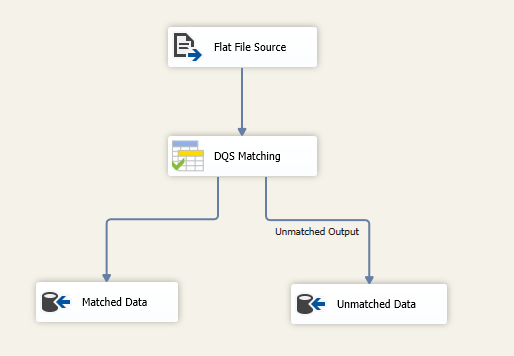
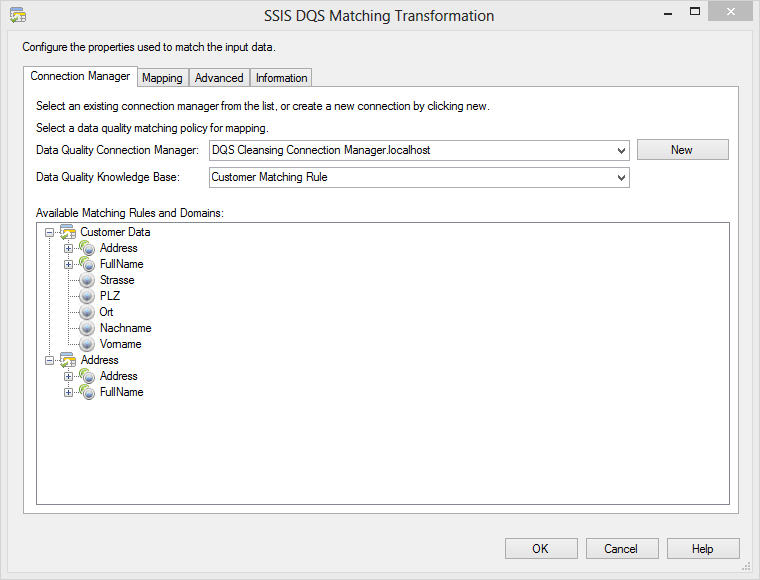
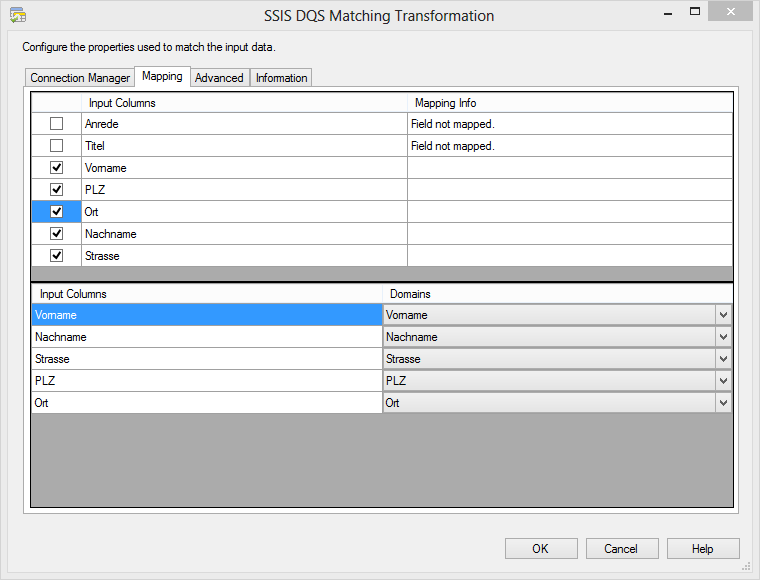
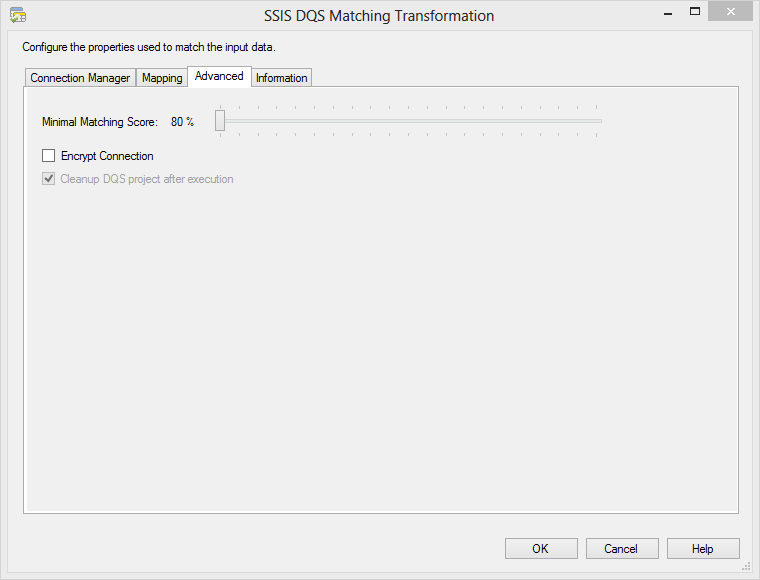
**Project Description**  
The SSIS DQS Matching Transformation uses Data Quality Services (DQS) to find duplicate data within the SSIS data flow.

**Now with support for SQL Server 2012, SQL Server 2014 and SQL Server 2016 CTP 2.3.**  
  
**A note beforehand: The DQS API is not officially supported by Microsoft.**   
  
The SSIS DQS matching component is developed and tested for SQL Server 2012 with DQS assembly version 11.0.3000.0. We try to support future versions of DQS and SQL Server. However, we cannot guarantee that the component will work with all future versions of DQS.  
  
The Data Quality Services (DQS) data matching process enables reducing duplicate data. The matching analyzes the degree of duplication in all records. It returns weighted probabilities of a match between each set of compared records.  
  
Similar to standard data quality processes in DQS, you have to perform the matching by building a knowledge base. With the SSIS DQS Matching component you can execute a matching activity in a SSIS package.  
  
You will find additional information about creating a knowledge base and the matching policy in the following MSDN articles: Data matching.  
  
With the SSIS DQS Matching component you could check duplicate data from different data sources - whether the data source is a SQL Server, an Excel file, an Oracle database or any other data sources.

A very good description on how to use the component can be found on the Data Quality Services Blog: Automating the data matching process in SQL Server Data Quality Services (DQS)  
  
[](http://ssis-components.net/wp-content/uploads/2013/06/001-Microsoft-Visual-Studio-2012.png)  [](http://ssis-components.net/wp-content/uploads/2013/06/002-SSIS-Connection-Manager.png)

[](http://ssis-components.net/wp-content/uploads/2013/06/003-SSIS-Mapping.png)  [](http://ssis-components.net/wp-content/uploads/2013/06/004-SSIS-Advanced.png)

##### Problem

Data Quality Service (DQS) in SQL Server 2012, which helps ensure data integrity and quality by data profiling, matching, cleansing, correcting and monitoring overall status of the data cleansing process. We also talked about using the interactive Data Quality Client tool and in this tip we will walk through how to do data cleansing in a automated mode.

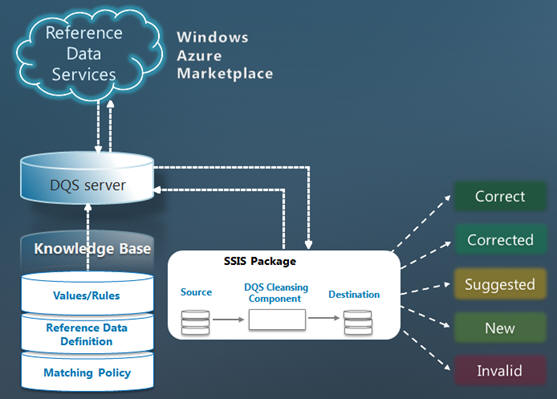
##### Solution

Data Quality Service (DQS) has a SSIS component which you can use for automating the data cleansing process or if you want to run it in batch mode (versus interactively in the Data Quality Client tool).

Before you can start using the Data Cleansing component in SSIS, you need to make sure you have already created and published a knowledge base for cleansing your source data. To learn more about the knowledge base or knowledge discovery and the process of creating a knowledge base, refer to my earlier tip in this series.

Please note, this data cleansing component is a SSIS component which can be installed and run on the same machine or on a different machine than the DQS server. This component works like a client and requires DQS Server for performing data cleansing operations. To learn more about DQS installation and configuration, please refer to my first tip in this series.

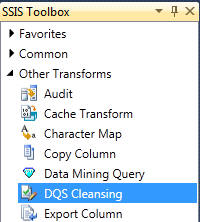
As you can see below, the SSIS component reads data from the source, sends it to the DQS Server for data cleansing and correction and writes the output to the destination. It also includes additional columns to indicate whether the value was already correct, was corrected by DQS, if any other value was suggested or the value was an invalid/unknown value.

  
(Image source is from Gadi Peleg's presentation available here)

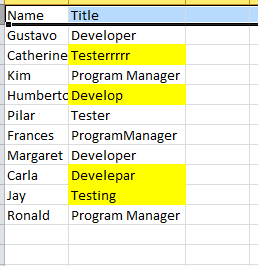
### GGetting started with the DQS cleansing component in SSIS

I am assuming you have basic understanding of SSIS and you are aware of how to create a simple package in SSIS, if not please refer to this tutorial.

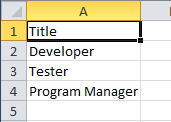
CCreate a SSIS package and drag a data flow task from the Toolbox on to the Control Flow; double click on data flow task to open the task in data flow pane. In the Toolbox of the data flow pane you will notice a new component (DQS Cleansing) appears as shown below:



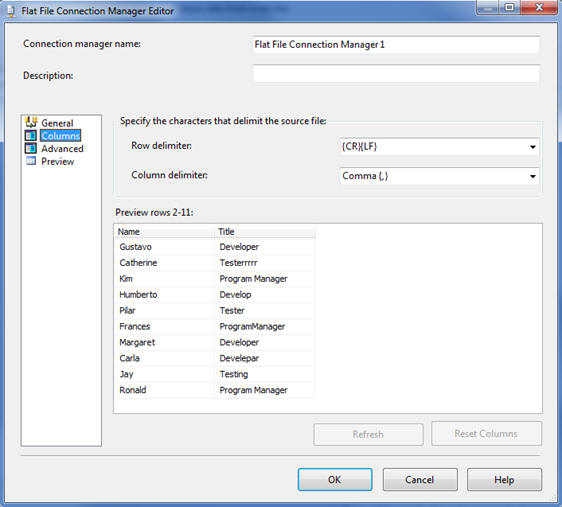
Before we begin using this component, we need to first select/specify the data source which needs to be cleansed, in my case I have a csv file (incorrect values have been highlighted here for clarity) with this data, which I want to cleanse, as shown below:



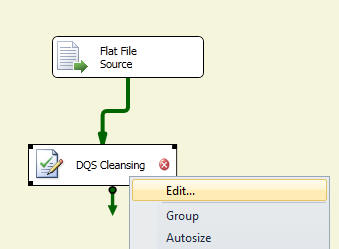
In this example I am going to use the same knowledge base (TitleDomain) which I created in my last tip of this series and which has these entries, as shown below, for valid Title domain:



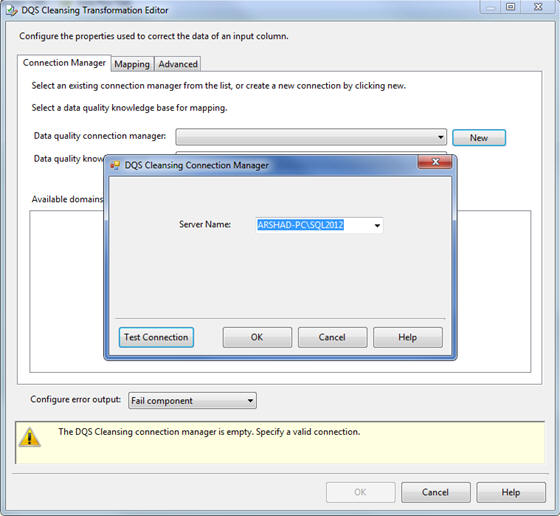
Now coming back to the data source for data cleansing, I have defined a flat file connection manager as shown below, for the above "csv" file:



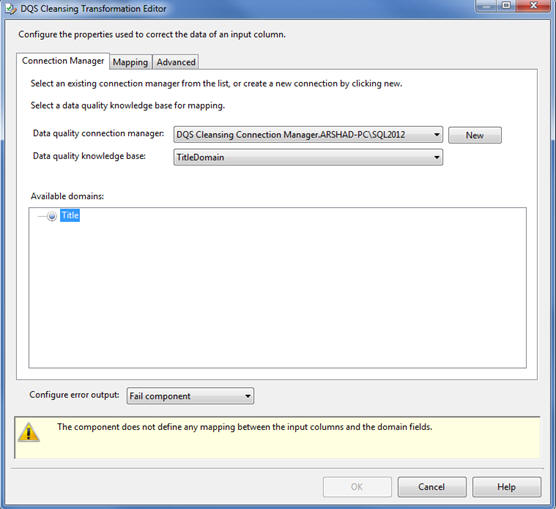
Now drag a Data Cleansing component from the Toolbox on to the data flow pane, drag the green arrow icon to the component from the flat file source and right click -> Edit on the Data Cleansing component as shown below:



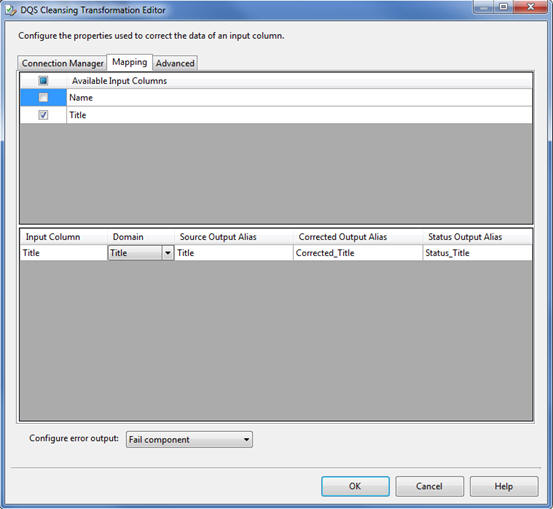
In the Connection Manager page of DQS Cleansing Transformation Editor, you need to either specify or create a new connection manager (a new type of connection manager introduced with this release to connect to DQS Server) that connects to DQS Server where you have created the knowledge base for use, you can click on the Test Connection button to verify the connection is valid:



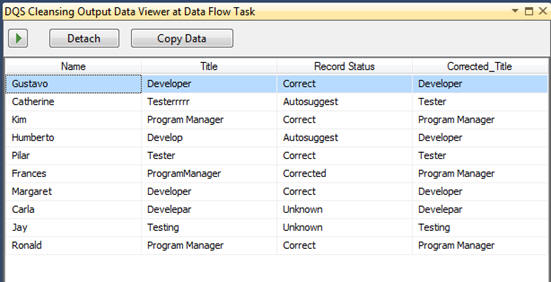
Once connected to the DQS Server, the data quality knowledge base combo-box will display all the published knowledge bases available on the DQS Server. Select the appropriate one and then go to the Mapping page:



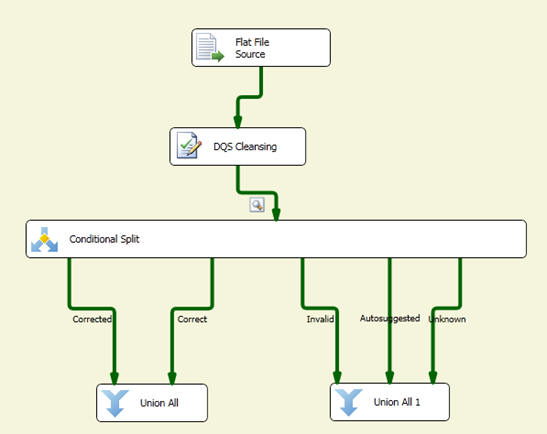
On the Mapping page, select the source/input columns to be cleansed and select the appropriate domain (will be contained in the knowledge base which you selected in last step) for each column in the bottom grid as you can see below.  I want to cleanse the Title column of the source and hence I have selected the Title domain (available in the knowledge base) to be used for cleansing the incoming titles:



The DQS Cleansing component adds a column called [Record Status] which indicates what operation has been taken by the DQS Server for each value. For example, if the value was already correct or corrected by DQS; if not then if there is any suggestion or it is invalid for the domain or an unknown value which is not in the domain and does not match with the domain rules:



You need to use a Conditional Split Transformation for splitting the output from the DQS Cleansing component based on the values of the [Record Status] column as shown below. Correct and Corrected values can be combined together (using Union All Transformation) and written directly to the destination whereas Autosuggested, Invalid or Unknown values can be combined and written to an intermediate table, so those values can be manually fixed before writing to the destination (or the knowledge base can be enhanced to include these additional values or updated for existing values so that manual fixing can be minimized):



When you run the above package, you will notice that 10 records come from the source, out of the 10 records, five are correct, one gets corrected, two are autosuggested and two values are unknown as shown below:

