

Introduction to the SQL Server Data Profiler Task - Part 1

 databasejournal.com/features/mssql/introduction-to-the-sql-server-data-profiler-task-part-1.html

[Database Journal](#) | [DBA Support](#) | [SQLCourse](#)
[SQLCourse2](#)

[Advertiser](#)
[Disclosure](#)

Featured Database Articles

MS SQL

Posted November 5, 2012

By Greg Larsen

If you need to analyze data in a SQL Server table one of the tasks you might want to consider is profiling your data. By profiling the data, I mean looking for data patterns, like the number of different distinct values for each column, or the number of rows associated with each of those distinct values, etc. What tools are you using to perform data profiling? In this article I will be exploring how to use the SSIS Data Profiling Task to perform data profiling.

Data Profiling Task

Microsoft introduced a new SSIS task to profile data. That task is called "Data Profiling". It was first introduced with SQL Server 2008 R2, and has been retained as an SSIS task in SQL Server 2012.

The Data Profiling task can be used to perform analysis of data patterns within a SQL Server table. This analysis is useful for examining data prior to loading it into a final destination, like a data warehouse. By analyzing data and determining the patterns in the data you can determine how clean the data might be, prior to loading it into the data warehouse. By performing a profiling task on incoming data you are able to verify your new data meets the quality you expect prior to loading the data into its final location. If the data doesn't meet the quality you normally expect data profiling allows you to reject the incoming data.

The Data Profile task within SSIS can look at your data using eight different profiles. Five of these profiles analyze data at an individual column level, while the other three profiles analyze multiple columns and/or relationships between columns. In Table 1 is a description of the eight different profiles as documented in Microsoft's Books Online.

The following five profiles analyze individual columns.

Profiles that analyze individual columns	Description
Column Length Distribution Profile	<p>Reports all the distinct lengths of string values in the selected column and the percentage of rows in the table that each length represents.</p> <p>This profile helps you identify problems in your data, such as values that are not valid. For example, you profile a column of United States state codes that should be two characters and discover values longer than two characters.</p>
Column Null Ratio Profile	<p>Reports the percentage of null values in the selected column.</p> <p>This profile helps you identify problems in your data, such as an unexpectedly high ratio of null values in a column. For example, you profile a Zip Code/Postal Code column and discover an unacceptably high percentage of missing codes.</p>
Column Pattern Profile	<p>Reports a set of regular expressions that cover the specified percentage of values in a string column.</p> <p>This profile helps you identify problems in your data, such as strings that are not valid. This profile can also suggest regular expressions that can be used in the future to validate new values. For example, a pattern profile of a United States Zip Code column might produce the regular expressions: <code>\d{5}-\d{4}</code>, <code>\d{5}</code>, and <code>\d{9}</code>. If you see other regular expressions, your data likely contains values that are not valid or in an incorrect format.</p>
Column Statistics Profile	<p>Reports statistics, such as minimum, maximum, average, and standard deviation for numeric columns, and minimum and maximum for datetime columns.</p> <p>This profile helps you identify problems in your data, such as dates that are not valid. For example, you profile a column of historical dates and discover a maximum date that is in the future.</p>

Column Value Distribution Profile	<p>Reports all the distinct values in the selected column and the percentage of rows in the table that each value represents. Can also report values that represent more than a specified percentage of rows in the table.</p> <p>This profile helps you identify problems in your data, such as an incorrect number of distinct values in a column. For example, you profile a column that is supposed to contain states in the United States and discover more than 50 distinct values.</p>
<p>The following three profiles analyze multiple columns or relationships between columns and tables.</p>	

Profiles that analyze multiple columns	Description
Candidate Key Profile	<p>Reports whether a column or set of columns is a key, or an approximate key, for the selected table.</p> <p>This profile also helps you identify problems in your data, such as duplicate values in a potential key column.</p>
Functional Dependency Profile	<p>Reports the extent to which the values in one column (the dependent column) depend on the values in another column or set of columns (the determinant column).</p> <p>This profile also helps you identify problems in your data, such as values that are not valid. For example, you profile the dependency between a column that contains United States Zip Codes and a column that contains states in the United States. The same Zip Code should always have the same state, but the profile discovers violations of this dependency.</p>
Value Inclusion Profile	<p>Computes the overlap in the values between two columns or sets of columns. This profile can determine whether a column or set of columns is appropriate to serve as a foreign key between the selected tables.</p> <p>This profile also helps you identify problems in your data, such as values that are not valid. For example, you profile the ProductID column of a Sales table and discover that the column contains values that are not found in the ProductID column of the Products table.</p>

Table 1: Different Profiles Available in SSIS

To profile your data you need to build an SSIS package. In this package you identify your data sources, an output file and the different profiles you want to run against your data sources. You can do this all through the Data Profiling Task. The output of the Data Profiling Task is an XML file. The XML file can be viewed graphically using the Data Profile Viewer. The Data Profile Viewer can be launched independently from an exe, or you can launch it from within the Data Profiling Task Editor with a click of a button. Using the Data Profile Viewer you can examine every profile you ran and then drill down on specific items within the profile output to review a set of records associated with a specific drill down request.

Rather than describe how this works, let me show you with an example.

Data to Analysis

As previously stated the Data Profiling Task only works against data loaded into a SQL Server table. The Data Profiling Task can be run against any SQL Server data table that resides in a SQL Server 2000 or above database. Therefore in order to demo the Data Profiler Task I will need some data to analyze. I will be using the AdventureWorks2008R2 database for my demonstration. If you want to follow along you can download the same database from the following location: <http://msftdbprodsamples.codeplex.com/downloads/get/478216>. Once you have downloaded this database you will have to attach it to your SQL Server database engine.

In my example I will be using the Visual Studio 2010 Shell that was installed with the SQL Server Data Tools, as part of my SQL Server 2012 installation to setup and run my Data Profiler Task. I will be running my data profiling against the AdventureWorks2008R2 database that resides within my SQL Server 2012 environment.

Defining Properties for a Data Profiling Task

To build my data profiling SSIS package I first need to open my Visual Studio 2010 shell and create a new Integration Services project. Once my new project opens up I can drag the "Data Profiling Task" from the toolbox, to the Control Flow area as I show in Figure 1.

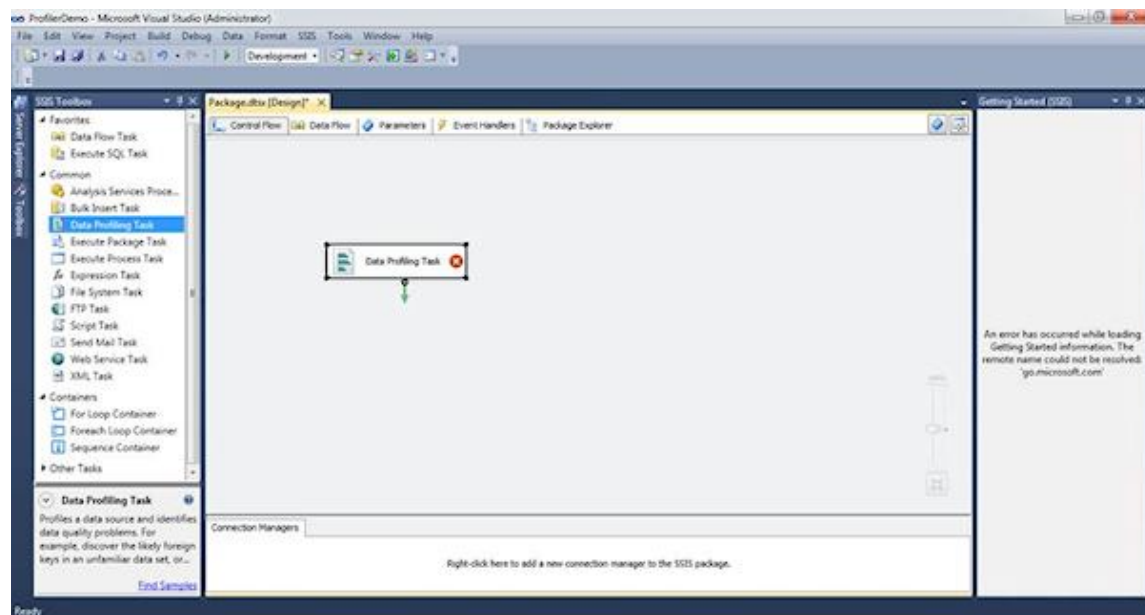


Figure 1: Data Profiling Task in Control Flow

The next step is to drill into the Data Profiling Task and start defining its properties. To start identifying the Data Profile Task properties I will double click with the left mouse button on the "Data Profiling Task". When I do this the Data Profiling Task Editor is displayed, as in Figure 2.

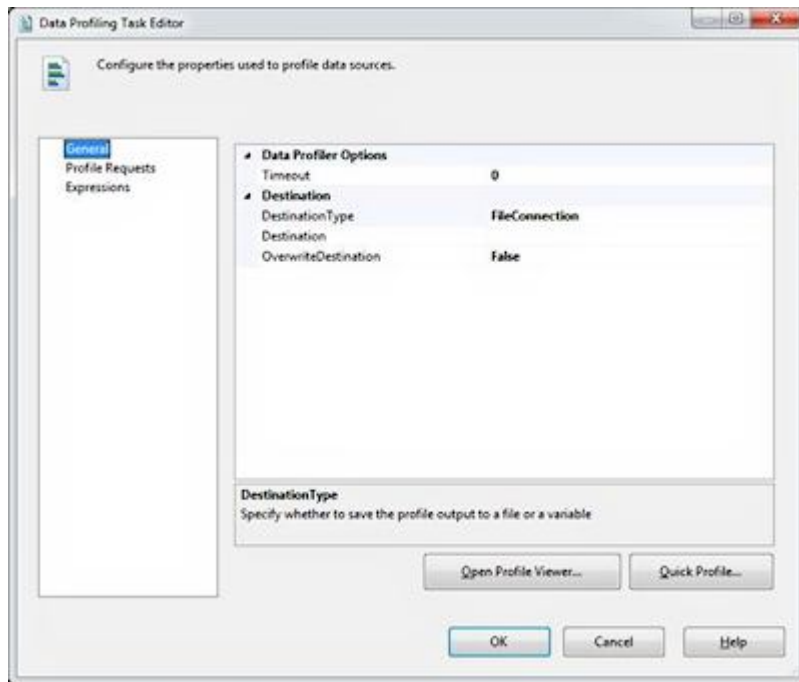


Figure 2: Data Profiler Task Editor

On the window displayed in Figure 2 you can see there are a number of different options in the left pane, and the properties for the General item is shown in the right pane of the Data Profiler Task Editor window. In the right pane I can set the General items properties to identify where to store the output of the Data Profiling Task. The output will be an XML file that contains profile information about the table I will be profiling. I want my profile information to go to a file named C:\temp\ProfileDemo1.xml. To identify this location to the Data Profile Task Editor I position my mouse over the cell next to the "Destination" label in Figure 2, and then click on the left mouse button. This will bring up a down arrow that I can select. Upon doing this a drop down window will appear and then I select the "<New File connection...>" item. When I do this a File Connection Manager Editor window is displayed. The File Connection Manager Editor defaults the "Usage Type" to "Existing File", since my XML file doesn't exist yet I need to use the drop down menu to select a usage type of "Create File", and then type the name of my file in the "File" text box. When I'm done specifying my location my File Connection Manager Editor window looks like the window show in Figure 3.

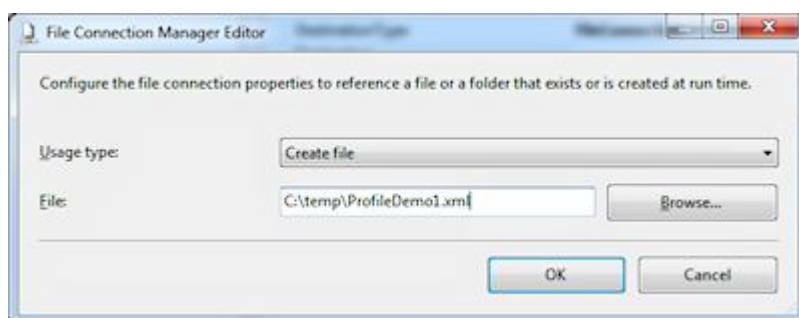


Figure 3: Identifying Connection for my XML file location

To finish creating my new connection to my XML file I just need to click on the "OK" button. When I do this, I will be taken back to the Data Profiling Task Editor window.

The next step in setting up my Data Profiling Task is to identify the table or tables I want to profile, and the profiles I want use to analyze those tables. There are two different ways to do this. One way to do this is to click on the "Profile Requests" item in the left pane and then identify the different profiles I want to run against a specific table or view, one by one. Or I can use the "Quick Profile" button to identify a number of profiles to run against a single table or view. In this article I will show you how to use the "Quick Profile..." button to identify the different profiles I want to run. When I click on this button the Single Table Quick Profile Form will be displayed. On this form I can identify the connection, table and profiles I want to run. Figure 4 has my completed form for the profile criteria I wish to run.

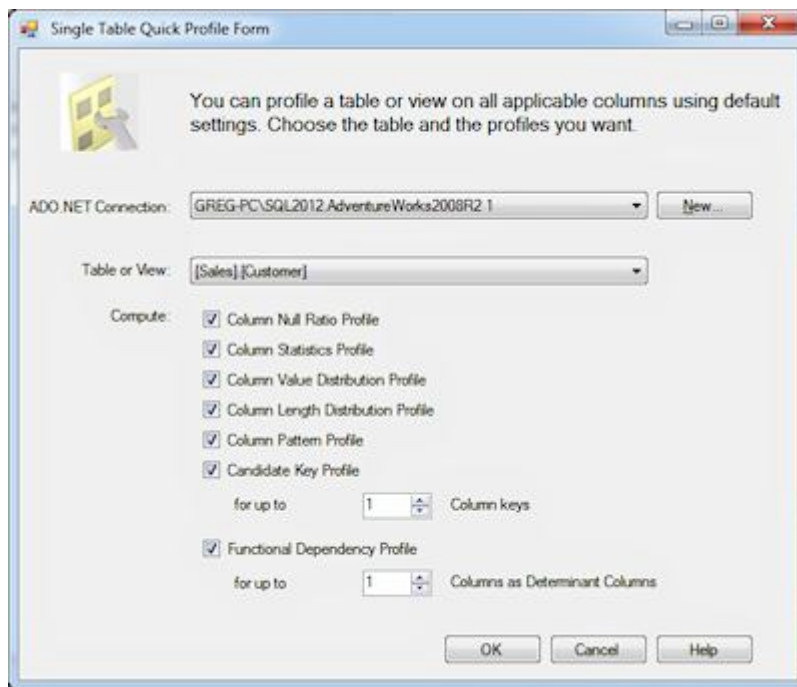


Figure 4: Completed Single Table Quick Profile Form

On this form, I created my connection to my Adventureworks2008R2 database by using the "New..." button dialog box to define this connection. As you can see, I want to profile the [Sales].[Customer] table. For this demo, I use the default profile settings that were identify by the quick profile process. Plus I added the "Column Pattern" and "Functional Dependency" profiles, which were not originally selected by default. I also took all the defaults for these different profiles. To get my selected profile settings associated with my Data Profiling Task all I need to do is click on the "OK" button. Upon doing this, the window in Figure 5 will be displayed.

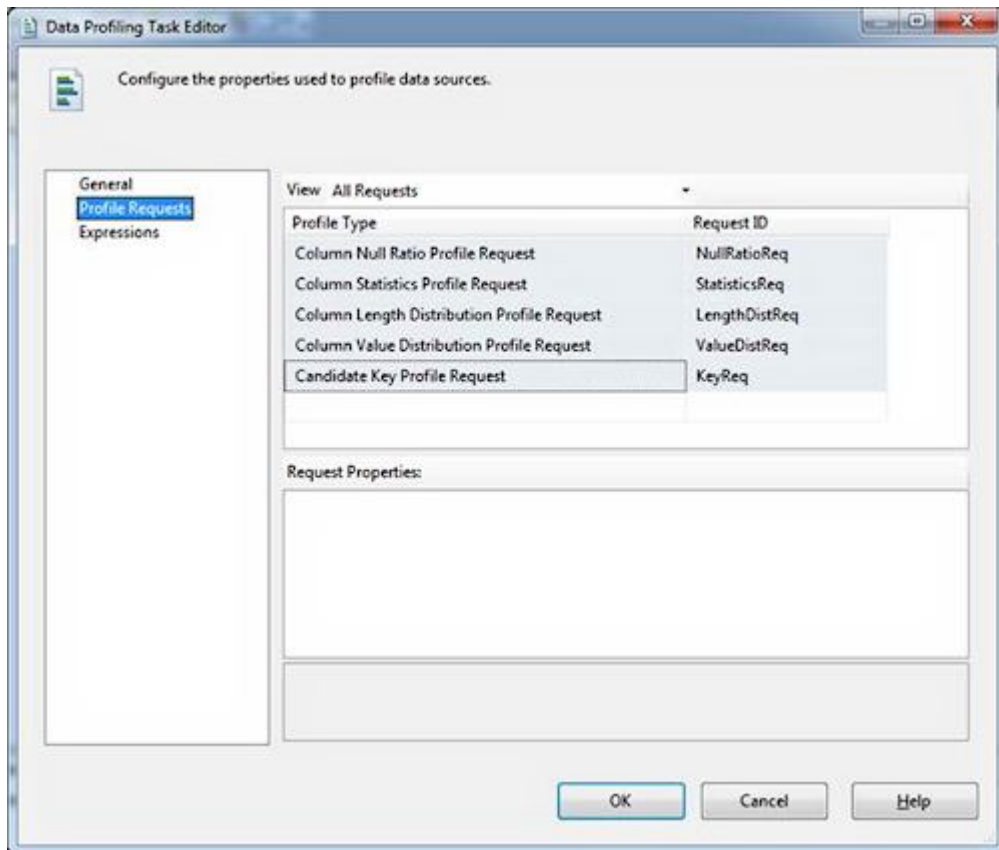


Figure 5: Profile requests

If you look at Figure 5 you will see that all the profiles that were identified during the quick profile process, are now listed as properties for the Profile Requests item in the left pane. I could click on any one of the Profile Type's to view or modify the properties. Since this is an introductory article I will not be discussing how to customize the different profiles. I suggest, if you are following along, that you consider reviewing each profile and understand what default properties were set and, and how you might be able to change these properties to customize the profile process.

One thing worth mentioning is there is a profile that is not available within the Quick Profile Form. That missing profile is the Value Inclusion profile. If you want to use this profile you will need to identify it by using the right pane when viewing the Profile Requests. This can be done by going to the Profile Type cell in the "View All Requests" pane and clicking on it, then selecting and configuring that profile.

If you'd like to profile multiple tables or views, you can do that with a single Data Profiling Task. In order to identify additional tables all you would need to do is click on the "Quick Profile" button multiple times and identify a different table or view that you want to profile each time you passed through the quick profile process. Alternatively, you could manually specify each profile you wanted to run against a table using the Profile Request item.

The Expressions item in the left pane of the Data Profiling task editor can be used to set expressions for the different Data Profile Task properties. You can use the expressions to dynamically control your data profiling properties.

Once I have set all my Data Profiling Task properties all that is left to do is run my Data Profiling Task, and then review the results. To do this I first click on the "OK" button on the "Data Profiling Task Editor" window, which will take me back to the Control Flow window of my SSIS package. Once there I can hit the F5 button to execute my Data Profiling Task.

Reviewing Data Profiling Task Output

As already mentioned when you run a Data Profiling Task the output is written to an XML file. To graphically view this file you need to use the Data Profile Viewer. There are two ways to launch the Data Profile Viewer. You can launch it with a button from the Data Profiling Task Editor window or manually launch the Data Profiler Viewer executable.

Once my Data Profile task has completed I can double click on the Data Profiling Task on the Control Flow to bring up the Data Profiling Task Editor where I can click on the "Open Profile Viewer..." button. Upon doing that the Data Profile Viewer will be displayed as found in Figure 6.

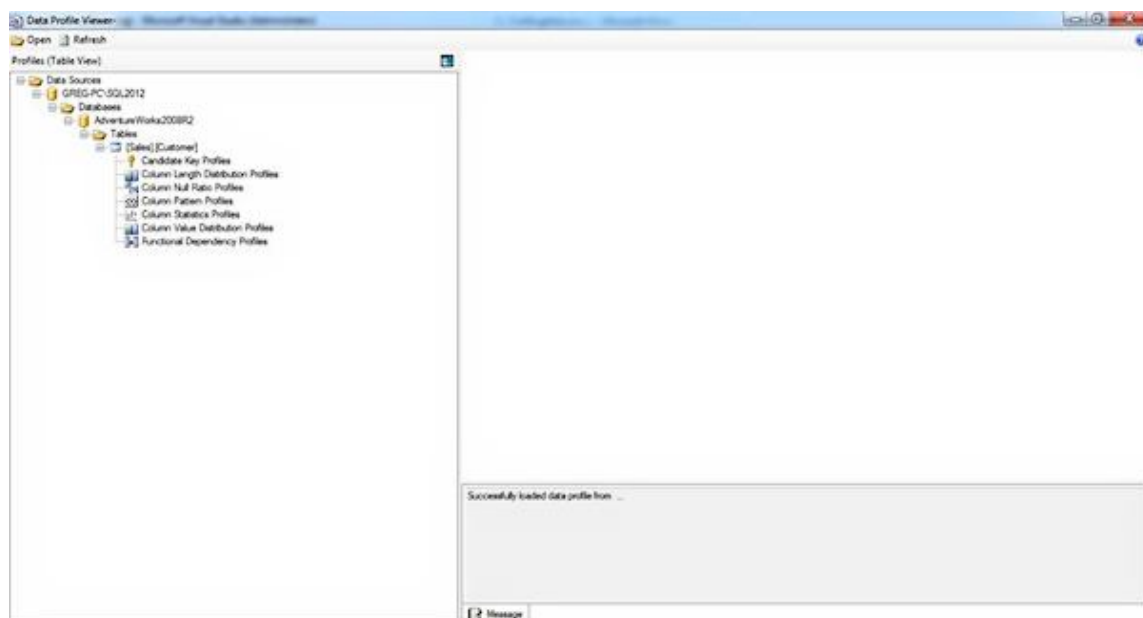


Figure 6: Data Profile Viewer

The left pane of the Data Profile Viewer shows a navigation tree that can be used to view the output of each of the different profiles I ran. To see the results of any one of these profiles all I need to do is click on the given profile. Let's review the Column Null Ratio Profiles item. To do that I just click on the Column Null Ratio Profiles item in the left pane and the window in Figure 7 is displayed.

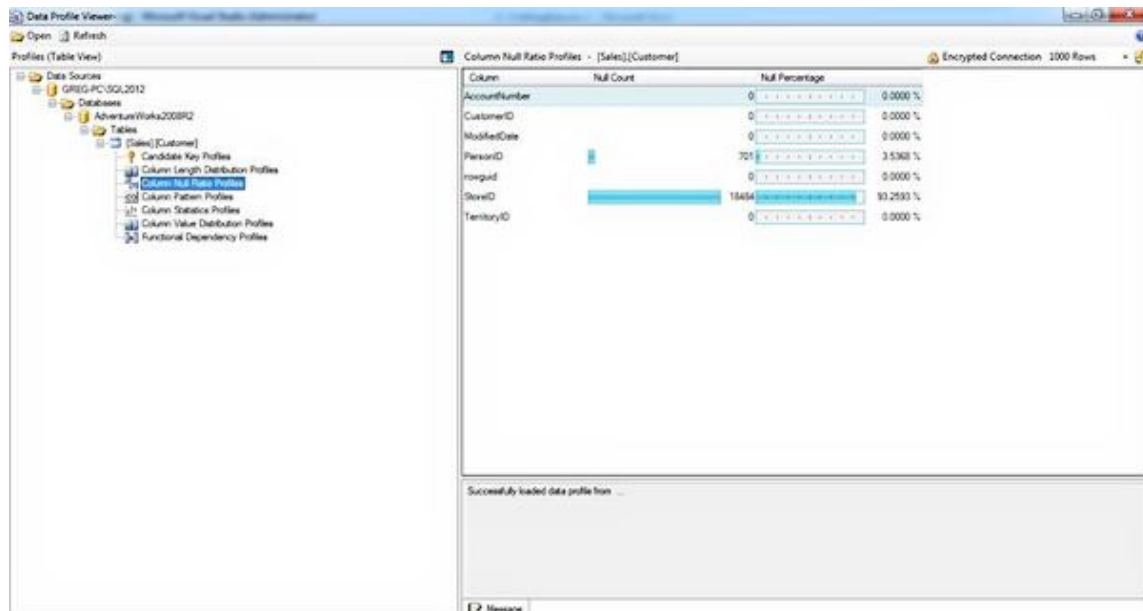


Figure 7: Column Null Profiles Results

By looking at the right pane you can see that the PersonID and StoreID have null values. A small amount of Customer records (3.5%) do not have a PersonID, and a little over 93% of the records do not have a StoreID. I could also sort this display by clicking on a particular column that I want to have in a sorted order. Clicking the column a second time will sort it in the inverse order.

This display also has drill down capabilities. If I wanted to drill down and see all the Customer records that have PersonID as null, I would just click on the PersonID row, and then clicking on the Drill down icon, which is in the upper right hand side of this pane next to the text that says "1000 rows". Or you can double click on the "PersonID" row. When I drill down on the PersonID column I get the results as show in Figure 8.

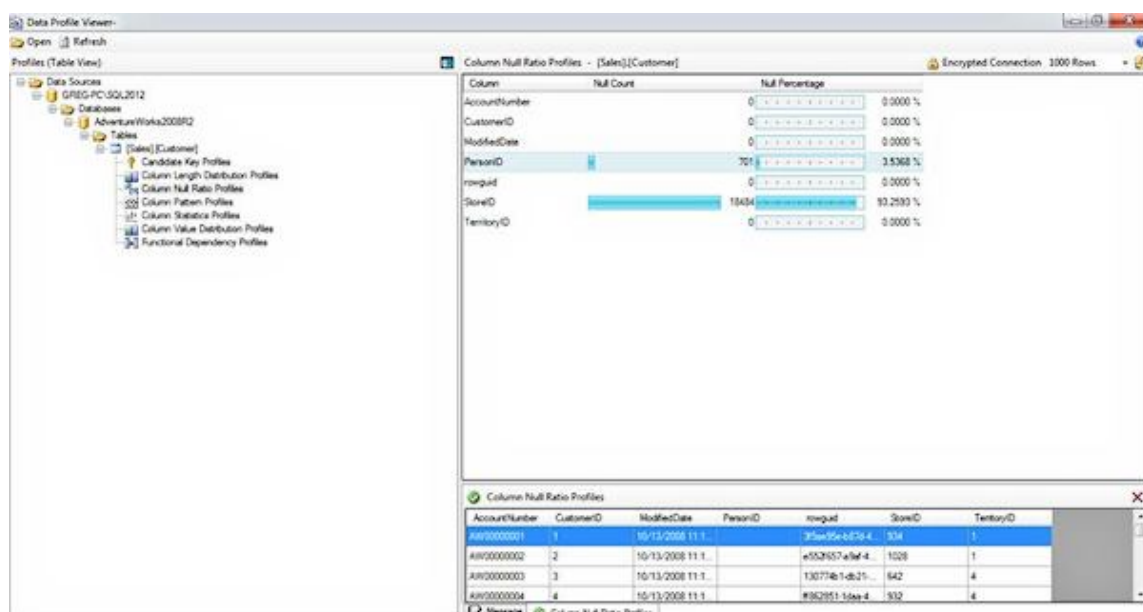


Figure 8: Rows with NULL PersonID

The drill down result is shown in the bottom part of the right pane. You can use the scroll bar on the right to scan all the rows where the PersonID is null. These detail results will help you determine if these are legitimate nulls, or there is a problem with the data in your table.

I will not be going through the rest of the profile output. But I suggest, if you are following along, you look at all different output for each profile that is displayed.

As mentioned earlier, the second way to view your Data Profiler XML output is to launch the DataProfilerViewer.exe executable manually. This exe is located in the Binn directory for DTS. On my machine I found it in the C:\Program Files (x86)\Microsoft SQL Server\110\DTS\Binn directory. Depending on your machine, and version of SQL Server, and where you installed SQL Server you may find it in a different location.

When manually running the Data Profile Viewer, it will come up with an empty profile view as seen in Figure 9.

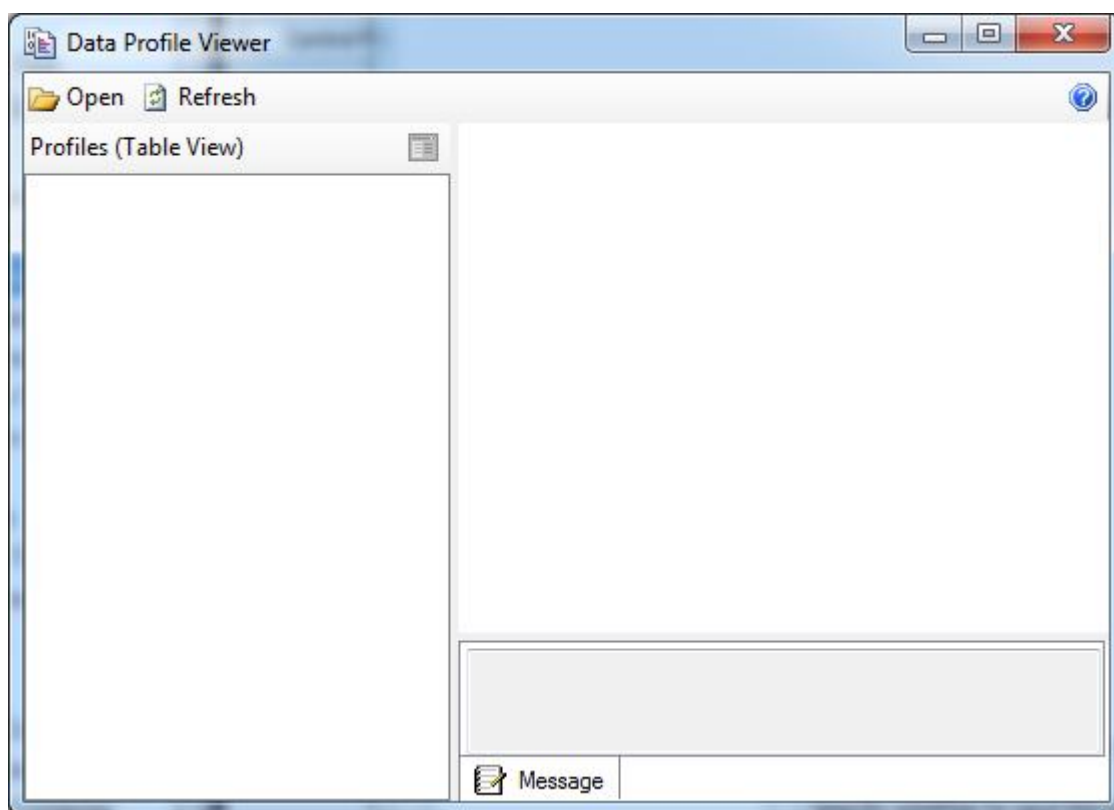


Figure 9: Data Profile Viewer

To view my profile information I would need to use the "Open" menu item to bring up my Data Profiler XML file results in the Data Profile Viewer. I find it much easier to use the "Open Profile Viewer..." button to launch the Data Profile Viewer to explore the data profiling results, especially when I am performing data profiling interactively. If you are running your profiling tasks as part of a batch process, then using the manual method of launching to Data Profiler Viewer would make more sense.

Profile Your Data

The Data Profiling Task is an excellent place to start profiling any incoming data prior to loading the data into a production environment. By profiling your input data you can make sure it meets acceptable quality levels. One of the drawbacks of the Data Profiling Task is it cannot profile flat files, or third party data sources. Even with this limitation it is a very good tool for exploring your SQL Server data. Next time you need to analyze some input data you might find that the Data Profiling Task is one of the best ways to accomplish profiling.

This article only showed you at a high level how the Data Profiling Task works. My next article will explore in more detail some of the different profiles. This more detailed explanation of the profiles will give you a better appreciation of the power of the Data Profiling task and how it can be used to improve the quality of the data in your database.

[See all articles by Greg Larsen](#)

[MS SQL Archives](#)

MS SQL Forum			
Topic	By	Replies	Updated
<u>SQL 2005: SSIS: Error using SQL Server credentials</u>	<u>poverty.</u>	<u>3</u>	<u>August 17th, 07:43 AM</u>
<u>Need help changing table contents</u>	<u>nkawtg</u>	<u>1</u>	<u>August 17th, 03:02 AM</u>
<u>SQL Server Memory confifuration</u>	<u>bhosalenarayan</u>	<u>2</u>	<u>August 14th, 05:33 AM</u>
<u>SQL Server ♦ Primary Key and a Unique Key</u>	<u>katty.jonh</u>	<u>2</u>	<u>July 25th, 10:36 AM</u>