the Execute SQL Task Component

Double-click the component you added just now, set its properties as follow.

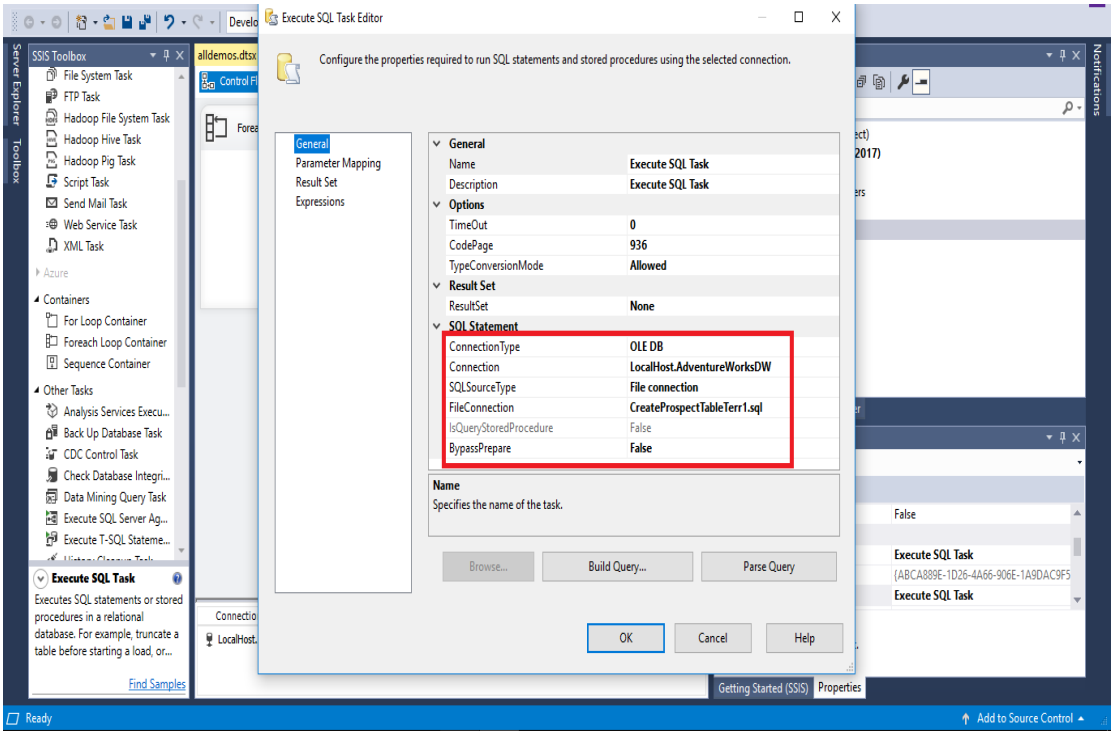
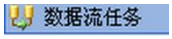



Figure 1.1.15 General Setting

(6) Choose the component 

in the pane of "Toolbox", drag it to the Control Flow.

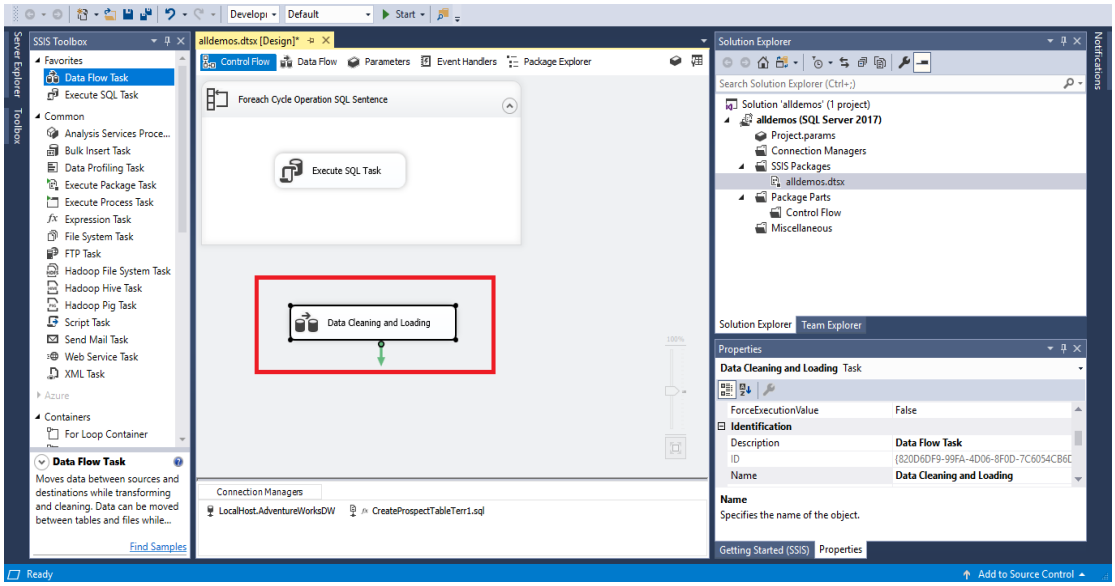



Figure 1.1.16 Drag the Data Flow Task Component

Rename it as "*Data Cleaning and Loading*". Then let precedence constraints from the Foreach Loop Container point to the data flow task (Right-click on "*Foreach Cycle operation SQL Sentence*" and choose "*adding Precedence Constraints*").

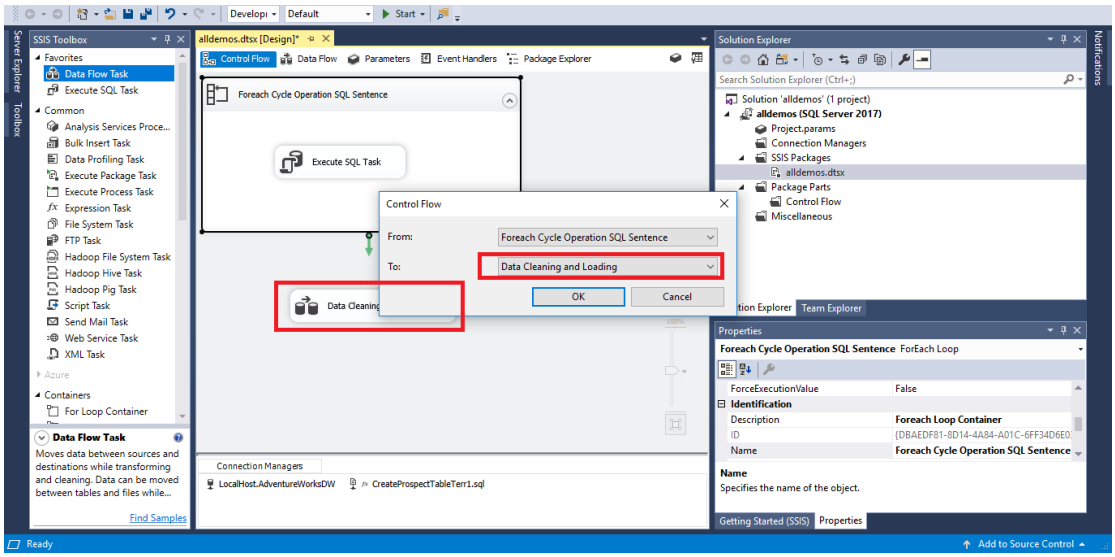
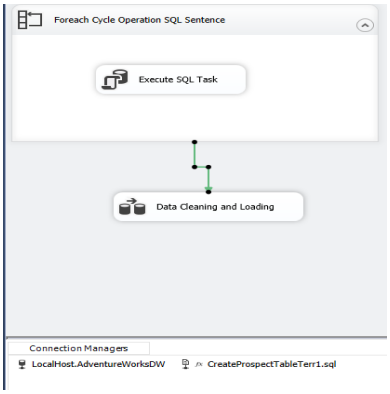



Figure 1.1.17 Connect Two Components



# 3. The Designment of Data Connection

(1) Create a flat file connection in the Connection Managers.

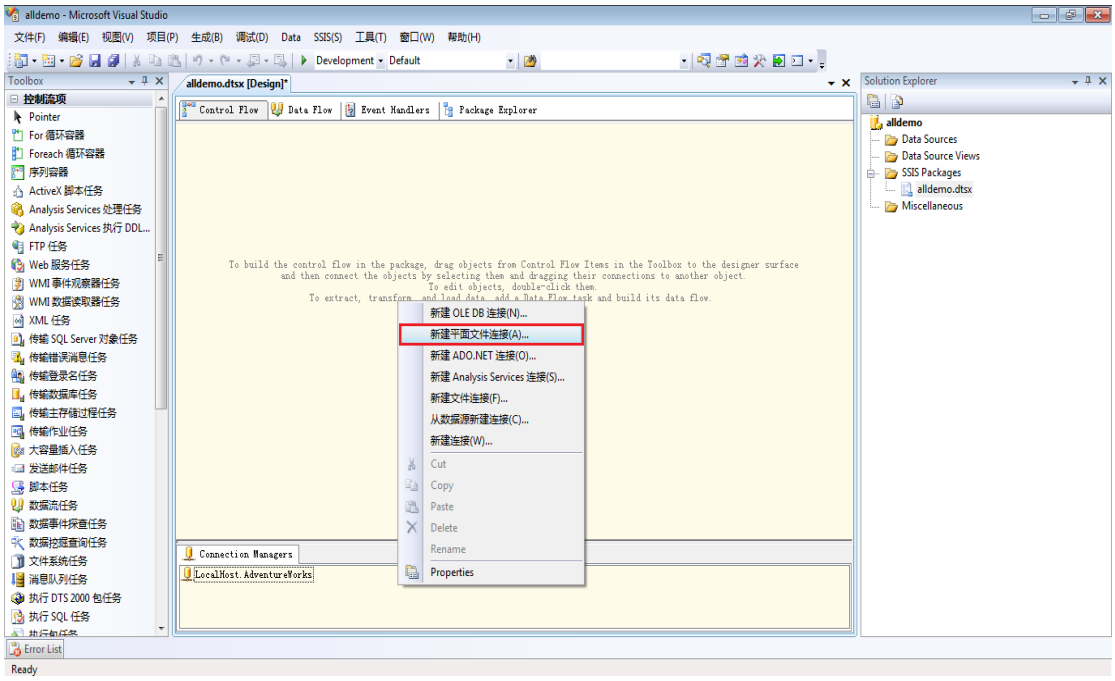



Figure 1.1.18 The Creation of Flat File Connection

You should do the same settings as follow.

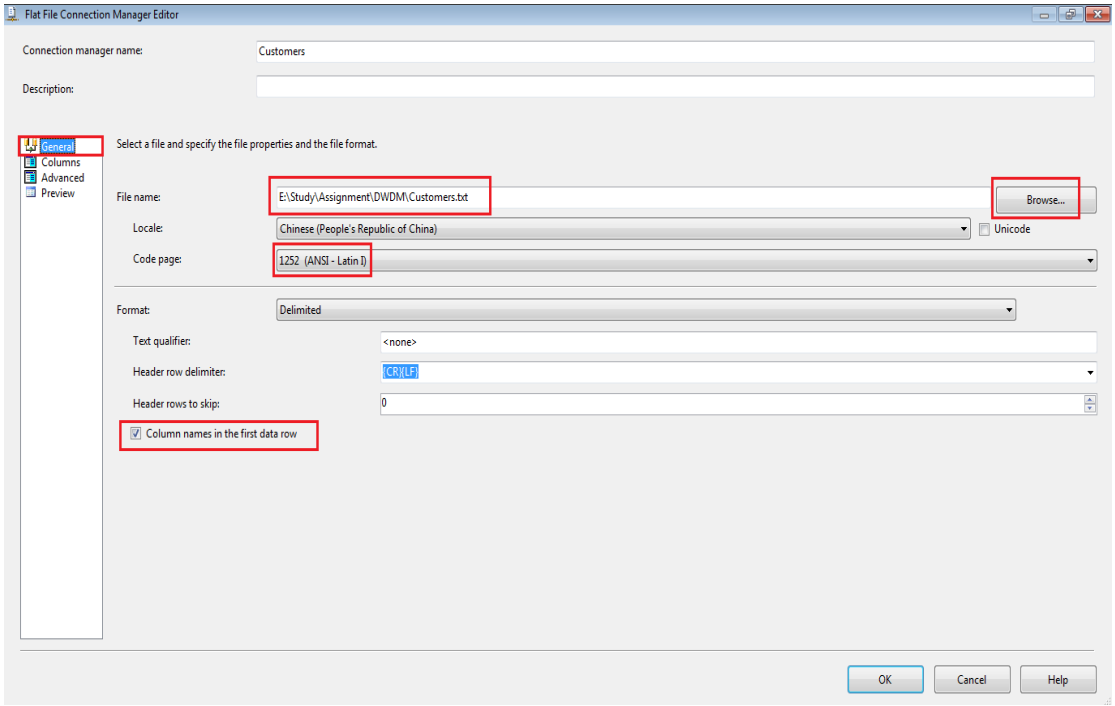



Figure 1.1.19 “Customers.txt” File’s General Setting

(2) Click "Advanced" item to configurate each column’s properties. Here you can set the name and width, of course you don’t need to modify every column. However, when you encounter warnings or errors in the following steps, you must return to this page for modifying. The parameters you must modify are shown as follows.

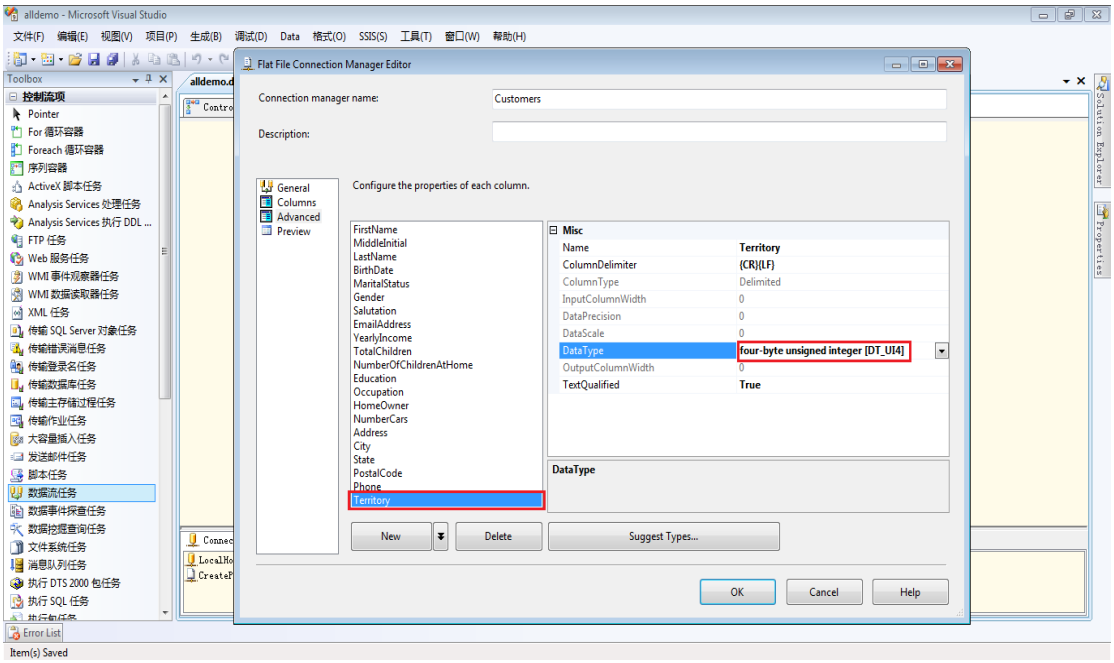



Figure 1.1.20 Territory Modification

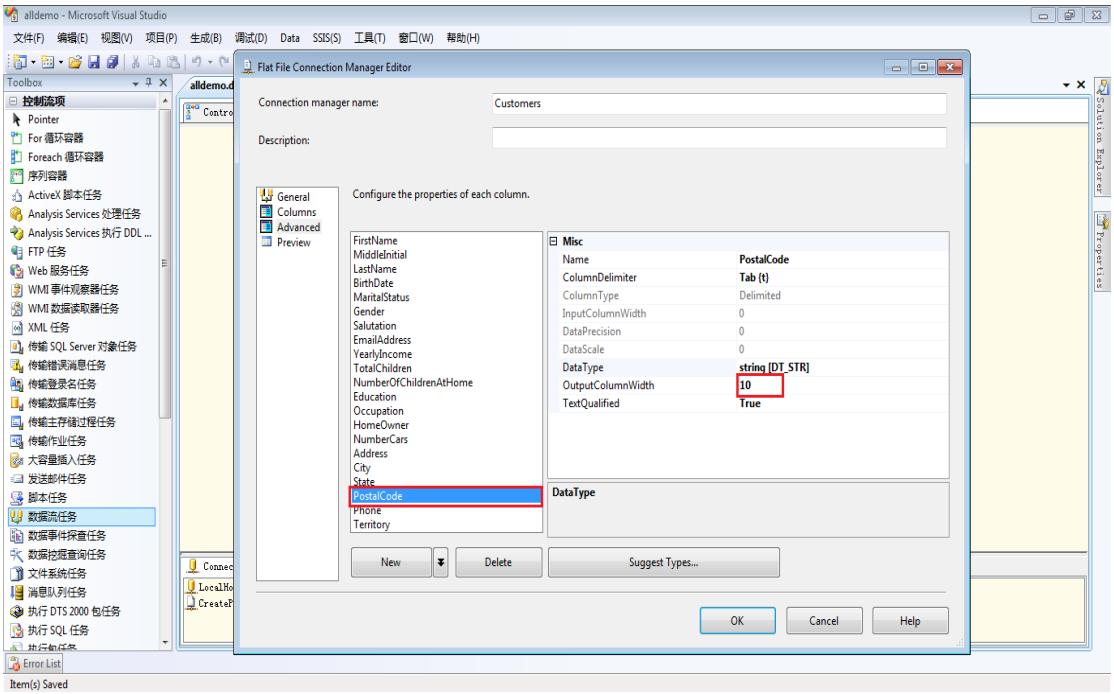



Figure 1.1.21 PostalCode Modification

It's neccessary for you to create another connection. It should point to a text file for saving invalid customer's data. After you right-click in Control Managers to create a "New Flat Connection", set its properties as follows.

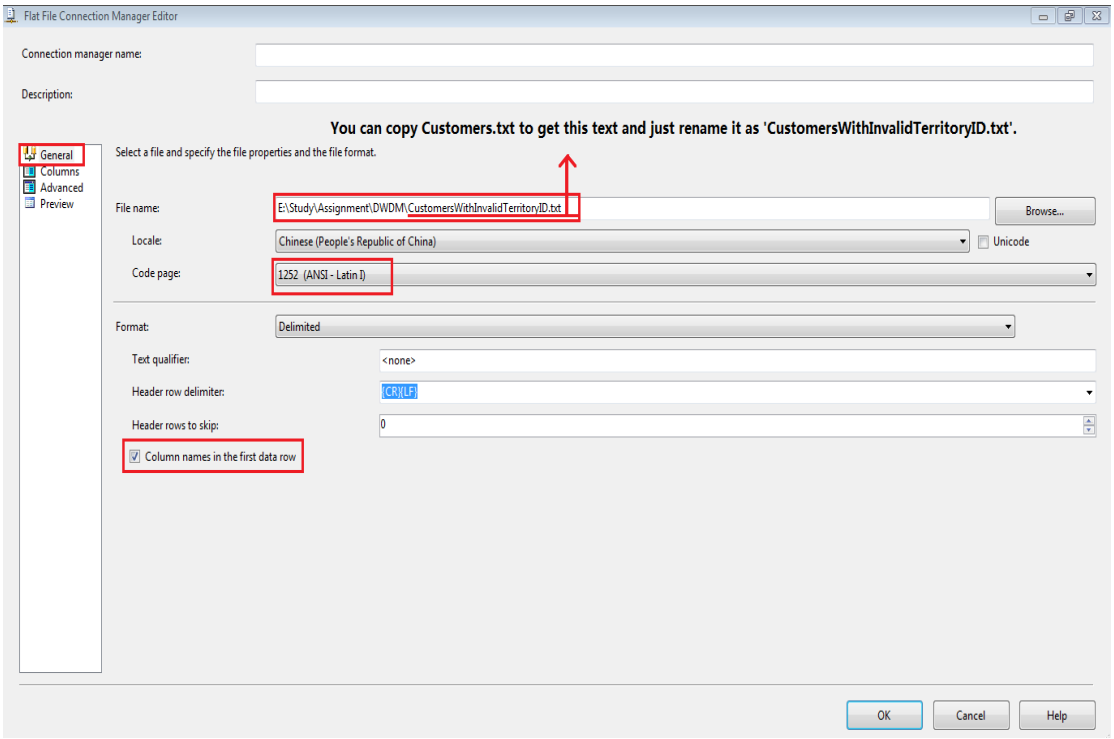



Figure 1.1.22 "CustomersWithInvalidTerritoryID.txt" File's General Setting

After you create the connections of the project, the Connection Managers panel should be the same as follow.

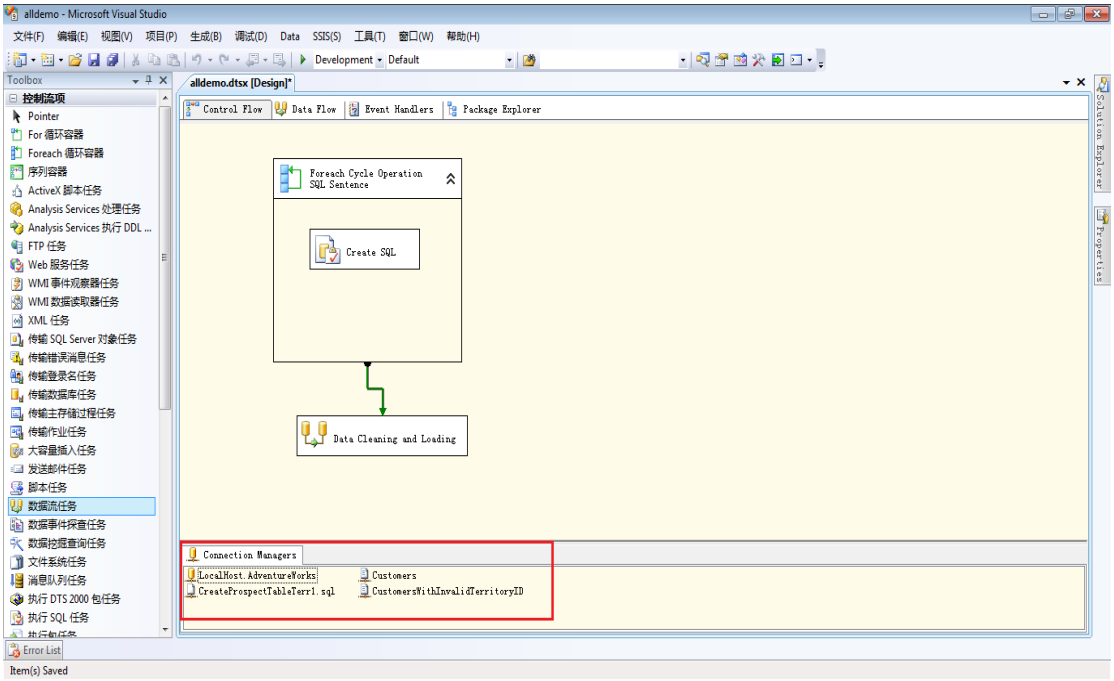



Figure 1.1.23 The Entire Connection Managers Panel

# 4. The Designment of Data Flow

Double-click the "*Data Cleaning and Loading*" component, then do it as follows:

(1) Drag the component 平面文件源 to the Data Flow panel.

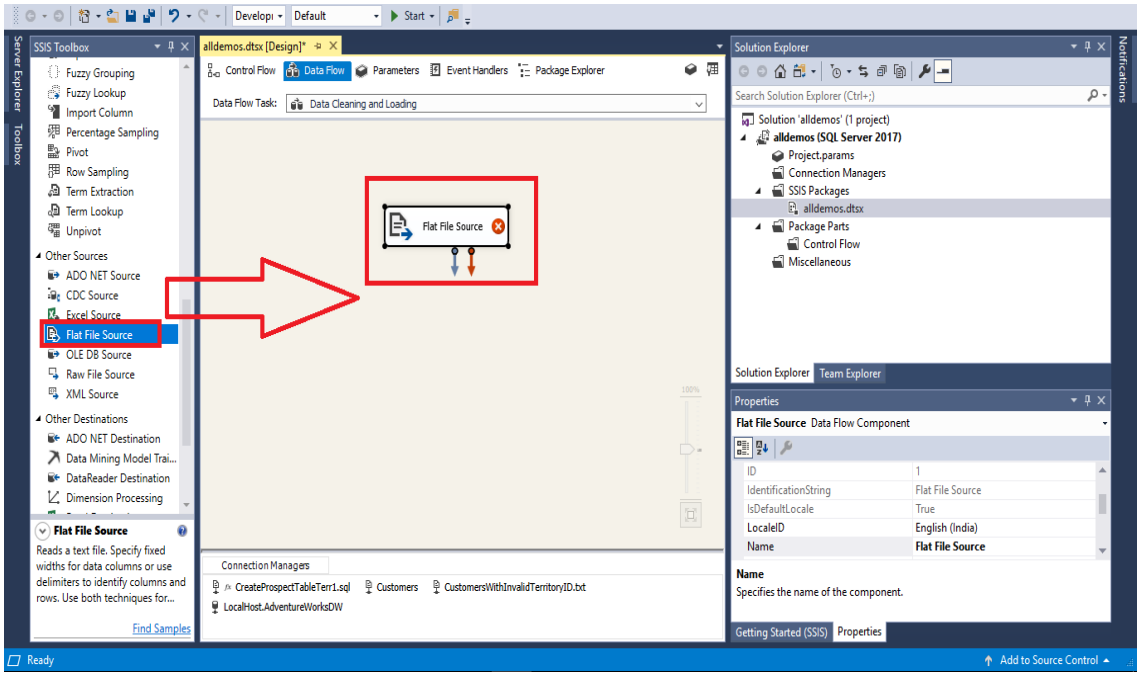



Figure 1.1.24 Drag the Flat File Source Component

Then rename it as "*Data Extraction*". Right-click it and choose to edit. Next, you should do the same settings as follow. Click【OK】finally.

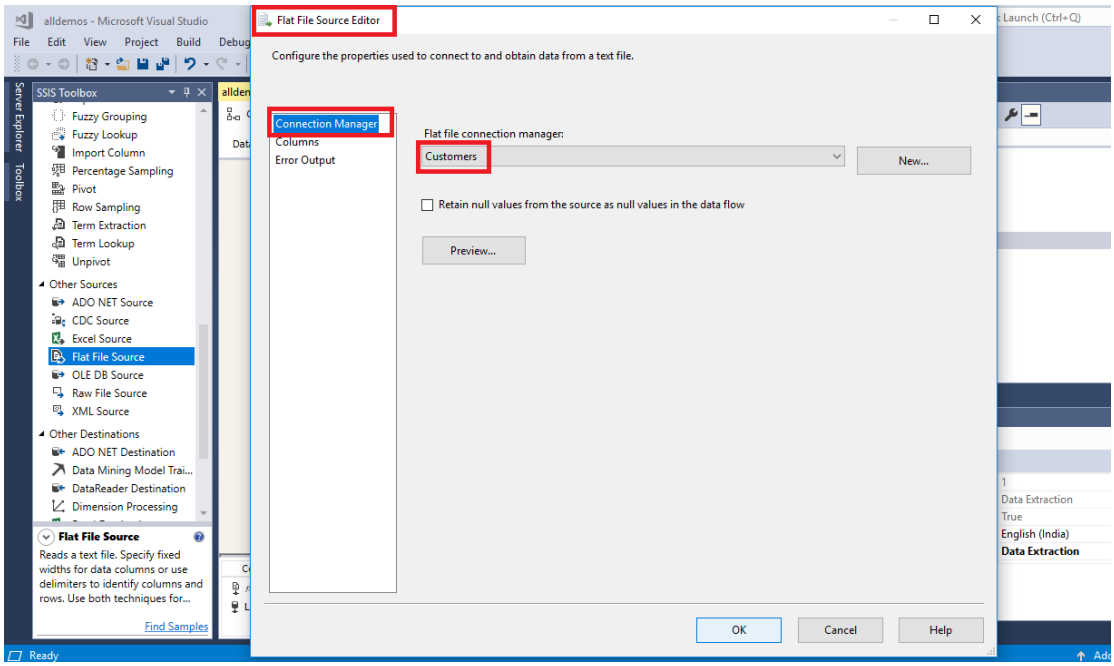



Figure 1.1.25 Specify a Manager to the Flat File Source

(2) Choose 条件性拆分 in the "Toolbox" and drag it to the Data Flow panel, then you can name it as "*Conditional Split by TerritoryID*".

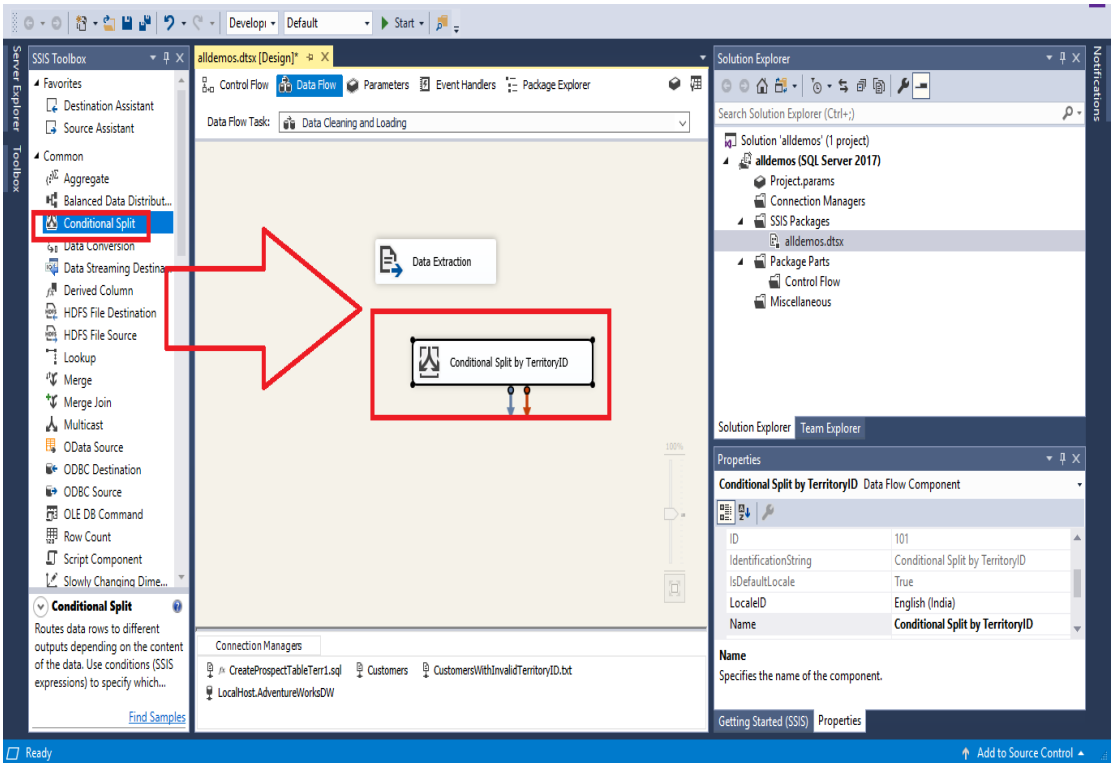



Figure 1.1.26 Drag the Condition Split Component

Finally, you should make "*Data Extraction*" point to "*Conditional Split by TerritoryID*".

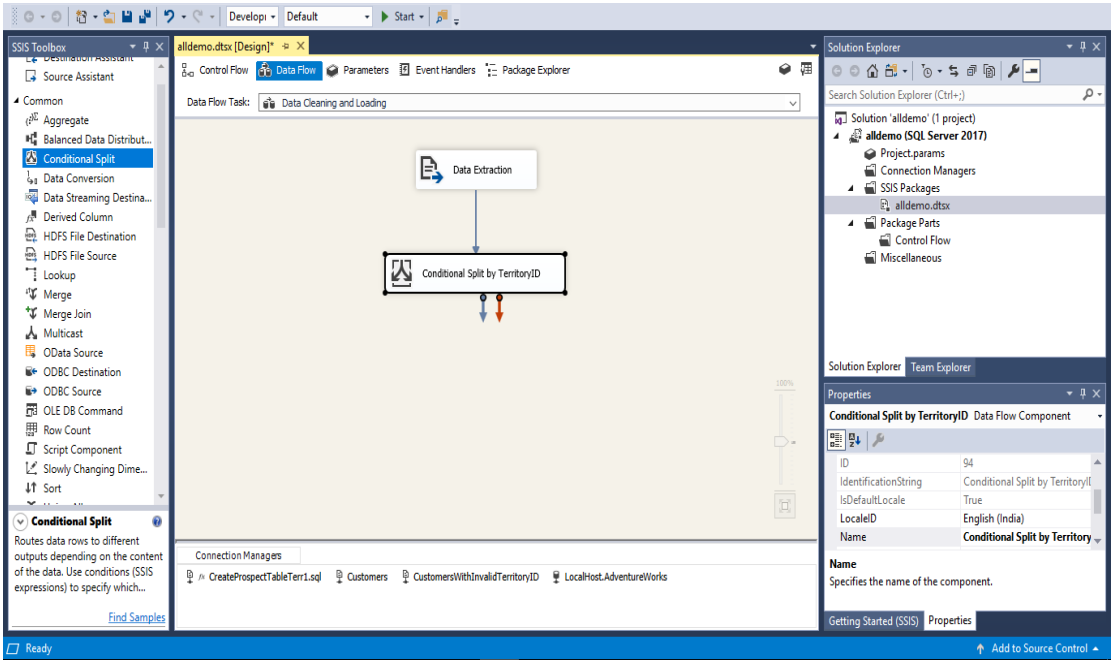



Figure 1.1.27 Connect Two Components

(3) Double-click the conditional split component. Expend the column object in the upper left corner of the "Conditional Split Transformation Editor", then you should drag the field "*Terrritory*" to the condition column in the grid below and edit the expression as the following figure. It's easy to recognize if you rename each output's name.

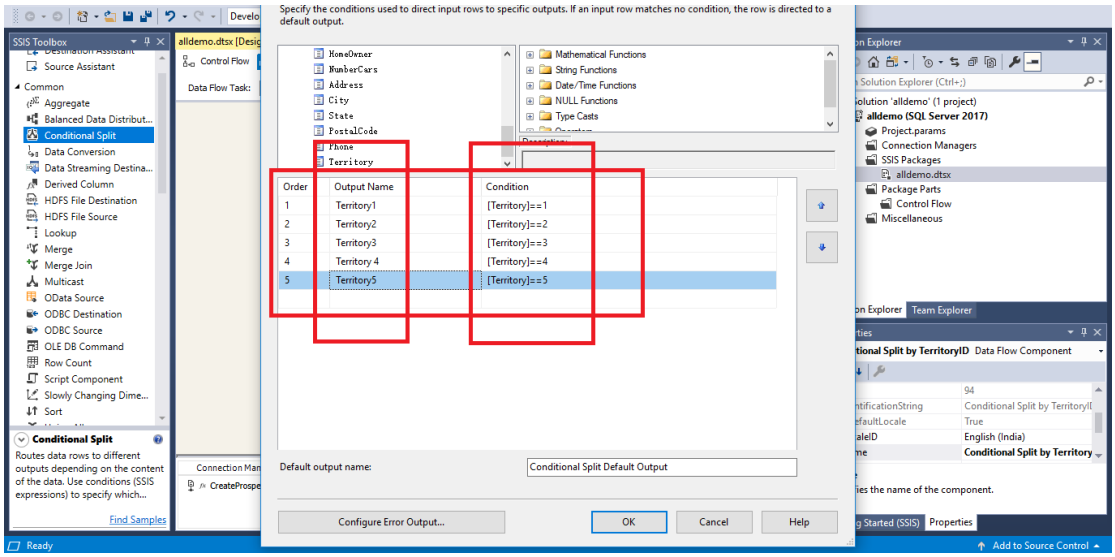



Figure 1.1.28 Edit Split Condition by Territory

(4) Drag four 

from "Toolbox" to the Data Flow panel. These “OLE DB Destinations” will become targets of the conditional split. And connect them to “Conditional Split by TerritoryID” component.

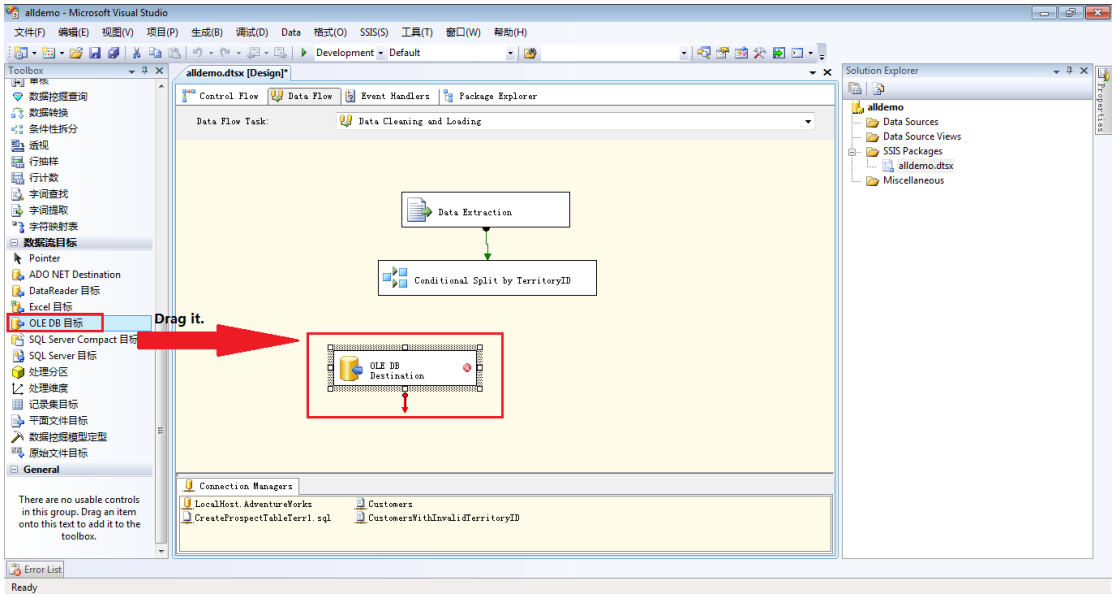



Figure 1.1.29 Drag the OLE DB Destination

Here is an example: you will see a dialog box "Choose input or output" when you make the data flow from the conditional split component which connects to the target data. Finally, you need to choose output and then click【OK】, until data flow is created.

(5) Edit OLE DB Destination and make it point to its target table. The detail operation is: first, you need to find corresponding objective table in the drop-down box and click "New" without changing the content, and then click "OK".

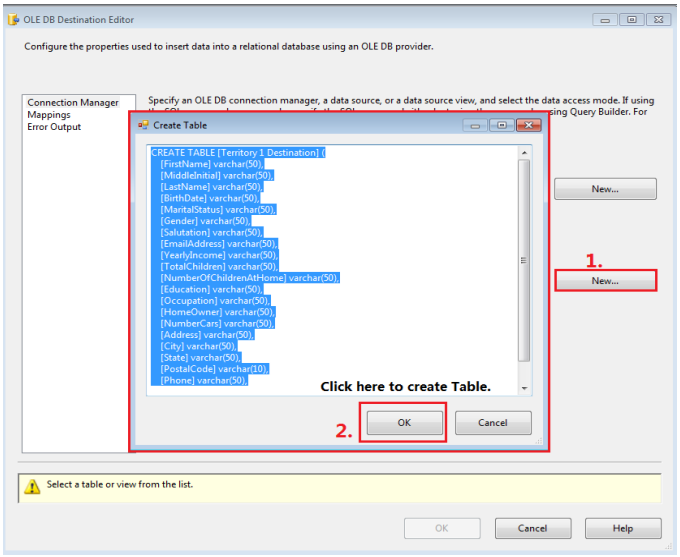



Figure 1.1.30 The Creation of Territory 1 Destination Table

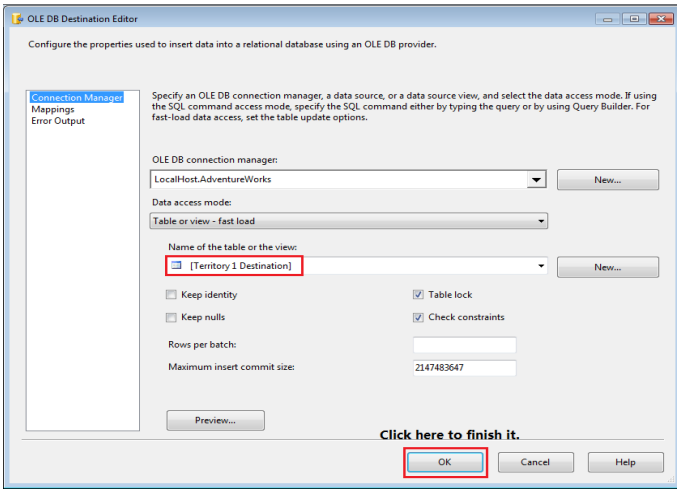



Figure 1.1.31 Creation Result

(6) Set the target file of invalid regions, which should point to the"*CustomersWithInvalidTerrritoryID*" flat file connection.

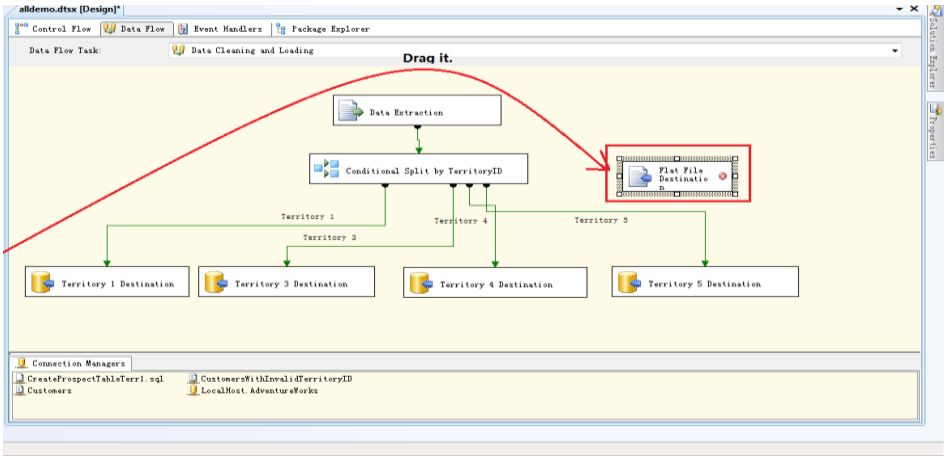



Figure 1.1.32 Drag the Flat File Destination Component

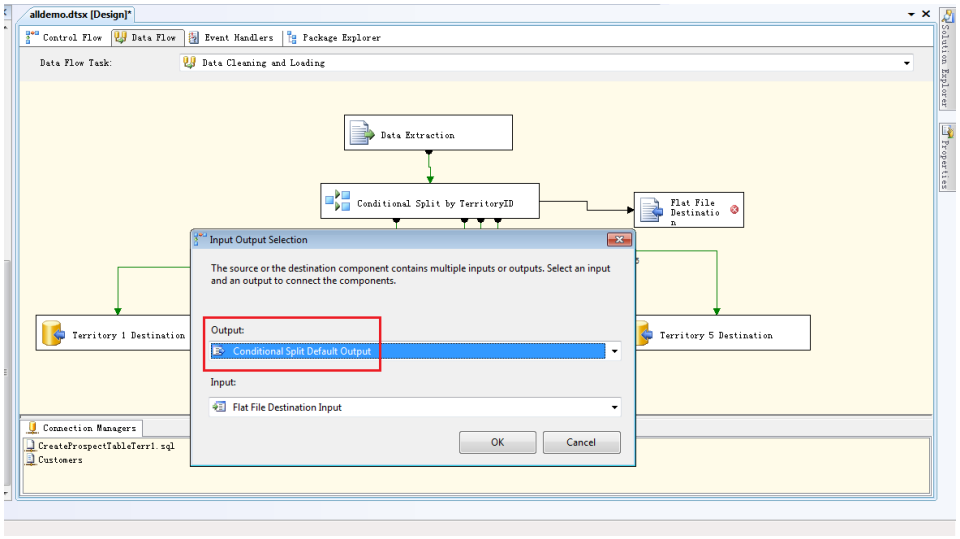



Figure 1.1.33 The Flat File Destination Output Setting

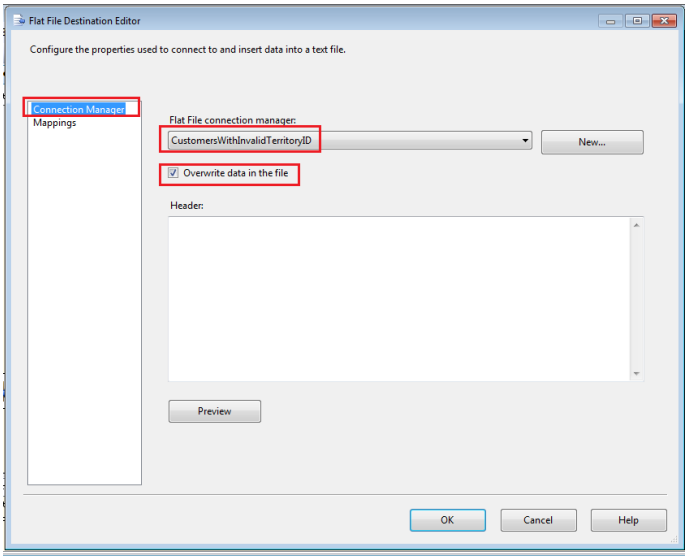



Figure 1.1.34 Specify a Manager to the Flat File Destination

The whole data flow panel should be like the follow figure after you have set it well.

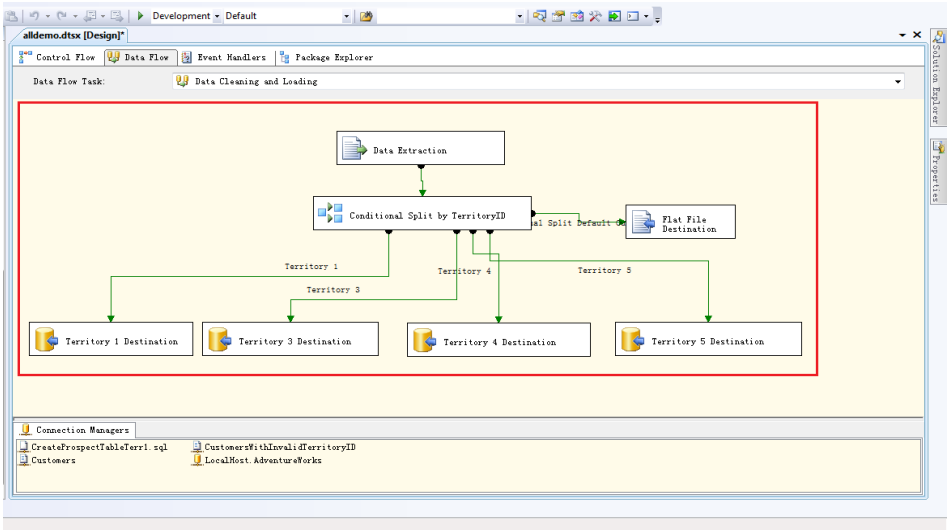



Figure 1.1.35 Current Data Flow Panel

(7) The data in region 2 has not targeted host in the conditional split, because the sales of this region have some problems in the process of inputting. The number of zip code should be five. But now it only has four digits, because it omits the front zero when a person input the data. Thus, you need to clean the zip code before inputting them into tables. Choose the component 派生列 in the "Toolbox" and drag it to the Data Flow panel.

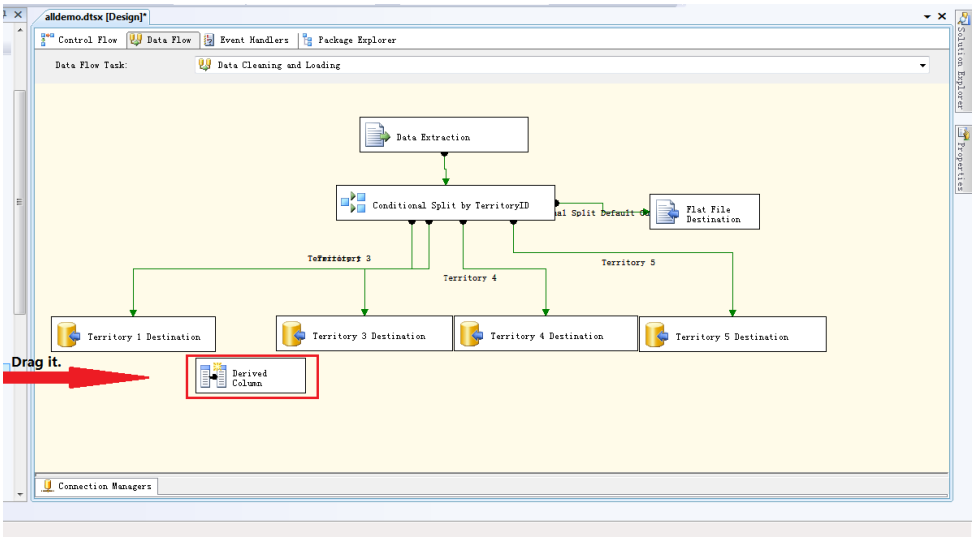



Figure 1.1.36 Drag the Derived Column Component

Double-click the component you added just now.

Attention: the expression "*LEN(PostalCode) == 4 ? "0"+PostalCode : PostalCode*", which means that you should add zero in the front of the postal code if the length of it is four, otherwise, it should not be changed.

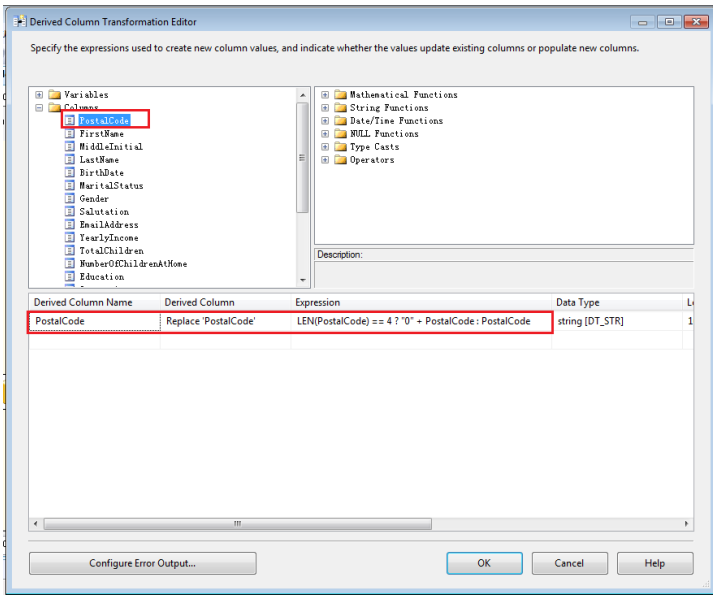



Figure 1.1.37 PostalCode Transformation

(8) You should add the target table from region 2 to the Data Flow panel and connect it by the derived columns component.

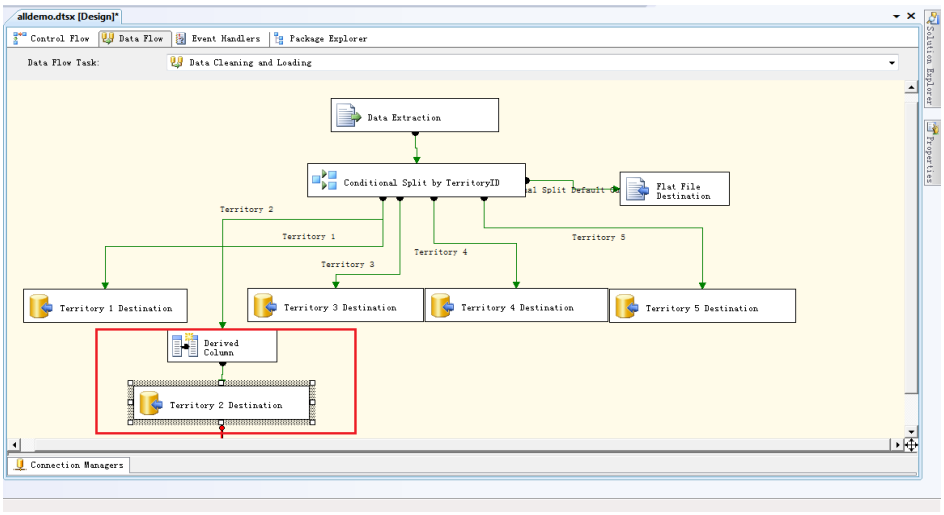



Figure 1.1.38 The Entire Data Flow Panel

(9) Debug the project.

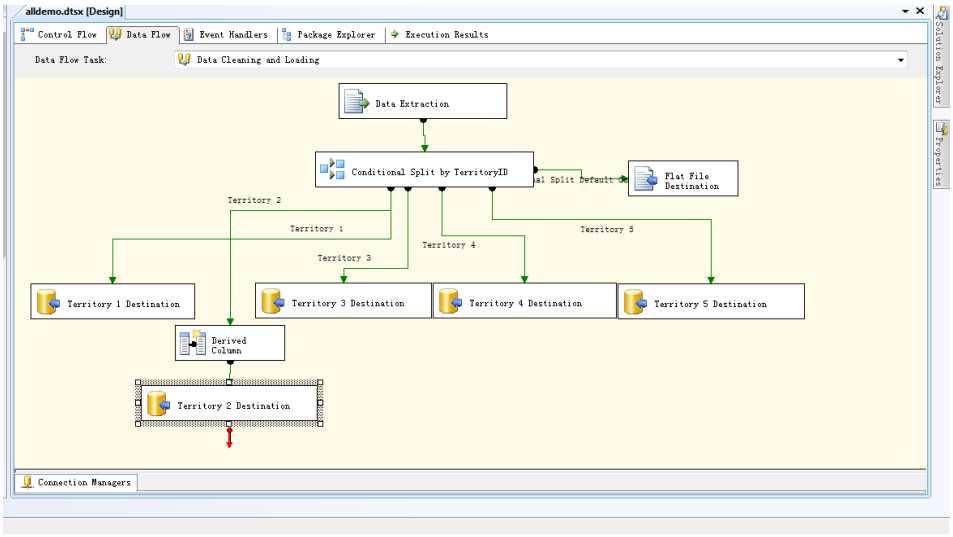



Figure 1.1.39 Start Debug

The final debugging result is as follows:

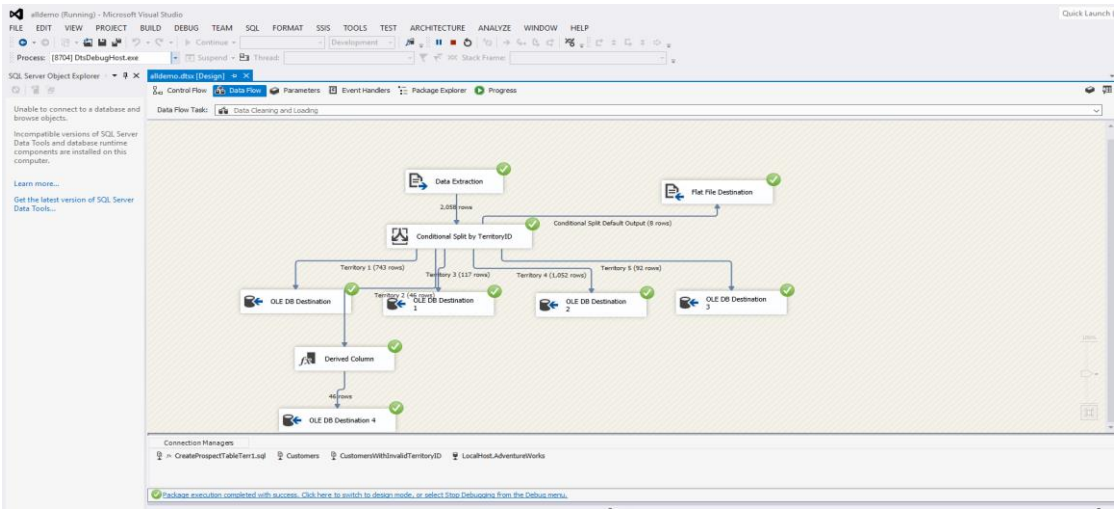



Figure 1.1.40 Successful Debug Result

# 5. The Configuration and Deployment of Package

## The Configuration for Package

(1) Open “Integation Sevices” project in the Business Intelligence Development Studio and double-click the “.dtsx” file in the Solution Explorer for Solution.

(2) Choose【SSIS】→【Configuration for Package】.

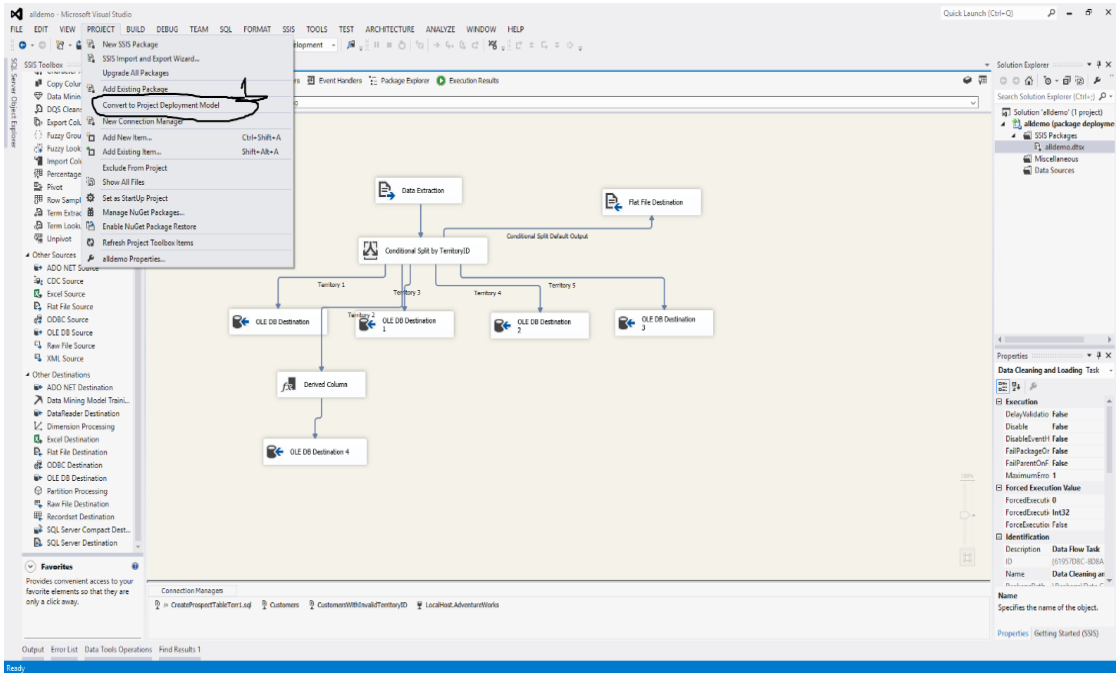


Figure 1.1.41 Start Package Configuration

(3) Do the following settings.

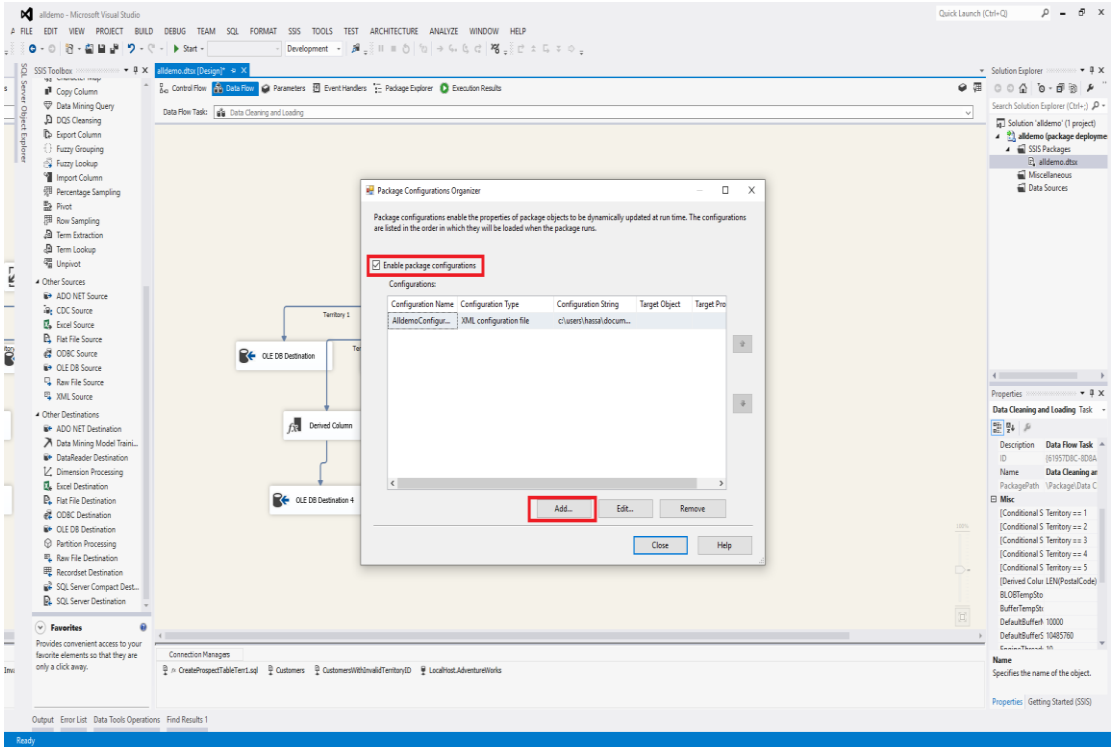


Figure 1.1.42 Organizer Setting

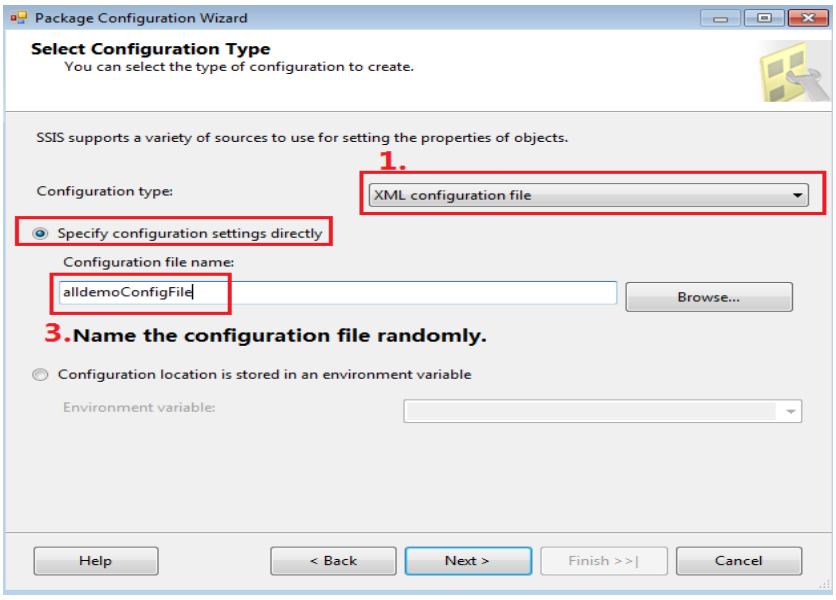



Figure 1.1.43 Select Configuration Type

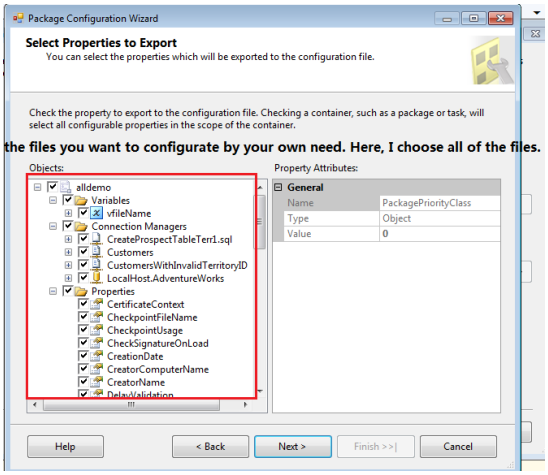



Figure 1.1.44 Select Properties to Export

(4) Click【Next】and input the configuration name after you have set the storage file and click【Finish】to quit the wizard finally.

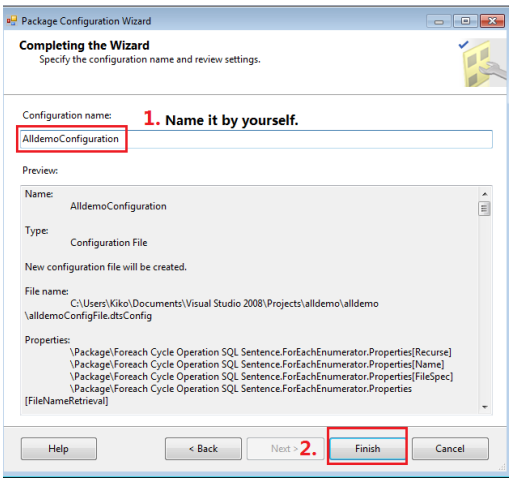



Figure 1.1.45 Finish the Wizard

## The Deployment for Package

(1) In Business Intelligence Development Studio, Right-click the project name and choose 【Properties】to open property pages.

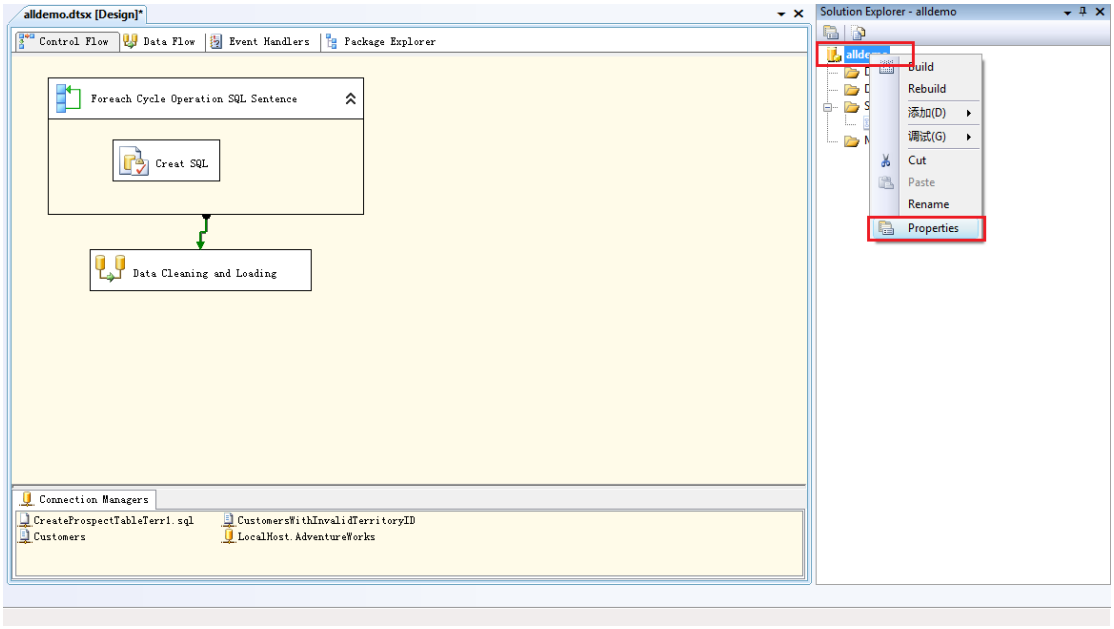



Figure 1.1.46 Start the Project's Properties Setting

Select "Deployment Utility" item. In this step, "AllowConfigurationChange" specifies whether the configuration can be updated when the project is deploying. "CreateDeploymentUtility" specifies whether the package can be recreated when the project is deploying, and you can create deployment while the attribute is "true". "DelopmentOutputPath" specifies the deployment tool's location.

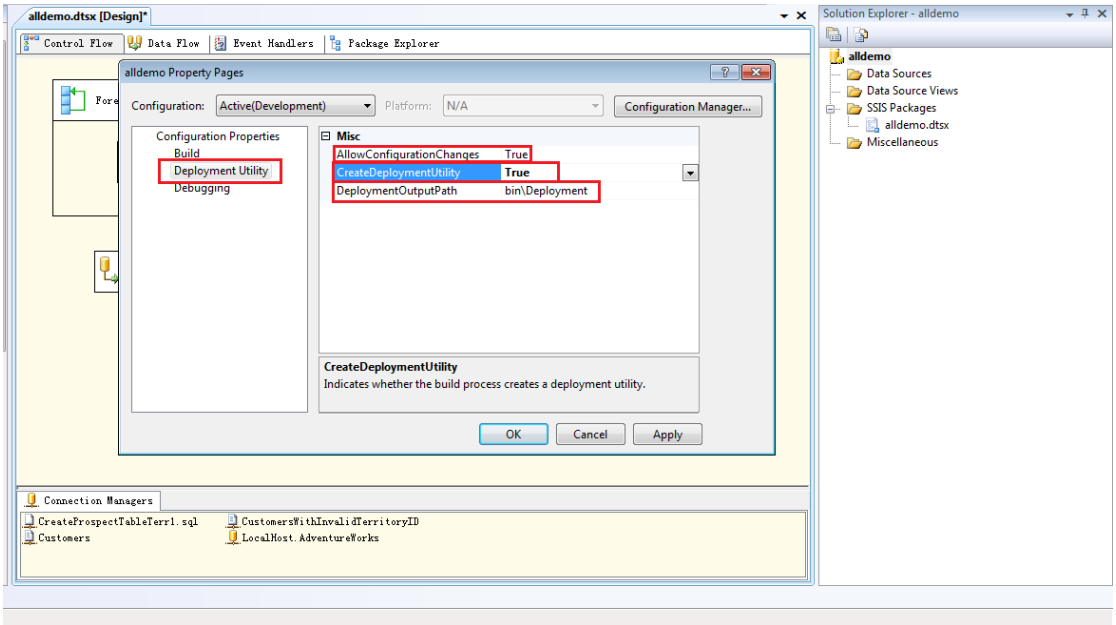



Figure 1.1.47 Deployment Utility Setting

(2) Next right-click your project name and click 【Build】 to build the project.

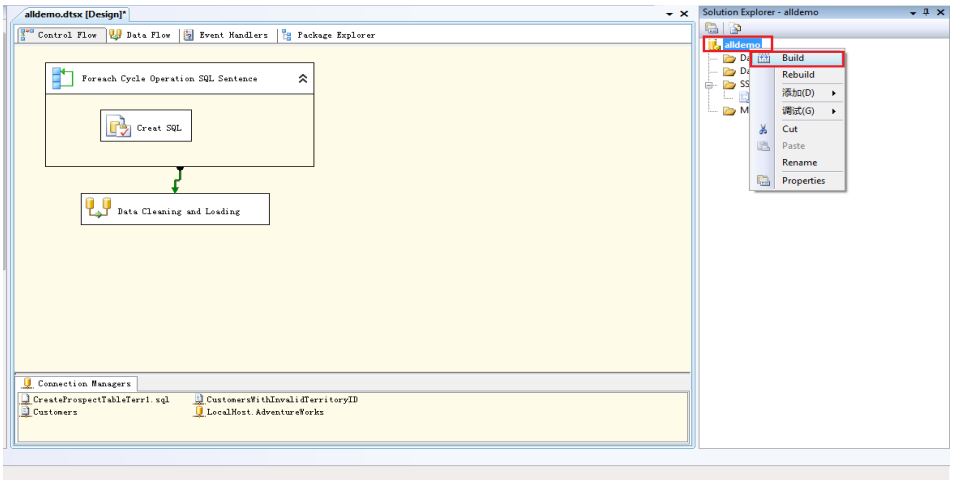



Figure 1.1.48 Start Build the Project

The successful result of building is as follow.

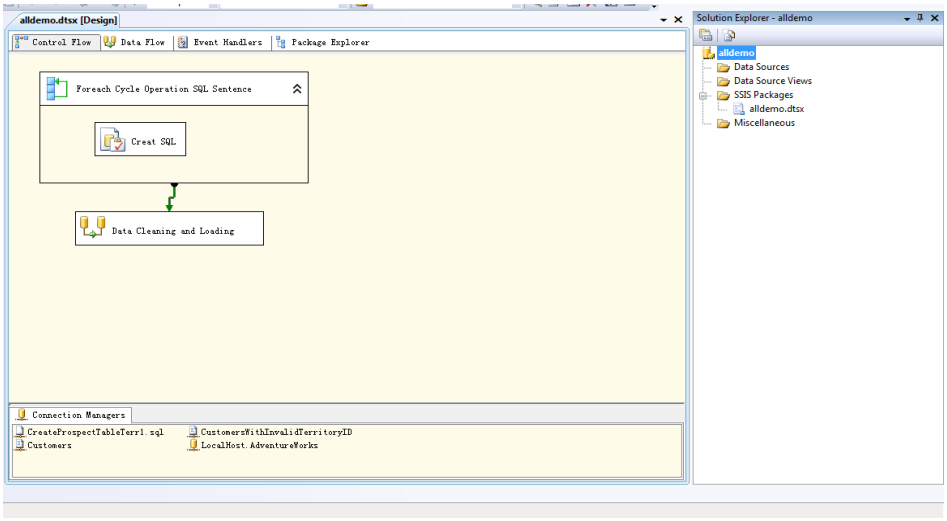



Figure 1.1.49 Successful Build Result

# 6. Deploying the Package to File System

(1) Open the deployment files, such as “alldemo.SSISDeploymentManifest”. Double-click it and start the wizard.

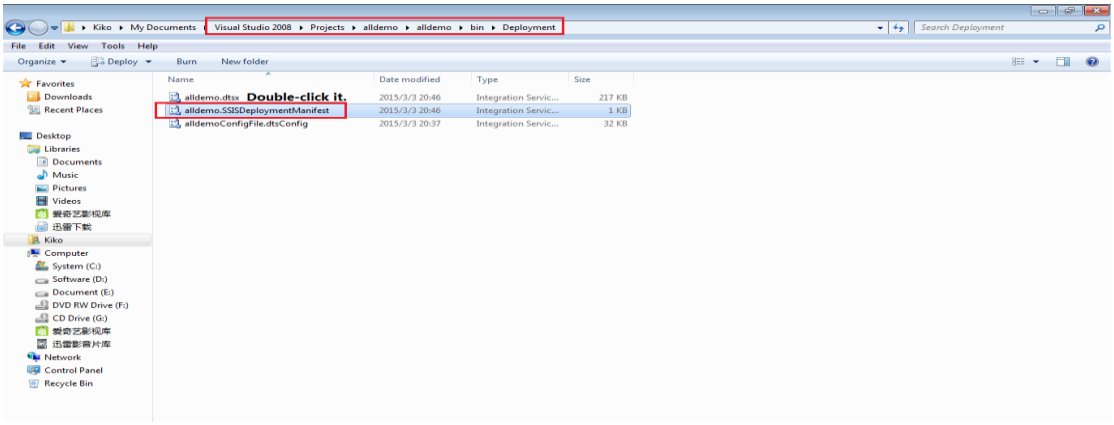



Figure 1.1.50 Find the Manifest File

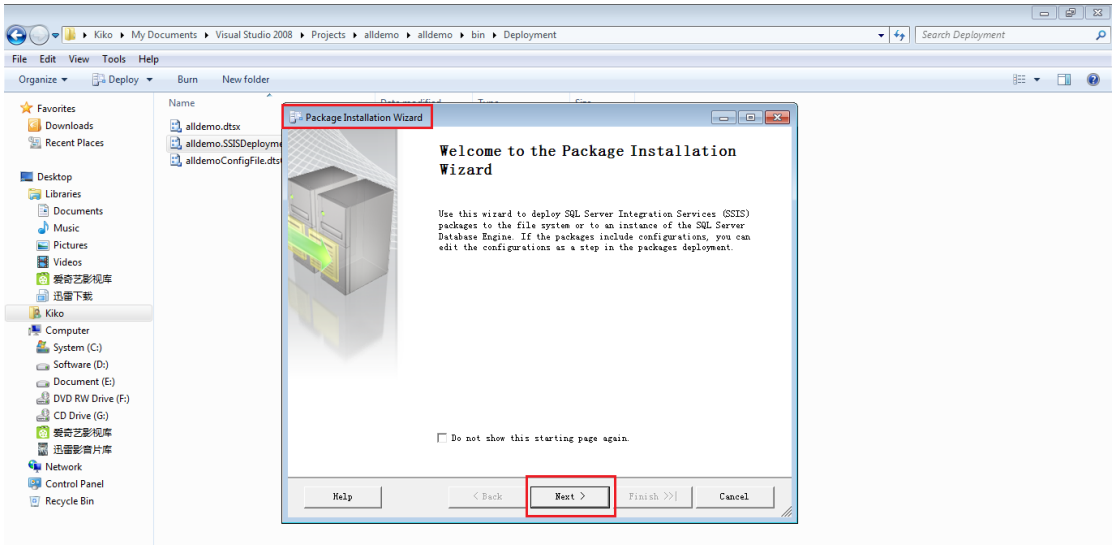



Figure 1.1.51 Open the Package Installation Wizard

(3) Choose "File system deployment". Click【Next】.

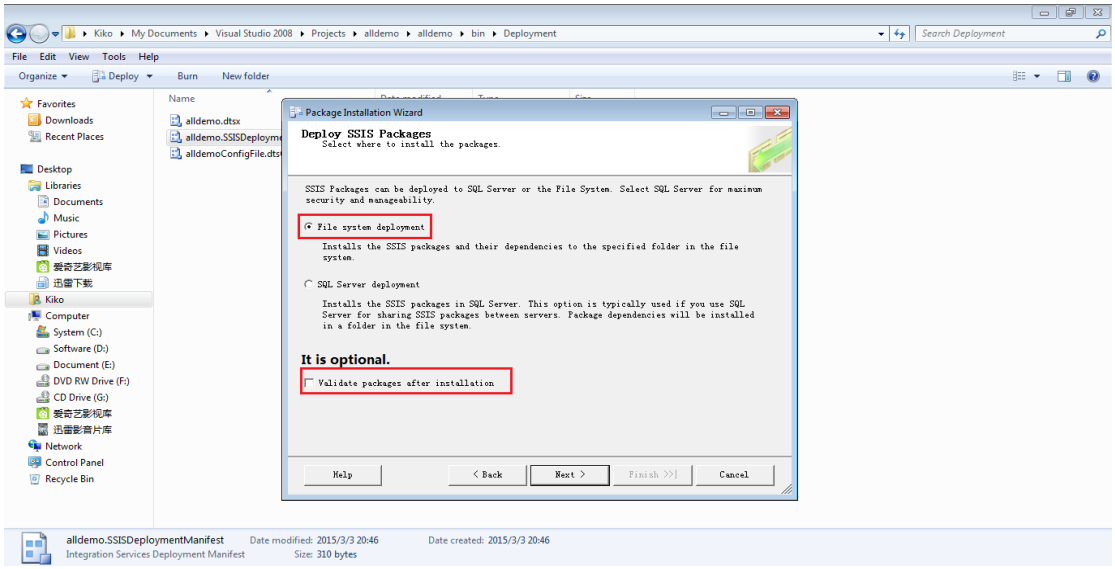



Figure 1.1.52 Select Deployment Destination

(4) Select installation folder.

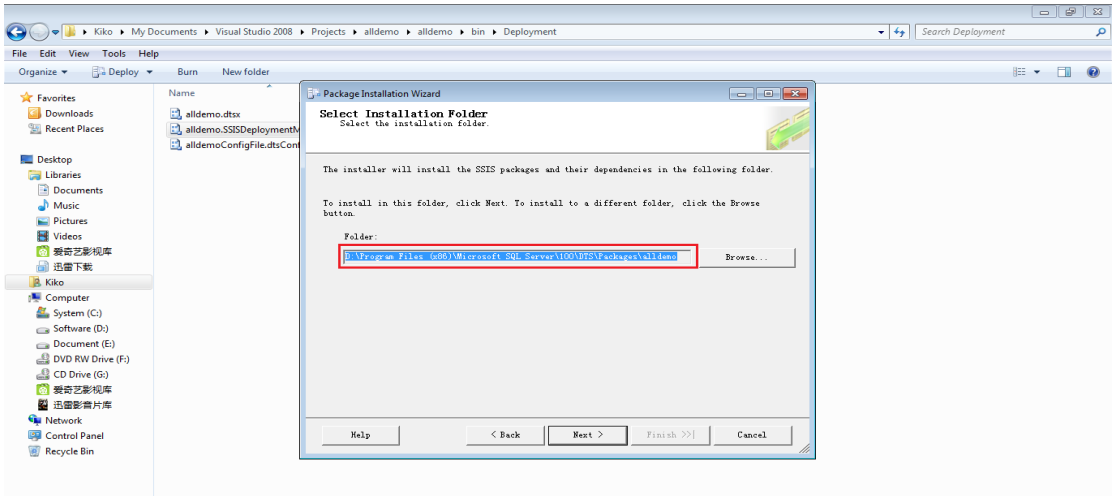



Figure 1.1.53 Select Installation Folder

Then click【Next】continuously until finish.

**Query with Problem (If You Faced Any Issue in deployment then Check it)**

While solving this we got an error, so we remove the semi column and use the same file name as in the file. Now you should add the target table from region 2 to the Data Flow panel and connect it by the derived columns component.

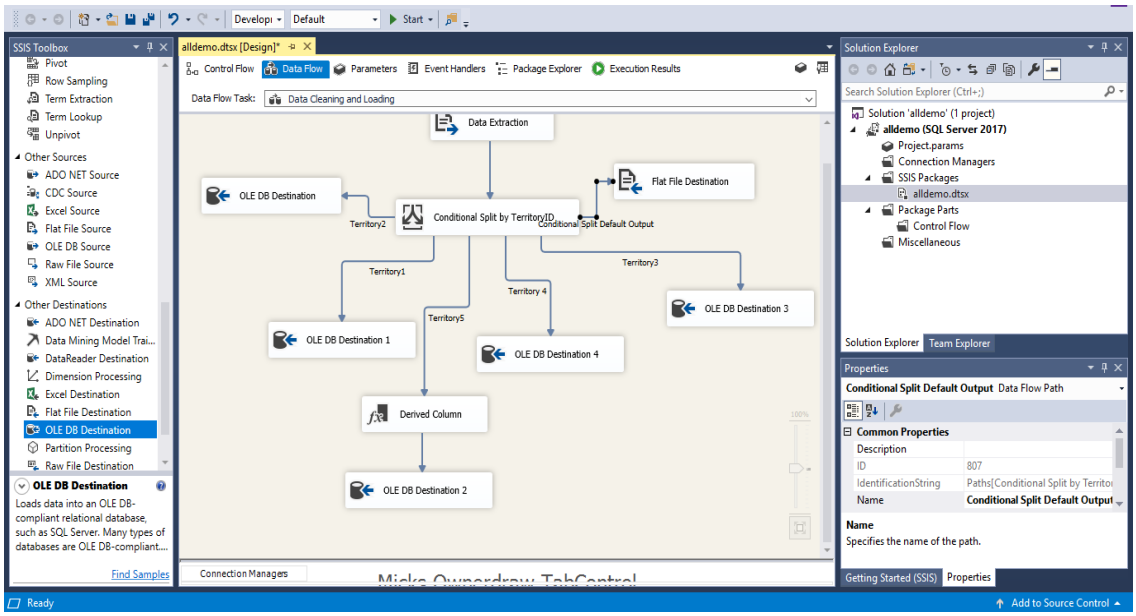



Figure 1.1.54 The Entire Data Flow Panel

Debugging result is as follows:

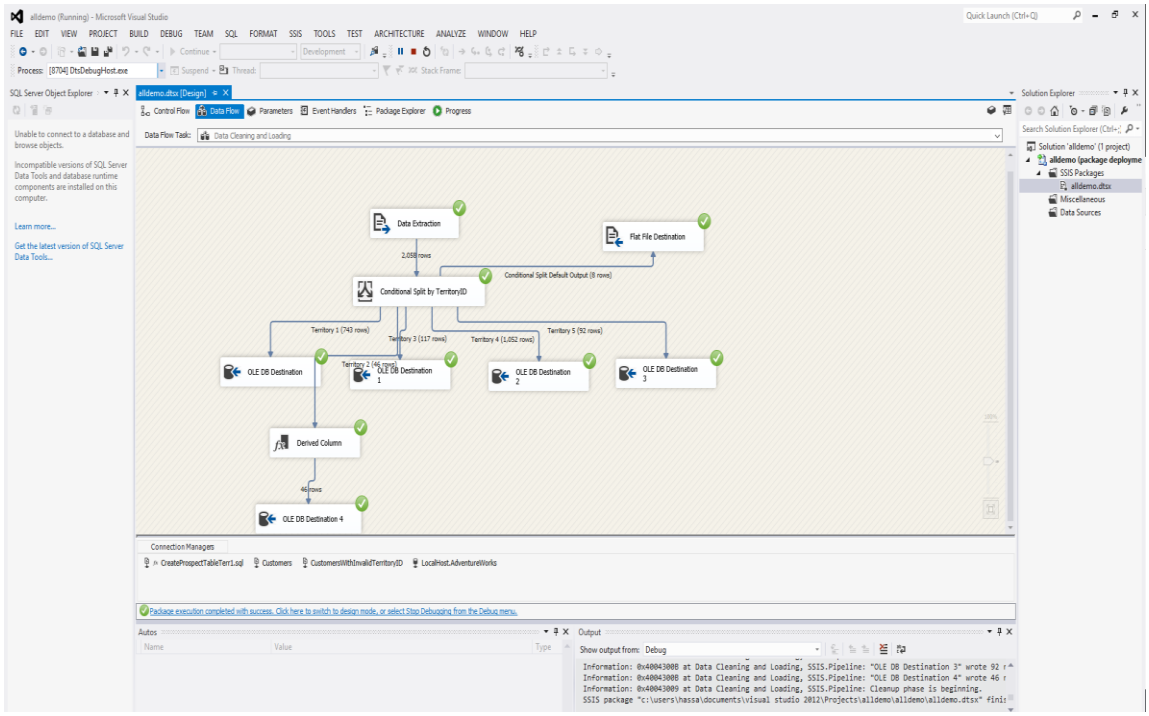



Figure 1.1.55 Successful Debug Result

## Deploying the Package to File System

Open project menu and select convert to Project Deployment Model as shown below.

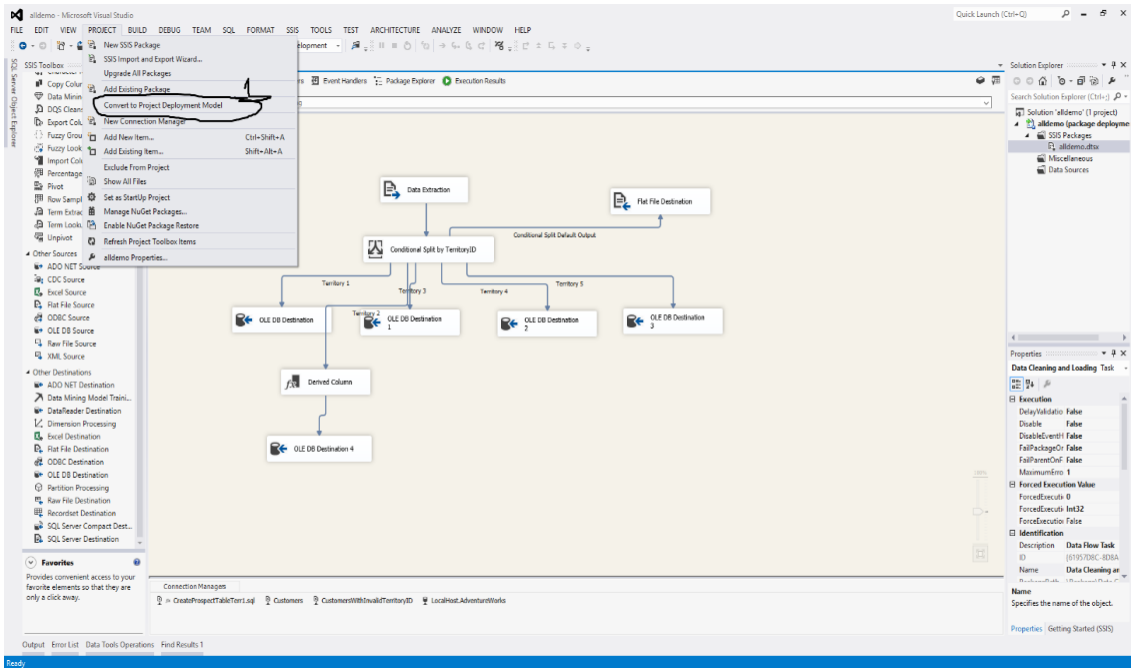



Figure 1.1.56 Project deployment model

The package configuration organizer shows the properties of the configuration objects. The configuration details are shown as follow.

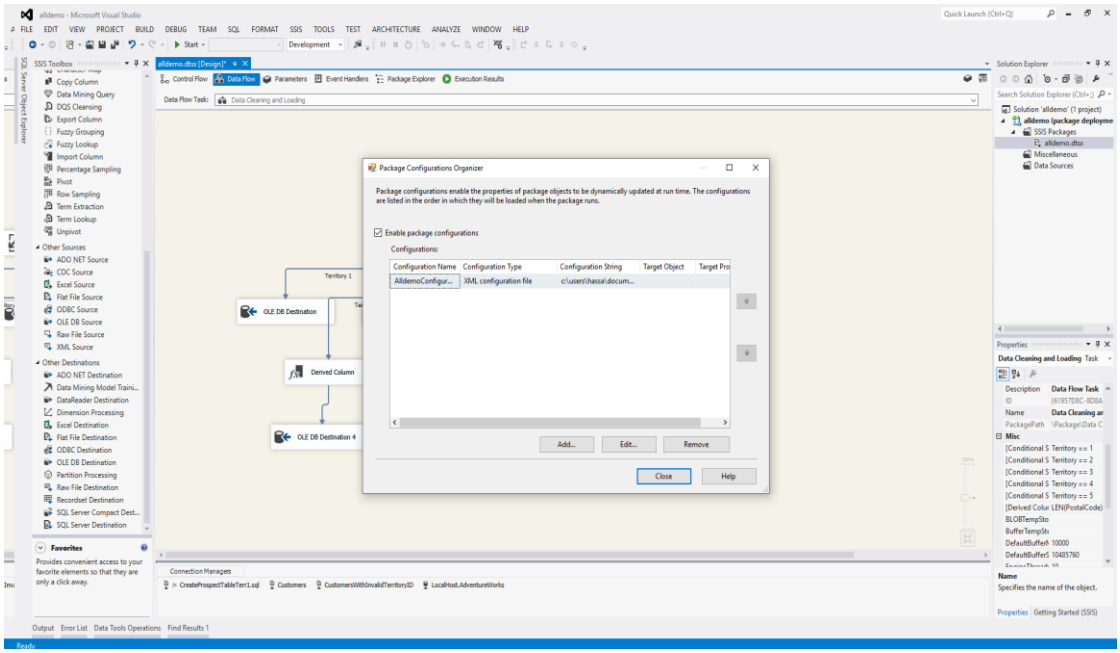



Figure 1.1.57 Package configuration Organizer

Next go to SSIS and select Package configuration.

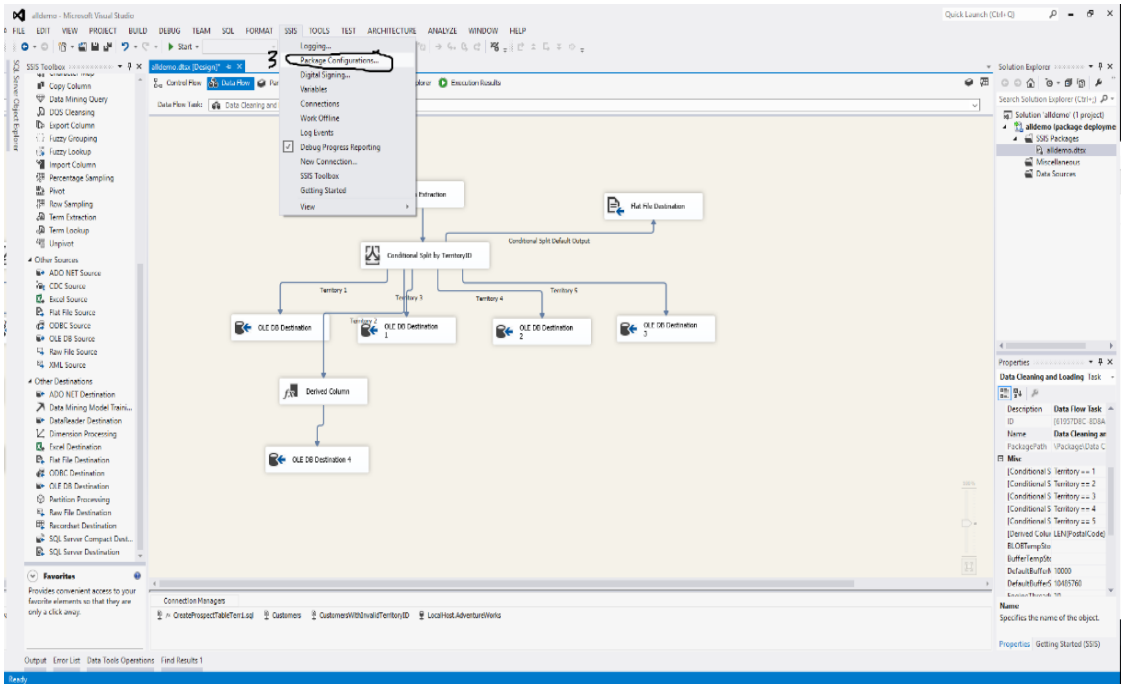


Figure 1.1.58 Package configuration

## Query with Solution

The problem found while doing the whole process is that the Package Installation Wizard does not run. For this we found the solution that we use the directory C:\\Program Files (x86) \\Microsoft SQL Server\\120\\DTS\\Binn. Then copy and paste it on Wizard file. Then run it and finally it successfully run.

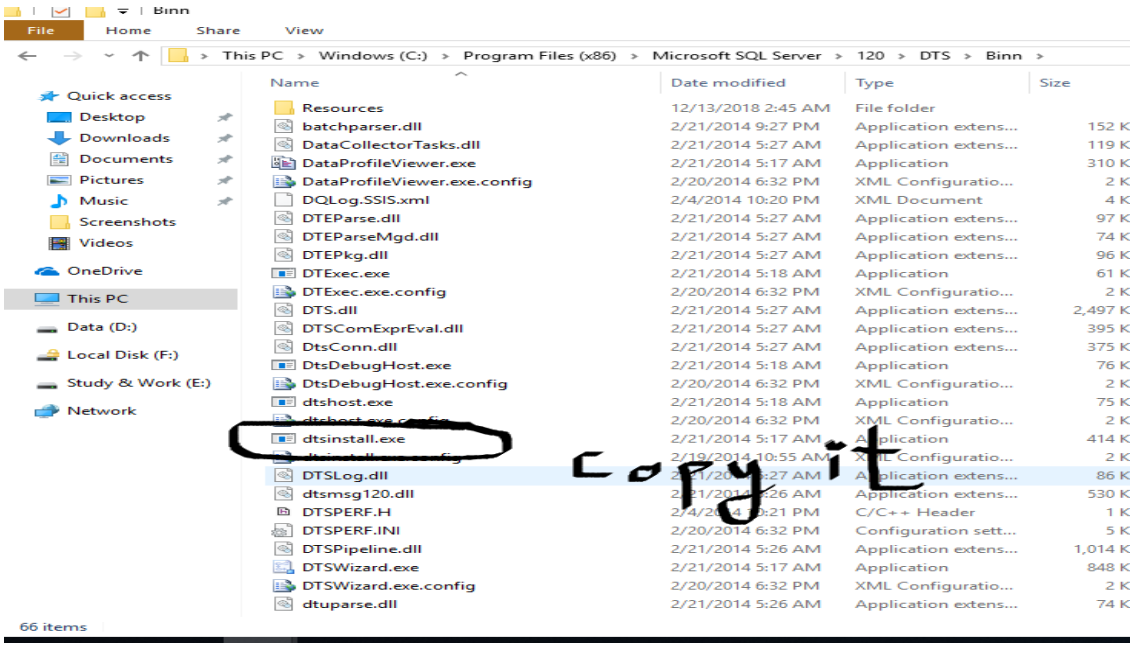



Figure 1.1.59 *Copy File "dtsinstall.exe"*

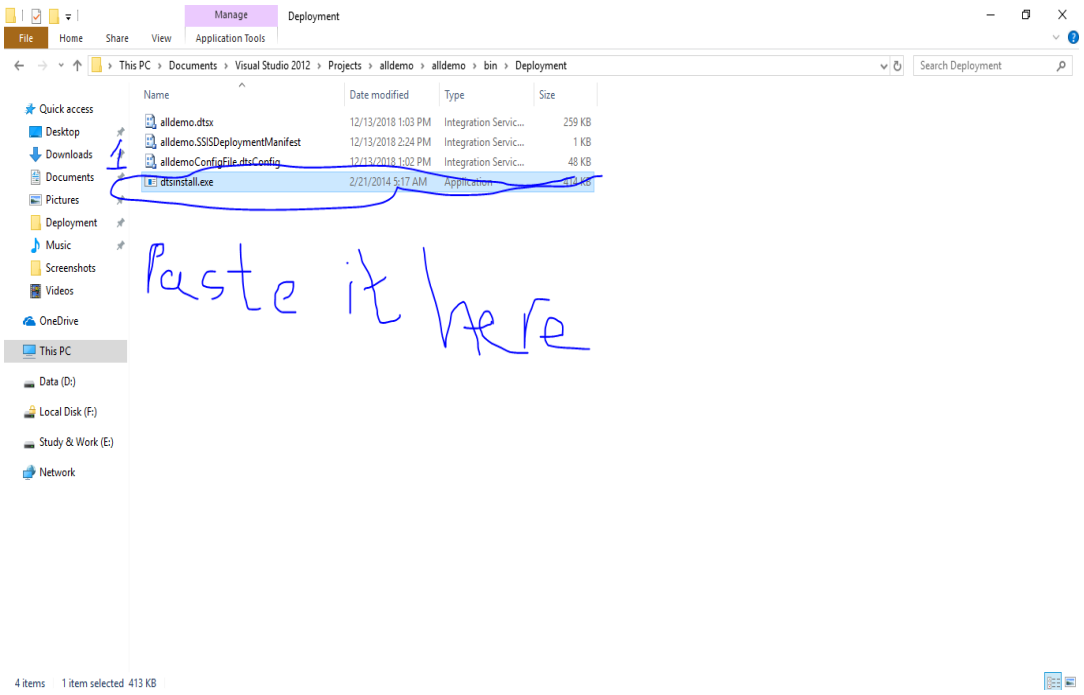



Figure 1.1.60 Paste File “dtsinstall.exe”

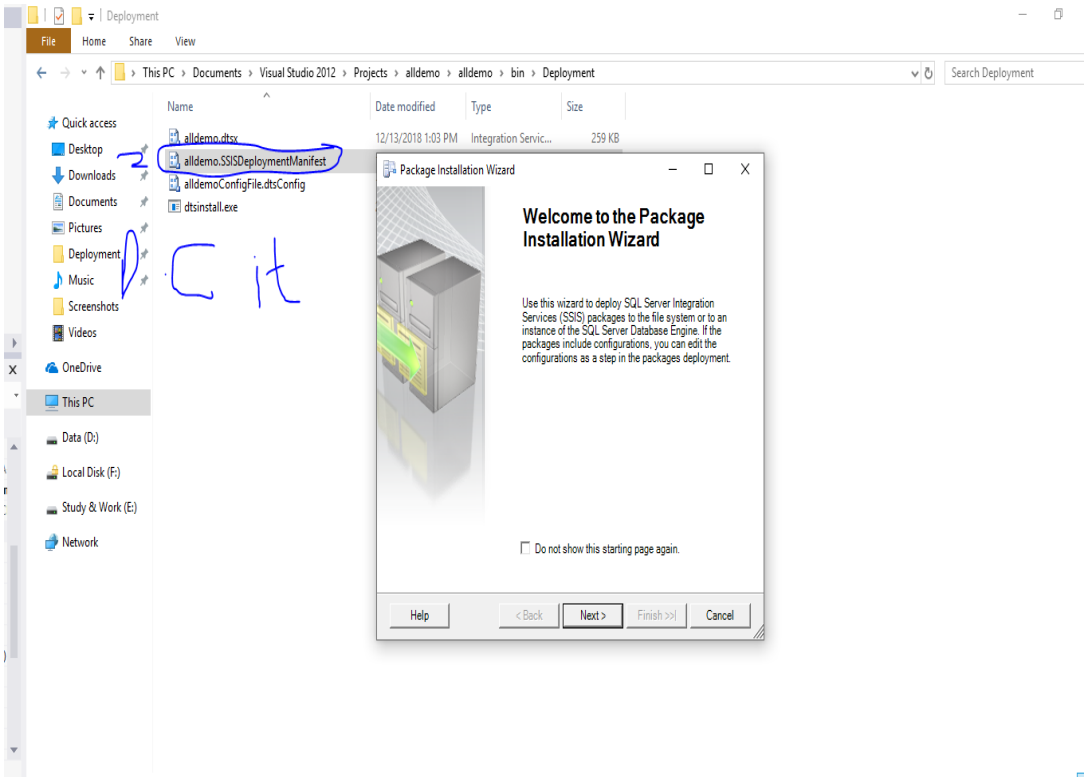



Figure 1.1.61 Start-up Setting

# 7. The Running of the Package

There are several ways for running the package and you will be introduced the most common method.

(1) Right-click the package that you will implement and choose【As Setup】.

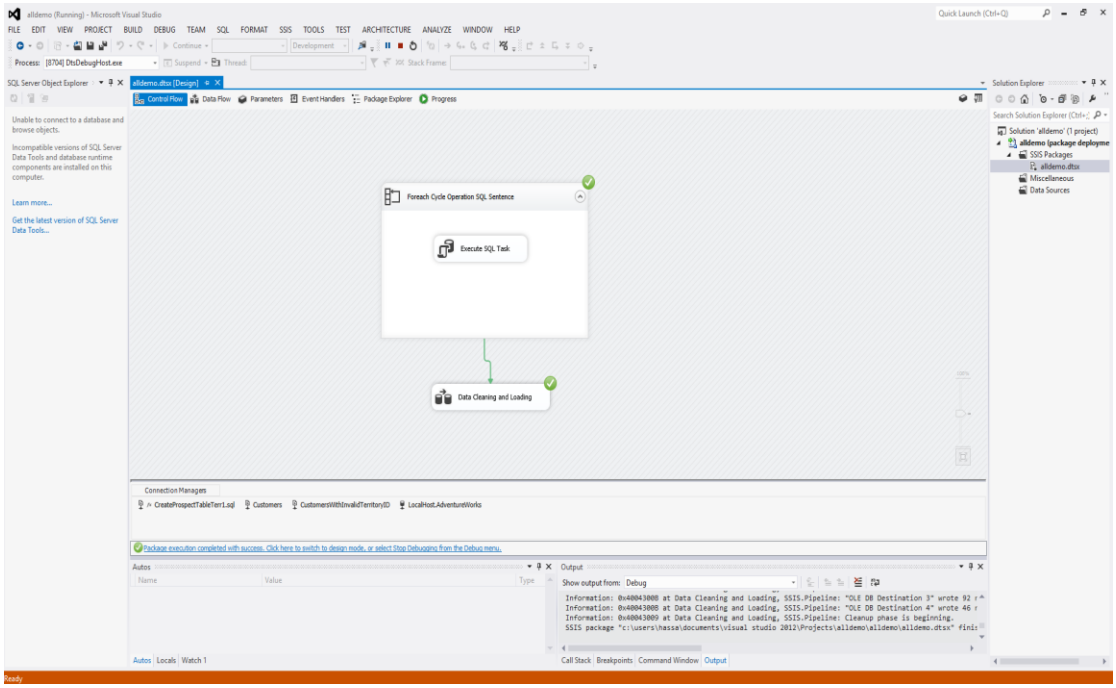



Figure 1.1.62 Startup Setting

(2) Click "Start Debugging" or press【F5】 to debug the project. After you have finished the debug, press【Shift+F5】to return the design pattern.

Until now you have finished operating ETL, then switch to SQL Server Management Studio and open “AdventureWorks” database. You will find the original data “customers.txt” which has already segmented to five tables (from Terrritory1 to Terrritory5) and the invalid data has also been imported into “CustomersWithInvalidTerrritoryID.txt” file.

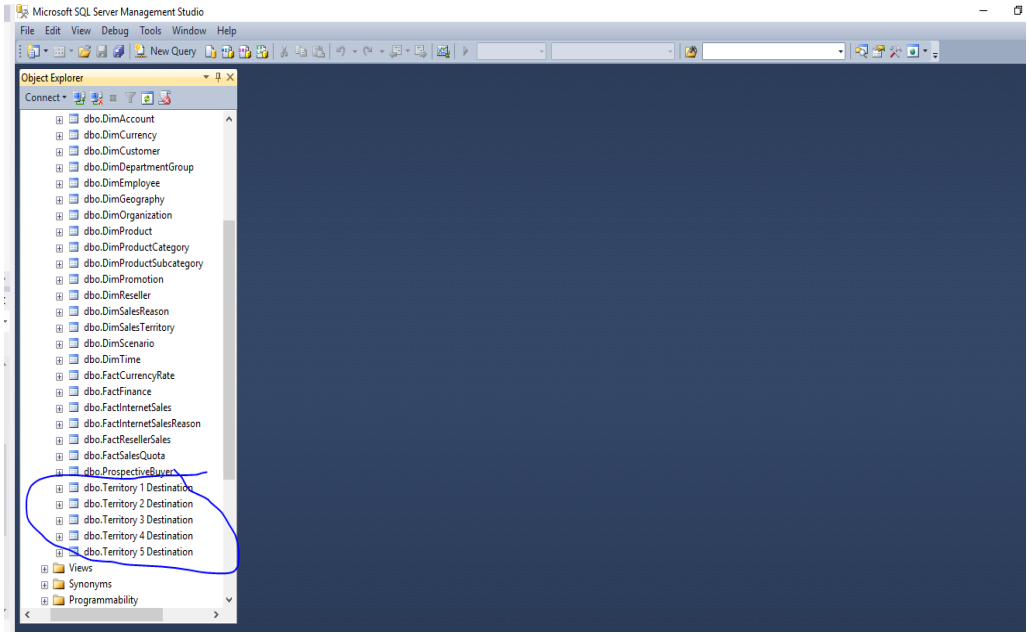



Figure 1.1.63 Result

# Experiment 1.2
# Building a Data Cube

## Section One: The Goal of the Experiment

(1) Familiar with the use of SSAS.

(2) Learn to use a bottom-up approach to create a data cube.

## Section Two: The Content of Experiment

The data source of this experiment is based on a sample database of SQL 2005—Adventure Works DW, which use a bottom-up approach to build cube.

## Section Three: The Procedure of Experiment

### 1. The definition of data source and the view of data source

(1) Build a "Analysis Services" project in SQL Server Business Intelligence Development Studio and name it as "Aworks".

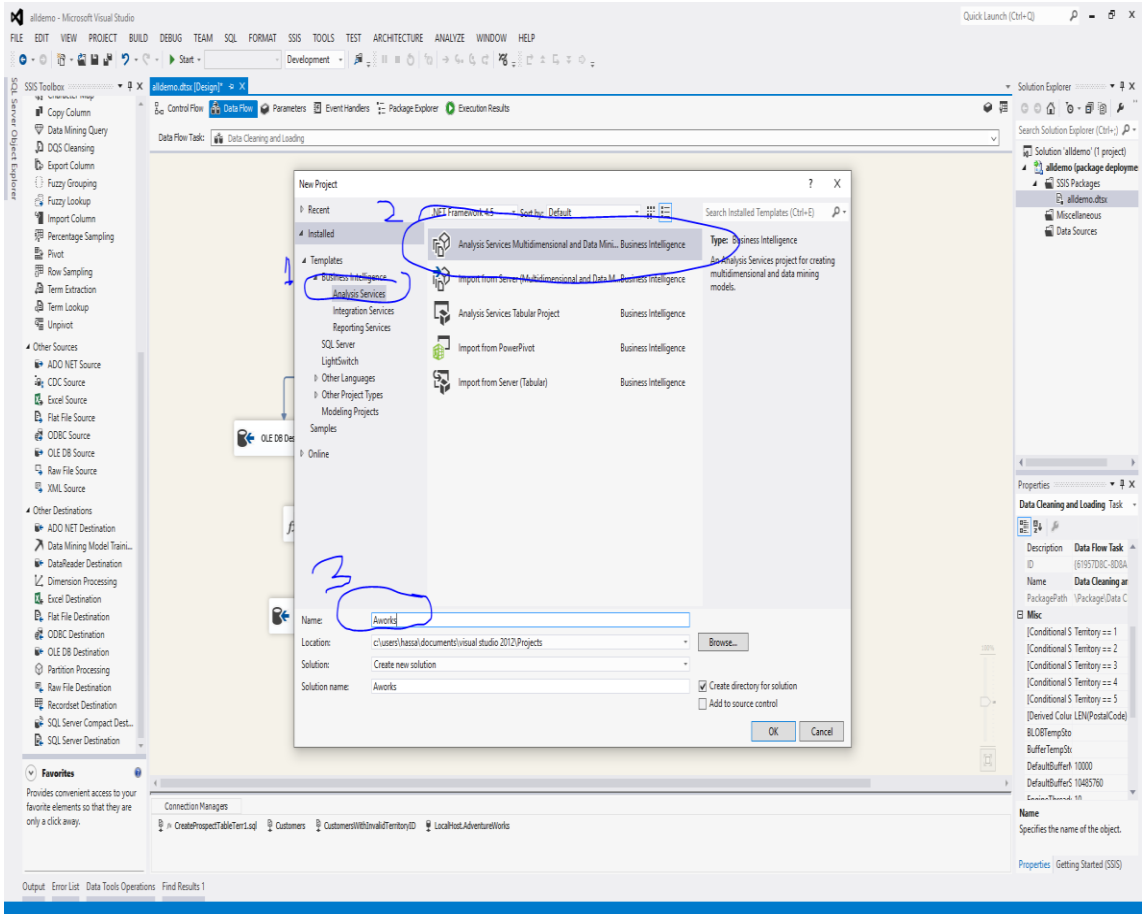



Figure 1.2.1The Creation of Aworks Project

In the Solution Explorer, right-click【Data Sources】 and choose 【New Data Source】.

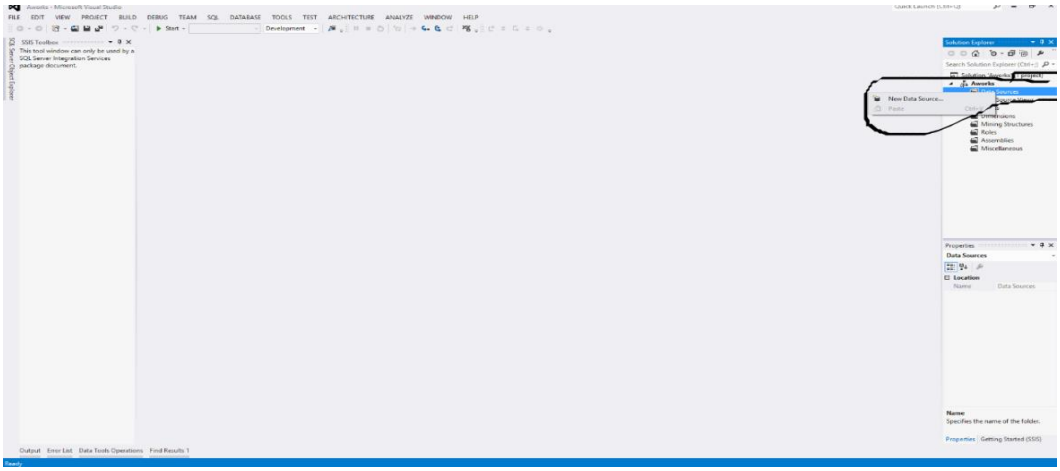


Figure 1.2.2 Create a New Data Source

Select “AdventureWorks” database.

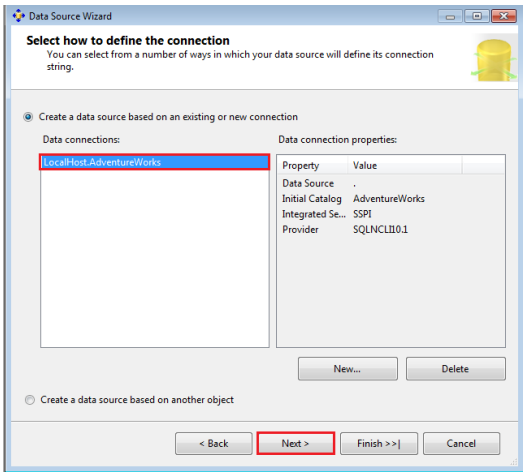



Figure 1.2.3 The Creation of a New Connection

Start the wizard. Click the button【Next】until you come to Impersonation Information page. Check “Inherit”.

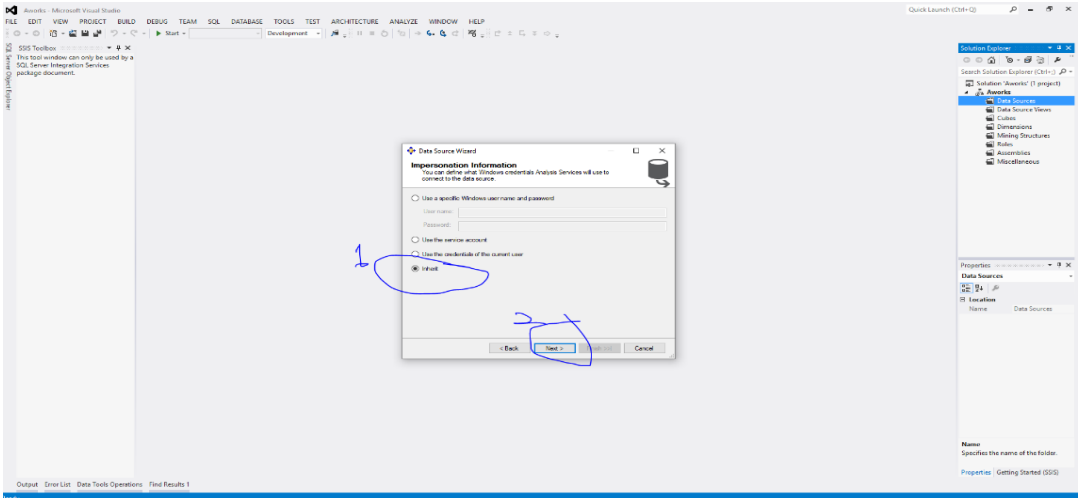



Figure 1.2.4 Select the Inherit Method

(2) Create the data source view. Right-click "Data Source Views" in the Solution Explorer panel and click【New Data Source View】.

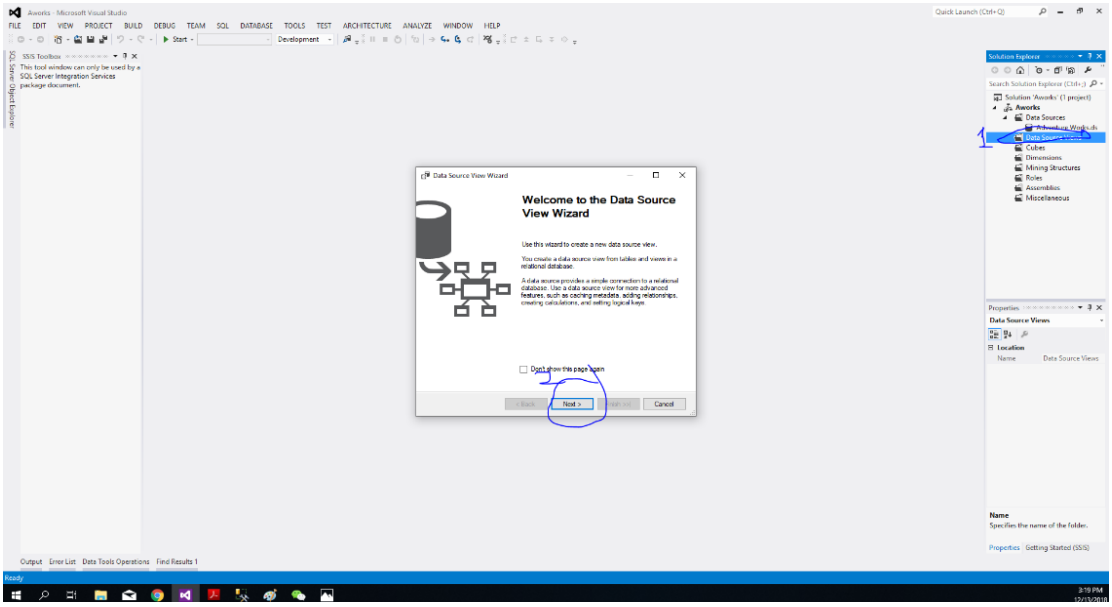



Figure 1.2.5 Create a New Data Source View

Start the wizard and choose five tables (DimCustomer，DimGeography，DimProduct，DimTime，FactInternetSales) from “Available objects” to “Included objects”.

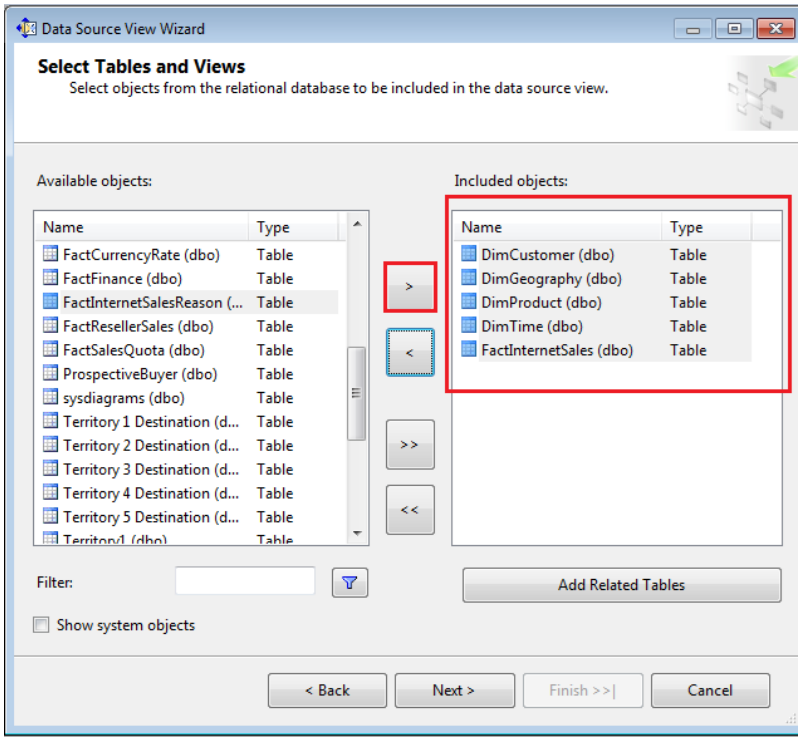



Figure 1.2.6 Select Tables

(3) Click【Next】until finish. It will generate a data source view of AdventureWorks database. You will see five tables connected to each other automatically.

## 2. Generate Cube

(1) Right-click “Cubes” and select “New cube”.

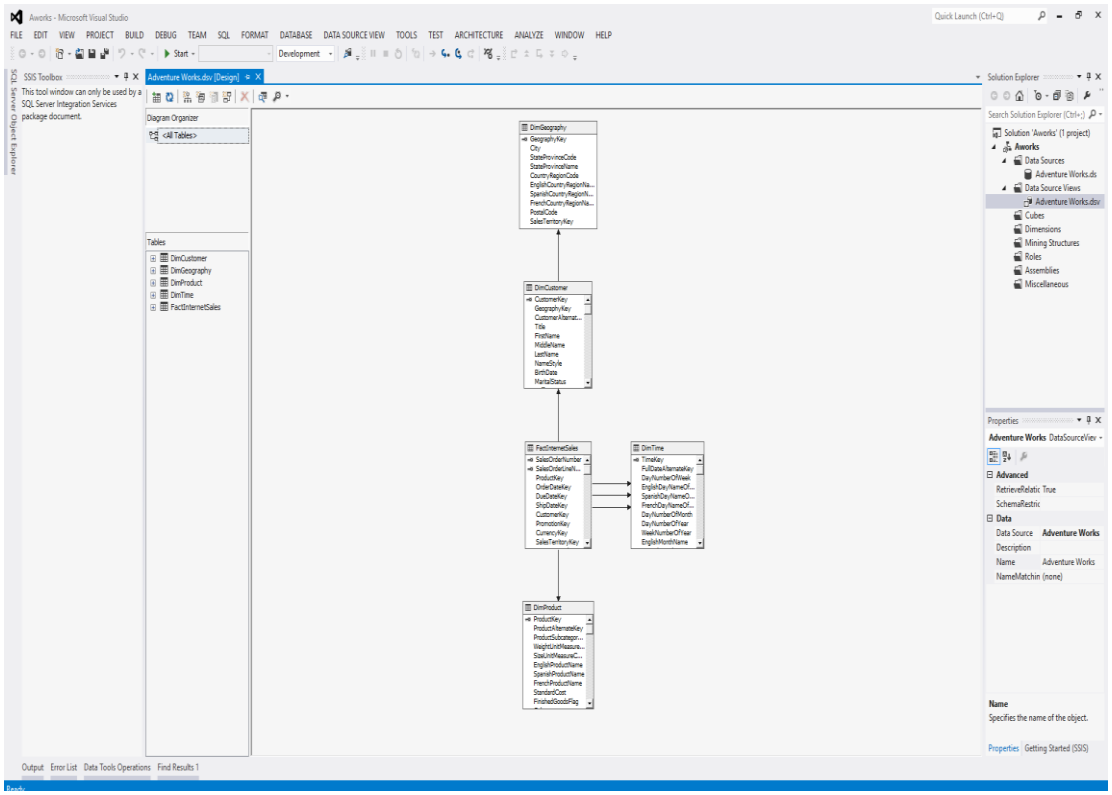


Figure 1.2.7 Create a New Data Cube

Select the “Adventure Works” data source view, and check “FactInternetSales” as the measure table.

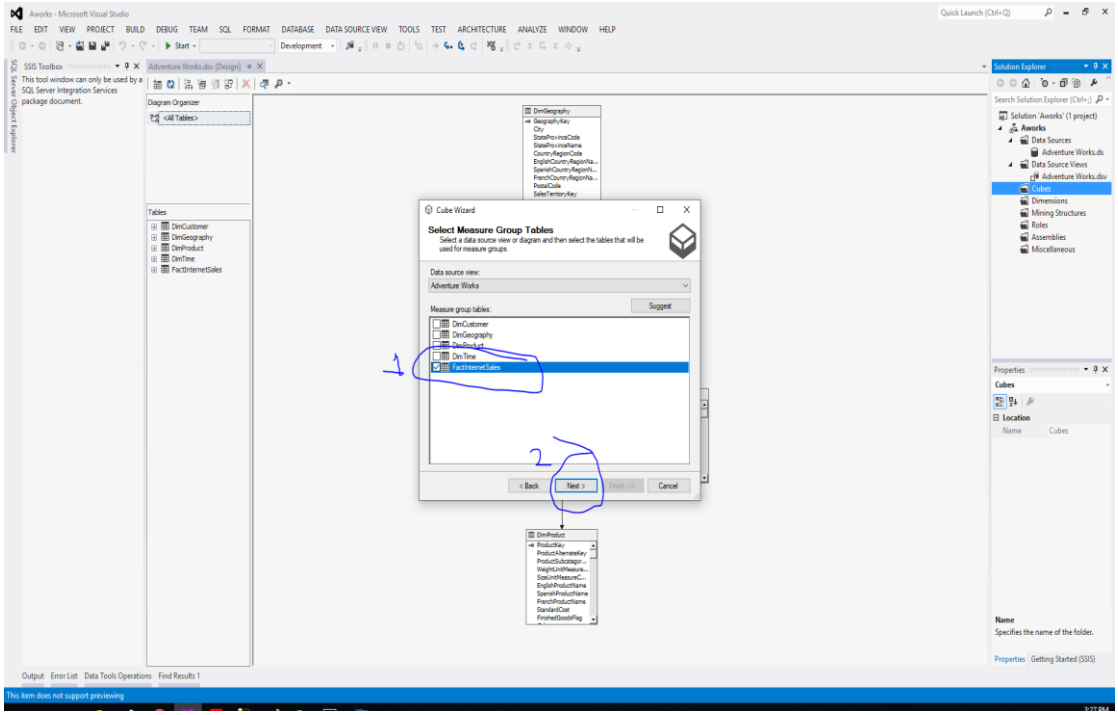



Figure 1.2.8 Select the Measure Group Table

In Select Measures page, uncheck four rows (Promotion Key, Currency Key, Sales Territory Key and Revision Number) because their units couldn’t be treated as measures.

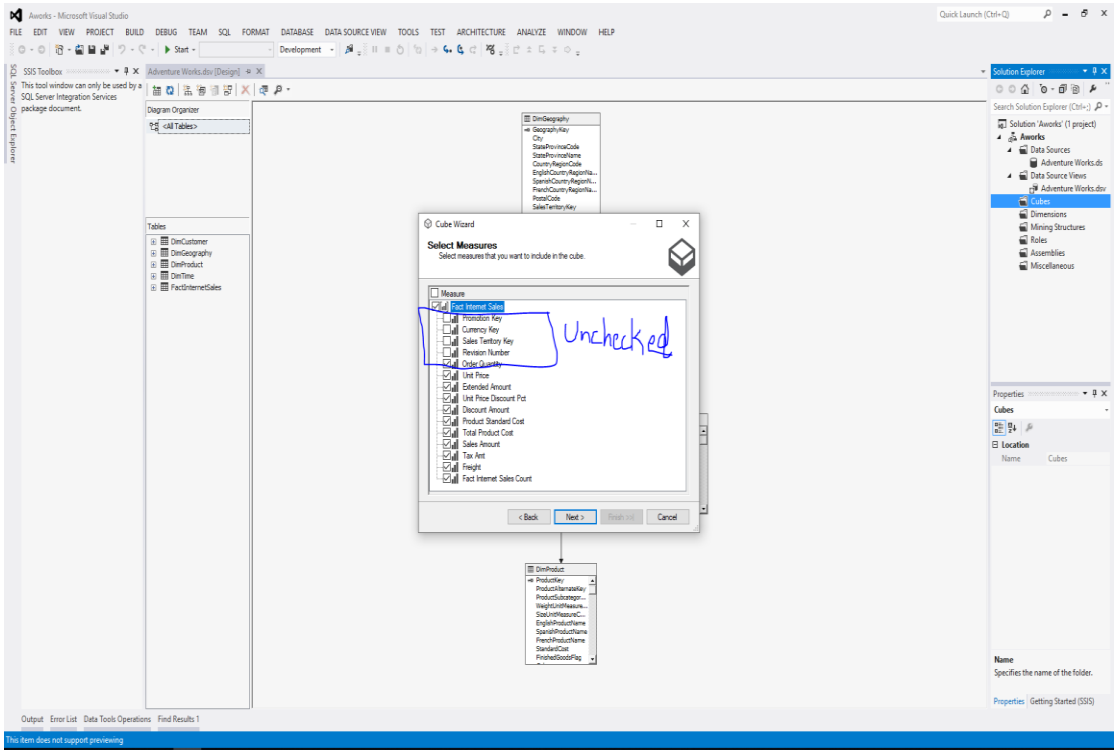



Figure 1.2.9 Uncheck Some Measures

Click 【Next】 continuously, until 【Finish】.

(2) Add a new attribute to "Dim Time" dimension. Double-click "Dim Time.dim".

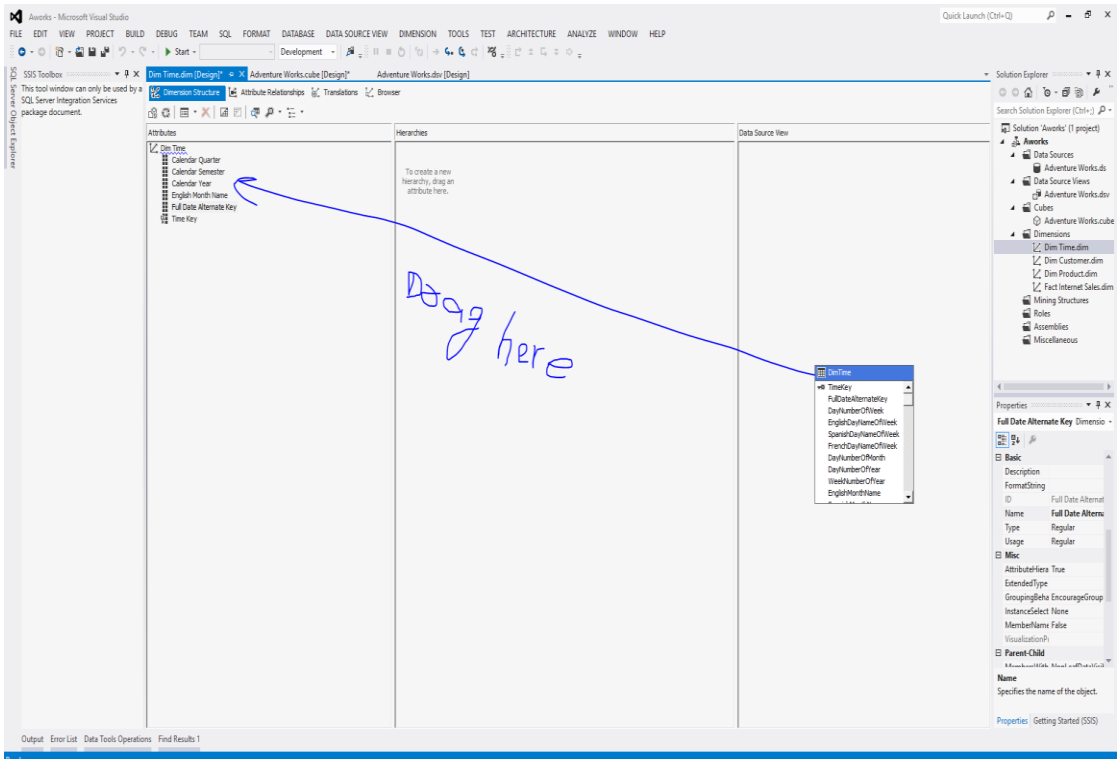



Figure 1.2.10 Time Dimension Modification

Open the dimension structure page, drag "Calendar Quarter" attribute from the right small window to the left panel.

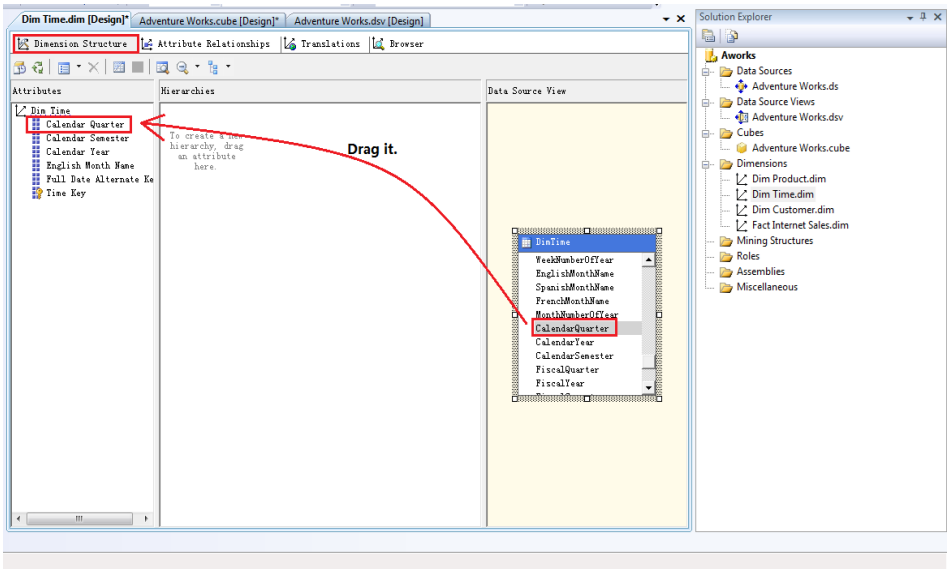



Figure 1.2.11 Drag the CalendarQuarter

Attention: You should do the same operation to add “Calendar Semester”, “Calendar Year”, “English Month Name”, “Full Date Alternate Key” attributes of “Dim Time.dim”, “Product Line” attribute of “Dim Product.dim”, “English Country Region Name” attribute of “Dim Customer.dim”. These attributes will be used in the following steps.

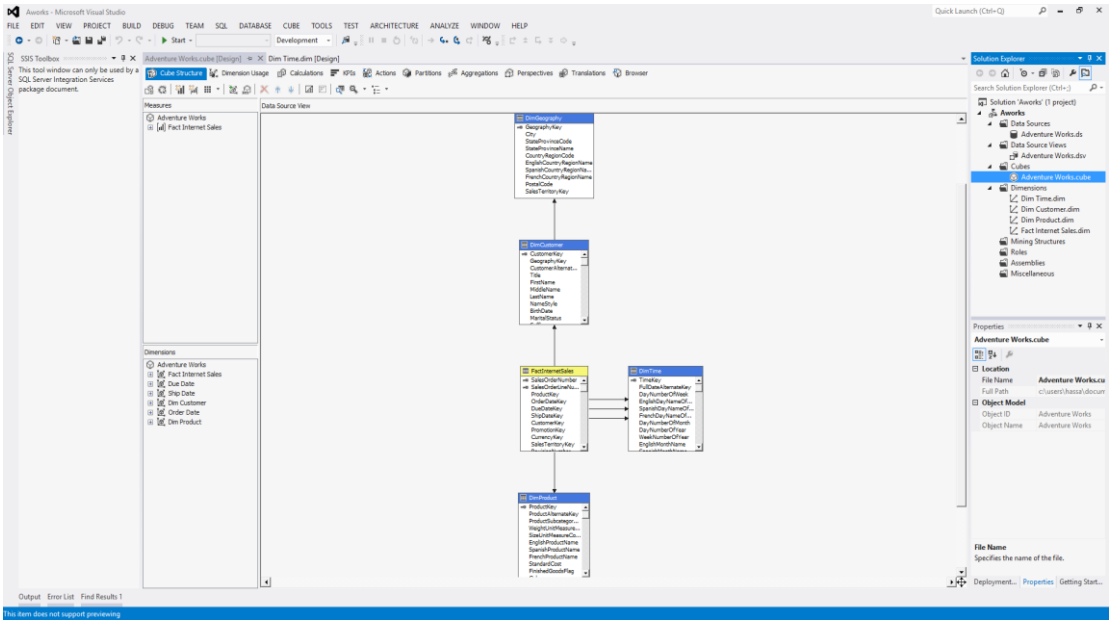


Figure 1.2.12 Deploy the Project

## Query with solution

We face a problem while deploying due to which it was not deployed and show different errors. In order to remove all those errors, we need to add user in SQL with login as same as in the error shown it is generic solution of this problem.

Go to SQL and click on new login

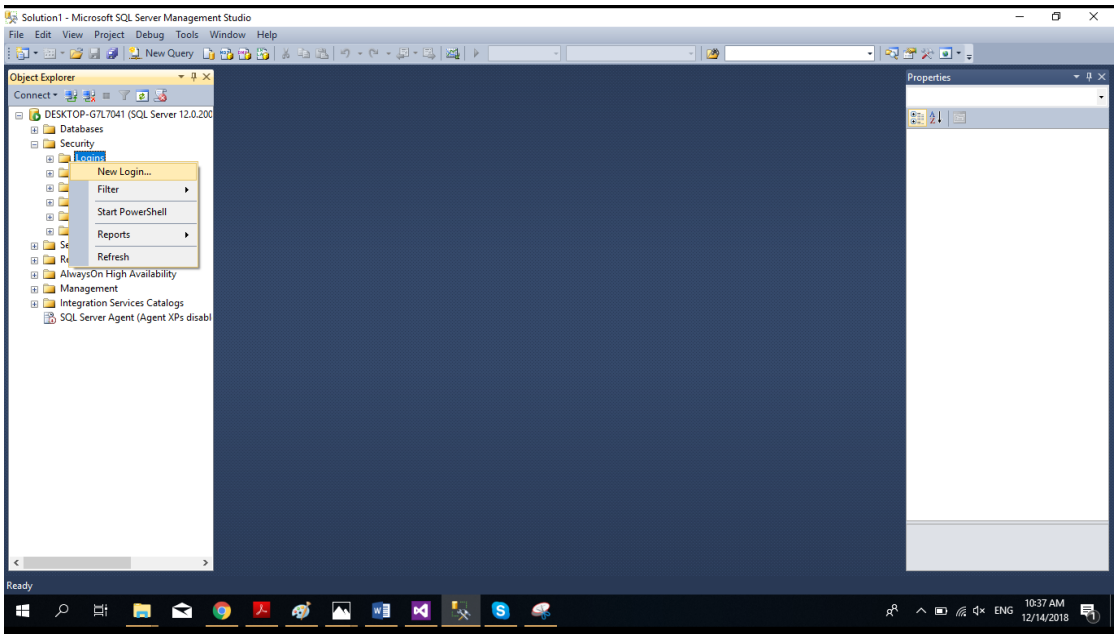



Figure 1.2.13 New login

It will show following logins, but we need to choose our desire one as shown in the figure.

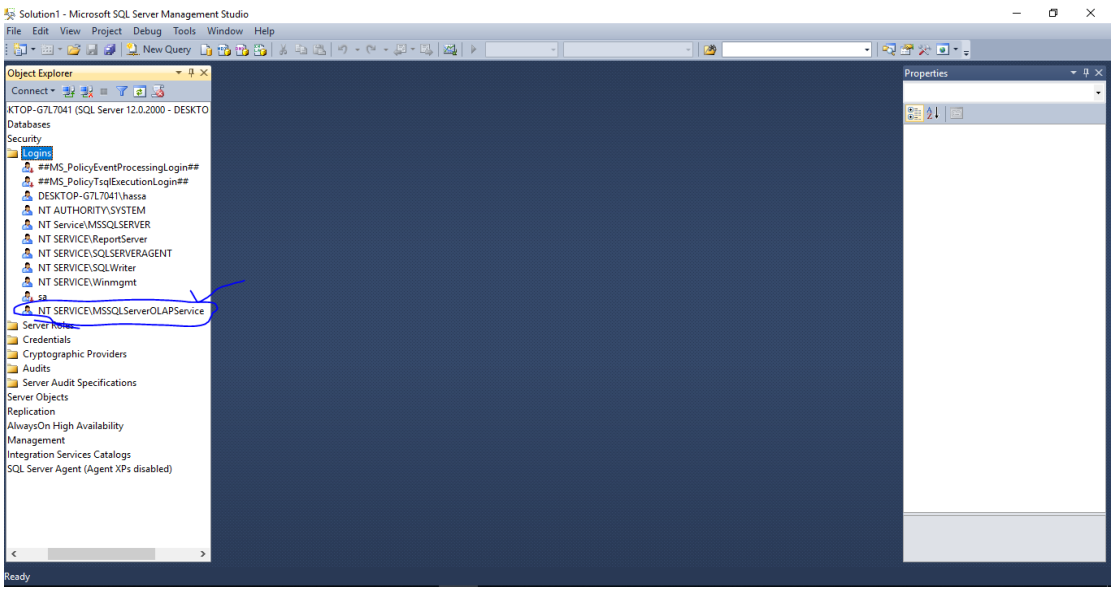



Figure 1.2.14 Solution

Now go to the properties of .NETSERVICE\MSSQLSERVEROLAPSERVICE then click on user mapping mark AdventureWorks as our file name is that you cn also add user according to your file then allow read and write after that click ok.

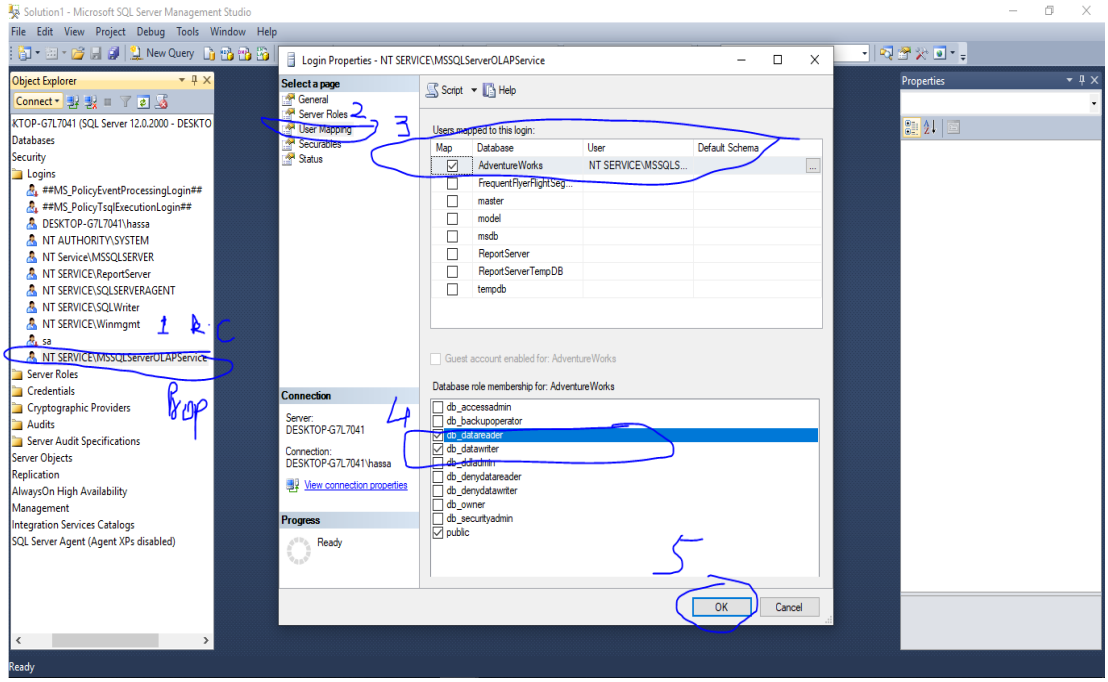



Figure 1.2.15 Complete Solution

(3) Deploy the project.

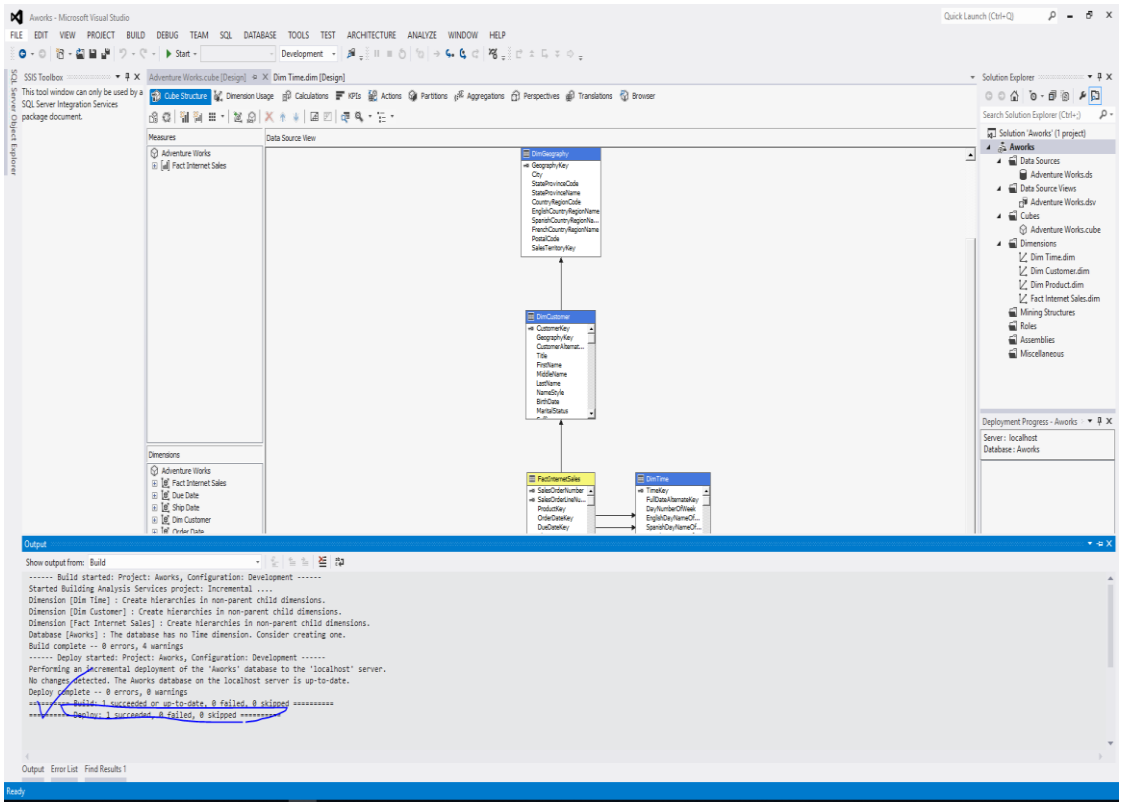



Figure 1.2.16 Successful Deploy Result

# Experiment 1.3
# The OLAP for Data Analysis

## Section One: The Goal of the Experiment

Learn to use the four operations of the OLAP to analyze the data warehouse.

## Section Two: The Content of Experiment

**Slice:** Here you use the data cube of Aworks as an example. First you should choose the dimension of customer and the dimension of product in this cube. Secondly you should choose one of the attribute members in the time dimension (example: 2012.01). Finally, you will get a slice of the cube of the sales of product in the dimension of product and customer.

**Dice:** Choose some or all members of the attribute in the three dimensions of the cube. You may obtain a dice if you set the value of the time dimension as an interval (example: 2003.01~2003.06) rather than a single attribute member and it can be regarded as six dices from 2003.01 to 2003.06.

**Drill:** It has contained two operations—drill up and drill down. The view from senior data to detail data is regarded as drill down and the view from detail data to senior data is regarded as drill up.

**Pivot:** pivot means that whether change the directions of the dimension of the report or the display of the page or not. You may get a data with different views through the process of pivot.

## Section Three: The Procedure of Experiment

Switch to Brower tab, click【Reconnect】in the tool bar of the designer.

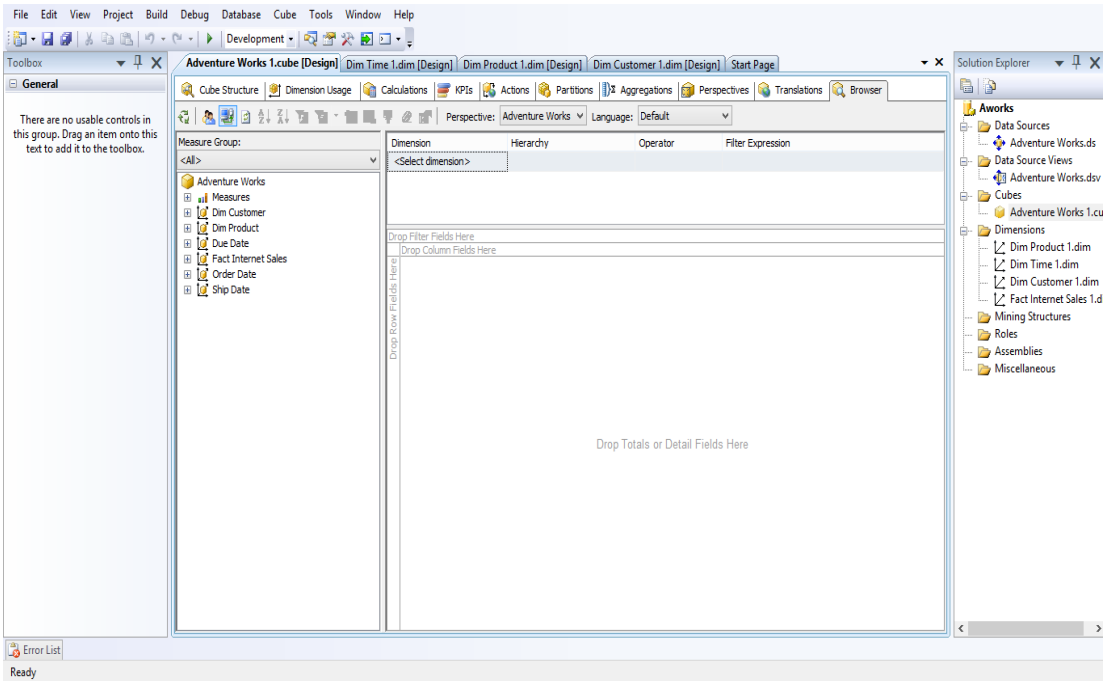



Figure 1.3.1 Open Browser Page and Reconnect

## 1.Slice

(1) Fold “Fact Internet Sales” and drag the "Sales Amount" to the center.

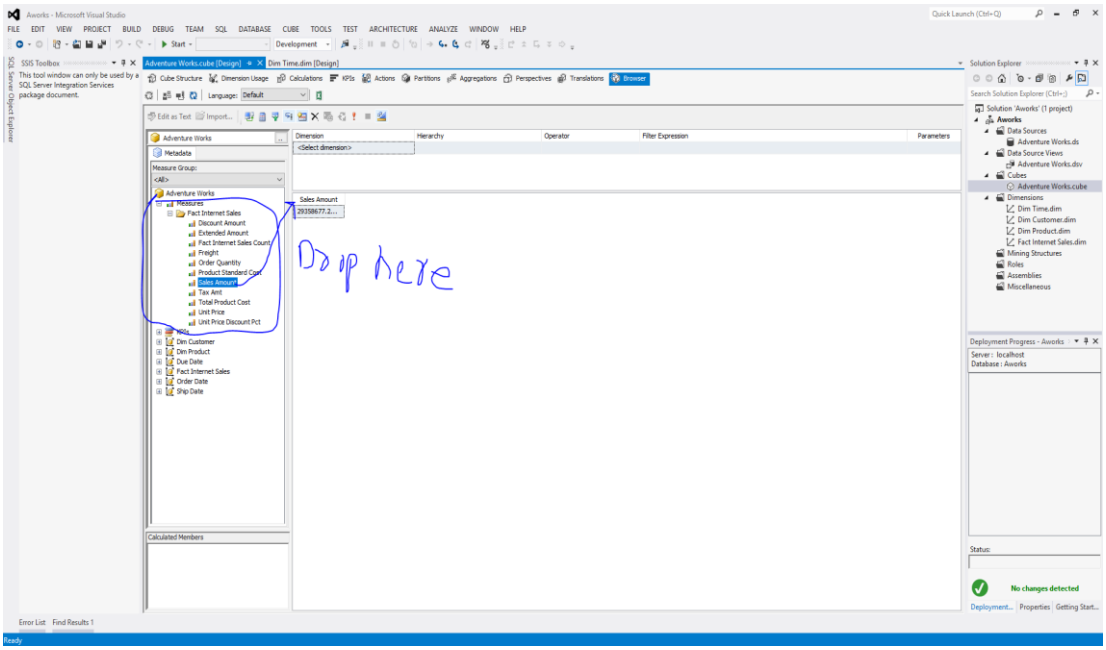



Figure 1.3.2 Drag the Sales Amount

(2) Fold “Dim Customer” and drag the “English Country Region Name “to the left column.
Attention: if the dimension tables don’t contain the dimension which you need, please consider the experiment 1.2 to reset your dimensions, then deploy and connect to browser.

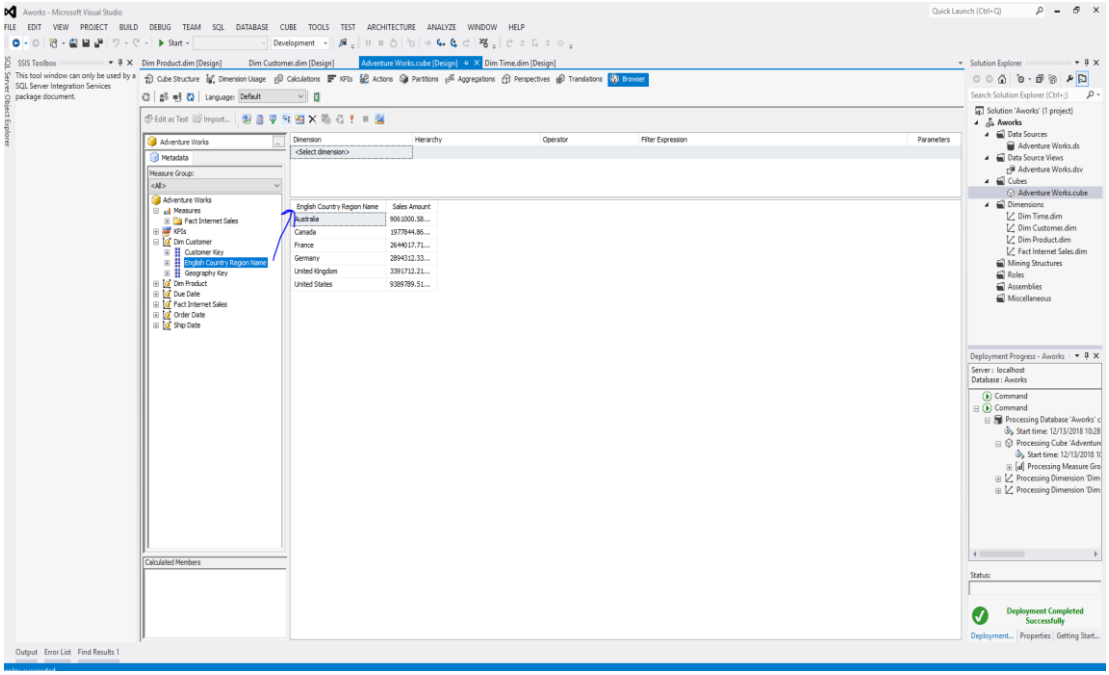


Figure 1.3.3 Drag the English Country Regions Name

(3) Unfold “Dim Product” and drag the " Product Line " to the second row.

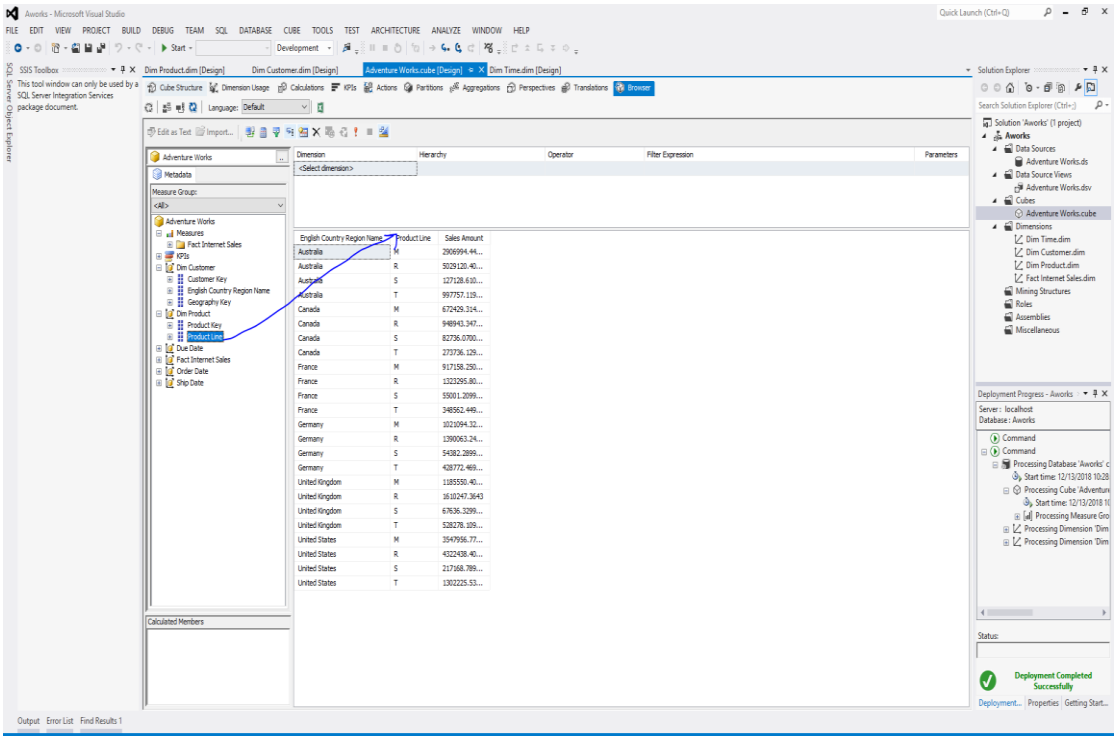


Figure 1.3.4 Drag the Product Line

(4) Unfold “Dim Date” and drag the “Due Date.Calendar Year” and “Due Date.English Month Name” to the first row.

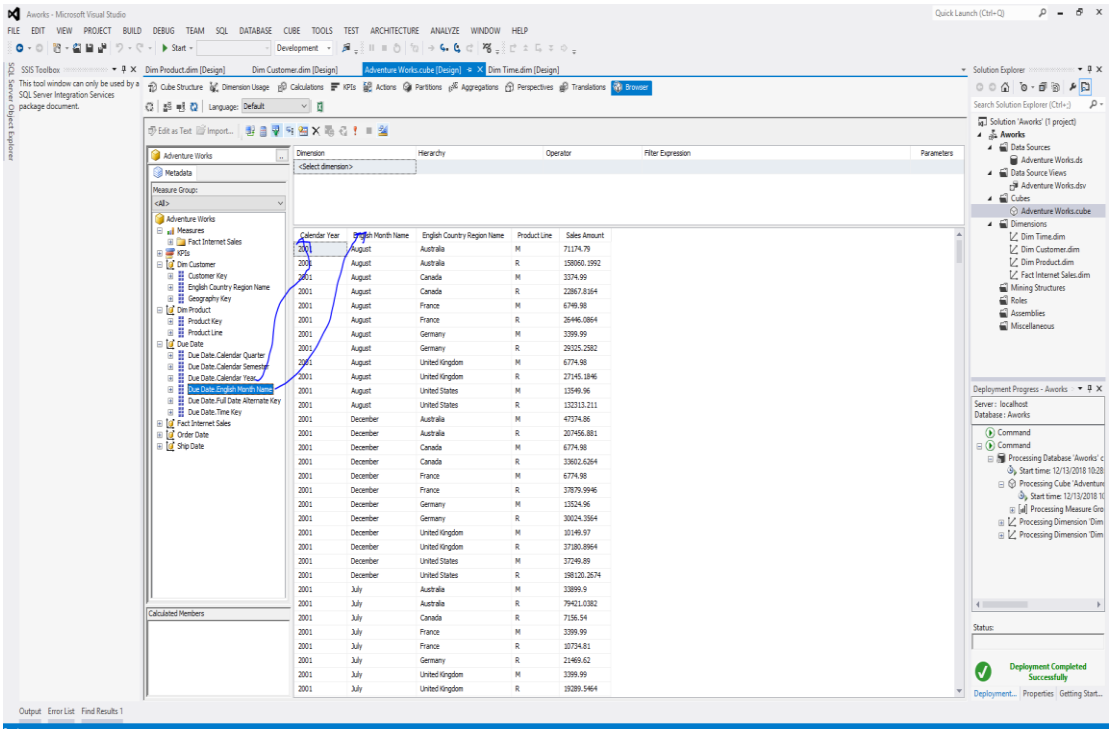


Figure 1.3.5 Drag Two Fields

(5) Select "2002" of “Due Date.Calendar Year” and "January" of “Due Date.English Month Name”. After those steps, you will get the network sales in January 2002.

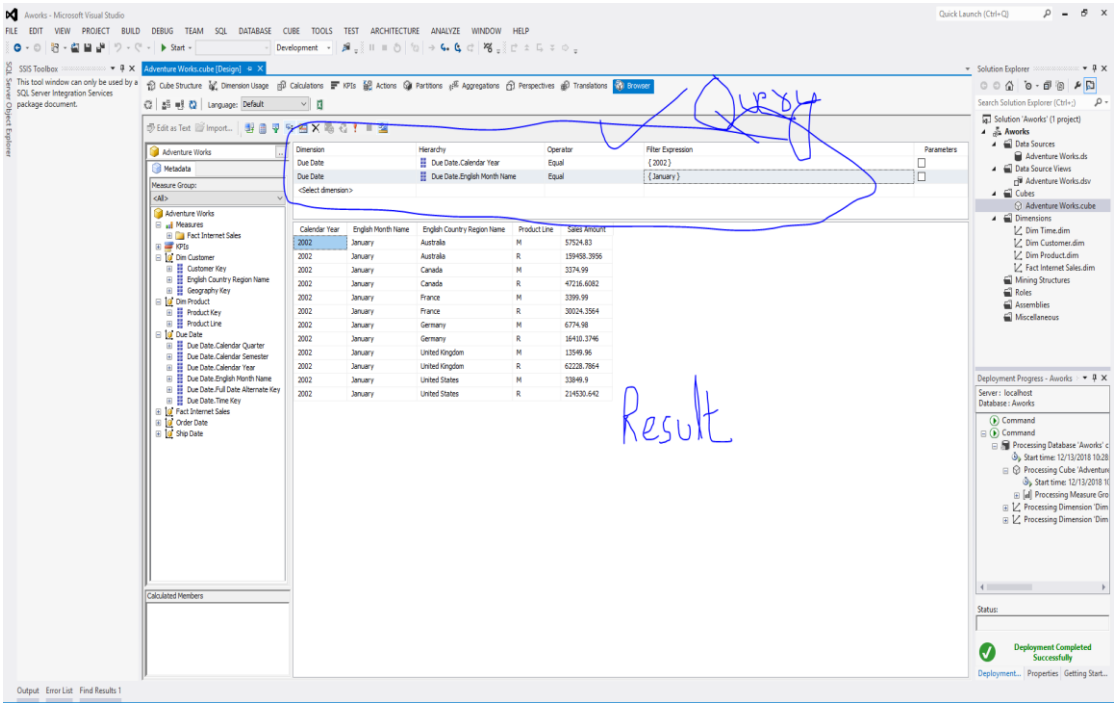



Figure 1.3.6 The Result of Slice

## 2.Dice

The operation of dice is to create a report by specified months of 2003, as is shown in the following figures.

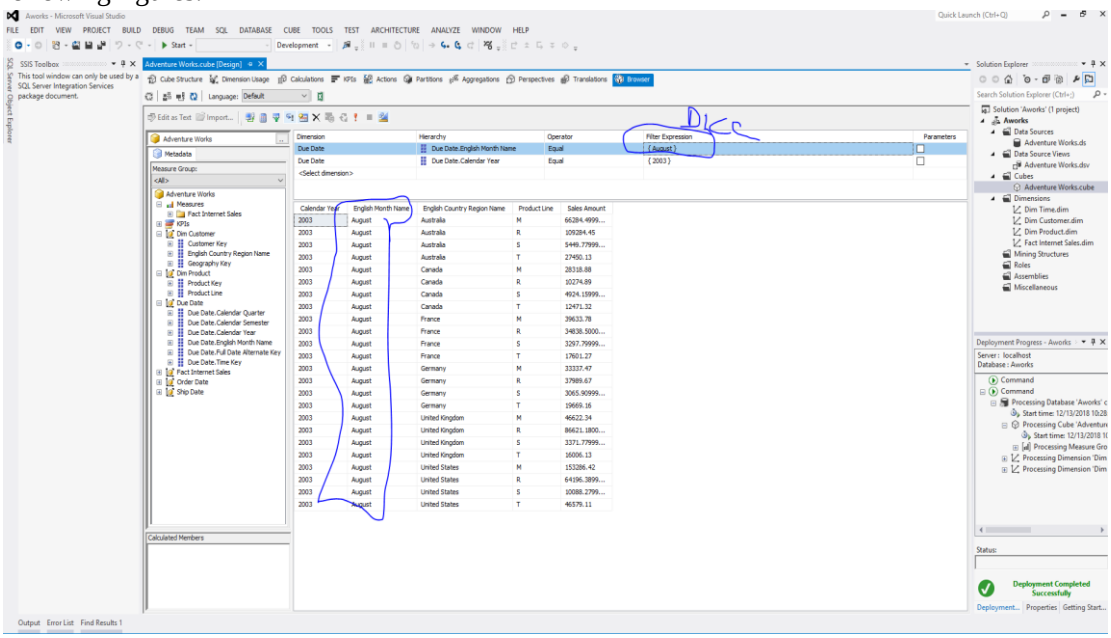


Figure 1.3.8 The Result of Dice

## 3.Drill

(1) Unfold “Fact Internet Sales” and drag the "Sales Amount" to the center.

(2) Fold “Dim Product” and unfold “Dim Date” and drag the “Due Date.Calendar Year”, “Due Date.Calendar Semester”, “Due Date.Calendar Quarter”, “Due Date.English Month Name” and “Due Date.Calendar Full Date Alternate Key” to the first row one by one.

(3) Unfold “Dim Customer” and drag the " English Country Region Name " to the left column.

You will get a report like the following figure after you have finished all the steps above. You can click the plus sign or the minus sign to drill up or drill down.

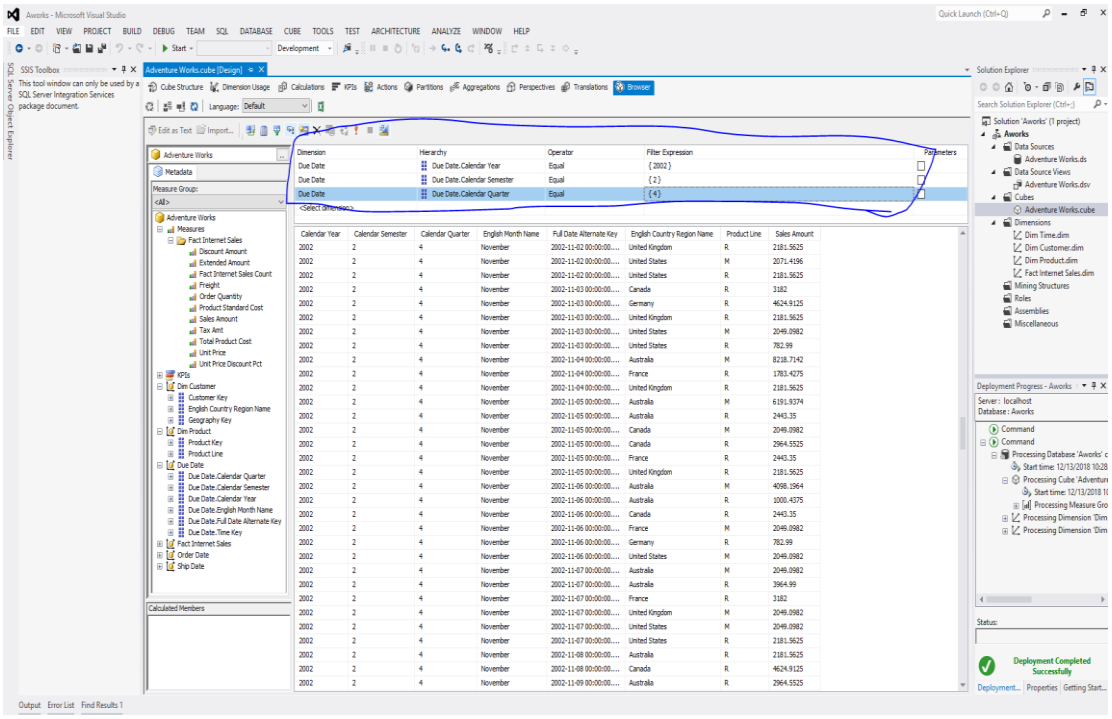


Figure 1.3.9 The Result of Drill Up or Down

# 4.Pivot

Pivot is to switch the row with the column of the report.

The original report is shown below:

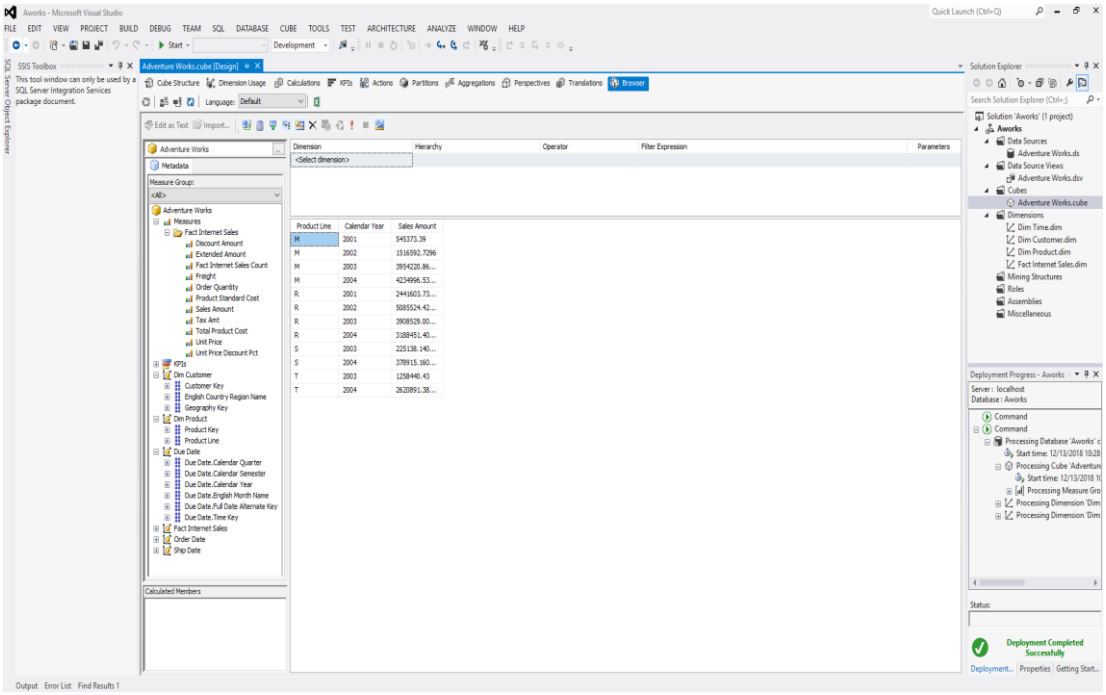


Figure 1.3.10 The Original Report

After pivoted:

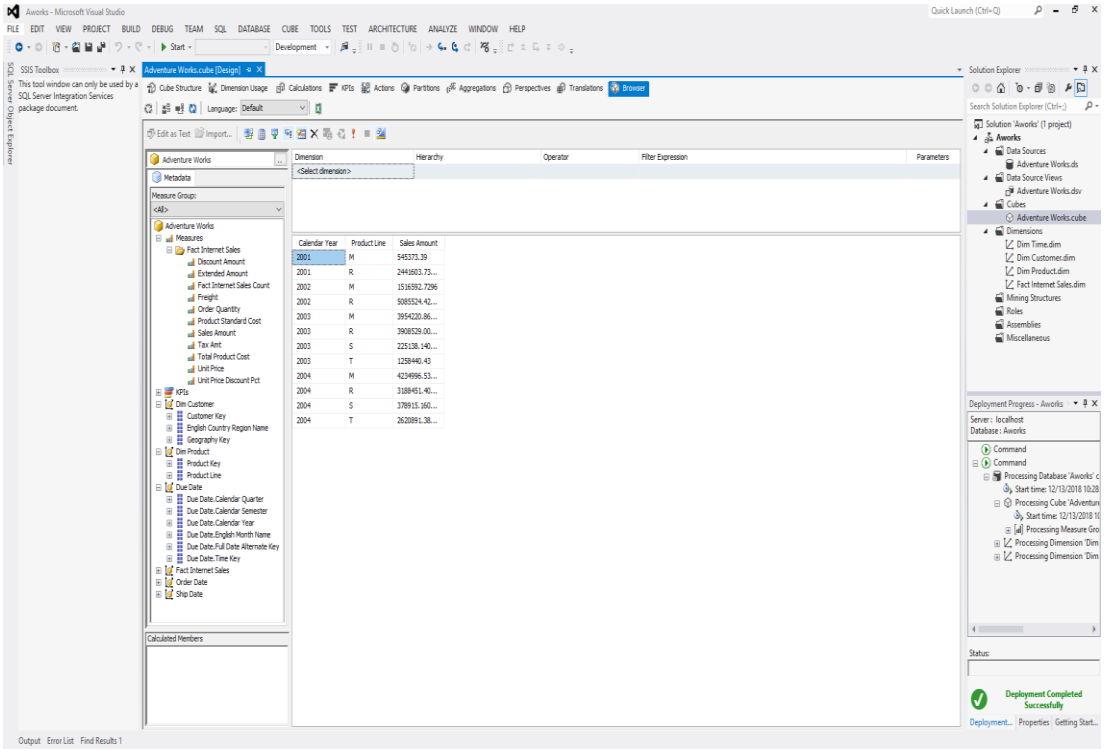



Figure 1.3.11 The Result after Pivot

# Experiment 1.4

## Build a data warehouse to manage 'Frequent-Flyer Flight Segment' information which is your homework in Chapter 4.

You should insert at least 40 data facts.

Your dimension tables should include 2 customers (one of them should be yourself), 2 years and at least one flight per month per person, 2 airline companies, 10 different airports.

# Experiment 1.5

## Do OLAP based on the data warehouse of Exp 1.4

1) Show the flight information of two passengers during Oct.-Dec. 2017 by taking Air China airplane.
2) Show the airport information that you have used for all flights in the whole year in 2017 or 2016.

# Experiment 2. Data Mining

## Section One: Several Important Concepts

### 1. Definition and Purpose of the Data Mining

Data mining: the process of efficient discovery of previously unknown patterns, relationships, rules in data warehouses and other data sets.

Goal: Help people analyze and understand the data.

### 2. Steps of a KDD Process:

(1) Data Cleaning: Fill in missing values, smooth noisy data, identify or remove outliers and resolve inconsistencies.

(2) Data Integration: Integrate multiple databases, data cubes or files.

(3) Data Selection: Extract the data which is useful for analysis tasks from database.

(4) Data Transformation: Normalize and aggregate data.

(5) Data Mining: Use intelligent methods to extract data modes.

(6) Pattern Evaluation: Identify the real interesting pattern which represents the knowledge.

(7) Knowledge Presentation: Use the technology of visualization and knowledge representation to provide users with the knowledge of mining.

## Section two: Techniques of Data Mining

### 1. Association Rule Discovery

There is a set of records which contains some numbers of items from a given collection. Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

### 2. Clustering

Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering.

### 3. Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a day will be sunny, rainy or cloudy. Popular classification techniques include decision trees and neural networks.

# Experiment 2.1 Mining Association Rules

## Section One: The Goals of Experiment

(1) Learn to mine the association rule from transaction database or data warehouse.

(2) Learn to use the SQL Server.

## Section Two: The Content of Experiment

According to a supermarket's customers" personal information and purchase data of goods, use Apriori algorithm to mining the association rules. The data file is a text file named *BASKETS.txt*. The data consists of two parts: one part includes the customer's personal information, the number of VIP Card(cardid), the amount of consumption(value), payment method(pmethod), sex, whether the head of household(howeown), age and income; another part is about the purchase data of goods including fruits and vegetables (fruitveg), fresh meat (freshmeat), dairy products (dairy), canned vegetables (cannedveg), canned meat (cannedmeeat), frozen food (frozenmeal),beer, wine, soft drinks (softdrink), fish and confectionery. All of these items are binary variables. Value T represents the purchase, and F indicates that no purchase. The target of analysis is to find what goods most likely be purchased at the same time.

## Section Three: The Procedure of Experiment

(1) Open SQL Server Management Studio, connect to your local server.

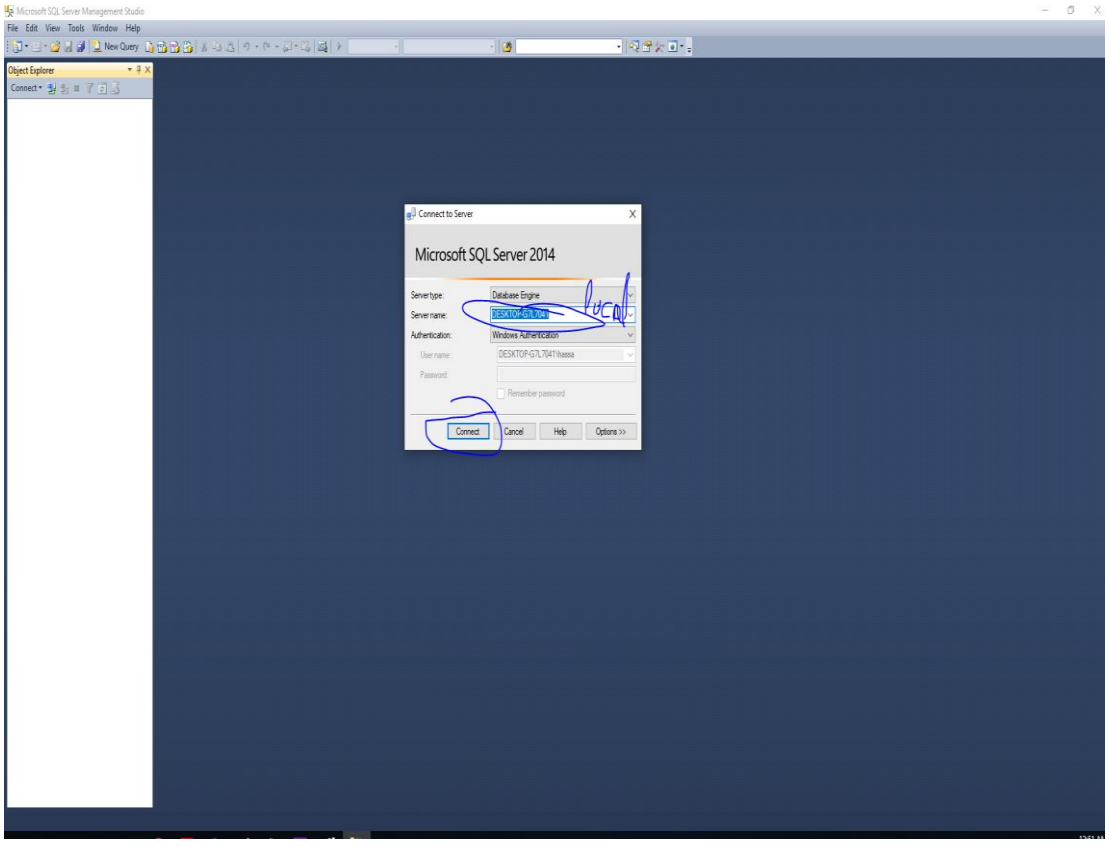



Figure 2.1.1 Connection to Local Server

(2) Create a new database.

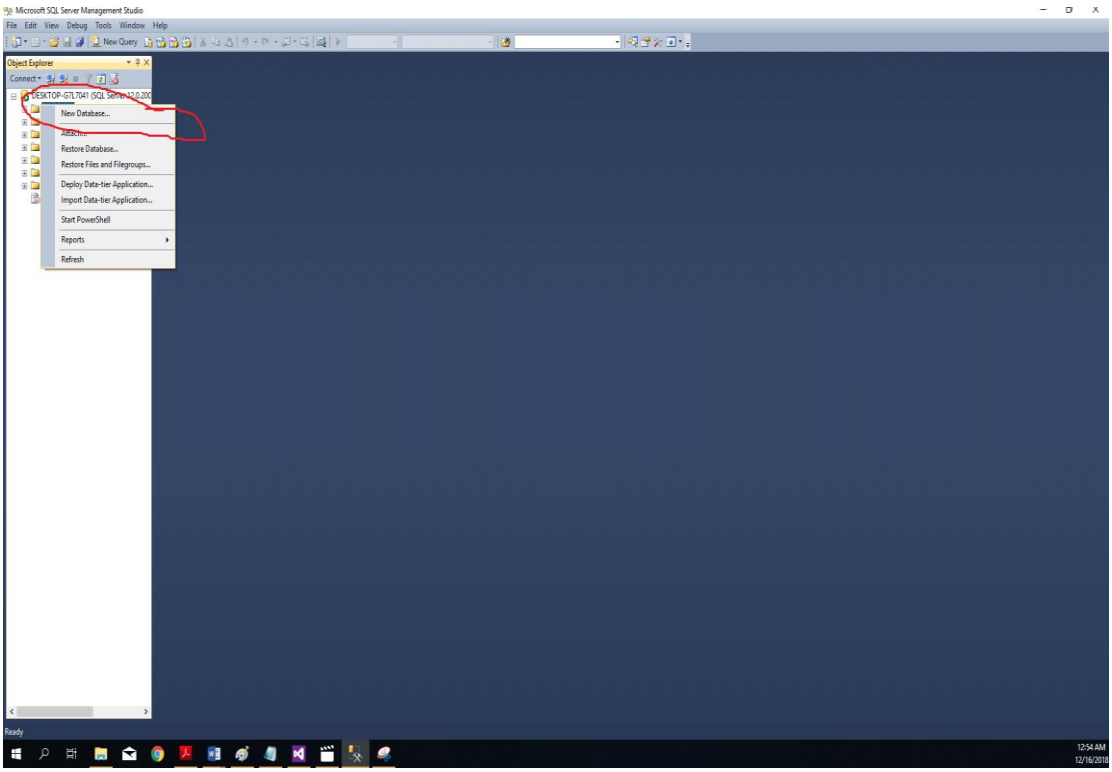



Figure 2.1.2 Create a New Database

Name it as “Aprior_sqlserver”.

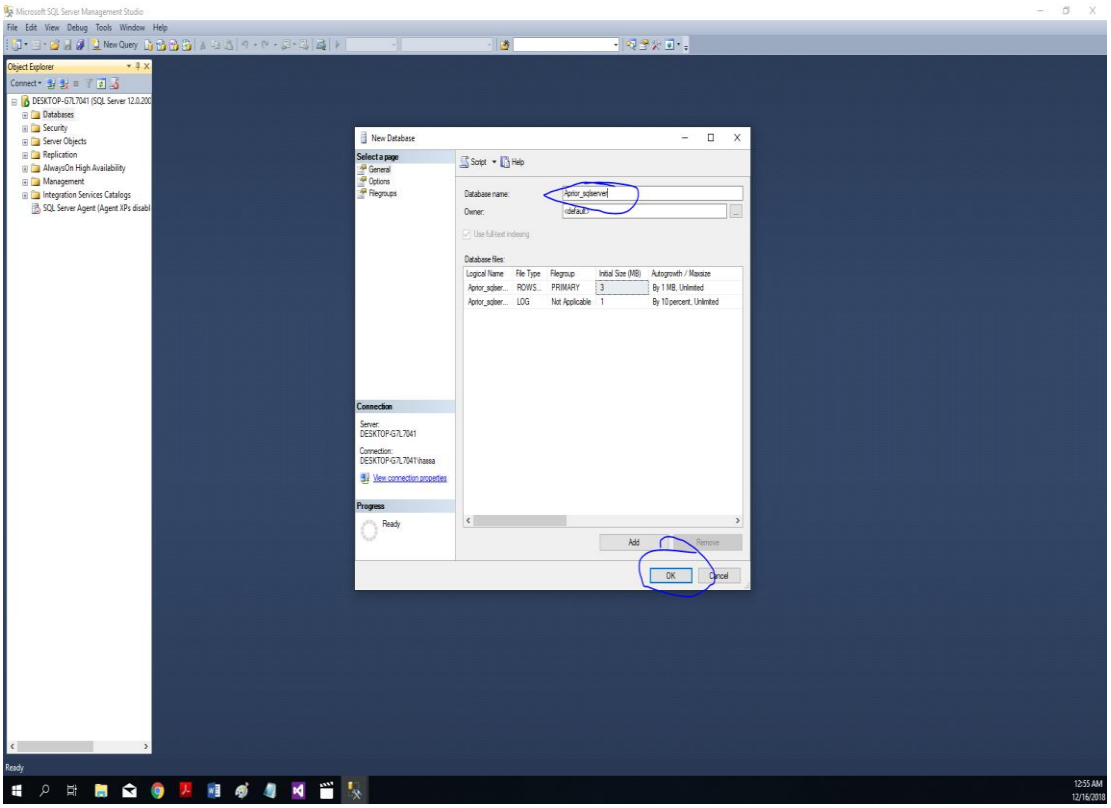



Figure 2.1.3 Rename the Database

(3) Right-click the database you just created. Click “Tasks” and select “Import Data…”

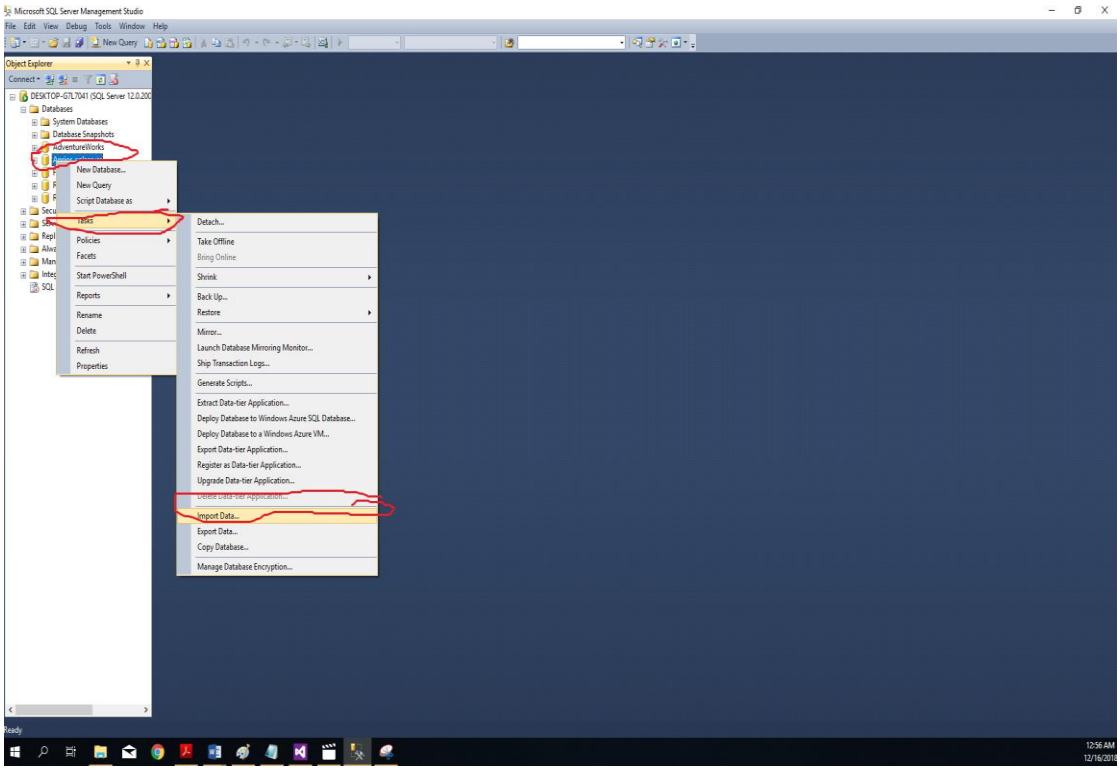



Figure 2.1.4 Import Data to “Aprior_sqlServer” Database

Find the text file, and do the same settings as follows.

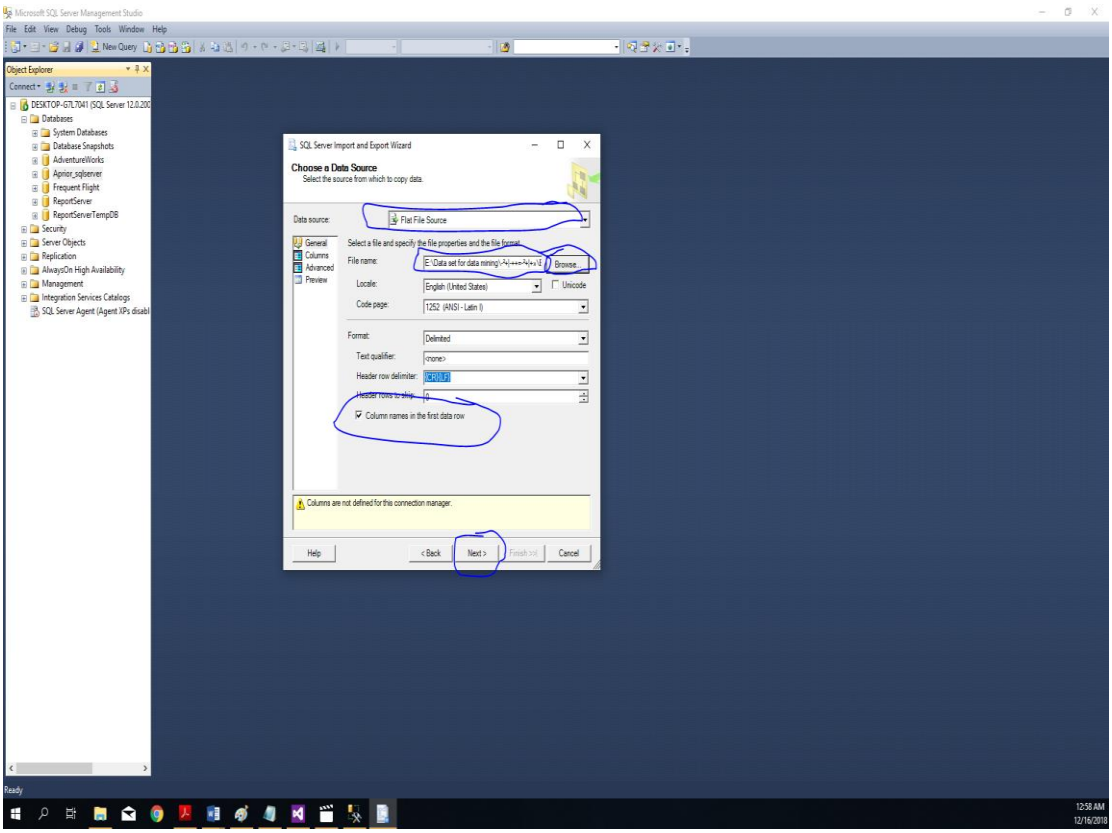



Figure 2.1.5 Import “BASKETS.txt” to “Apriori_sqlserver” Database

(4) After you importing data successful, you can find the table.

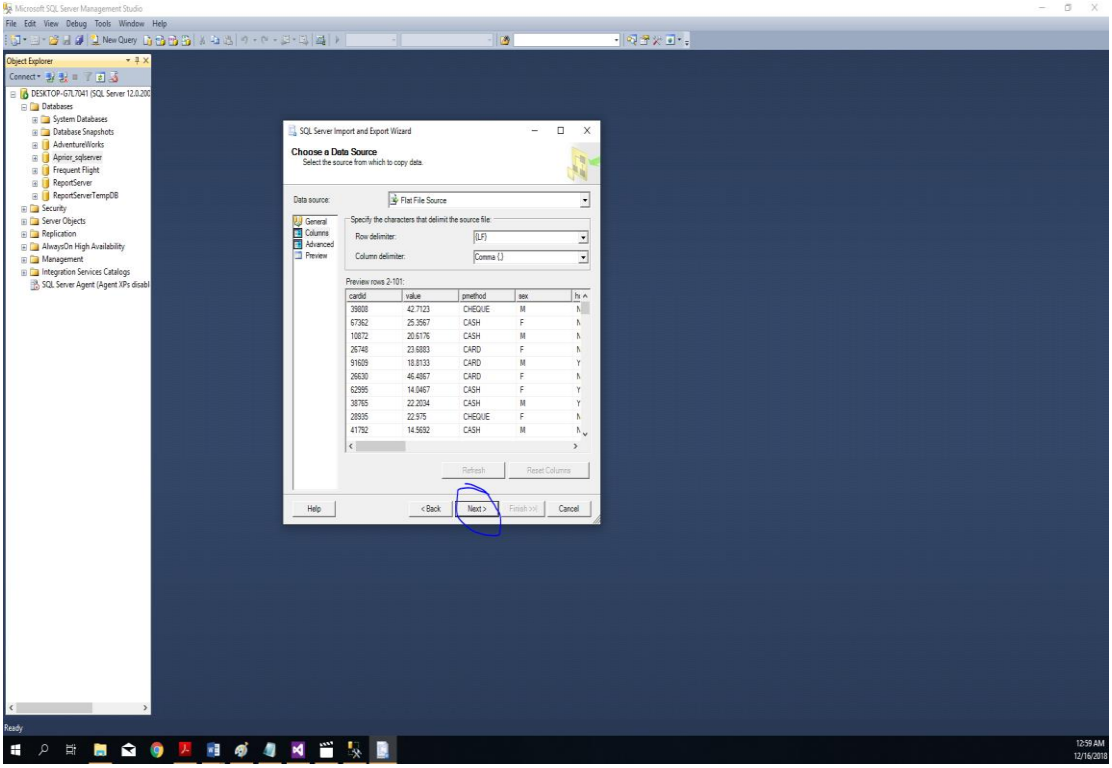



Figure 2.1.6 BASKET Table Data

(5) Open SQL Server Business Intelligence Development Studio. Create an "Analysis Services" project and name it as "Aprior".

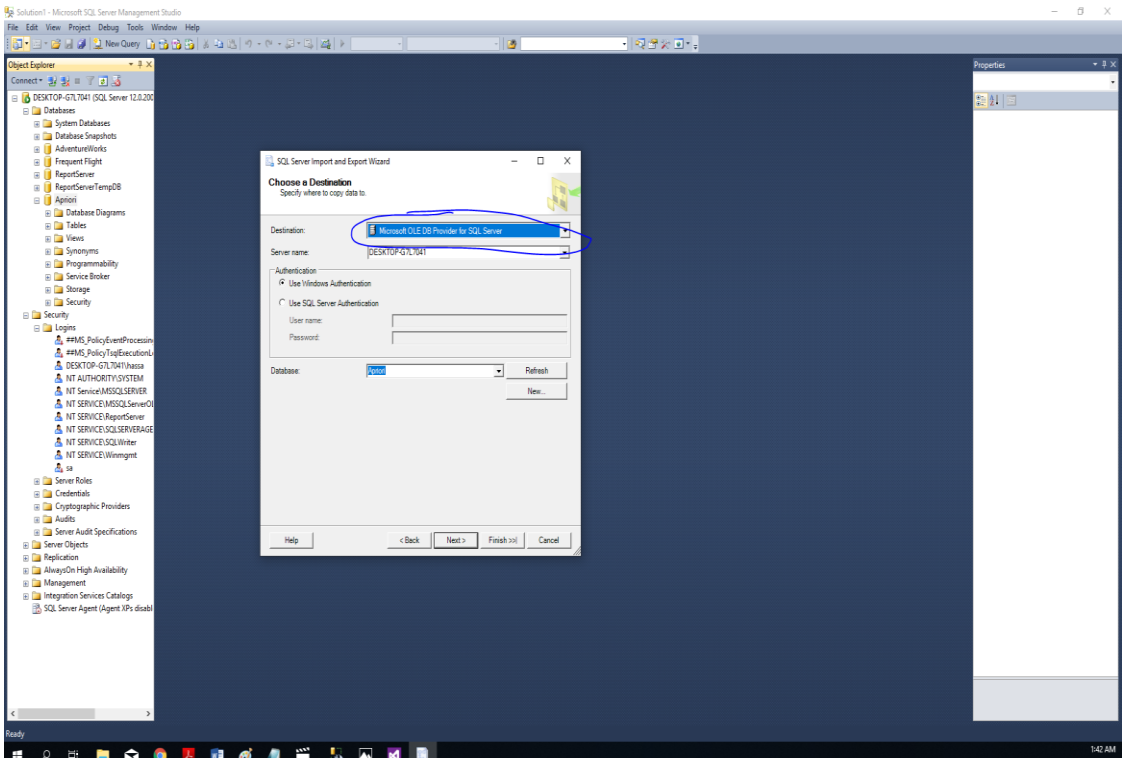



Figure 2.1.7 Create "Aprior" Analysis Services Project

(6) Create a data source.

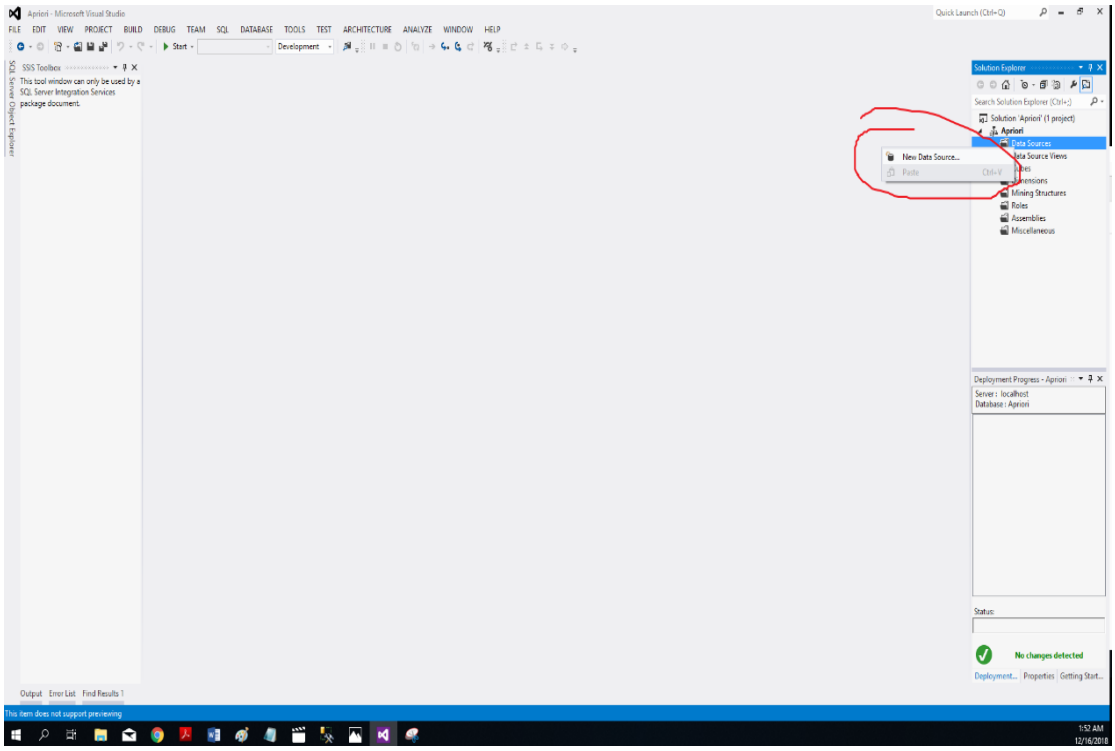



Figure 2.1.8 Create a New Data Source

In the wizard, create a data source based on "Apriori_sqlserver" database you just created.

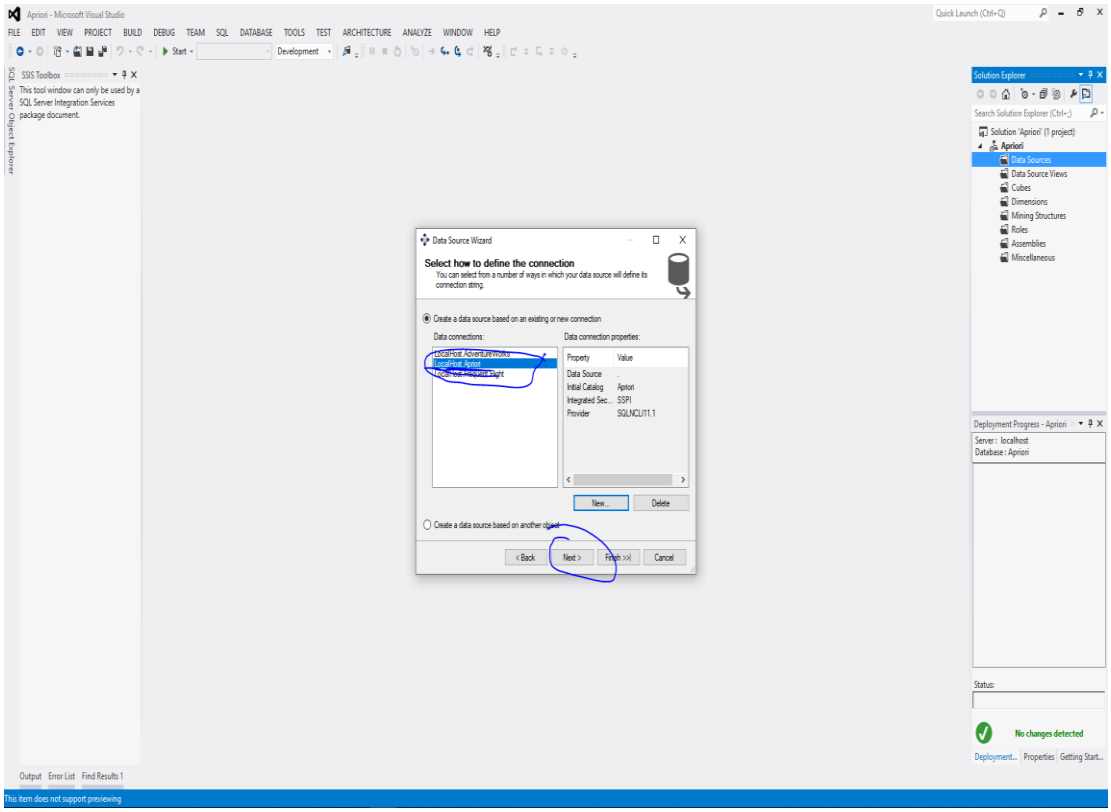



Figure 2.1.9 Create a New Database Connection

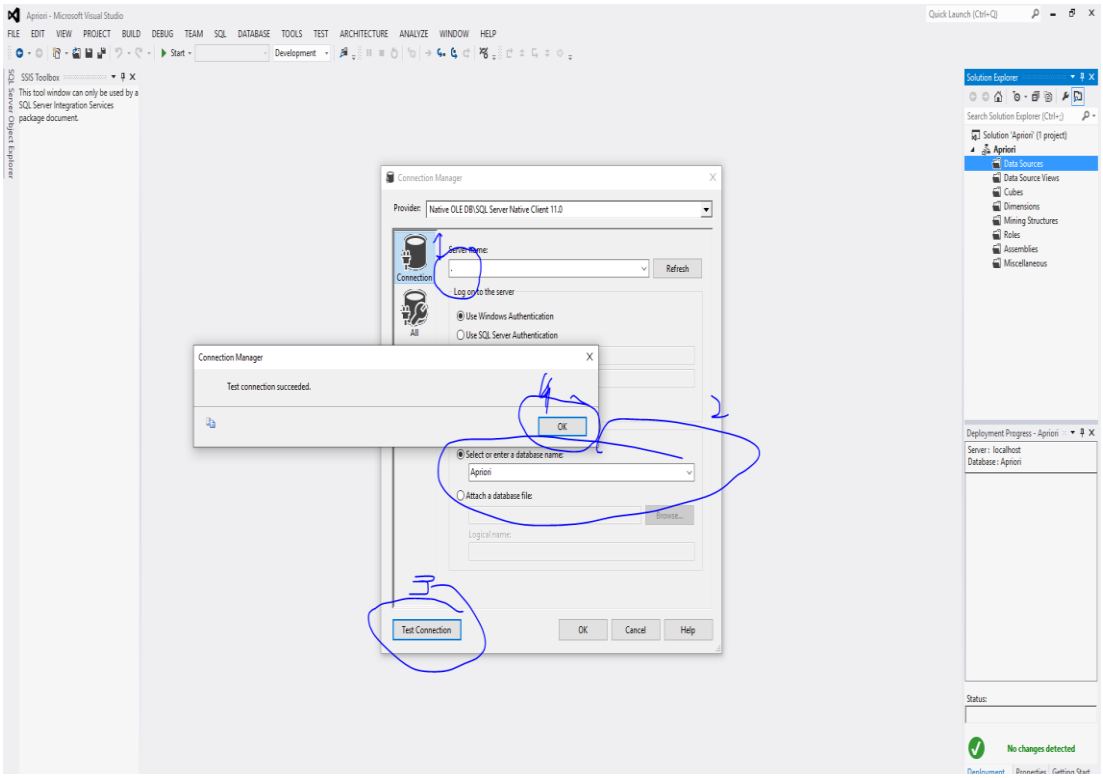



Figure 2.1.10 Connect to "Apriori_sqlserver" Database

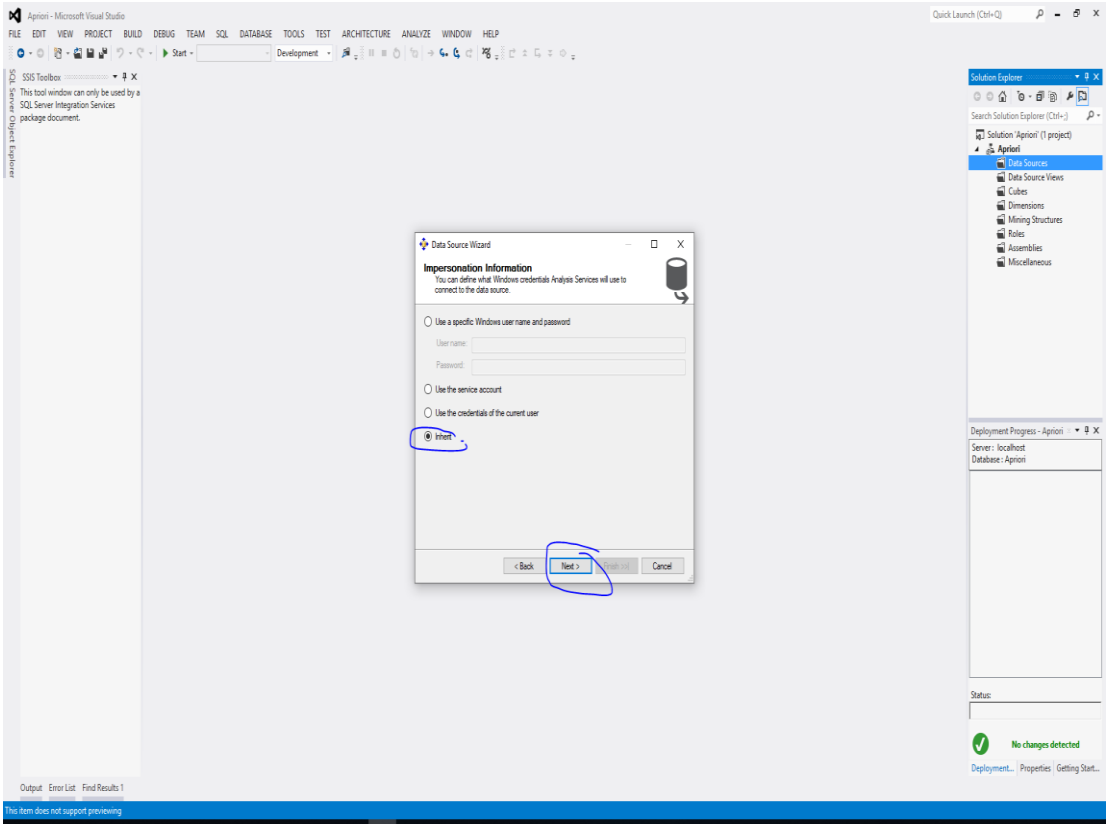



Figure 2.1.11 Select the Inherit Method

(7) Create a data source view.

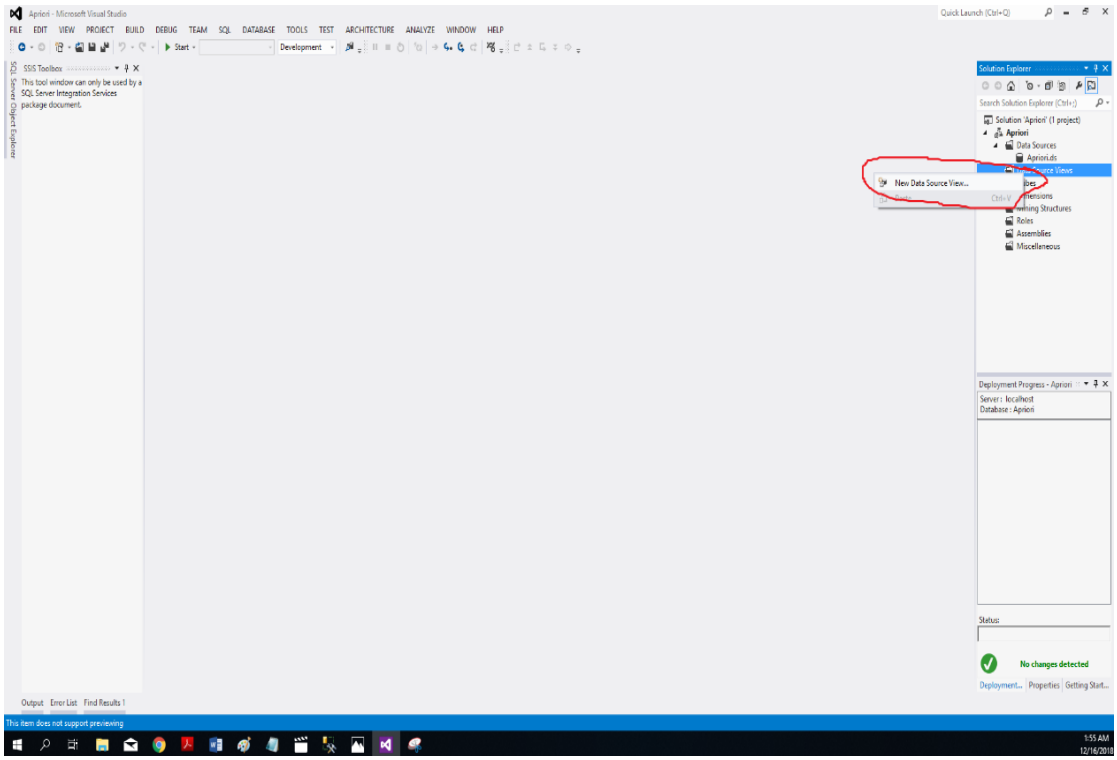



Figure 2.1.12 Create a New Data Source View

(8) Create a mining structure.

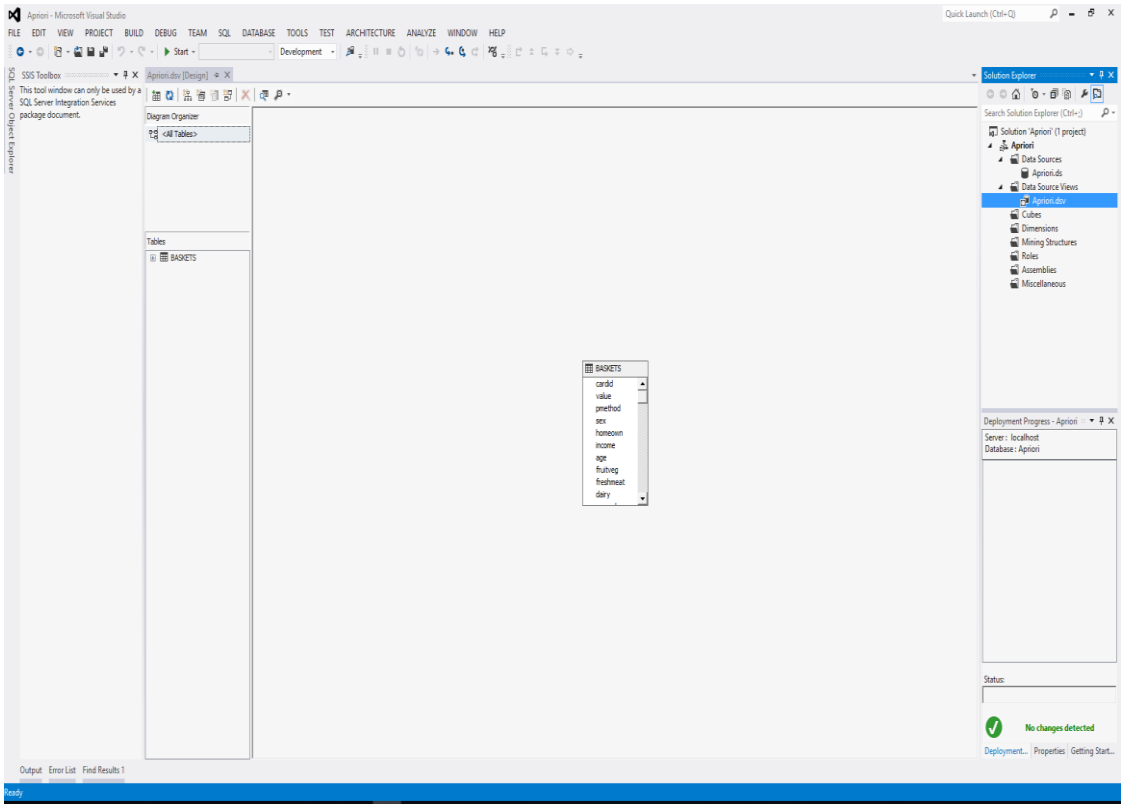



Figure 2.1.13 Create a New Data Mining Structure

In the wizard, select “Microsoft Association Rules” technique.

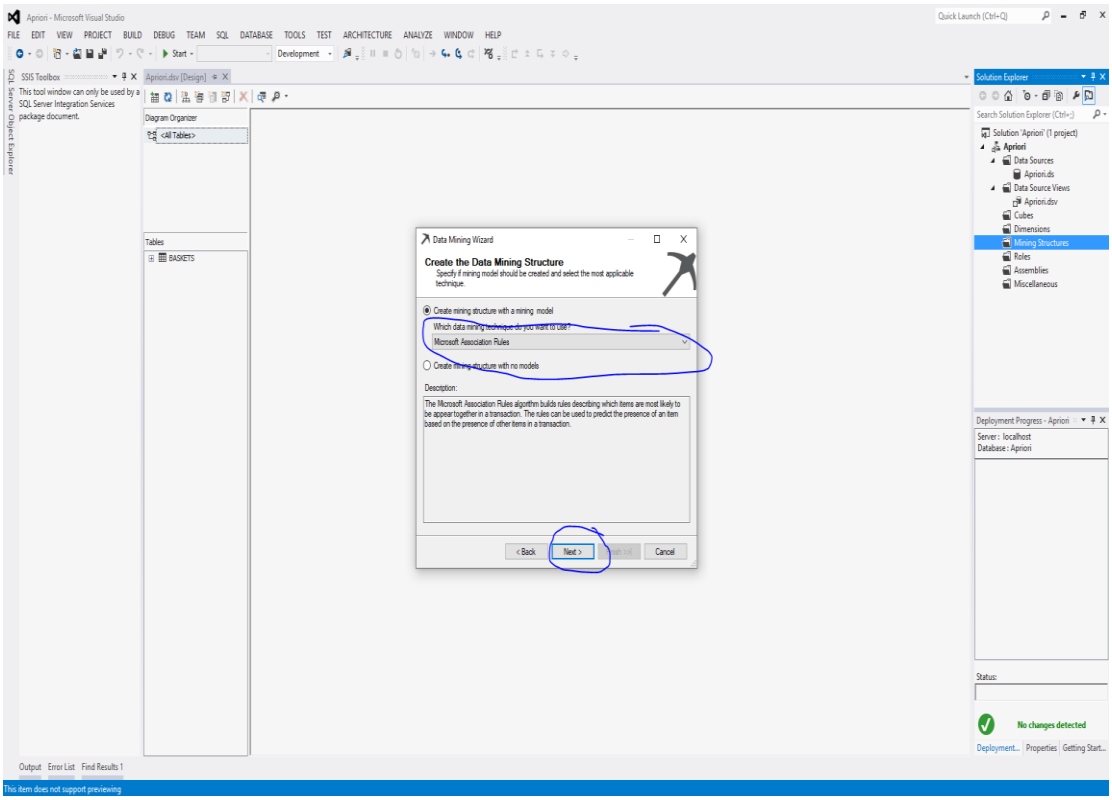



Figure 2.1.14 Select Microsoft Association Rules technique

Set “cardid” as the key and set all of the food as input and predictable parameters.

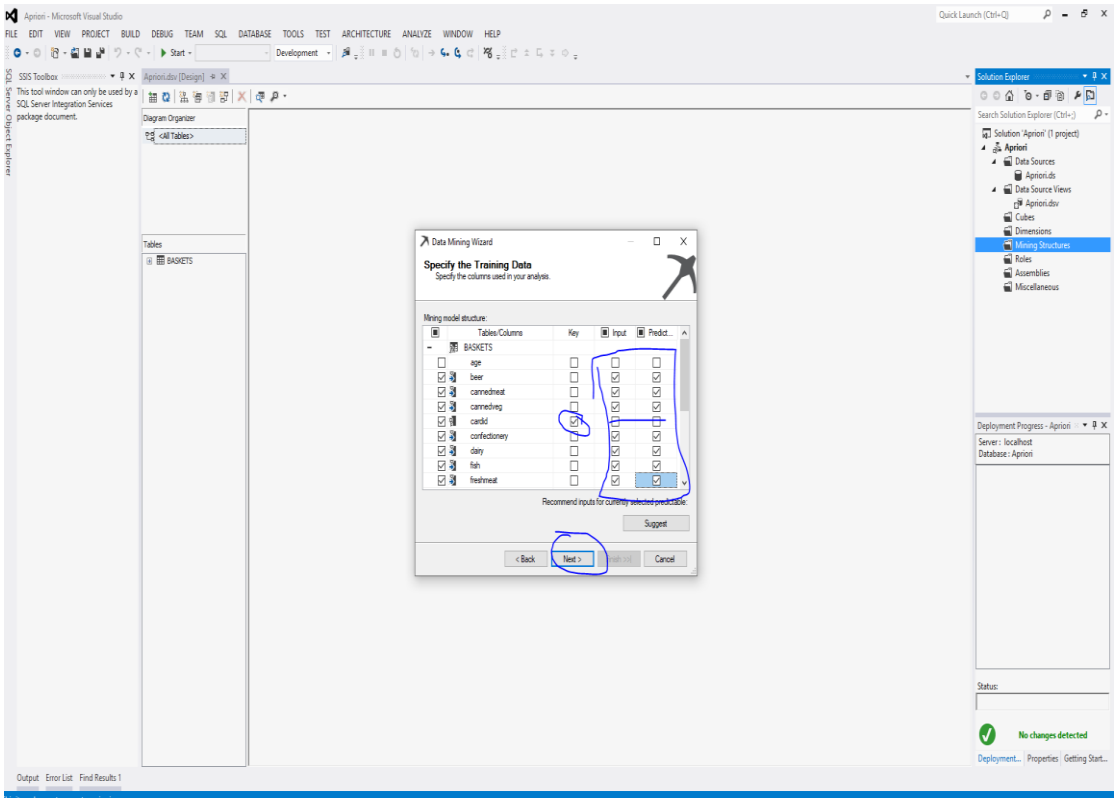



Figure 2.1.15 Specify the Training Data

(9) Deploy the project.

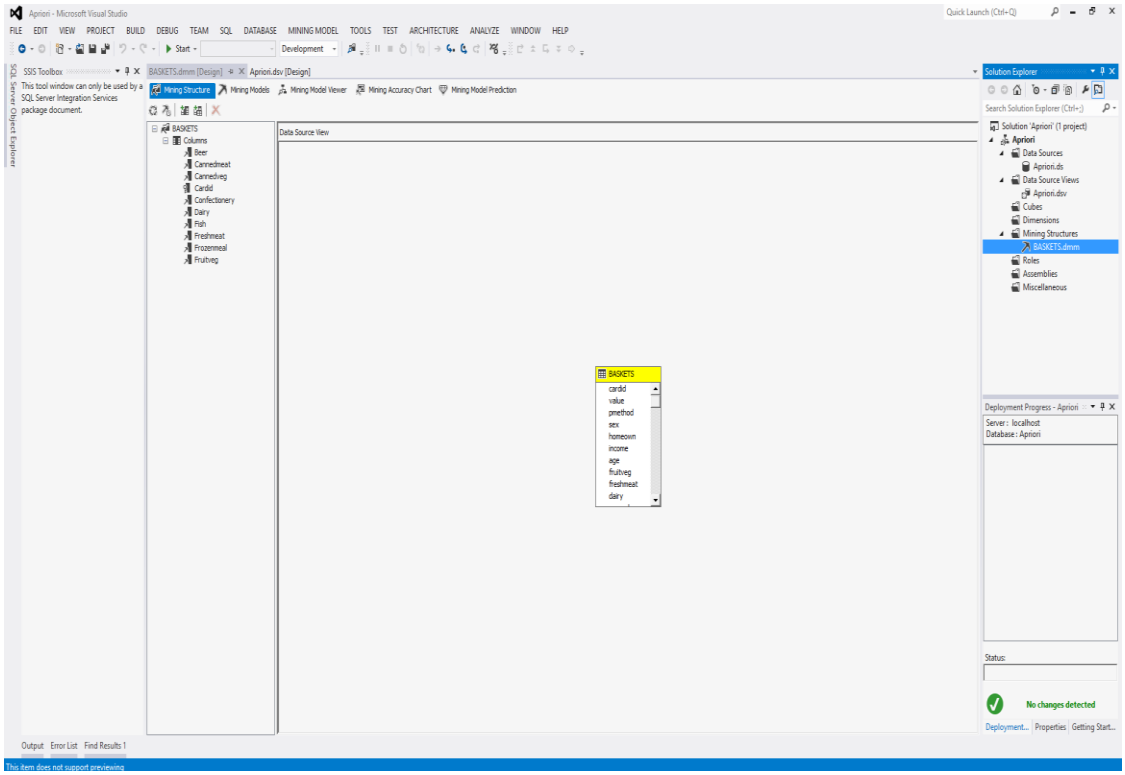


Figure 2.1.16 Deploy the Project

(10) Analyze the result.

The following figure shows all item-sets of the data set. The “Support” value stands for the possibility that all items happen at the same time. Thus, contrary to large “Support” value, the size of “Itemset” is small. When the size of “Itemset” is only one, the “Support” value is the largest.

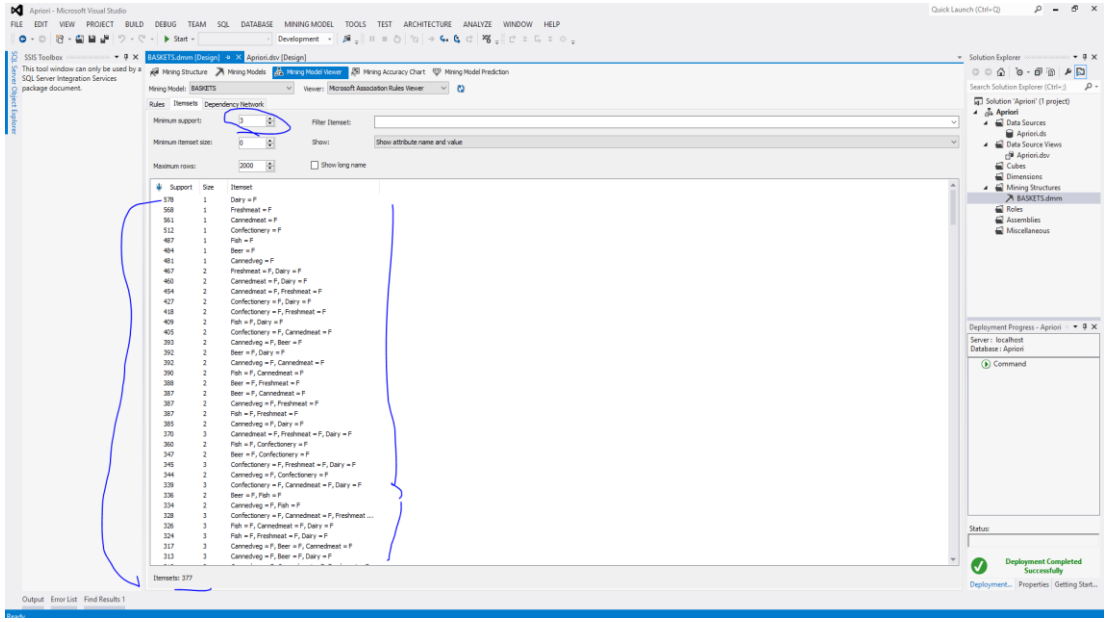


Figure 2.1.17 Itemsets Diagram

The following figure shows the relevance of buying different products simultaneously. From the figure, you will know when “Frozenmeal” and “Beer” were brought, “Cannedveg” would be brought too. When “Frozenmeal” and “Cannedveg” were brought, “Beer” would also be brought.

That means the possibility of buying “Beer”, “Cannedveg” and “Frozenmeal” at the same time is very great.

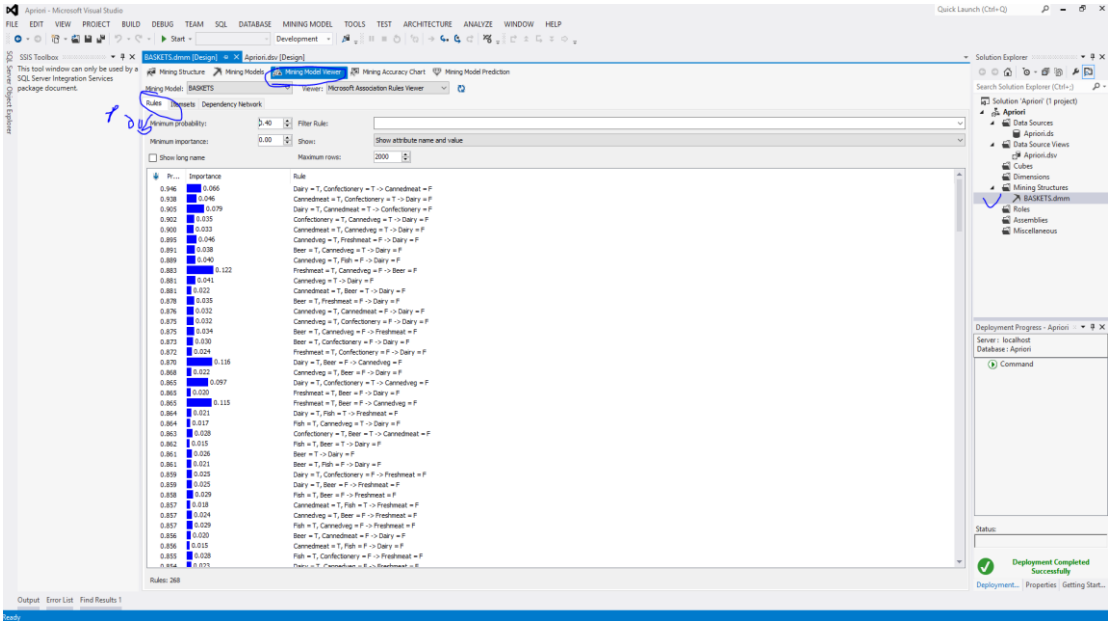


Figure 2.1.18 Rules Diagram

The following figure shows the dependencies of buying different products simultaneously.

Figure 2.1.19 The Dependency Network

In the following figure, move through the visualization by dragging the left slider. When the slider is much lower, the sales of “Beer”, “Frozenmeal” and “Cannedveg” is very great.

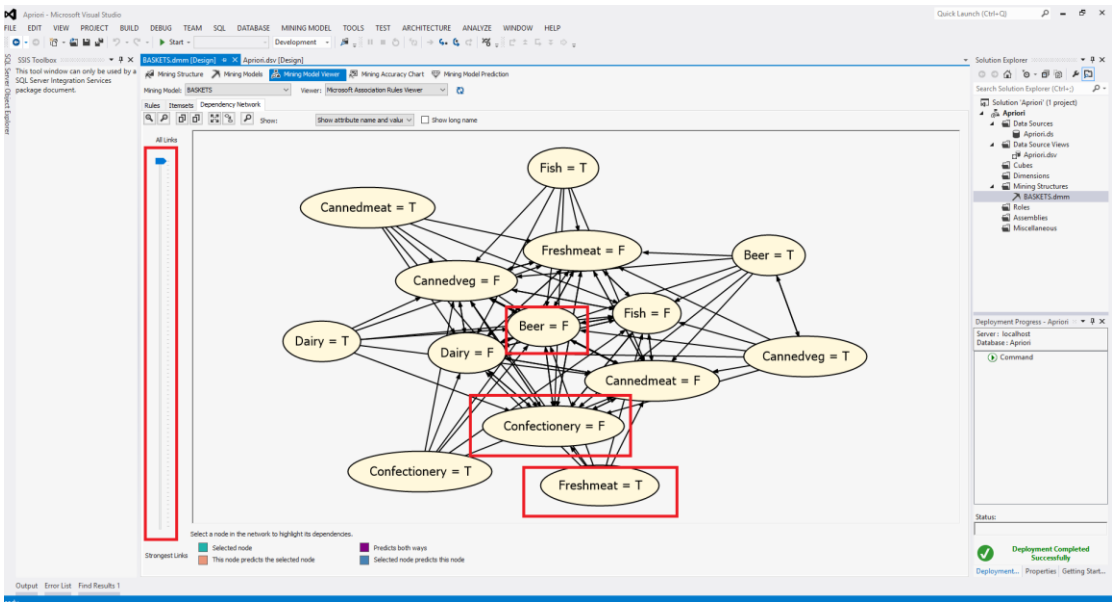



Figure 2.1.20 The Simple Dependency Network

# Experiment 2.2 K-Means Clustering Algorithm

## Section One: The Goal of Experiment

(1) Learn to use K-Means clustering algorithm for mining the information.

## Section Two: The Content of Experiment

The experiment data set comes from the book "Medical Statistics and Computer Experiment", edited by Fang JiQian for a plastic surgery hospital surgical. It collected 300 cases of patients" healthy ear shape measurement data. The three variables are Ear length (EC), Ear breath (EK) and Ear outreach distance (EZ). This experiment is to find four types of ear by different variables using K-Means clustering algorithm.

## Section Three: The Procedures of Experiment

(1) Create a new database named "Kmeans_sqlserver", and import the data by "data17-2".

Attention: After you import the "data17-2", you should add a row named "id". The new row will be the primary key and be used in the following steps.

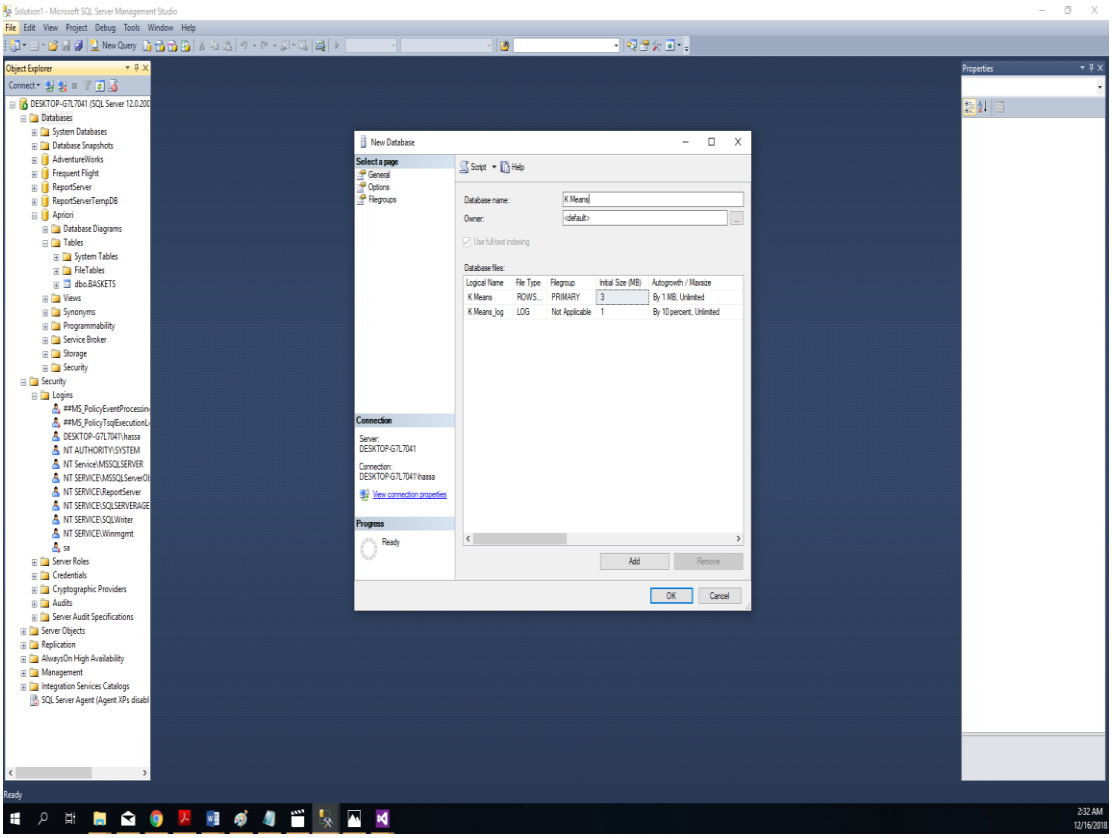



Figure 2.2.1 Add a new row as the Key

(2) Create an “Analysis Services” project and name it as “K_Means”.

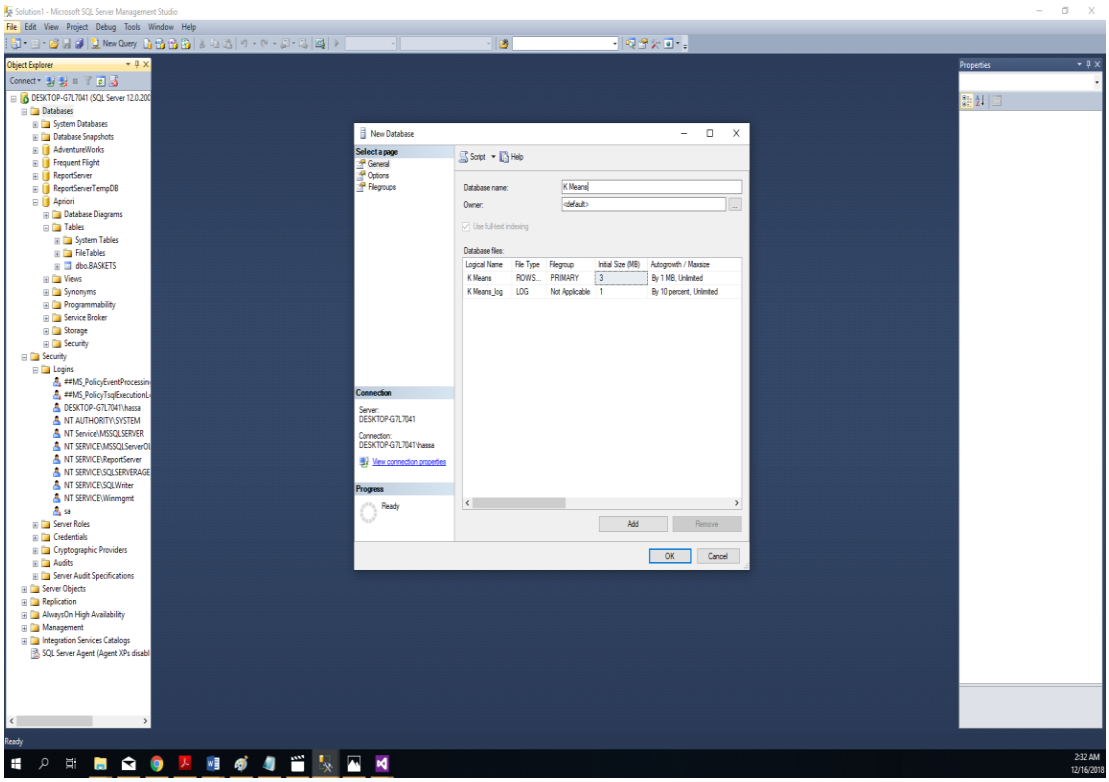



Figure 2.2.2 Create “K_Means” Analysis Services Project

(3) Create a data source and name it as “Kmeans Sqlserver”.

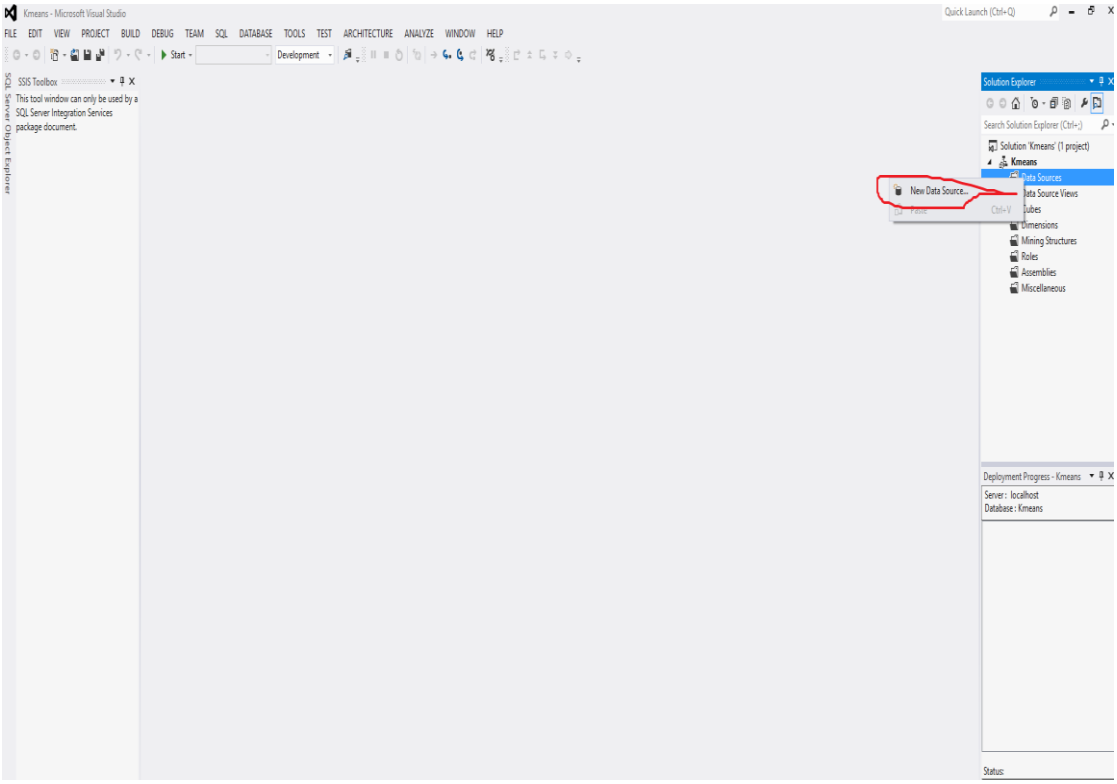



Figure 2.2.3 Create a New Data Source

In the wizard, create a data source based on “Kmeans_sqlserver” database you just created.

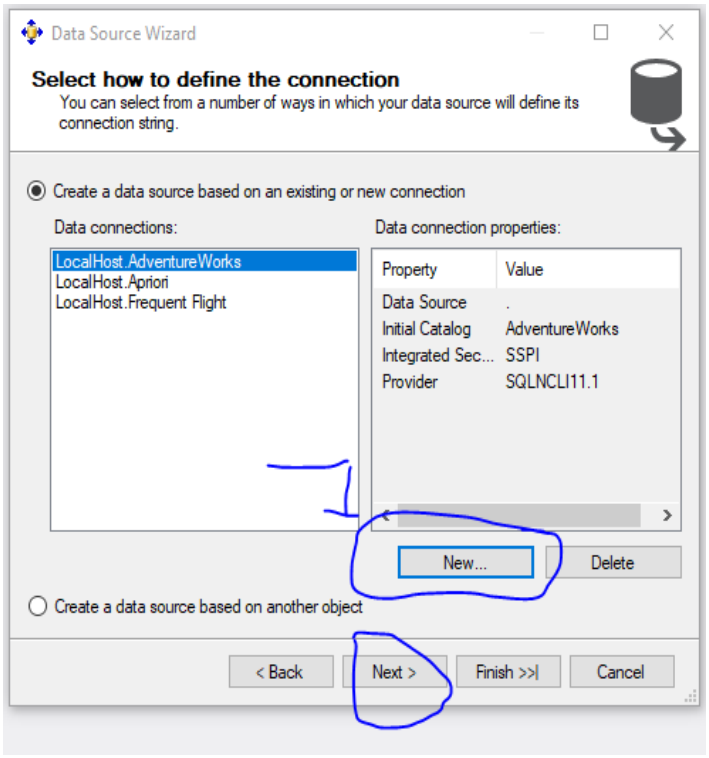



Figure 2.2.4 Create a New Database Connection

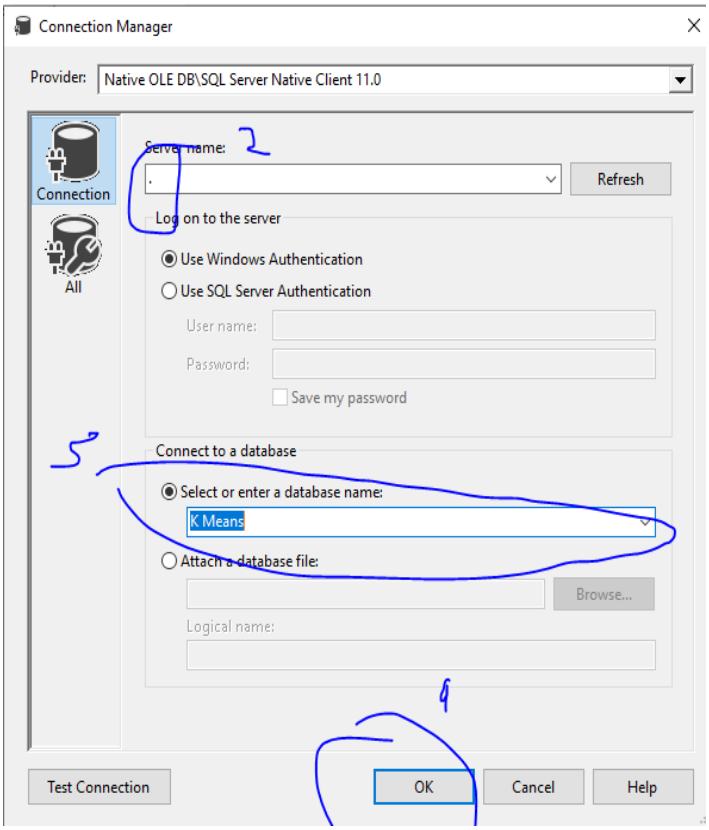



Figure 2.2.5 Connect to “Kmeans_sqlserver” Server

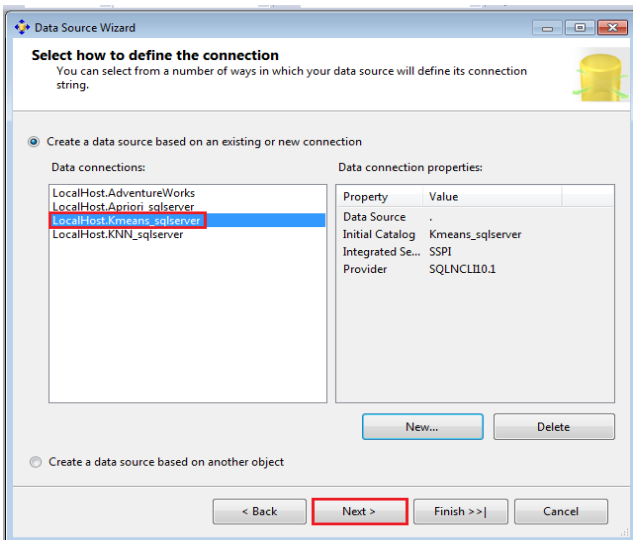



Figure 2.2.6 Select the New Database Connection

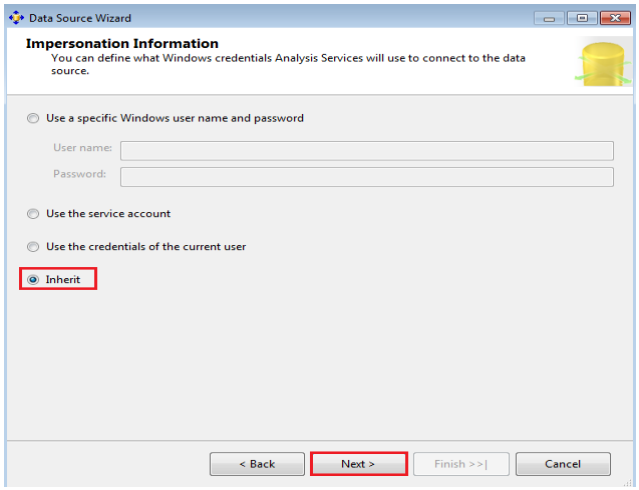



Figure 2.2.7 Select the Inherit Method

(4) Create a data source view same as experiment 2.1

In the wizard, right-click to build a new file and select “Microsoft Clustering Analysis” technique.

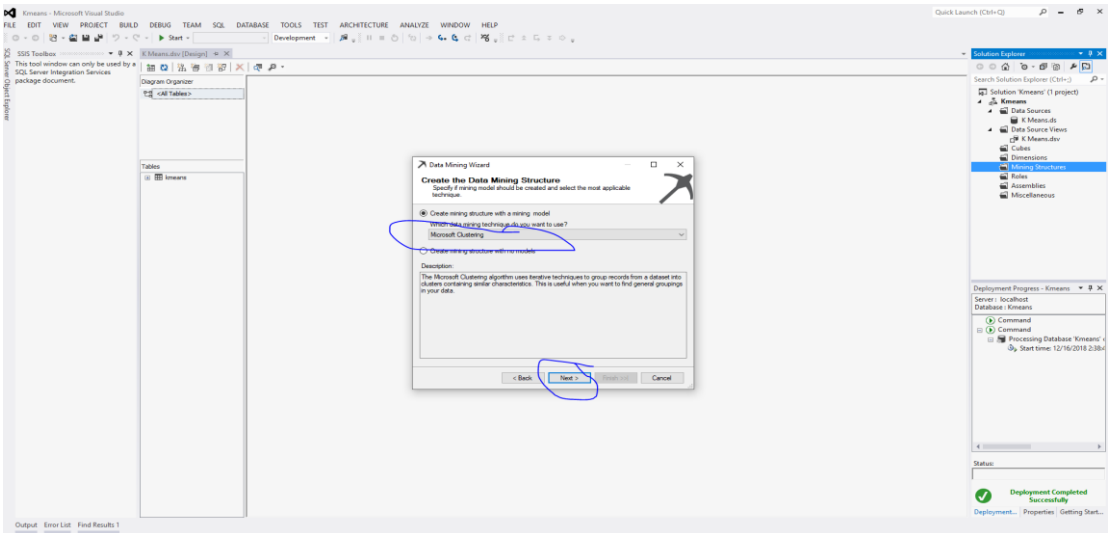



Figure 2.2.9 Select Microsoft Clustering Analysis Technique

Set “id” as the key, and set “EC”, “EK” and “EZ” as inputs and predictable parameters.

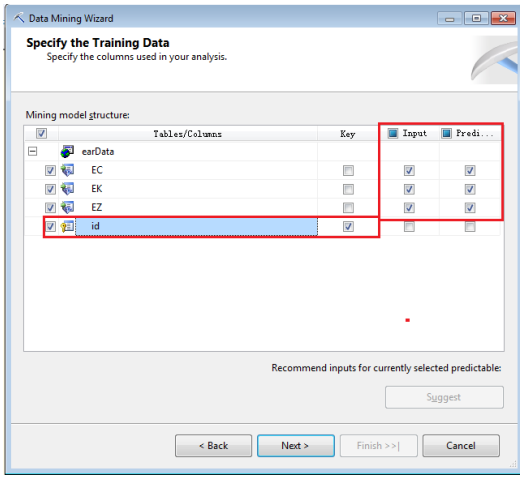



Figure 2.2.10 Specify the Training Data

(5) Deploy the project.

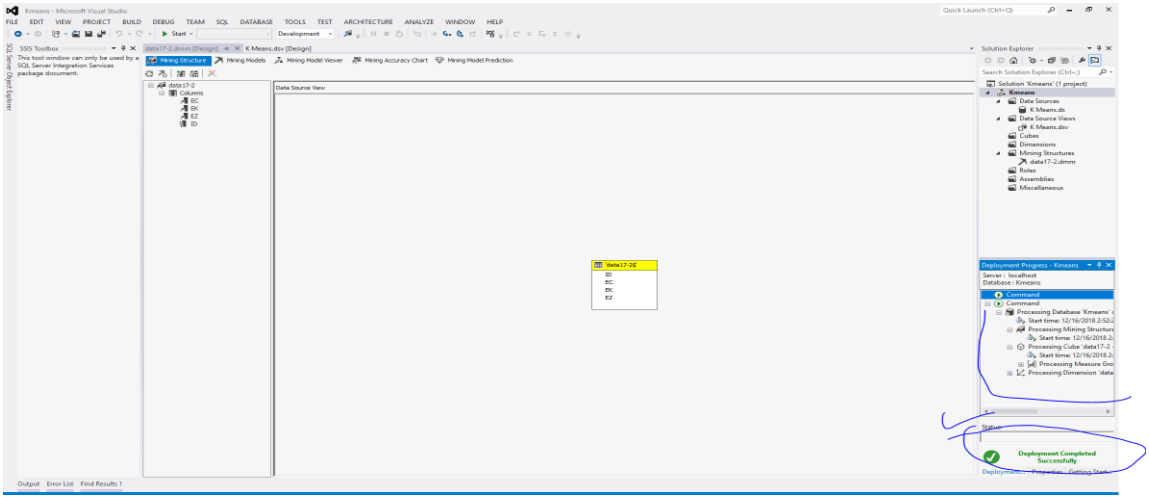


Figure 2.2.11 Deploy the Project

(6) Analyze the result.

The following figure shows the ear data was classified into ten clusters.

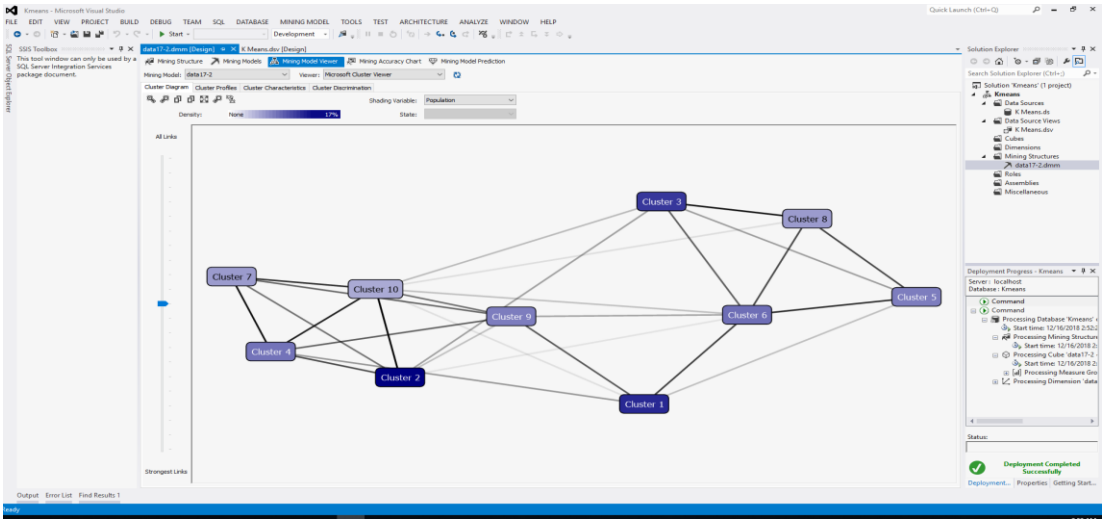



Figure 2.2.12 Cluster Diagram

The following figure shows ear variable in each cluster. For example, the size of clusters including “EC” variable is 210 and the size of “EC” variable is 33 in cluster 1, 32 in cluster 2.

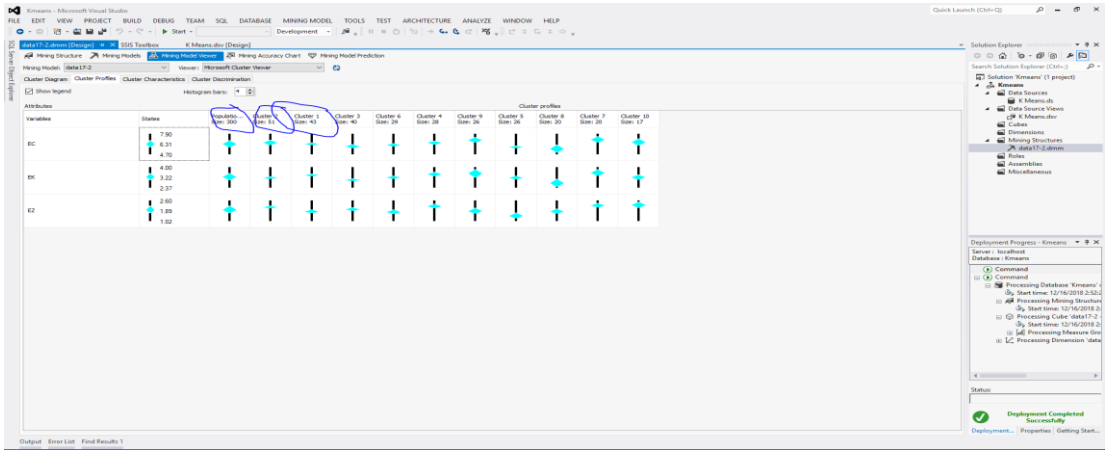


Figure 2.2.13 The Result of Cluster Profiles

The following figure shows all ear variables in a certain cluster. When you choose cluster 4, you will see the possibilities of each ear variable in different variable bounds and you will find the minimum and maximum easily.

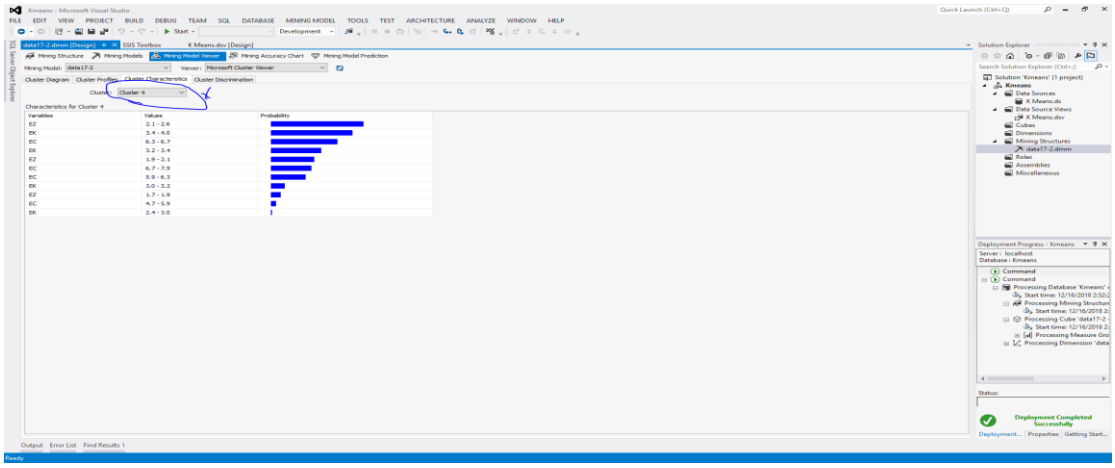


Figure 2.2.14 The Result of Cluster Characteristics

The following figure shows the comparison of cluster 2 and cluster 4. From the graph, we can know that “EZ” variable is likely to be a member of cluster 2 in the 1.0-2.0 range. While in 2.0-2.6 range, it tends to be in cluster 4.

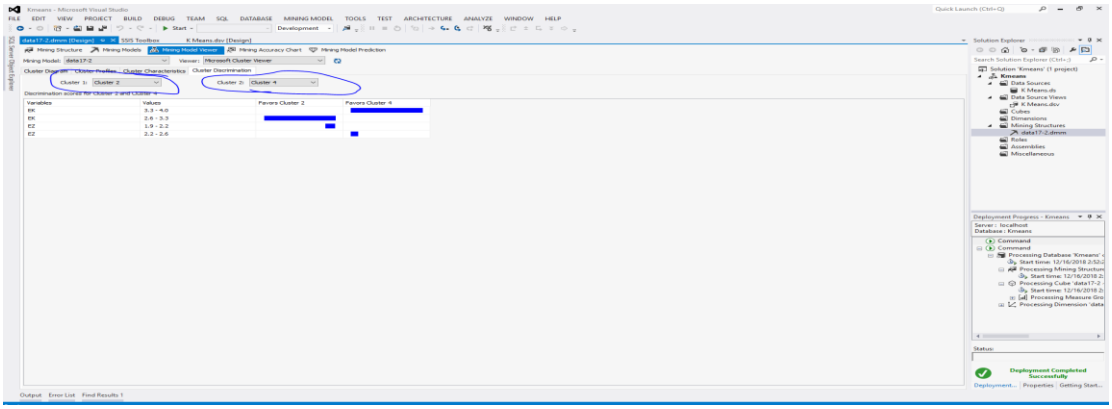


Figure 2.2.15 The Result of Cluster Discrimination

# Experiment 2.3 Classification Algorithm

## The Goal of Experiment

(1) Learn to use classification algorithm (Decision tree or Naïve Bayes for mining the information.

(2) You should learn SQL server to do this experiment.

# Requirements of Your Experiment Report

You need to organize the experiment data and analyze the results after you have finished the experiment, finally you should summarize the conclusions and write an experiment report.

1. The outline of your experiment report:

(1) The title and the purposes of the experiment;

(2) The content of the experiment;

(3)The principle of the experiment algorithm or the introduction of the experiment model that you designed;

(4) The analysis of the experiment results (graph, table, screenshots, MDX sentences, etc.);

(5) After data warehousing，OLAP， and data mining 3 parts, you should write **what you have learned from the experiments of this part and what main problems you've encountered in the experiment and how you solved them**. (there are 3 discussions based on 3 parts)

2. The writing requirements of your experiment report:

(1) The first page (cover page) of the report should include: the title of course, student's ID, student's name, teacher's name, date;

(2) The content of the experiment (you need refer to the outline above);

(3) You should **use MS words, Times New Roma, 11 font size**. **Figures should have a caption below them,** and **tables should have a title above them** (please refer to this guide book).