## Experiment No. : 2

**Title:** Demonstration of Data Profiling using SQL Server Integration services

**Objectives**: 1. To perform various data profiling operations using SQL Data Profiler

**Key concepts:** Data Profiling

**Theory:**

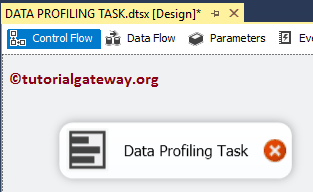
Data Profiling Task in SSIS

The Data Profiling Task in SSIS used to computes various profiles that help us to become familiar with the data source and to identify the problems in the data (if any) that have to fix.

The Data Profiling Task in SSIS will work only with the data present in SQL Server. The SSIS Data Profiling Task doesn’t support the data present in the file system, or the third-party data.

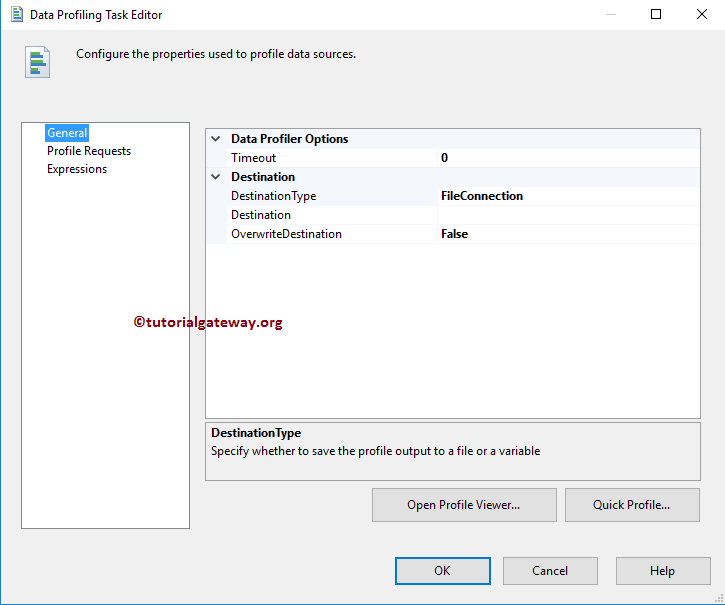
## Data Profiling Task in SSIS Example

Drag and drop the SSIS Data Profiling Task into the Control Flow region as we showed below

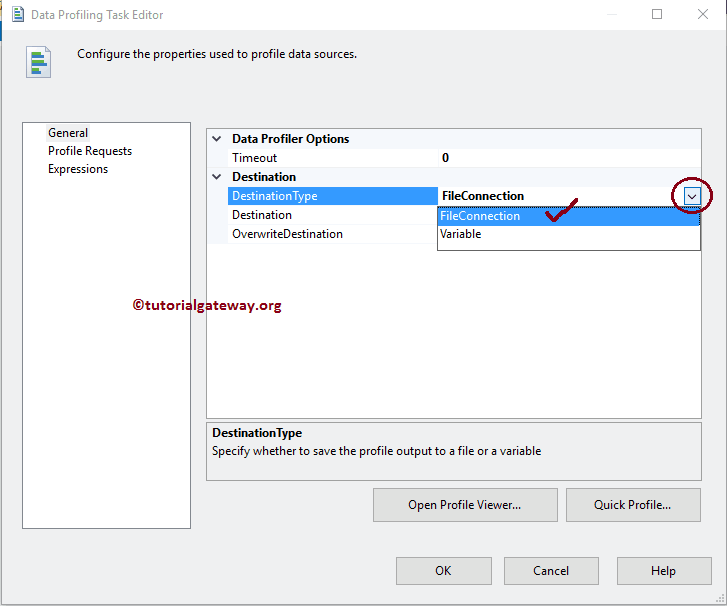


Double click on it will open the SSIS Data Profiling Task Editor to configure it.

* **Time-out (in seconds):** Please specify the connection time out in seconds. If the connection takes more than this time, the connection will fail.
* **OverwriteDetination:**This SSIS Data Profiling Task property has two options: True and False. If we set this property to true, the File System Task will overwrite the existing files in the Destination path.
* **Open Profile Viewer:** This button shows the profiling data after you run the integration service package.

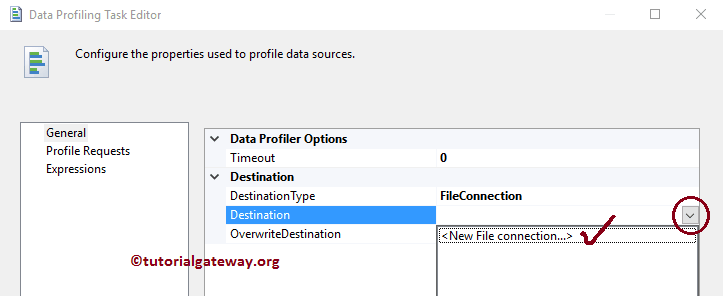


**Destination Type:**This SSIS Data Profiling Task property has two options: File Connection and Variable. If we set this variable to true, Destination data stored in a variable if we set to File Connection, select the Destination file manually using File Connection Manager.



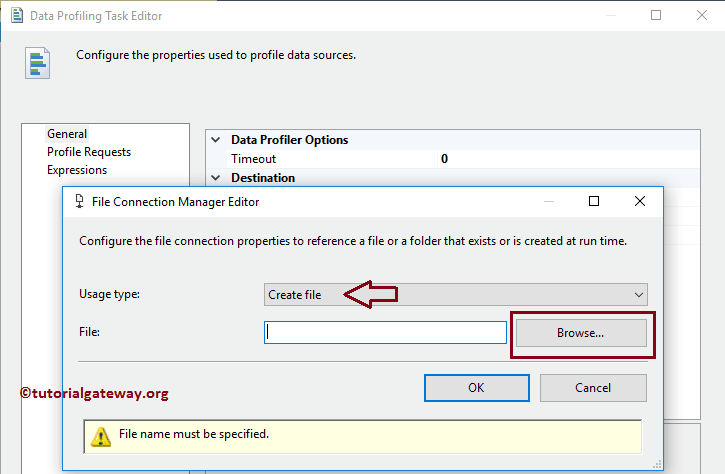
When you set the **DestinationType** to File Connection, we have to configure the Destination Connection using **Destination**Property. If you already created the [File Connection Manager](https://www.tutorialgateway.org/file-connection-manager-in-ssis/), you can select it from the drop-down list.

If you haven’t created any connection Manager before, You have to create by selecting **<New Connection..>**.

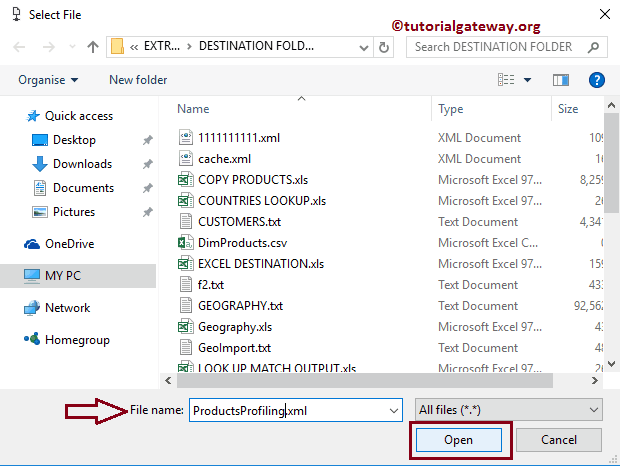


Once you click on the **<New Connection..>** option, the File Connection Manager Editor will open to configure the Destination Connection.

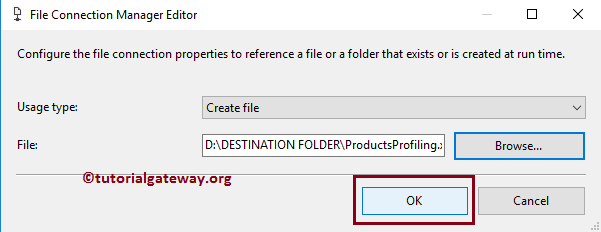
If you have any existing file, select the Existing File option from the**Usage Type.**Otherwise, select the Create File option and Click on the Browse button to select the Existing File from the file system or create a new file.



From the above screenshot, you can observe that we created the ProductsProfiling.xml file inside the Destination Folder

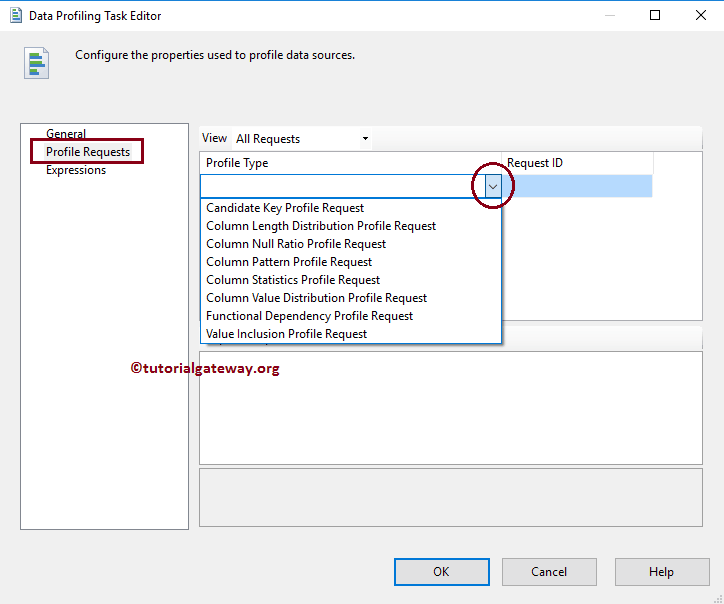


Click OK to finish configuring the Source connection. If you find any difficulty understanding, please refer [File Connection Manager](https://www.tutorialgateway.org/file-connection-manager-in-ssis/) article.

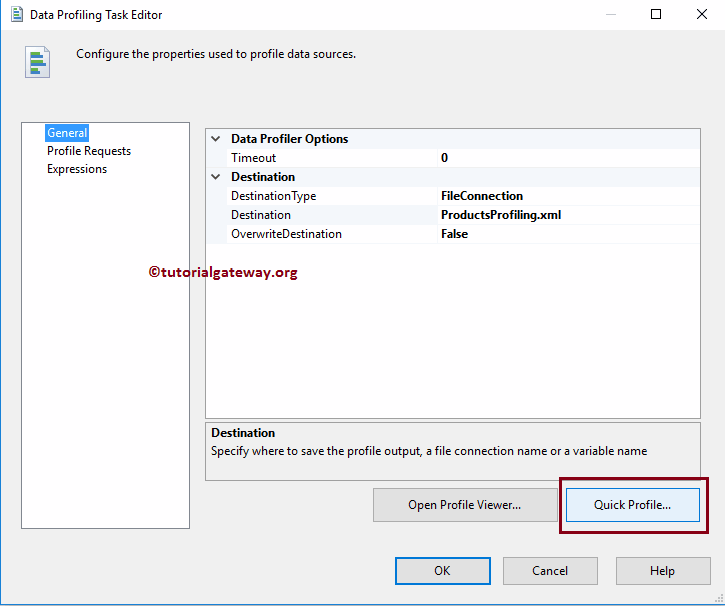


The Data Profiling Task in SSIS computes eight different data profiles. The following table will show you the list of available profiles in the SSIS Data Profiling Task.

|  |  |
| --- | --- |
| **PROFILES IN A DATA PROFILING TASK IN SSIS** | **DESCRIPTION** |
| Candidate Key Profile | This SSIS Data Profiling Task will report whether a column or set of columns is an approximately Key, or a key for the selected data. It is helpful to identify the problems (such as duplicate data in key columns) in your data. |
| Column Length Distribution Profile | It reports all the distinct string lengths available in our selected columns, and the percentage of rows that represent the same length. It is beneficial to identify whether the column data is valid or not. For example, if we select the Zip Code of the UK, then it should be a combination of letters and numbers of length 6 and discover the values longer than 2 |
| Column Null Ratio Profile | This SSIS Data Profiling Task profile will report the percentage of Null Values in a Column. Helpful to check which column is holding highest Nulls (analyze the data) |
| Column Pattern Profile | This will report the set of RegExp (regular expressions) that cover the specified percentage of values in a string column. |
| Column Statistics Profile | This SSIS Data Profiling Task profile reports the statistics, such as Minimum value, Maximum Value, Mean and Standard Deviation of every Numeric Column, and Minimum value and Maximum Value for the Datetime columns. Useful to check whether the Date column is holding correct data or not. |
| Column Value Distribution Profile | Reports all the distinct values available in our selected column and the percentage of rows that each value represents. Crucial to identify whether the column data is valid or not. For example, if your column is supposed to store states in the United States, and if you discover more than 50, your data is incorrect. |
| Functional Dependency Profile | Report the extent to which the values in the dependent column depends upon the values in the determinant column (it may be one or set of columns). Handy to identify whether the column data is valid or not. For example, if you profile the dependency between a column that contains India Zip Codes and columns that contain states in India. If your dependence finds multiple states for the same zip code, then your data is not valid. |
| Value Inclusion Profile | This SSIS Data Profiling Task profile will compute the values overlap between two columns or two sets of columns. Recognize whether the column is appropriate to serve as the foreign key between two columns or not. |



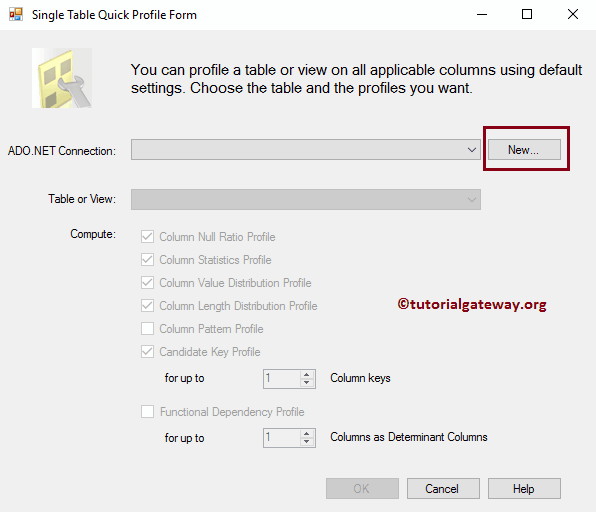
Please go to **General** Tab and click on the Quick Profile button to create a new profile.



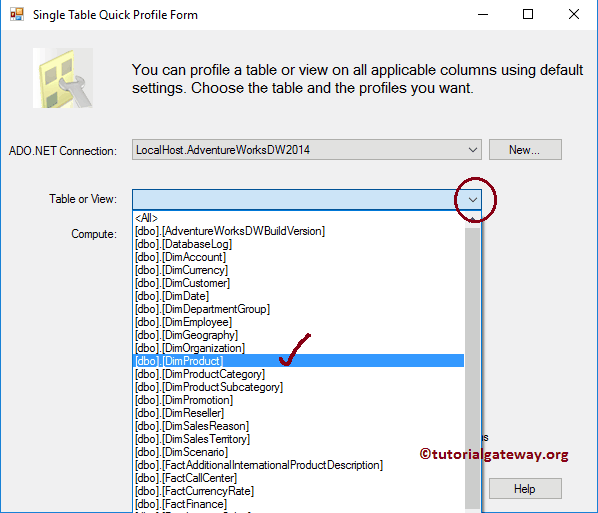
Once you click on the **Quick Profile** button, a new window called **Single Table Quick Profile Form** opened.

Click on the new button opens another window called **Connection Manager** to select the Provider, Server Name, and Database Name. If we created any connection managers before then, select it from the drop-down list.

Here, we are selecting the already created ADO.NET connection. If you find any difficulty in understanding the steps, please refer [ADO.NET connection manager in the SSIS](https://www.tutorialgateway.org/ado-net-connection-manager-in-ssis/) tutorial.



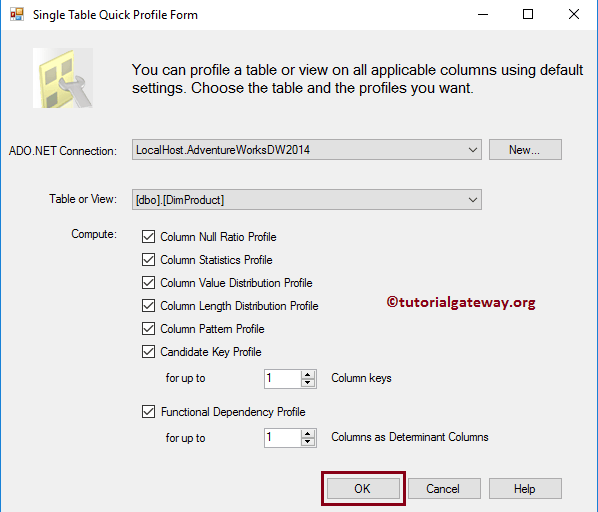
Here, we are selecting the DimProduct table.



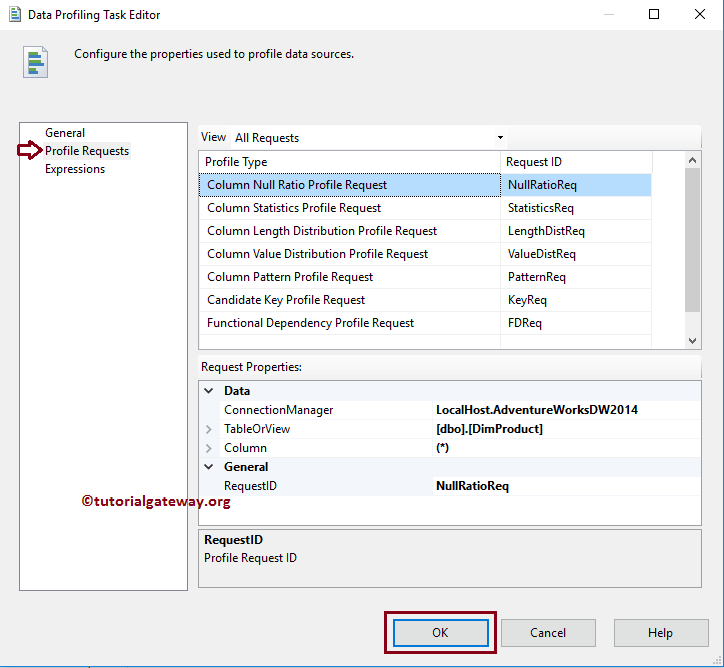
From the above screenshot, you can observe that we are using our local host windows account as server name and [AdventureWorksDW2014] as the database name.

**NOTE:** In real-time, you have to select the **Use SQL Server Authentication** option and provide the valid credentials given by your Admin person.

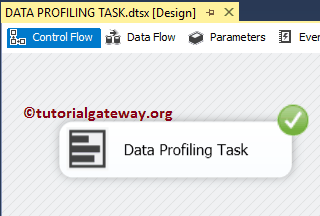
Here, we are selecting all the available options



Once you click on OK button, SSIS Data Profiling Task Editor will navigate to **Profile Requests** Tab

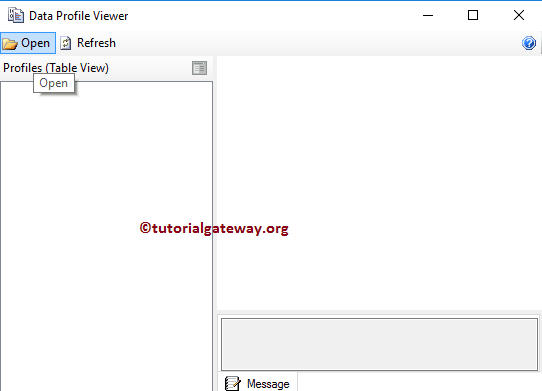


Click OK to finish configuring and closing Data Profiling Task in SSIS Editor. Let us run the package and see.

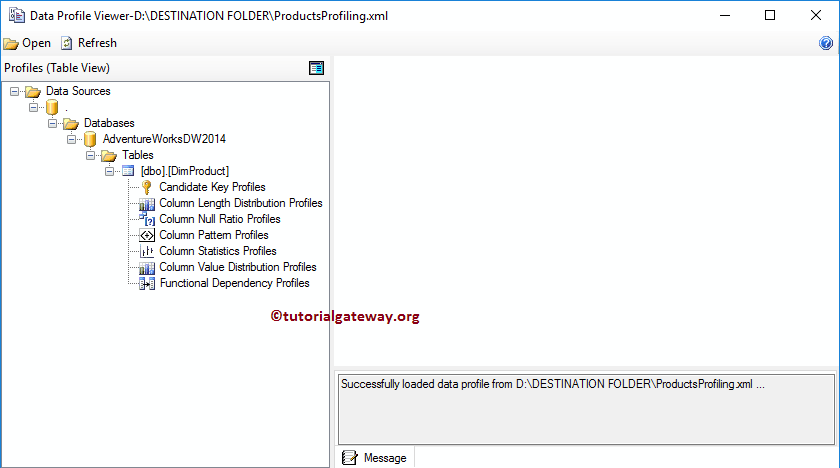


As you see from the above screenshot, our SSIS package executed successfully. To see the SSIS Data Profiling Task generated Data, double click and select the Data Profile Viewer or alternatively use the search bar and type Data Profile Viewer or navigate to C:\ProgramData\Microsoft\Windows\Start Menu\Programs\Microsoft SQL Server 2014\Integration Services and double-click on the Data Profile Viewer will open the following window.

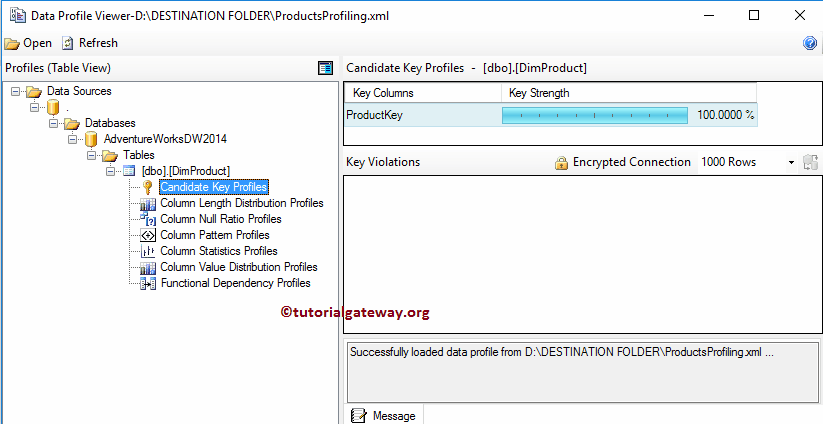
To view the data, please click on the Open folder and select the XML file that was generated by the Data Profiling Task in SSIS.



From the below screenshot, see the list of profiles that we selected while configuring the SSIS Data Profiling Task package.



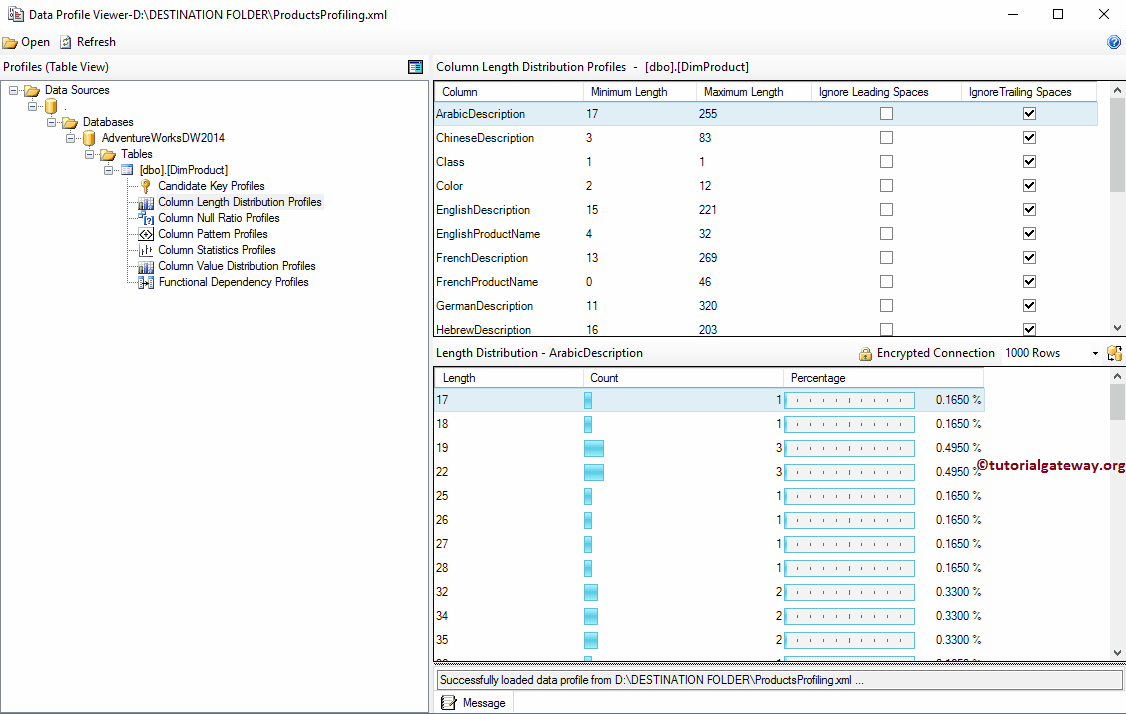
**Candidate Key Profiles:** This will report whether a column or set of columns is an approximately Key, or a key for the selected data. In our table chosen (DimProducts) Product key is the key column, and its key strength is 100%, which means data is valid.



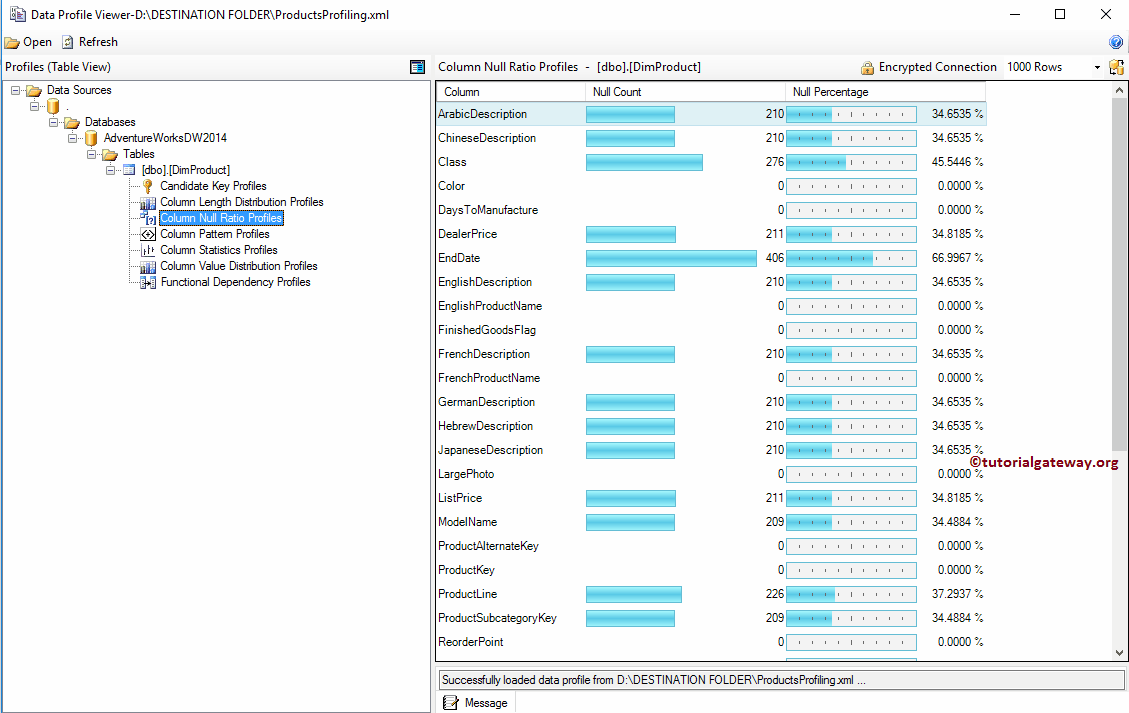
**Column Length Distribution Profiles:** This SSIS Data Profiling Task profile report has two sections:

* **Column Length Distribution Profiles:**In this section, the report will display the minimum and maximum length of every column present in our selected table.
* **Length Distribution:** This will report all the distinct string lengths available in our selected columns, and the percentage of rows that represent the same length.

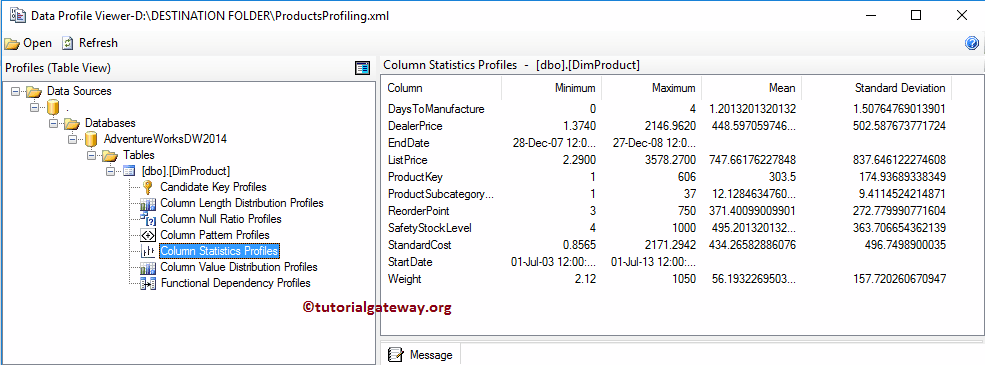
Here, we selected the Arabic Description Column, and you can see it has distinct lengths of 17, 18, 19 …., and 19, 22 length holds the highest percentage. It means that while transferring data, we can assign the destination string length as 25 (rather than giving 255).



**Column Null Ratio Profiles:**This will report the percentage of Null Values in a Column. From the below image, see that the End Date has the highest percentage of NULLS, and Color, Days to manufacture, English and French Product Name, Product Key, Product Alternate Key, Photo, and Finished Good Flag has no Nulls.



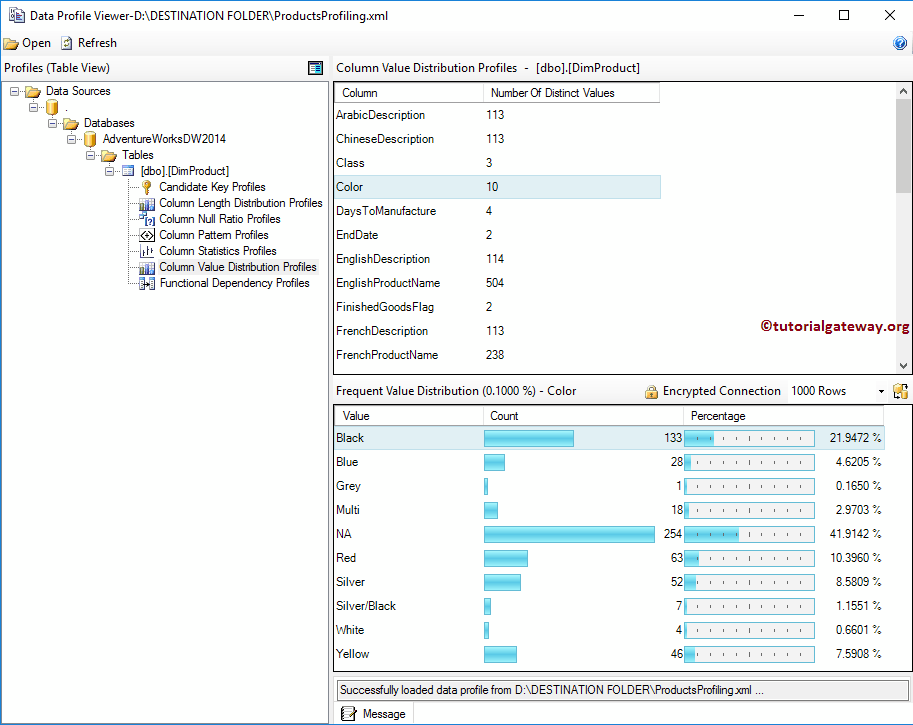
**Column Statistic Profiles:** This will report the statistics, such as Minimum value, Maximum Value, Mean and Standard Deviation of every Numeric Column, and Minimum value and Maximum Value for the Datetime columns



**Column Value Distribution Profiles:**This SSIS Data Profiling Task report has two sections:

* **Column Value Distribution Profiles:**In this section, the report will display the Number of Distinct Values available in each column present in our selected table.
* **Frequent Value Distribution:** This SSIS Data Profiling Task option will report all the distinct values available in our selected columns, and the percentage of rows that represent the same value.

From the below screenshot, you can see that we selected the Color Column, and it has distinct values of 10. Here, NA row holds the highest percentage, followed by Black, etc.



**Functional Dependency Profiles:** This will report the extent to which the values in the dependent column depends upon the values in the determinant column (it may be one or set of columns).

