helpIT systems
CLEANER DATA. BETTER DECISIONS.

What to Expect from DQS in SQL Server 2012

Will the next generation of SQL Server meet your data quality needs?

As a data technologist, you are no doubt anticipating the Data Quality Services (DQS) in the upcoming release of SQL Server (codenamed "Denali"). You are also probably wondering how these new DQS features are going to impact your approach to data cleansing.

- Will SQL Server 2012 DQS help solve the data quality issues you experience in your existing database?
- Should you be looking to build your data quality rules directly into SQL Server using just DQS, or can third party SQL Server components add real value?
- What is the quickest and most effective way to establish a Single Customer View – build using DQS, or buy?
- Will DQS be extensible enough to provide a complete data quality firewall to keep your database clean?

To help you visualize the capabilities and limitations of DQS, we present a checklist of **DQS cleansing and matching capabilities**.

Take a look...

DQS Capabilities	DQS Limitations
IMPLEMENTATION/EXTENSIBILITY	
DQS uses a Data Quality server and client to connect to your SQL data source. The server is implemented through three new SQL Server DQS Catalogs, which contain the DQS Engine, DQS Stored Procedures and DQS Knowledgebase. The client software can be installed on the same physical machine as your SQL Server installation, or setup and run from a remote computer. DQS offers a data cleansing component for SSIS. DQS can utilize third party services e.g. for address verification using USPS data.	DQS is not available via Table Valued Functions or Triggers to deal with transactional changes to your database for batched or single record updates. Once a database is clean, you need to protect it from bad data entering the database from new data feeds, including data captured via web services.
DATA STANDARDIZATION	
DQS uses domains to map specific database fields to measure and standardize data based on entries within a knowledgebase, which is built from the data that you are processing. DQS will build up a profile on each domain to reassign values according to the current knowledgebase. A DQS knowledgebase can be built from manual update or analysis of a sample of your data.	Each new data source requires manual review and expansion, development or acquisition of suitable knowledgebases to correct and standardize data within database columns. This can increase turnaround times, or worse, result in problems that are not realized until the data is used. DQS requires rigid domain mapping, but sometimes data values can be in the wrong columns and difficult to standardize fully. DQS cannot determine the type of data within each column.

DQS Capabilities

DQS Limitations

MATCHING

DQS can match data using a probability algorithm to determine similarity of complete data elements mapped to domains using a user defined matching policy. The results for the domain mapping improve as the associated reference tables or accumulated knowledgebase improves.

DQS returns similarities between individual data elements as percentages.

Through standardization, DQS can match entries with typos and casing/punctuation differences - and even more significant differences such as first name in one record compared with just initial in other records as in the example below:

Name	Email	Telephone
Mr. J. R. Smith	johnsmith@aol.com	
J.R. Smith	JohnSmith@AOL.com	(211) 456 8352
John Smith		211-456-8352
JR Smith	john.smith@aol.com	211 4568352

Another example is I.B.M., I B M, IBM and International Business Machines.

Through standardization using a knowledgebase, DQS can match quite different entries such as University of Pennsylvania and UPenn and even allow for companies that change their name or are acquired.

DQS requires the user to train the software and to improve and refine the results. New or different data sources will require analysis prior to getting effective results.

Percentages often don't correctly reflect the similarity of, or difference between, two values e.g. based on this measure, Roy Hamilton and Ros Hamilton are 91% the same but Tom Hood and Tom Good are only 86% the same - but the latter pair is more likely to be the same person. Additionally, the percentage score doesn't allow for phonetic equivalences e.g. Deighton and Dayton, or phonetic similarities such as Hannah and Hammer.

Sometimes a view of the similarity of data on a column by column basis doesn't allow for the interplay between groups of fields in two records e.g.

First Name	Last Name
Pat	Murphy
Patricia	Murphy

MAY BE THE SAME PERSON

First Name	Last Name	Suffix
Patricia	Murphy	
Pat	Murphy	Jr

ARE NOT THE SAME PERSON

Additional matching passes are required to match data at multiple levels – users often require both organization and contact level grouping, or individual level matching and householding.

Work is required to merge data and manage files and tables to create a single view prior to implementing a matching policy.

DQS Capabilities

DQS Limitations

OUTPUT

DQS allows a user to output the matching results and the survivorship results as two separate outputs.

DQS uses a manual review process prior to output to determine the effectiveness of the results.

DQS allows review of overlapping matching clusters. This allows manual amalgamation of clusters that share a common record, which typically happens when some of the records in the two clusters have different items of data missing or incomplete. Even on small databases, it is not sufficient to match on name alone, so you might get the following example:

URN	Name	Email	Telephone
101	John Smith	johnsmith@aol.com	
144	John Smith	johnsmith@aol.com	211-456-8352
298	John Smith		211-456-8352
144	John Smith	johnsmith@aol.com	211-456-8352

There are two clusters here, one containing three records with the same email address and another one containing three records with the same phone number.

Grading of matches is not sufficiently granular to allow the user to focus on a small subset of the matching pairs, which can lead to overmatching or undermatching, if ample time is not spent on the review.

Overlapping clusters arise as shown on the left, but also happen when apparently different records are bridged by a record that shares data in common with each of the other records e.g.

URN	Name	Email	Telephone
101	Juan Marcos	jmarcos487@aol.com	646-498-3055
144	Juan Marcos	jmarcos487@aol.com	211-456-8352
298	Juan Marcos	juanmarcos@gmail.com	646-498-3055
144	Juan Marcos	jmarcos487@aol.com	211-456-8352

The clusters based on email address and the clusters based on phone number have to be amalgamated manually, if the duplication is to be fully resolved.

PRODUCT SUPPORT

DQS will be available with SQL Server 2012 BI and Enterprise editions. The support is available through the normal Microsoft Software Assurance Program, MSDN and through a peer based network of web forums and user groups.

As with any Microsoft product, whether using Microsoft's own or community-based support, it is not always possible to get immediate answers or have access to someone familiar with your business rules and configurations.

So what can DQS do for you?

For data technologists working within SQL Server, the new features of DQS are a long-awaited opportunity to get a handle on data quality issues that may have previously gone unresolved. The good news is that for some of these scenarios, DQS looks to be capable of providing adequate data cleansing within certain data sets. However, while the post-launch months and years will see additional DQS data cleansing strategies made available through the broad network of SQL developers, there will always be situations in which DQS will not be the ideal solution and will even present significant challenges for the data quality tasks at hand. A quick look at the **two main data quality scenarios** faced by developers, showcases the strengths and weaknesses of DQS and when you can plan to build it yourself.

IDEAL DQS SCENARIO. *Go Ahead and Build It!*

DQ Problem: Deduplication and standardization within proprietary data sets such as product names, inventory details, other limited scope field data.

Solution: Use DQS to create data associations and equivalents and enforce appropriate standardization for data entry and new data feeds going forward.

Results: The quality of the results will be driven by the coverage and breadth of the sources you analyze, so expect this to be an iterative process which will steadily improve - but you should be able to get good results quickly.

POOR DQS SCENARIO. *Invest in a Targeted Solution.*

DQ Problem: Cleanse, standardize and deduplicate contact data including name, company, address, etc.

Solution: Repeatable rules that leverage extensive knowledgebase(s) of names data and can perform matching across columns within and across datasets.

Results: Quick, effective, out-of-the-box matching without the need to build and maintain the knowledgebase(s) critical for quality results. The higher the data volumes, the faster the return on investment.

When it comes to your data quality initiatives, the primary focus should be on finding the right product or toolset for the job by testing alternative solutions and comparing the results with what you can achieve easily using DQS. While the initial project may zero in on batch cleansing of existing data, controlling the quality of contact data as it enters the database from different sources – web leads, call centers, new business, customer moves, etc. - is **the next challenge**. It is far easier for an operator to get data right in real time at the point of entry than to review batches of corrections later.