



DQM Data Quality Manager

Technical paper Scoring process

This paper describes the design behind the scoring process used in the Cleanse and Match subsystem in DQM.

BACKGROUND

Matches conducted in DQM may result in different degree of matches. Scoring is a process that occurs after the matching step, where the users can fine tune the degree of matches by adjusting the relative weighting of these matches. The score is then used to decide which is the 'best of breed' record in the Survive process.

CONCEPTS

DQM operates on both deterministic and probabilistic matching. Records are always matched between a set - of target and source. Typically more than one column participate in the Match process. In DQM, each match column is graded with a match result, which is a score between +10 and -10.

Match result										Scoring	
10	9	8	7	6	5	4	3	2	1		+2
Exact	Lookup	Partial	Fuzzy90	Fuzzy80	Fuzzy70	Fuzzy60	Fuzzy50	Not used			
0										X	
No contest											
-1	-2	-3	-4	-5	-6	-7	-8	-9	-10		
No match											-2

The scoring process adjusts the Match result by applying a weighting scale - a value between +2 and -2.

The result of this is match score weighted by the importance of that column.

DQM will aggregate these weighted score together to form a composite match score for that record. This aggregated score is then used in the survivorship process, to select the highest quality record into the result set.

OPERATIONS

You can set the Score for each of the Match column, by adjusting a slider. This mechanism also provide you with a visual representation of the relative importance of each of the match column.



EXAMPLE

Assuming DQM is matching two customer records using full names, and addresses. (See below). Further assume that the business rule is to find customers with the same name, and living in the same suburb are very likely to be the same person. The Scoring process could be configured to have Name, Surname and Locality set to 2 or close to 2. Where street no, name etc is of little significance in this match, it could be set to zero or -0.5. In the example below, the Surname, Name, Suburb are exact or lookup match. This combined with the score setting, produced an overall score of 10.4. This score will then be compared against other match sets, and the highest score set survived into the final result.

	Name	Initial	Surname	Unit	Street	Type	Locality	State	Country	
Target	William	S	Freeman	1	Hilltop	Drive	Bestown	Sydney	Australia	10.39
Matchtype	Lookup	Partial	Exact	No contest	No match	No match	Exact	Exact	Exact	
Match result	9	8	10	0	-10	-10	10	10	10	
Source	Bill	Simon	Freeman		City	Road	Bestown	Sydney	Australia	
Scoring	2	1.5	2	0	-0.5	-0.2	1.5	0.75	1.4	

CONCLUSION

The Scoring process is an integral component of the match and survive process. It gives the users a method to fine tune the match result and places specific bias on individual matching components.