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

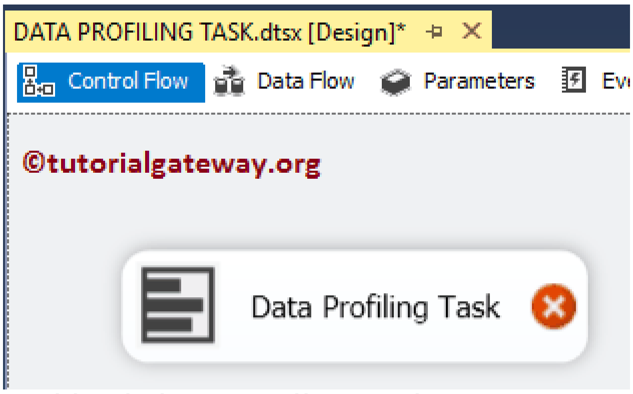
Theory:

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.

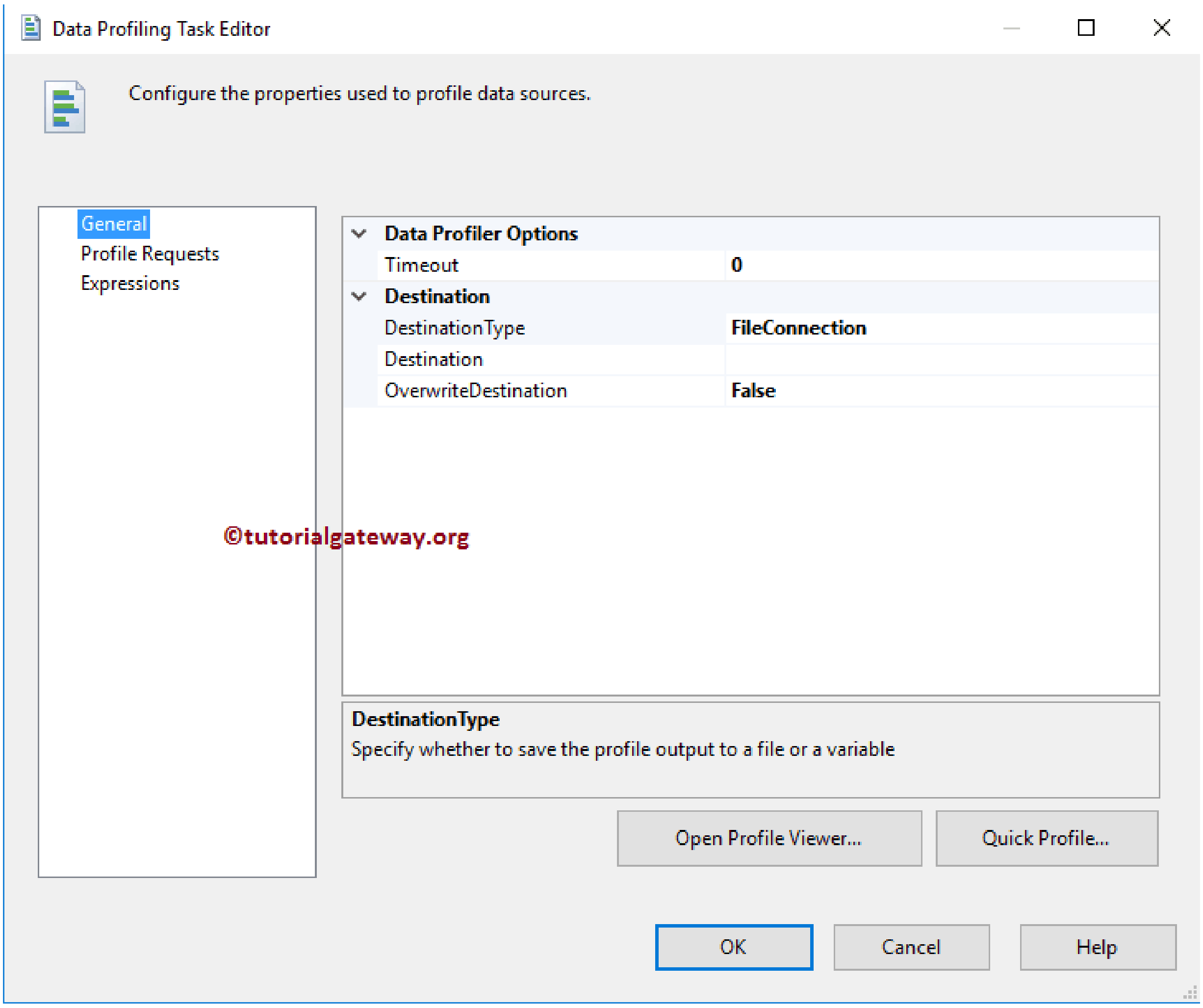
Data Profiling Task in SSIS Example

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.

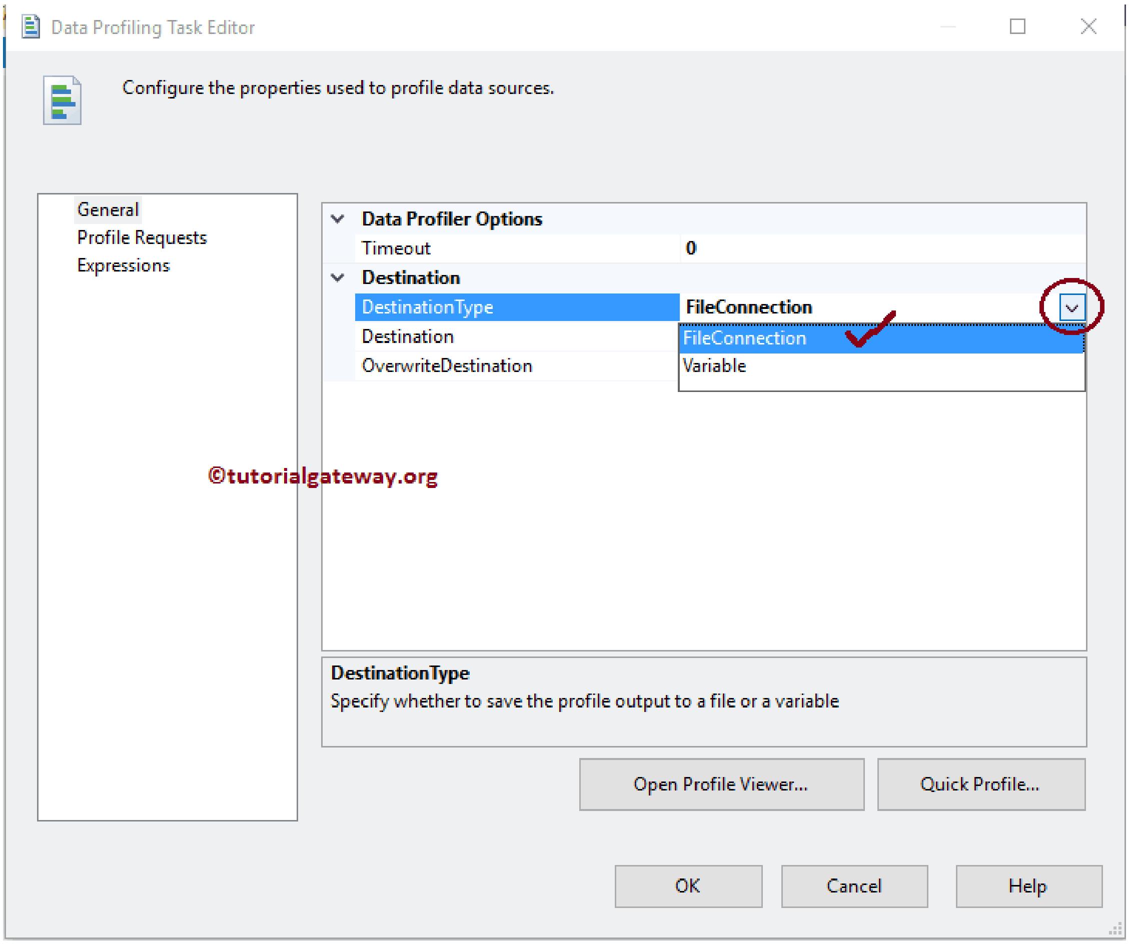


Configure the properties used to profile data sources.

Data Profiler Options	
Timeout	**0**
Destination	
DestinationType	**FileConnection**
Destination	
OverwriteDestination	**False**

DestinationType
Specify whether to save the profile output to a file or a variable

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.


Data Profiling Task Editor

Configure the properties used to profile data sources.

Data Profiler Options
Timeout 0
Destination
DestinationType FileConnection
Destination
OverwriteDestination

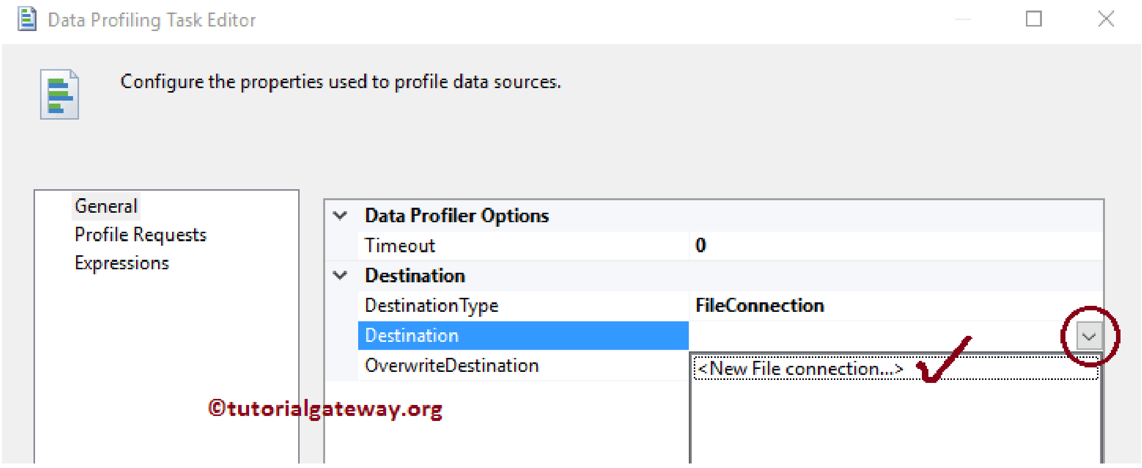
FileConnection
Variable

©tutorialgateway.org

DestinationType
Specify whether to save the profile output to a file or a variable

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, 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..>**.


Data Profiling Task Editor

Configure the properties used to profile data sources.

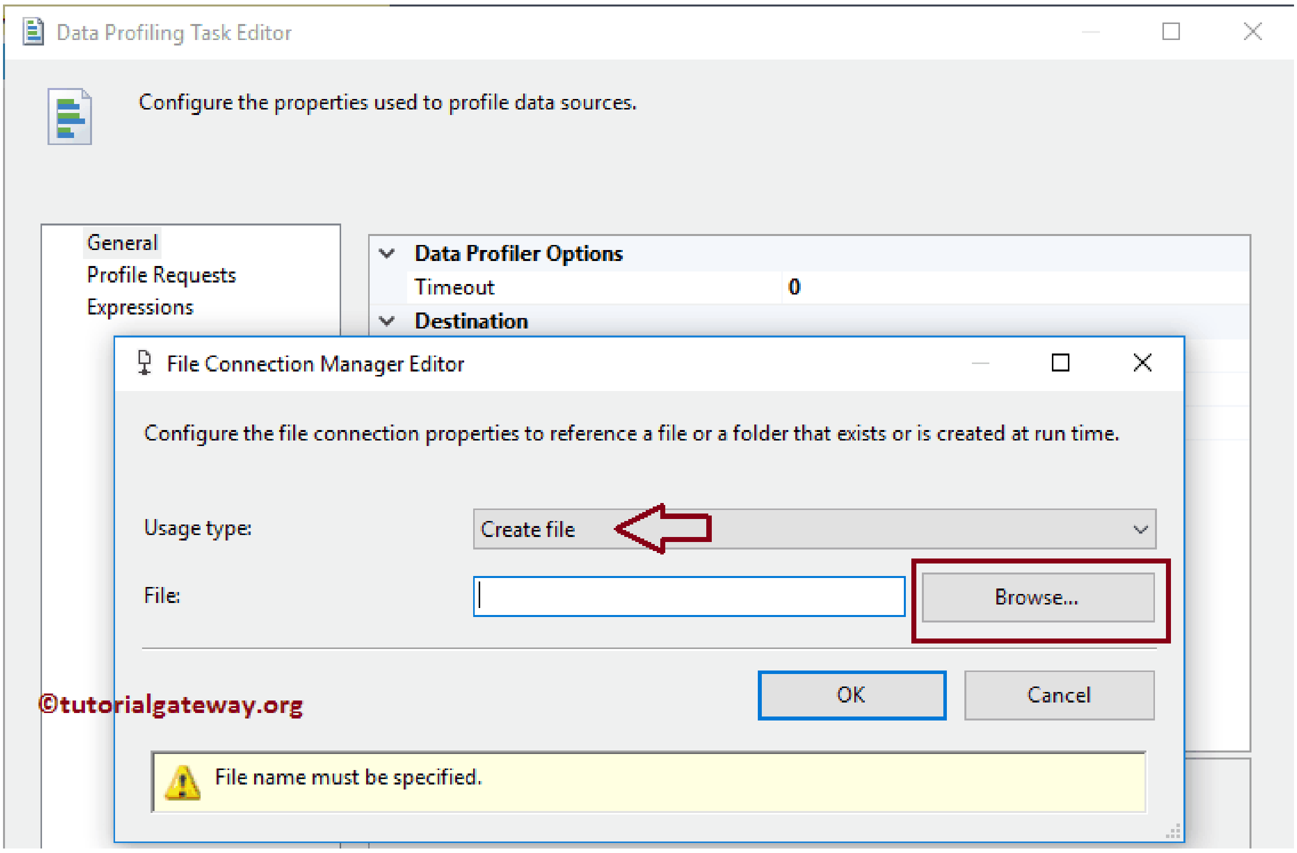
Data Profiler Options
Timeout 0
Destination
DestinationType FileConnection
Destination
OverwriteDestination

<New File connection...>

©tutorialgateway.org

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.



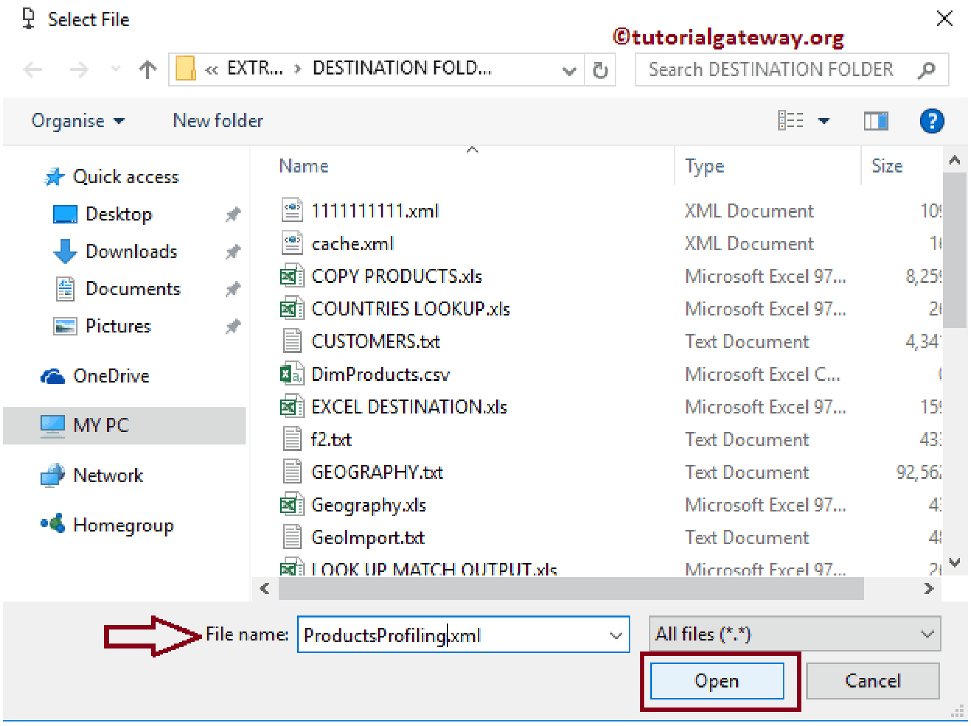
Configure the file connection properties to reference a file or a folder that exists or is created at run time.

Usage type: Create file

File:

File name must be specified.

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



Select File

©tutorialgateway.org

« EXTR... › DESTINATION FOLD...

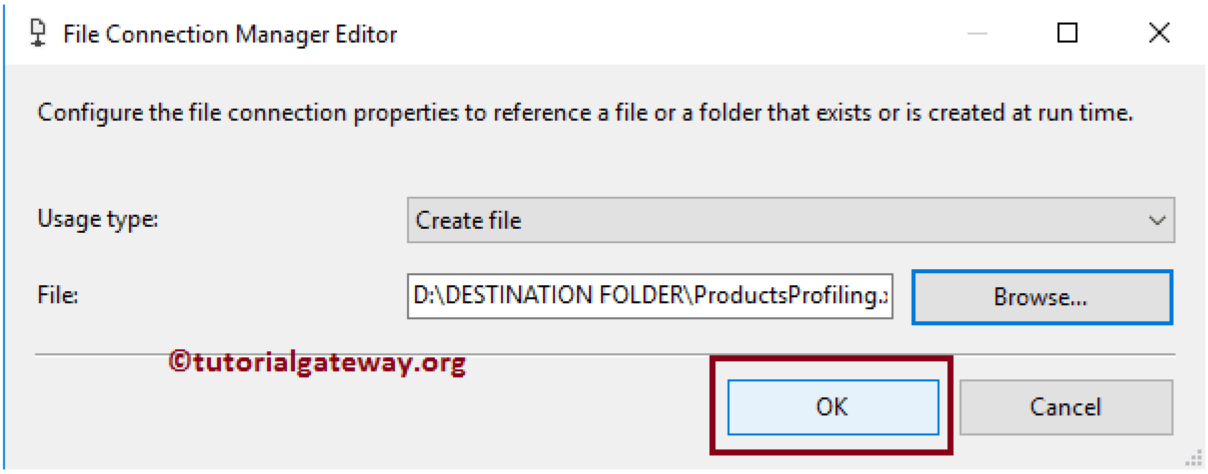
Search DESTINATION FOLDER

Name	Type	Size
1111111111.xml	XML Document	10
cache.xml	XML Document	1
COPY PRODUCTS.xls	Microsoft Excel 97...	8,25
COUNTRIES LOOKUP.xls	Microsoft Excel 97...	2
CUSTOMERS.txt	Text Document	4,34
DimProducts.csv	Microsoft Excel C...	
EXCEL DESTINATION.xls	Microsoft Excel 97...	15
f2.txt	Text Document	43
GEOGRAPHY.txt	Text Document	92,56
Geography.xls	Microsoft Excel 97...	4
GeoImport.txt	Text Document	4
LOOK UP MATCH OUTPUT.xls	Microsoft Excel 97...	2

File name: ProductsProfiling.xml

All files (*.*)

Click OK to finish configuring the Source connection. If you find any difficulty understanding, please refer File Connection Manager article.



File Connection Manager Editor

Configure the file connection properties to reference a file or a folder that exists or is created at run time.

Usage type: Create file

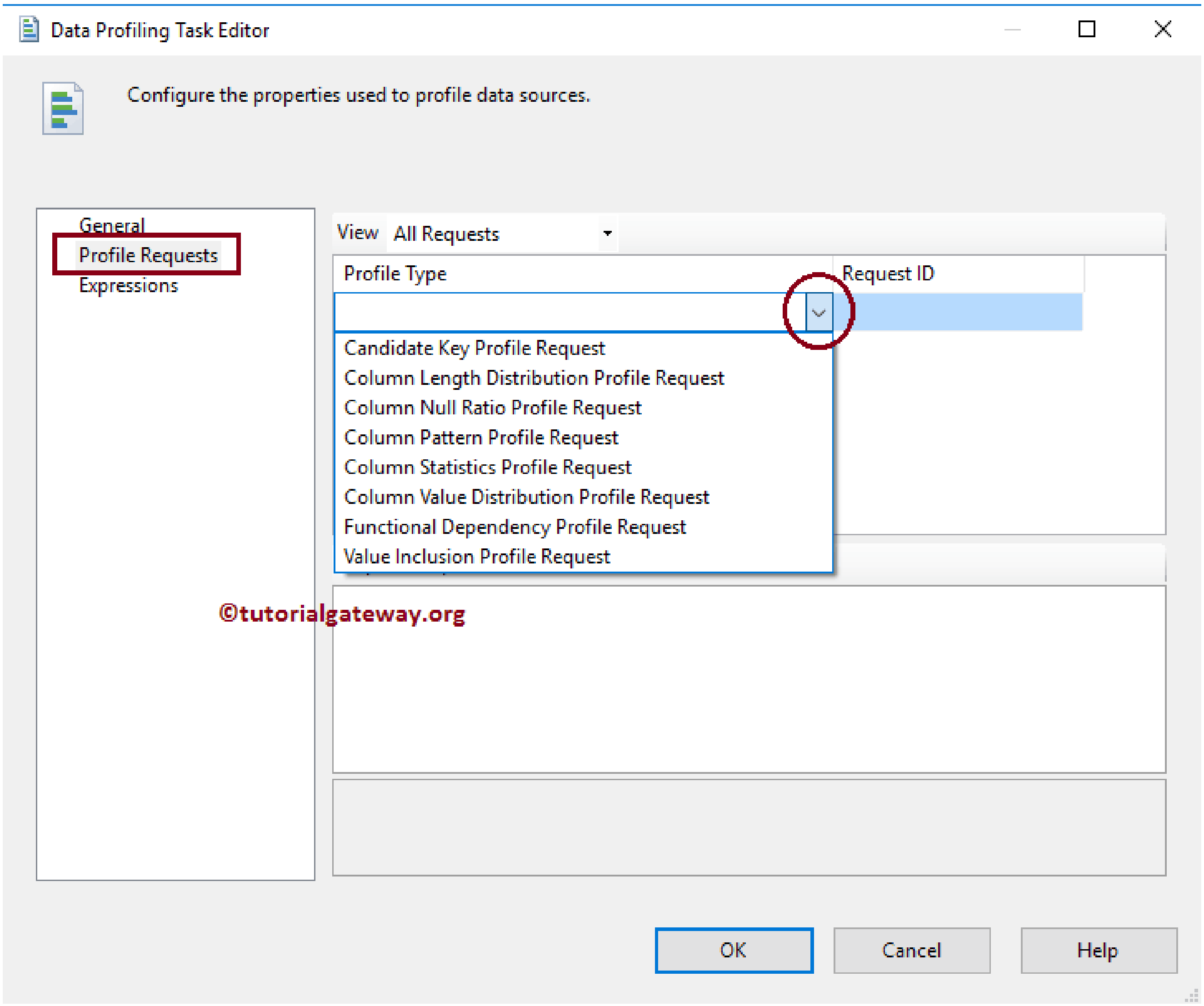
File: D:\DESTINATION FOLDER\ProductsProfiling.x

©tutorialgateway.org

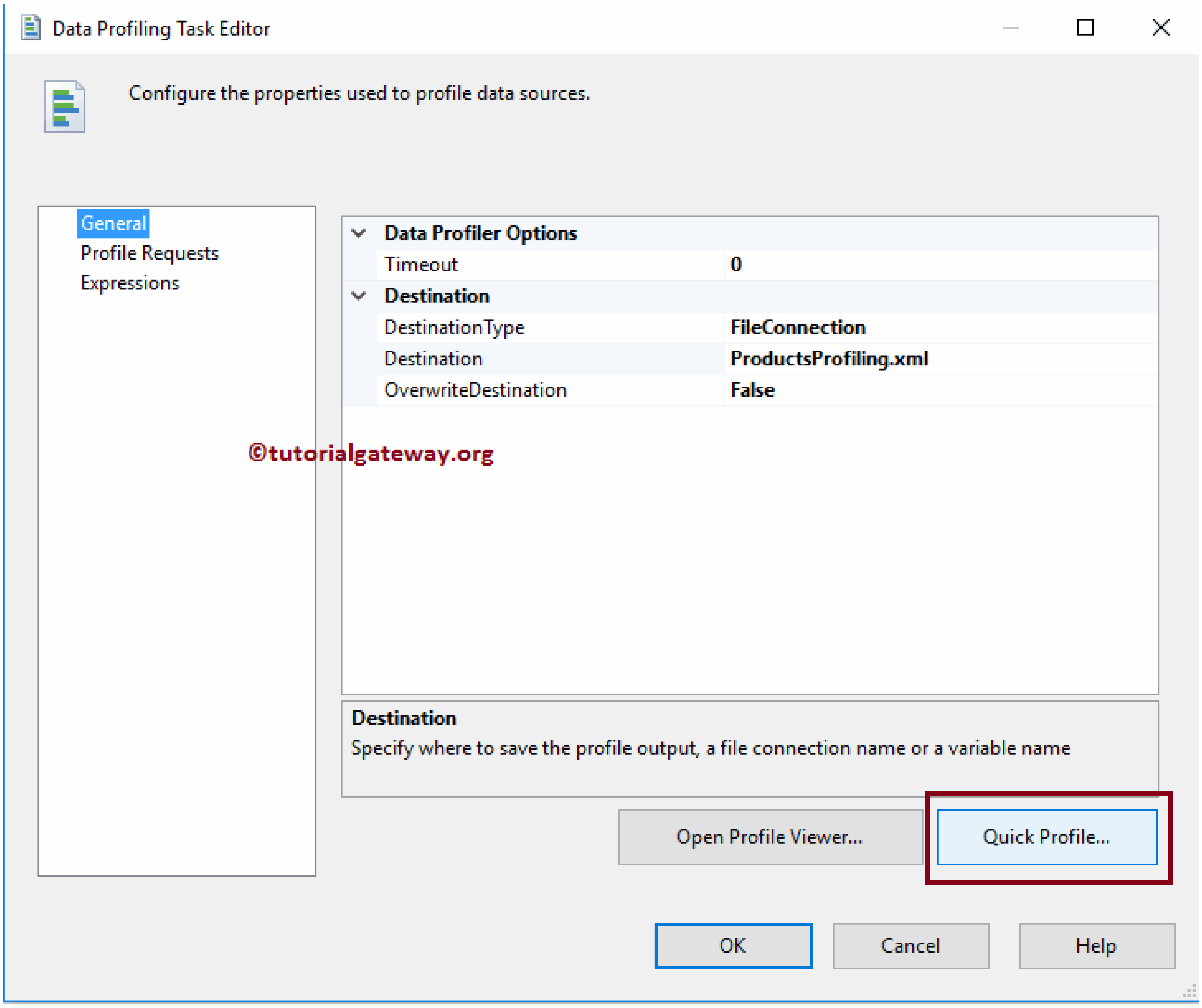
The Data Profiling Task in SSIS computes eight different data profiles. The following table will show you the list of available profiles in the SSIS Data Profiling Task.

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	This SSIS Data Profiling Task profile reports the statistics, such as

Profile	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.


Data Profiling Task Editor
Configure the properties used to profile data sources.
General
Profile Requests
Expressions
View All Requests
Profile Type
Request ID
Candidate Key Profile Request
Column Length Distribution Profile Request
Column Null Ratio Profile Request
Column Pattern Profile Request
Column Statistics Profile Request
Column Value Distribution Profile Request
Functional Dependency Profile Request
Value Inclusion Profile Request
©tutorialgateway.org
OK
Cancel
Help

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



Configure the properties used to profile data sources.

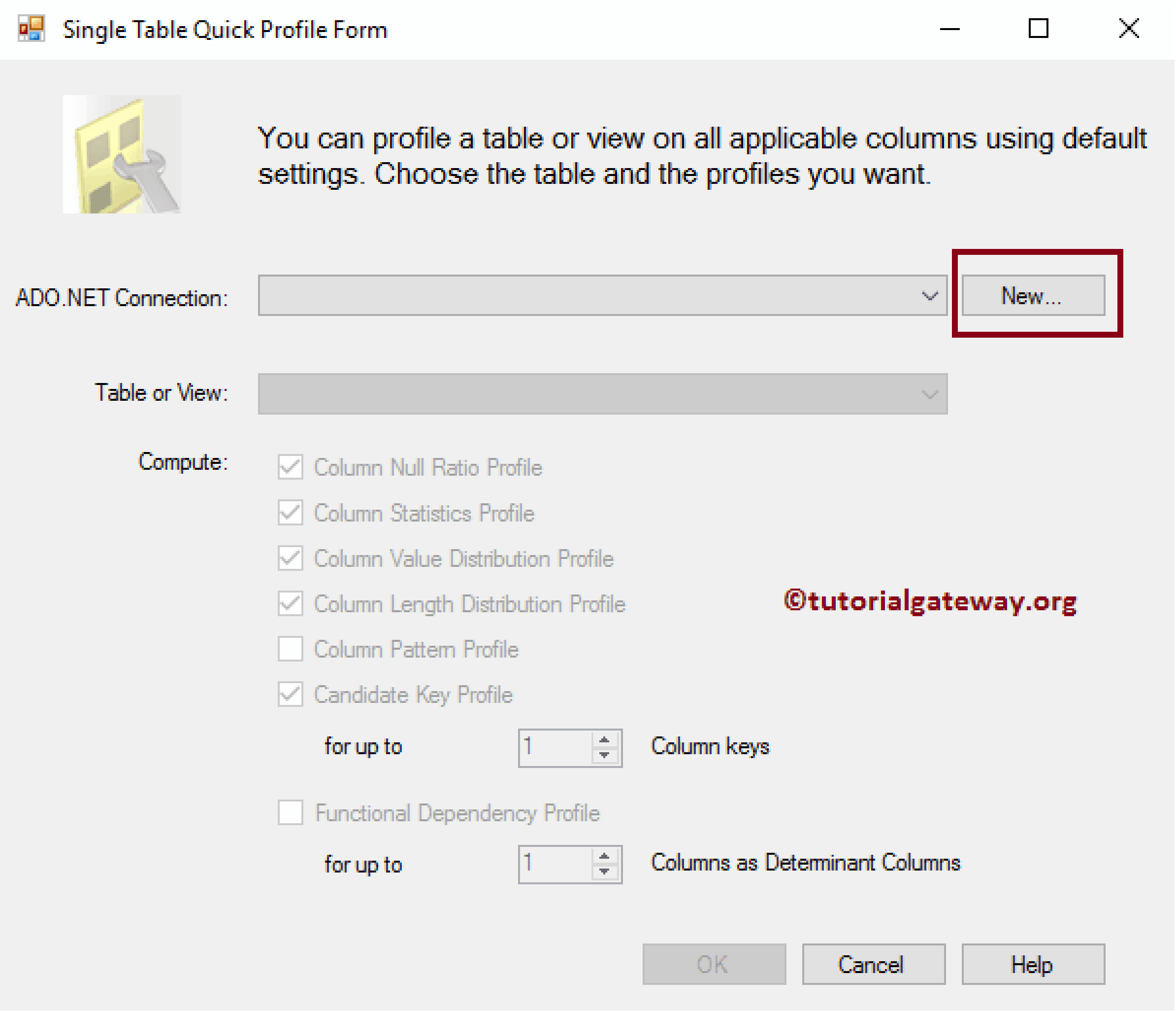
Data Profiler Options	
Timeout	**0**
Destination	
DestinationType	**FileConnection**
Destination	**ProductsProfiling.xml**
OverwriteDestination	**False**

Destination
Specify where to save the profile output, a file connection name or a variable name

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 tutorial.



Single Table Quick Profile Form

You can profile a table or view on all applicable columns using default settings. Choose the table and the profiles you want.

ADO.NET Connection:

Table or View:

Compute:

☑ Column Null Ratio Profile

☑ Column Statistics Profile

☑ Column Value Distribution Profile

☑ Column Length Distribution Profile

□ Column Pattern Profile

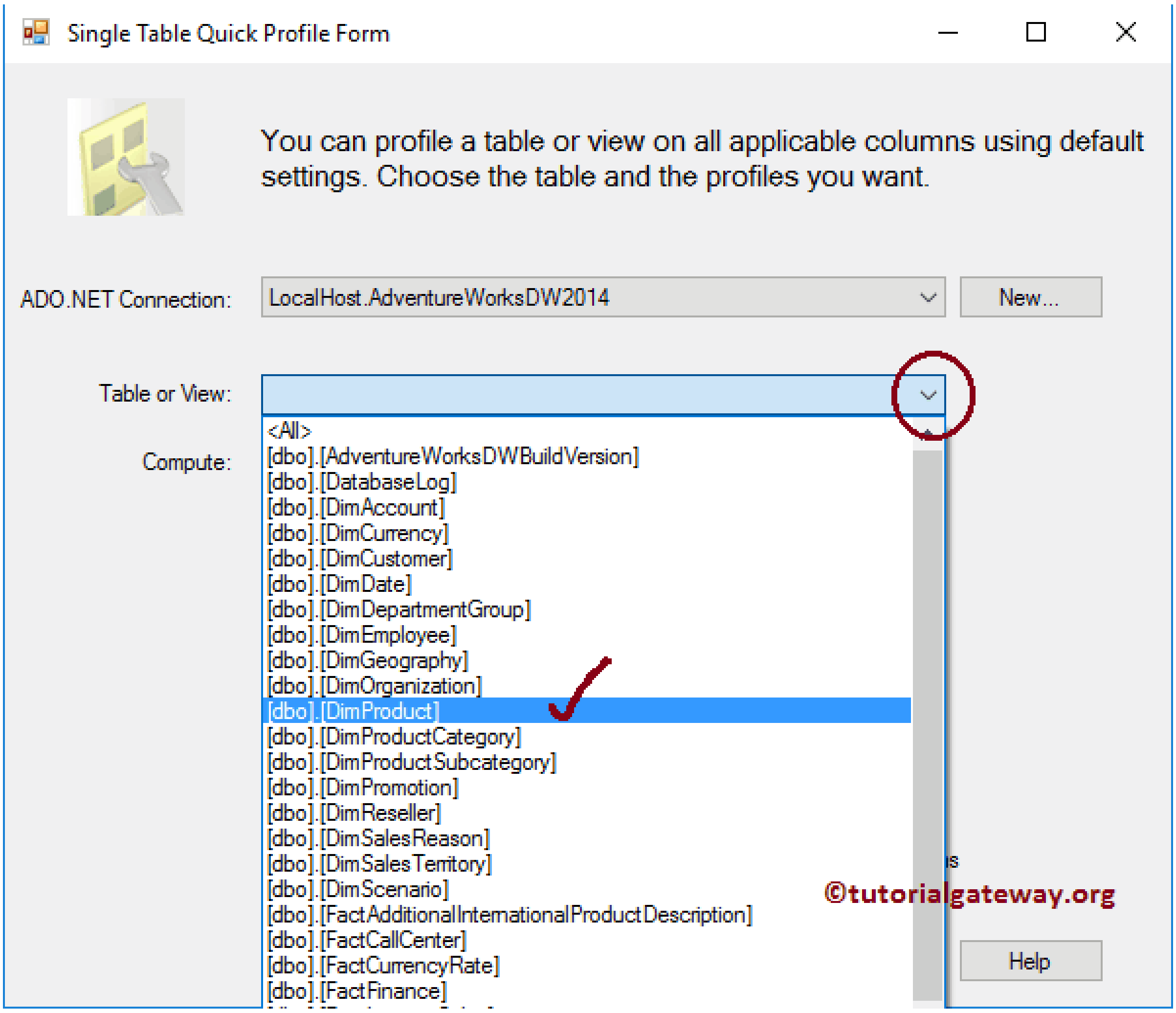
☑ Candidate Key Profile

for up to 1 Column keys

□ Functional Dependency Profile

for up to 1 Columns as Determinant Columns

Here, we are selecting the DimProduct table.



Single Table Quick Profile Form

You can profile a table or view on all applicable columns using default settings. Choose the table and the profiles you want.

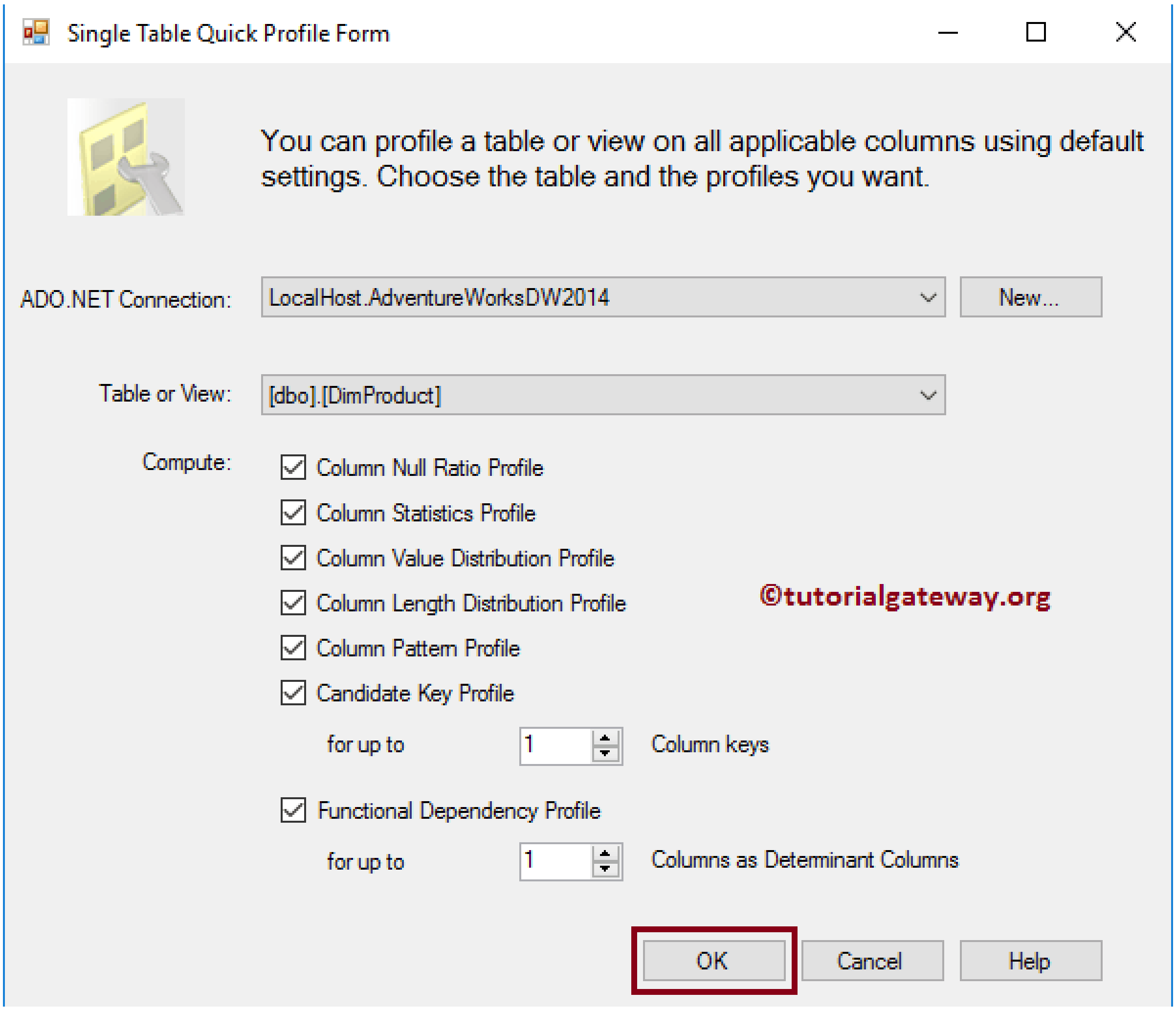
ADO.NET Connection: LocalHost.AdventureWorksDW2014

Table or View: [dbo].[DimProduct]

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



Single Table Quick Profile Form

You can profile a table or view on all applicable columns using default settings. Choose the table and the profiles you want.

ADO.NET Connection: LocalHost.AdventureWorksDW2014

Table or View: [dbo].[DimProduct]

Compute:

☑ Column Null Ratio Profile

☑ Column Statistics Profile

☑ Column Value Distribution Profile

☑ Column Length Distribution Profile

☑ Column Pattern Profile

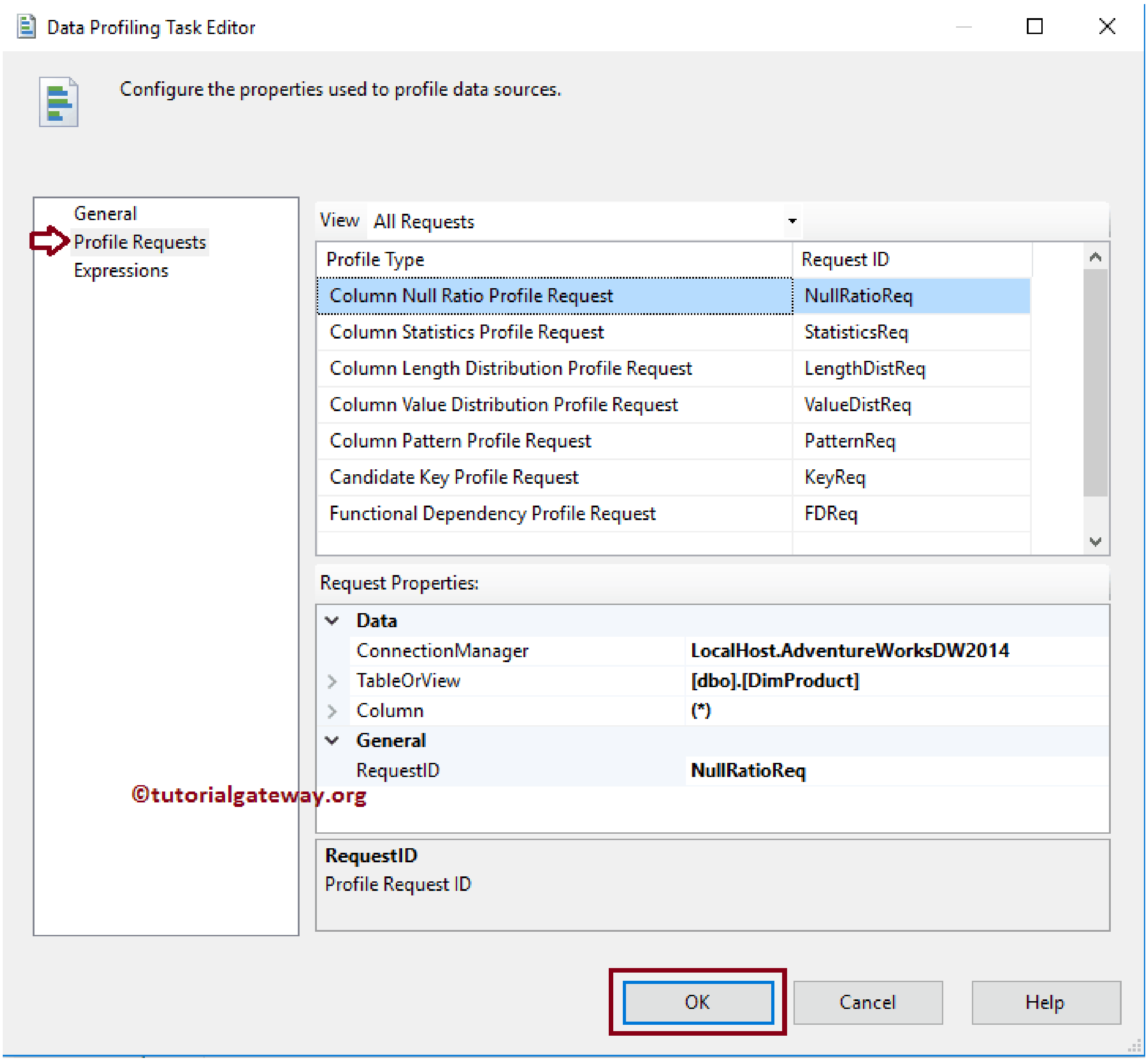
☑ Candidate Key Profile

for up to 1 Column keys

☑ Functional Dependency Profile

for up to 1 Columns as Determinant Columns

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



Data Profiling Task Editor

Configure the properties used to profile data sources.

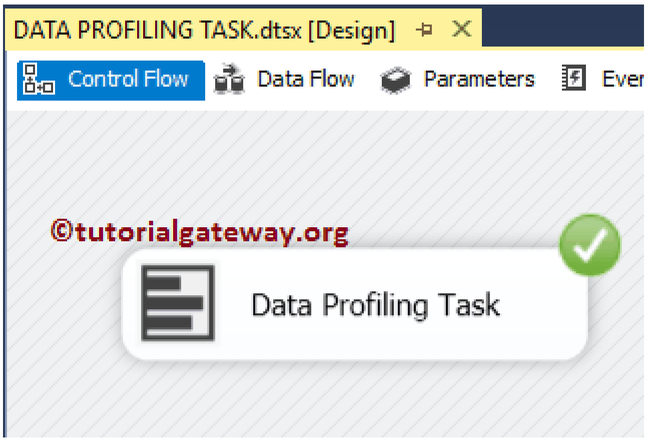
Profile Type	Request ID
Column Null Ratio Profile Request	NullRatioReq
Column Statistics Profile Request	StatisticsReq
Column Length Distribution Profile Request	LengthDistReq
Column Value Distribution Profile Request	ValueDistReq
Column Pattern Profile Request	PatternReq
Candidate Key Profile Request	KeyReq
Functional Dependency Profile Request	FDReq

Request Properties:

Data	
ConnectionManager	**LocalHost.AdventureWorksDW2014**
TableOrView	**[dbo].[DimProduct]**
Column	**(*)**
General	
RequestID	**NullRatioReq**

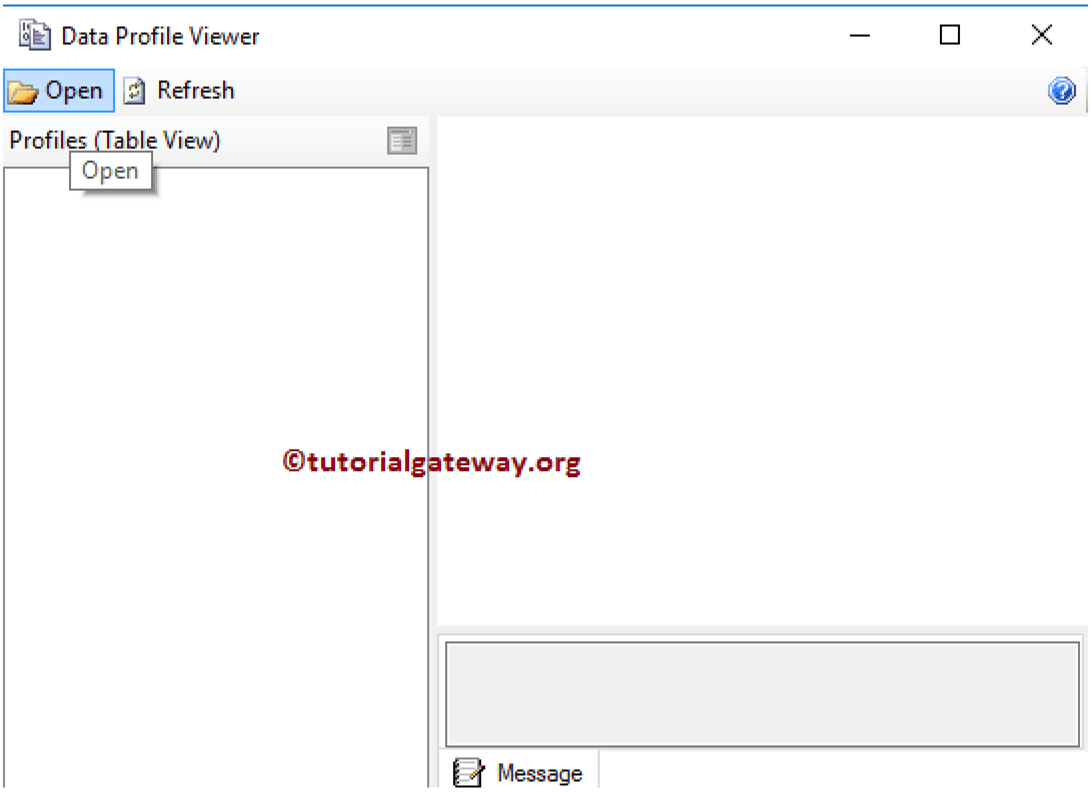
RequestID
Profile Request ID

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

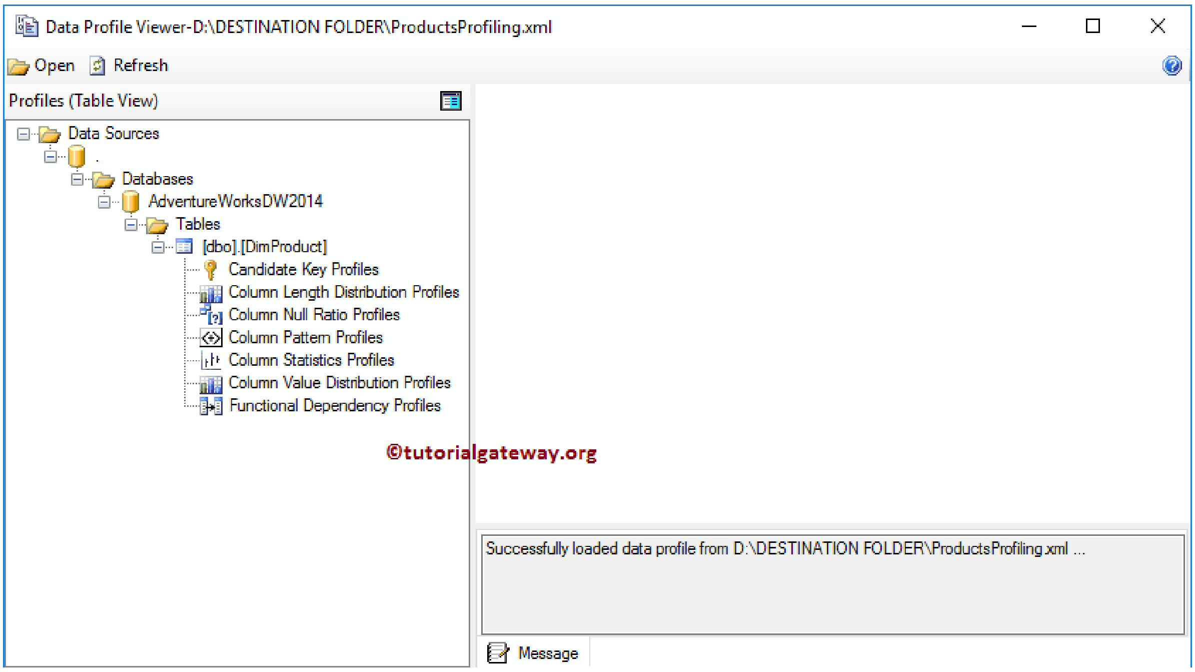


As you see from the above screenshot, our SSIS package executed successfully. To see the SSIS Data Profiling Task generated Data, double click and select the Data Profile Viewer or alternatively use the search bar and type Data Profile Viewer or navigate to *C:\ProgramData\Microsoft\Windows\Start Menu\Programs\Microsoft 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.

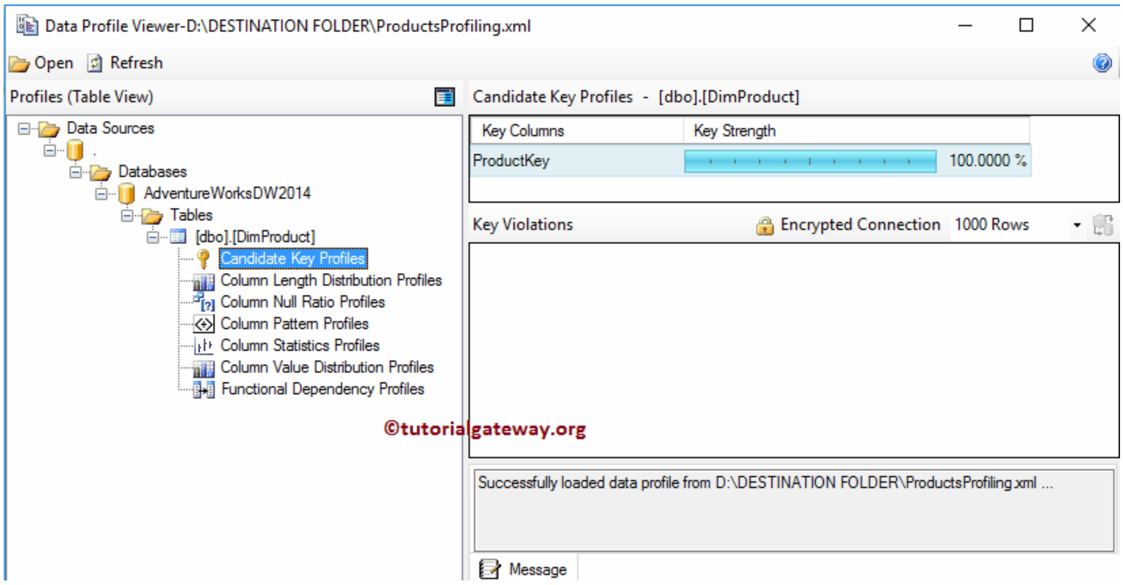


From the below screenshot, see the list of profiles that we selected while configuring the SSIS Data Profiling Task package.


Data Profile Viewer-D:\DESTINATION FOLDER\ProductsProfiling.xml
Open
Refresh
Profiles (Table View)
Data Sources
.
Databases
AdventureWorksDW2014
Tables
[dbo].[DimProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
©tutorialgateway.org
Message

Successfully loaded data profile from D:\DESTINATION FOLDER\ProductsProfiling.xml ...

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.


Data Profile Viewer-D:\DESTINATION FOLDER\ProductsProfiling.xml
Open
Refresh
Profiles (Table View)
Data Sources
.
Databases
AdventureWorksDW2014
Tables
[dbo].[DimProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
Key Violations
Encrypted Connection
1000 Rows
©tutorialgateway.org
Message

Candidate Key Profiles - [dbo].[DimProduct]

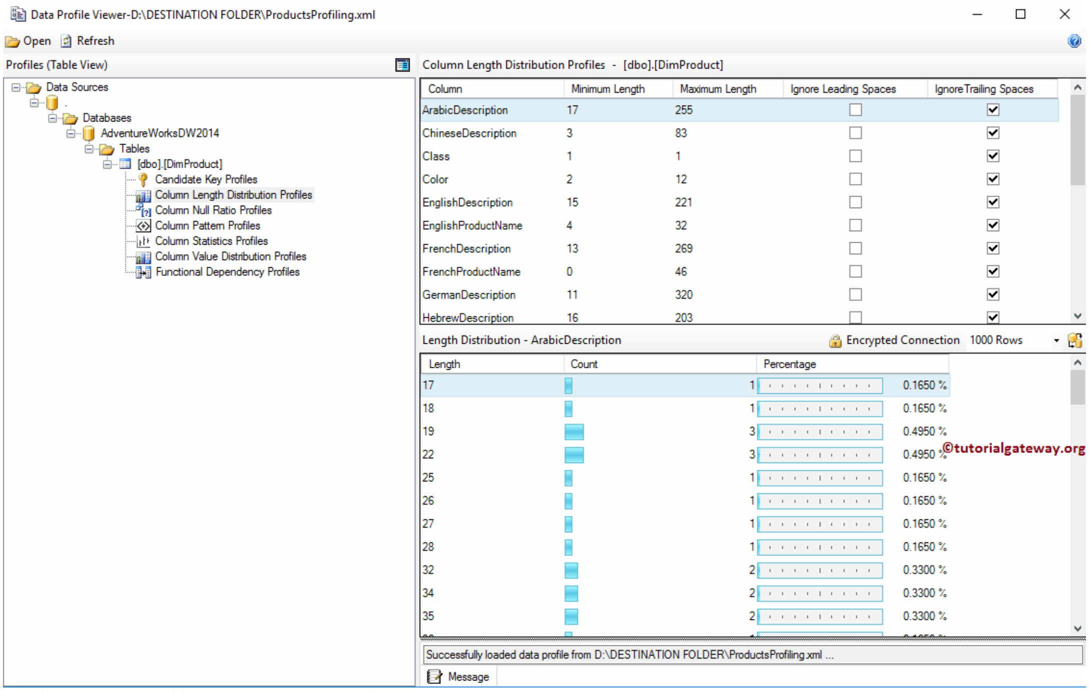
Key Columns	Key Strength
ProductKey	100.0000 %

Successfully loaded data profile from D:\DESTINATION FOLDER\ProductsProfiling.xml ...

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).


Data Profile Viewer-D:\DESTINATION FOLDER\ProductsProfiling.xml
Open
Refresh
Profiles (Table View)
Data Sources
.
Databases
AdventureWorksDW2014
Tables
[dbo].[DimProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
Encrypted Connection
1000 Rows
Message
©tutorialgateway.org

Column Length Distribution Profiles - [dbo].[DimProduct]

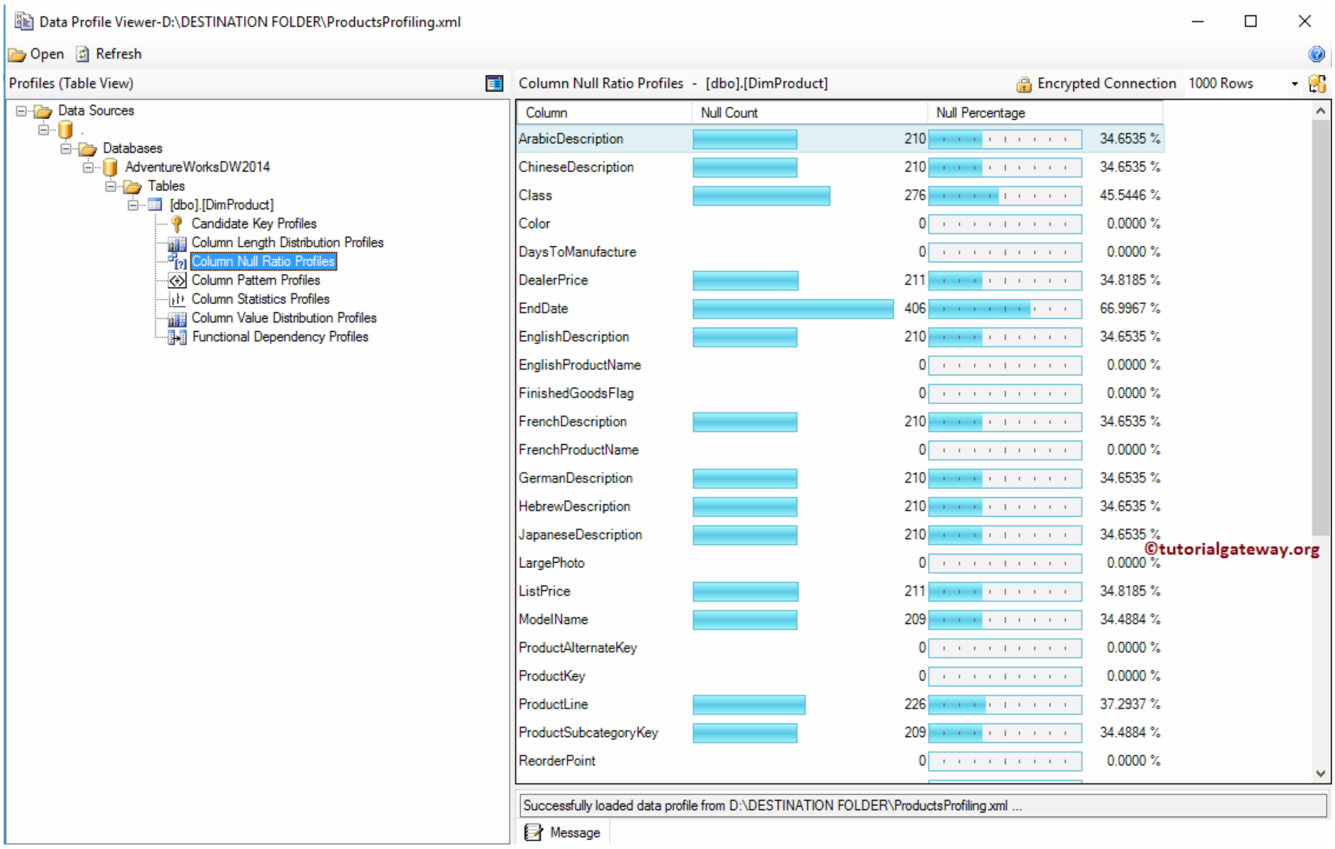
Column	Minimum Length	Maximum Length	Ignore Leading Spaces	IgnoreTrailing Spaces
ArabicDescription	17	255	☐	☑
ChineseDescription	3	83	☐	☑
Class	1	1	☐	☑
Color	2	12	☐	☑
EnglishDescription	15	221	☐	☑
EnglishProductName	4	32	☐	☑
FrenchDescription	13	269	☐	☑
FrenchProductName	0	46	☐	☑
GermanDescription	11	320	☐	☑
HebrewDescription	16	203	☐	☑

Length Distribution - ArabicDescription

Length	Count	Percentage
17	1	0.1650 %
18	1	0.1650 %
19	3	0.4950 %
22	3	0.4950 %
25	1	0.1650 %
26	1	0.1650 %
27	1	0.1650 %
28	1	0.1650 %
32	2	0.3300 %
34	2	0.3300 %
35	2	0.3300 %

Successfully loaded data profile from D:\DESTINATION FOLDER\ProductsProfiling.xml ...

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.



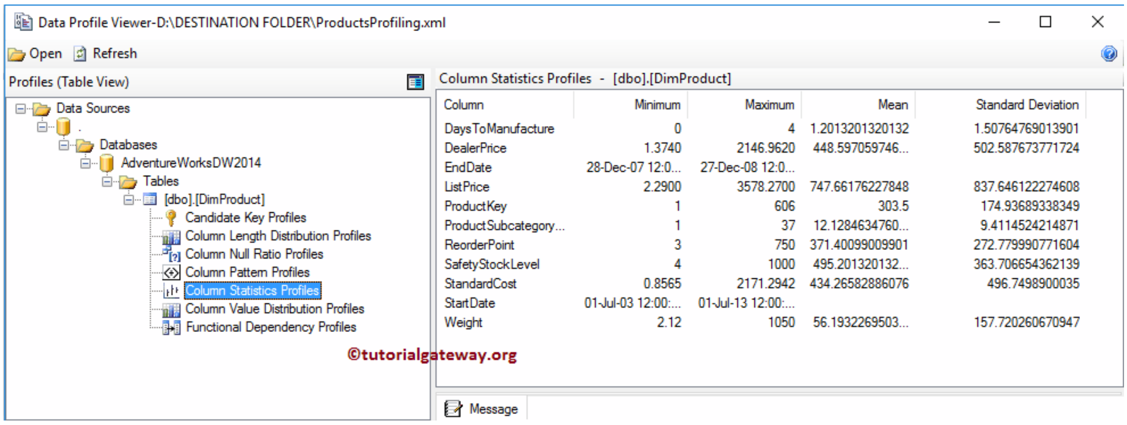
Data Profile Viewer-D:\DESTINATION FOLDER\ProductsProfiling.xml
Open
Refresh
Profiles (Table View)
Data Sources
Databases
AdventureWorksDW2014
Tables
[dbo].[DimProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
Encrypted Connection
1000 Rows
Message
©tutorialgateway.org

Column Null Ratio Profiles - [dbo].[DimProduct]

Column	Null Count	Null Percentage
ArabicDescription	210	34.6535 %
ChineseDescription	210	34.6535 %
Class	276	45.5446 %
Color	0	0.0000 %
DaysToManufacture	0	0.0000 %
DealerPrice	211	34.8185 %
EndDate	406	66.9967 %
EnglishDescription	210	34.6535 %
EnglishProductName	0	0.0000 %
FinishedGoodsFlag	0	0.0000 %
FrenchDescription	210	34.6535 %
FrenchProductName	0	0.0000 %
GermanDescription	210	34.6535 %
HebrewDescription	210	34.6535 %
JapaneseDescription	210	34.6535 %
LargePhoto	0	0.0000 %
ListPrice	211	34.8185 %
ModelName	209	34.4884 %
ProductAlternateKey	0	0.0000 %
ProductKey	0	0.0000 %
ProductLine	226	37.2937 %
ProductSubcategoryKey	209	34.4884 %
ReorderPoint	0	0.0000 %

Successfully loaded data profile from D:\DESTINATION FOLDER\ProductsProfiling.xml ...

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



Data Profile Viewer-D:\DESTINATION FOLDER\ProductsProfiling.xml
Open
Refresh
Profiles (Table View)
Data Sources
Databases
AdventureWorksDW2014
Tables
[dbo].[DimProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
©tutorialgateway.org
Message

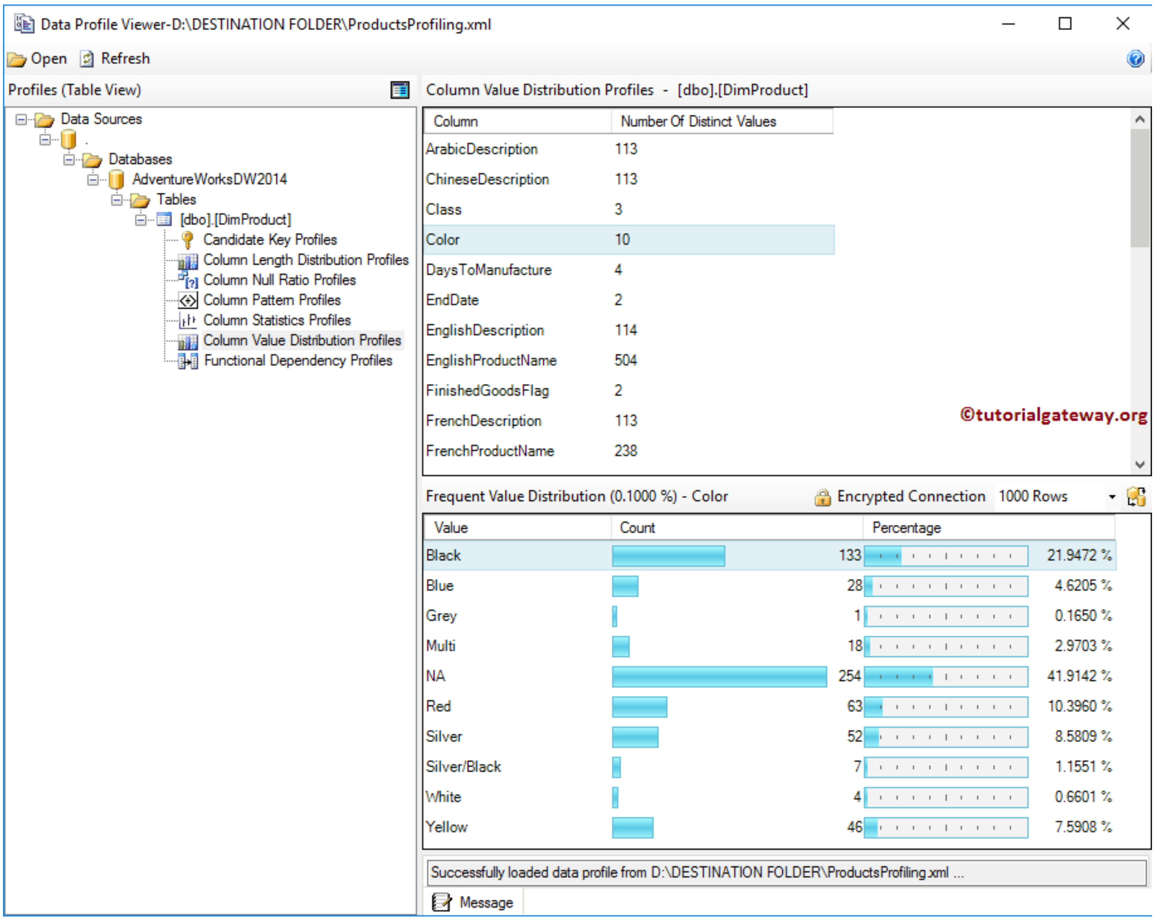
Column Statistics Profiles - [dbo].[DimProduct]

Column	Minimum	Maximum	Mean	Standard Deviation
DaysToManufacture	0	4	1.2013201320132	1.50764769013901
DealerPrice	1.3740	2146.9620	448.597059746...	502.587673771724
EndDate	28-Dec-07 12:0...	27-Dec-08 12:0...		
ListPrice	2.2900	3578.2700	747.66176227848	837.646122274608
ProductKey	1	606	303.5	174.93689338349
ProductSubcategory...	1	37	12.1284634760...	9.4114524214871
ReorderPoint	3	750	371.40099009901	272.779990771604
SafetyStockLevel	4	1000	495.201320132...	363.706654362139
StandardCost	0.8565	2171.2942	434.26582886076	496.7498900035
StartDate	01-Jul-03 12:00:...	01-Jul-13 12:00:...		
Weight	2.12	1050	56.1932269503...	157.720260670947

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.


Data Profile Viewer-D:\DESTINATION FOLDER\ProductsProfiling.xml
Open
Refresh
Profiles (Table View)
Data Sources
.
Databases
AdventureWorksDW2014
Tables
[dbo].[DimProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
©tutorialgateway.org
Encrypted Connection
1000 Rows
Message

Column Value Distribution Profiles - [dbo].[DimProduct]

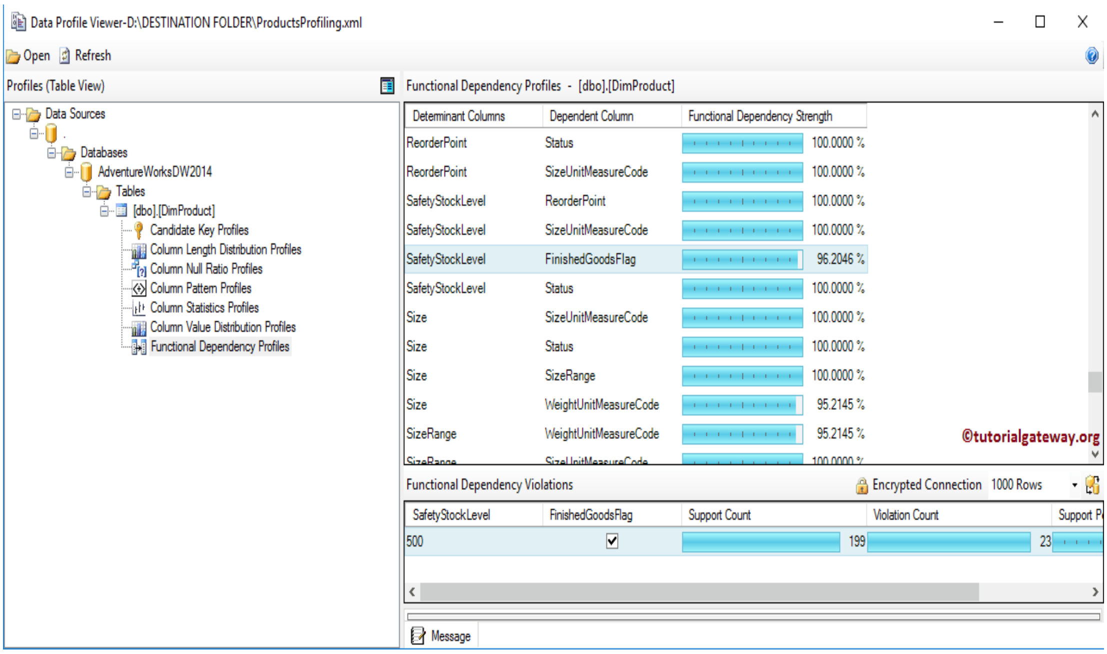
Column	Number Of Distinct Values
ArabicDescription	113
ChineseDescription	113
Class	3
Color	10
DaysToManufacture	4
EndDate	2
EnglishDescription	114
EnglishProductName	504
FinishedGoodsFlag	2
FrenchDescription	113
FrenchProductName	238

Frequent Value Distribution (0.1000 %) - Color

Value	Count	Percentage
Black	133	21.9472 %
Blue	28	4.6205 %
Grey	1	0.1650 %
Multi	18	2.9703 %
NA	254	41.9142 %
Red	63	10.3960 %
Silver	52	8.5809 %
Silver/Black	7	1.1551 %
White	4	0.6601 %
Yellow	46	7.5908 %

Successfully loaded data profile from D:\DESTINATION FOLDER\ProductsProfiling.xml ...

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).


Data Profile Viewer-D:\DESTINATION FOLDER\ProductsProfiling.xml
Open
Refresh
Profiles (Table View)
Data Sources
.
Databases
AdventureWorksDW2014
Tables
[dbo].[DimProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
©tutorialgateway.org
Encrypted Connection
1000 Rows
Message

Functional Dependency Profiles - [dbo].[DimProduct]

Determinant Columns	Dependent Column	Functional Dependency Strength
ReorderPoint	Status	100.0000 %
ReorderPoint	SizeUnitMeasureCode	100.0000 %
SafetyStockLevel	ReorderPoint	100.0000 %
SafetyStockLevel	SizeUnitMeasureCode	100.0000 %
SafetyStockLevel	FinishedGoodsFlag	96.2046 %
SafetyStockLevel	Status	100.0000 %
Size	SizeUnitMeasureCode	100.0000 %
Size	Status	100.0000 %
Size	SizeRange	100.0000 %
Size	WeightUnitMeasureCode	95.2145 %
SizeRange	WeightUnitMeasureCode	95.2145 %
SizeRange	SizeUnitMeasureCode	100.0000 %

Functional Dependency Violations

SafetyStockLevel	FinishedGoodsFlag	Support Count	Violation Count	Support P
500	☑	199	23	