Delivering information you can trust
December 2006

IBM **Information Management** software

IBM WebSphere QualityStage: Superior technology produces superior results

Contents

High-quality data is strategic

No matter how carefully the interface is designed, when critical data pours into your information systems with every business transaction, problems can occur. Users enter nicknames, transpose digits, leave blanks, add "extra" information and misspell words. As a result, incomplete, inaccurate and inconsistent data enters your system, impacting operations, analysis and integration with downstream applications and databases.

What happens? Without a dependable customer ID, it is difficult to locate an order or account balance. Because of name variations, existing customers are recorded as new customers. In addition, you are unable to link a transaction to relevant information to leverage a marketing opportunity, and your database now contains redundant records, skewing business analysis and planning. Your organization needs information it can trust to effectively meet business goals and achieve a competitive edge.

Lack of reliable identifiers affects your understanding of suppliers, materials, parts, products and employees. Without high-quality data, you can no longer count on an accurate picture of your enterprise or the expected ROI from your critical business applications. The solution calls for a product that can automatically reengineer and match all types of customer, product and enterprise data—in batch or at the transaction level in real time. That solution is IBM® WebSphere® QualityStage™—an integral part of IBM Information Server.

Drivers for record matching and linkage

Record matching and linkage represent the ability to automatically determine with the highest accuracy possible that one new record with customer, location and/or product data is (or is not) the same as one of the millions of records in a reference file. This has long been a complicated computer science problem. A superior result—the degree to which records are matched (and not matched) when called for by the business—is a function of the software and methodology used to solve the problem.

Generally three major functional steps need to be performed on input data to produce a match result:

1. ***Strongly type:*** *Identification and standardization of all the attributes necessary to evaluate whether one record "matches" another record. Strongly typed attributes are the "fuel" that feeds a match process. The more definitive the attribute definition and the higher the percentage of legitimate business values, the better the match result.*

2. ***Block:*** *Flexibility to break the match problem down into discrete sets. Even with today's technology, it is not feasible to compare each input record to every record on a reference file. Most record linkage applications use hash or match keys that pull critical characters from the input record's attribute set to read into an index and return only those records that agree on the match key for more extensive evaluation.*

3. ***Score:*** *Rigorous comparison of the attributes associated with an input record against the attributes for each record returned in the "candidate set" where the match keys are all the same.*

Superior results with WebSphere QualityStage

This white paper examines why the WebSphere QualityStage implementation of probabilistic matching and linkage produces matching results that are consistently superior to other deterministic approaches.

IBM pioneered and continues to deliver the most flexible solution for strongly typing data—the first functional step in producing an accurate match result. In order to strongly type data, business information must be broken down into discrete pieces that can be used to determine the business meaning of all the attributes associated with an organization's master data, including customer, vendor, location, materials and parts.

The next step of the matching process creates blocking keys that provide the flexibility to break the match problem down into discrete sets. Even with parallel processing, today's technology does not support the detailed scoring of each input record to each record on a master file. So before the record-to-record scoring comparisons can begin, blocking keys are created to enable the selection of candidate sets for detailed comparison to each input record. This limits the number of record pairs being examined and increases computational efficiency.

WebSphere QualityStage performs this task by first considering only records that agree on a blocking key composed of portions of one or more variables. For example, to match individuals by location, a blocking key may take the first three characters of zip code, the first character of the street name and the first character of the last name to create a five-character block key. All records containing the same value in the blocking field are eligible for probabilistic match scoring.

Those records that do not contain the specified value can be addressed in subsequent blocking iterations. Adding further blocking variables to a blocking key will reduce the number of records for comparison, just as using a smaller block key will return larger candidate sets. WebSphere QualityStage not only incorporates optimal blocking keys in matching templates, but it also allows end users to modify the content of the blocking key and the number of blocking keys to execute a particular matching strategy.

To uncover the maximum number of matched pairs, multiple match passes (blocking iterations) may be executed. Each blocking iteration seeks to use different blocking keys to help ensure that no potentially matched pairs are omitted from the overall match process.

Scoring is the last step in the three-phase process, where records retrieved from the blocking iteration are subjected to rigorous field-by-field evaluation. A weight is calculated for each field comparison based on the statistical properties of the individual field values. All the weights for all the points of comparison are combined into a single score that represents the probability that those two records represent the same business entity.

WebSphere QualityStage uses probabilistic record linkage that determines the likelihood that two records are a true matched pair, given all observed field agreements and disagreements. When the record-matching process is executed, those record pairs with a high-match probability are retained, and those with a low-match probability are ignored or tagged for review.

The key to record matching is to set matching criteria that allow the greatest number of accurately matched pairs to be uncovered. If the matching criteria are too tightly defined, there is a risk of dropping record pairs that are, in fact, matches. Matching criteria that are too loosely defined can result in false record matches.

Critical "statistical" role of information content

Central to the probabilistic matching technique that WebSphere QualityStage employs is the calculation of a match weight based on the amount of information content contributed by each compared variable. Match weight has been statistically proven to provide the best method of discriminating between matched and unmatched pairs.

Two statistical properties of each match variable—reliability and discriminating power—determine the information content and hence the resulting match weight:

- ***Reliability:*** *Defines how reliably the data field is typically recorded. Variables with low cardinality typically have higher reliability since there is less likelihood that a value will be coded incorrectly. For example, gender (M/F) has a higher inherent reliability than Tax ID because gender has a very low cardinality and Tax ID has a very high cardinality.*
- ***Discrimination:*** *Defines how useful the match variable is to the matching process. Consequently, variables with high cardinality are far more discriminating than variables with low cardinality. When compared with reliability, therefore, discrimination functions in an inverse way. For example, gender is not a highly discriminating variable for matching records since it has a 50 percent chance of random agreement. However, Tax ID is highly discriminating because of its high cardinality and uniqueness by person or company.*

Each of these properties is automatically measured through algorithms that then allow comparison between fields to be scored relative to the amount of information content contributed to the overall match. As might be expected, rare or highly unusual field values are much more useful for matching than common values. WebSphere QualityStage is unique in that its weighted scoring process dynamically adjusts not only for variations between record fields in general, but also for individual field values to determine the precise amount of significance to assign to each agreement and disagreement.

Thus, the final score for each matched record pair reflects the relative amount of information supporting the probability that a match is true. Relative scoring calculations draw heavily on the mathematics of information theory. The WebSphere QualityStage computation of a value's information content, or entropy, follows accepted principles from the published literature of statistics and record linkage. Fortunately, the user is not required to be acquainted with these details. On the other hand, many WebSphere QualityStage customers, such as organizations from public health, justice, census and medical outcomes research, have sought out the technology precisely for its statistically justifiable basis.

The difference is probabilistic matching

What truly differentiates WebSphere QualityStage from other competitive offerings are three characteristics of its probabilistic matching process:

- ***Frequency analysis:*** *This analyzes how often a field's data values appear in a set of data. A value that occurs often within the same field (such as Smith for Last Name), will not have as much strength as a matching criterion compared to a value that appears rarely (such as Zweibel for Last Name).*
- ***Numeric match weight:*** *Competitive products that use deterministic match algorithms typically assign an alpha match grade to each pair of matching field values. Often these alpha grades are restricted to an A–F range. Conversely, WebSphere QualityStage assigns numeric match weights, allowing for far greater discrimination on the quality of the match. The higher the numeric weight assigned to a set of values, the greater the probability of a truly matched pair.*

- ***Significance:*** *WebSphere QualityStage matching algorithms assign significance to fields of data within a file record, which is critical to the success of the match process. If an organization is trying to sort records by geographic location, a match on street address may be highly significant. However, if an organization is trying to find all records of persons who have had emergency room treatment, then street address has very little significance in that match. A higher match value is assigned to fields that have more significance to the objective being sought.*

WebSphere QualityStage distinguishes itself from its competitors through its unique record-matching techniques, a task that frequently qualifies as a nondeterministic problem because of the growing business requirement for extracting high-quality information from noisy and incomplete data. However, the more significant difference between WebSphere QualityStage and other products stems from the way in which records are evaluated and successful matches selected. Most competitive offerings use a decision table methodology in which valid matches are filtered by a table of rules representing the vendor's best attempt at historical "best practices."

IBM probabilistic scoring yields more matches

The typical strategy is to rank or classify each field comparison by assigning a code (for example, A through F) that identifies the quality of the match, or "degree of closeness." The field comparisons are represented as a string or pattern of letters/numbers that collectively express how well each field matched. IBM also assesses degrees of closeness, but it uses that only as a coefficient or multiplier for adjusting the value's net informational contribution to the statistical confidence.

Competitors use the string of comparison codes generated by their field evaluations as a lookup key into the decision table. Each row of the decision table is a "rule" that specifies how to handle that particular pattern of field results. While not statistically based, this strategy can generate a fair assessment of data matches for data that is largely high quality and contains few missing values. In the interest of manageability, however, the technique compromises completeness and accuracy. A decision table representing 10 fields, where each field could have six states (A through F), would have more than 60 million rows of decision rules. This is impossible to audit or maintain for practical purposes. In practice, the decision tables must limit the number of fields being evaluated and greatly constrain the number of ranks or categories assigned to field evaluations (see Figure 1).

Figure 1: Probabilistic scoring yields more matches and less under matching

In the following household match, the deterministic pattern ABBCB is a non-match (fail), but the probabilistic cutoff score for 95% certainty is any weighted score > 21.

	L-Name	***Hse#***	***Street***	***Apt#***	***Zip***	
Rec-1	SMITH	123	BEECH	18A	02112	
Rec-2	SMITH	132	BEACH	18	02111	
Pattern	A	B	B	C	B	ABBCB
Weight	5	2	7	1	4	19

Reject

Rec-3	YUSKA	5401	VETCH	818A	02112	
Rec-4	YUSKA	5410	VEECH	81A	02111	
Pattern	A	B	B	C	B	ABBCB
Weight	7	3	8	2	4	24

Erroneous reject

Pass

Deterministic decisions tables apply the same “rule” regardless of the difference in information content; to be safe, decision tables must forgo many good matches.

But probabilistic linkage “sees” the difference between these two pairs. Rare values can compensate for missing and conflicting fields. The second pair is a good household match; the first is not.

For example, bad and missing data values can have a profound effect on the accuracy of the decision table method. Fuzzy data values and missing data typically receive a score, such as “C,” that has a positive connotation as a means of staying within the scoring pattern. However, this may mean a record with a blank receives the same score as one with match data, leading to poor or incorrect matches and potentially questionable results. With the probabilistic approach, missing values simply contribute zero weight to the composite score unless the user explicitly wishes to override the evaluation by assigning business significance (penalty or positive score) to the condition.

The decision table has no ability to distinguish rare from commonly occurring values; all comparisons receive the same treatment. The net result is that decision tables only attempt to derive an average, indicating the significance of a field's contribution to the matching assessment. Moreover, to ensure that false matches are not accepted, the decision table must reject pattern codes that on average do not have enough strength to justify matching. Using probabilistic scoring, WebSphere QualityStage is able to extract more good matches from otherwise noisy or incomplete record groups.

Dynamic weighting strategies used in a probabilistic implementation result in far fewer clerical review cases than traditional decision-table matching software. Whereas WebSphere QualityStage can treat all data consistently and account for missing or bad data, decision-table products cannot offer the same level of consistency through the record-matching process. The disadvantage of a decision table arises when business users need to customize or extend the "as-delivered" rules. Adding another field to the matching process increases the number of rules exponentially and potentially requires the manual review of the entire rule set. What was originally a straightforward implementation has now become a time-consuming, high-risk effort requiring extensive business analysis and testing.

Similar risks arise when users try to modify the decision rows of the existing vendor-supplied decision table. To save space and reduce rule volume, some vendors have eliminated rules by establishing wildcards and processing order dependencies. Thus, a hierarchy of rule precedence exists that can cause simple insertions or discrete modifications to produce unanticipated and erroneous results further downstream.

Probabilistic matching delivers more accurate results

Probabilistic matching suffers from none of these constraints. With WebSphere QualityStage, additional fields can easily be added to a match process. The more supplementary matching fields considered, the greater the statistical confidence. Decision tables cannot leverage additional matching fields, especially when the field is sparsely or inconsistently populated. Since probabilistic methodology compensates for a value's statistical properties, these secondary or marginal fields can safely add great value to increasing overall match quality results.

Probabilistic methodology is sometimes criticized because it does not provide a table of static or persistent rules that predefine the outcome of each potential record pairing. This is more of a misunderstanding than a true shortcoming. By its very nature, probabilistic record matching is data driven and therefore has the ability within user-defined limits to identify all relevant matches, not just those that comply with predefined business rules.

All things being equal, WebSphere QualityStage will produce a 3 percent to 6 percent minimum improvement in match accuracy over competing approaches. The reason that slight increases in matching accuracy matter is that each unmatched record has the potential to taint the integrity of any records to which it should have been directly matched as well as any records that are subsequently matched to these records. It is also the case that an organization's highest value customers, contacts and products are more susceptible to error because they occur more frequently. Therefore, since there can be instances where even one unmatched record is responsible for propagating significant errors, a small increase in methodological matching accuracy drives dramatic, positive benefits.

As an integral part of IBM Information Server, WebSphere QualityStage can help produce the high-quality, strategic data that offers a more accurate picture of the enterprise.

IBM Information Server delivers information you can trust

Organizations face an information challenge. Where is it? How do I get it when I need it in the form I need? What does it mean? What insight can I gain from it? Can I trust it? How do I control it? The challenges continue to grow if businesses cannot ensure that they have access to authoritative, consistent, timely and complete information.

IBM Information Server is a revolutionary new software platform that helps you derive more value from the complex, heterogeneous information spread across your systems. It enables your organization to integrate disparate data and deliver trusted information wherever and whenever needed, in line and in context, to specific people, applications, and processes. It helps business and IT personnel to collaborate to understand the meaning, structure, and content of any type of information across any sources. It provides breakthrough productivity and performance for cleansing, transforming, and moving this information consistently and securely throughout the enterprise, so it can be accessed and used in new ways to drive innovation, increase operational efficiency, and lower risk.

For more information

To learn more about WebSphere QualityStage, IBM Information Server or information integration solutions from IBM, contact your IBM marketing representative or IBM Business Partner, or visit **ibm.com**/software/data/integration

IBM Software Group
Route 100
Somers, NY 10589

Printed in the United States of America
December 2006

TAKE BACK CONTROL WITH **Information Management**

IMW11806-USEN-00