A discussion of IBM master data management offerings
January 2006

WebSphere® software

Successful MDM deployments using IBM WebSphere Customer Center and IBM WebSphere Data Integration Suite.

Contents

Introduction

This document has been created to show the complementary nature of IBM WebSphere® Customer Center™ (formerly known as DWL Customer), a Master Data Management Information Accelerator, and the IBM WebSphere Information Integration platform, which includes the IBM WebSphere Data Integration Suite, formerly from Ascential.

Both of these companies were acquired by IBM in 2005 and were former business partners. The DWL business and DWL Customer offering is now part of IBM's Enterprise Master Data Solution business segment, and the Ascential business and offerings are now part of the IBM Information Integration Solutions (IIS) business segment. With the addition of Ascential, IBM created the IBM WebSphere Information Integration platform, which includes the WebSphere Data Integration Suite and the existing IBM related products such as IBM WebSphere Information Integrator, to solve a broader set of data centric challenges present in strategic IT initiatives.

IBM WebSphere Customer Center™ (WCC) is a Master Data Management offering for managing party information, including customer, also known as Customer Data Integration Solution (CDI). CDI enables organizations to achieve their goal of becoming a more customer centric corporation by enabling the synchronization of their key party data across multiple applications and line of business channels. With WebSphere Customer Center, IBM markets the industry's leading, most robust and mature customer master data hub. It's unique set of more than 480 pre-built 'business services' enables Fortune 1000 to build a solid foundation of 'actionable' customer data for consumption by front and back-office systems in real-time or in batch. The party-centric data model provides robust support for managing customer information for both individuals and organizations.

IBM WebSphere Data Integration Suite provides a comprehensive set of data integration functions, which provide complementary functionality that enables faster deployment, greater data integrity, and on-going data monitoring of any WebSphere Customer Center investment. The suite offers capabilities centered on the ability to understand, cleanse, transform, and federate data in both batch and real-time via a service oriented architecture. The suite also includes fully integrated metadata management and a high volume GRID-enabled parallel processing environment.

Industry analysts including Gartner, 5000+ customers, system integrators and industry pundits agree, IBM's WebSphere Customer Center, and the IBM WebSphere Information Integration platform respectively are the industry's leading solutions in their field. This document summarizes the core capabilities of each product set, highlights the complementary nature of the combined solution scenarios and provides additional background on the capabilities of WebSphere Customer Center and the WebSphere Data Integration Suite.

High level overview

As described in more detail later in the document, both WebSphere Customer Center and the WebSphere Data Integration Suite are highly complementary with very little overlap between the offerings. WebSphere Data Integration Suite provides a consistent foundation for overall data quality in WCC. WebSphere Data Integration Suite provides the following core capabilities that will be utilized by WCC:

- *IBM WebSphere ProfileStage™ and AuditStage™*
    - *Data profiling and auditing*
- *IBM WebSphere QualityStage™*
    - *Data standardization*
    - *Probabilistic matching & record linkage*
    - *Survivorship*
- *IBM WebSphere DataStage®*
    - *Data Extract, Transform & Load (ETL)*
    - *Pre-packaged adapters/interfaces for major applications and databases*

- *Platform capabilities*
    - *Unlimited scalability via parallel processing*
    - *Support for GRID computing*
    - *Data integration lifecycle metadata management*

WebSphere Customer Center consists of a party-centric data model with service oriented architecture accessible business services. These services include large grain customer data processes, such as adding customers, and hundreds of fine grain services that represent customer data functions, such as adding an address for a customer. WCC is a headless solution designed to be a transaction processor with system of record capabilities by utilizing multiple interfaces for real-time, batch and middleware connectivity. WCC provides enterprise insight customer knowledge by storing and maintaining information on:

- *Parties*
- *Location*
- *Relationship/hierarchy*
- *Products/account*
- *Billing*
- *Interactions*
- *Campaigns*
- *Privacy preference*
- *Customer identity and directory*

WCC also provides business processes for:

- *Customer knowledge management*
- *Duplicate suspect processing*
- *Event notifications*
- *Rules of visibility*
- *Integrated business rules*

Data integration is critical for MDM success

As with the creation of any new solution such as Master Data Management, a Data Warehouse, or SAP application, the successful implementation of WebSphere Customer Center is directly dependent upon the success of the data integration processes that supports it. In order to reduce project risk, speed time to value, and provide maximum business benefit, the WCC data hub requires data to be initially loaded that is of the correct format, data quality, and fit-for-purpose.

The adoption of a best-practice tools-based-approach to using data integration helps to ensure the success of both the initial implementation of WCC and the ongoing data integrity of transactional updates. The IBM WebSphere Data Integration Suite provides the market-leading suite of tools for this type of project, especially when combined with a best-practices methodology (such as IBM Iterations®) that has been built-up over hundreds of similar projects.

Both WCC and WebSphere Data Integration Suite are market leaders in their own right. The ability to use WebSphere Data Integration Suite with other data-centric projects provides an increasingly compelling reason to look to use both technologies to help create the best CDI (Customer Data Integration) solution that mitigates project risk, costs and ensures maximum data quality.

From a high level architectural perspective and taking into account aspects such as enterprise standards, re-usability and shared services, it is preferable to use consistent data quality techniques and tools throughout an organization. For example, WebSphere Data Integration Suite capabilities can be used and reused across multiple projects such as data warehousing projects, mainframe migrations, and application rationalization and consolidations, as well as MDM deployments. WebSphere Data Integration Suite has capabilities such as data profiling, validation, data standardization, cleansing, de-duplication and data quality monitoring that can represent substantial time and cost savings to an organization compared to traditional, more manual approaches to these tasks.

WCC manages the processing of customer data (i.e., when to cleanse and match customer data) and WebSphere Data Integration Suite does the cleansing and matching. Customers are encouraged to find out more about the IBM Center of Excellence for Data Integration, and the best practices and methodologies available to make data integration an indispensable and repeatable asset to any IT organization.

WCC & IIS integration scenarios

There are seven primary integration scenarios ("touch points") where IBM WebSphere Customer Center can leverage the capabilities of the broader IBM WebSphere Information Integration portfolio:

1. *Profile source system data*
2. *Load WebSphere Customer Center*
3. *Synchronize WCC across applications and channels*
4. *Data quality*
5. *Provide party matching*
6. *Enable federated query across WCC and external content sources*
7. *Add new sources and WCC "targets"*

The explicit usage of the IBM Information Integration platform to enriching MDM solutions is called Master Data Integration. Please refer to the following links for more information:

http://ibm.ascential.com/solutions/master_data_management.html
http://www-306.ibm.com/software/data/masterdata/launch.htm
http://www-306.ibm.com/software/data/masterdata/customer

1. Profile source system data

Source system analysis – or data profiling – concerns the systematic assessment of source systems' data and using that information in the design of new target systems. Without a data profiling tool, it is a laborious, manual process to fully assess an application's data; instead companies rely on out-of date (or nonexistent) source system documentation or the knowledge (sometimes folklore) of the people maintaining the source systems. In all cases, the data within the system itself does not conform to the rules provided in documentation, or application metadata, or often people's knowledge of the business. But source system analysis is crucial to understanding what data is available, and understanding its current state.

With a complete analysis of your data behind you, this phase of a data integration project helps you to define exactly what you need to do to your data in order to integrate, consolidate and harmonize prior to loading into WCC. Source system analysis drives the definitions of the business rules that you need to develop in the subsequent data integration project phases.

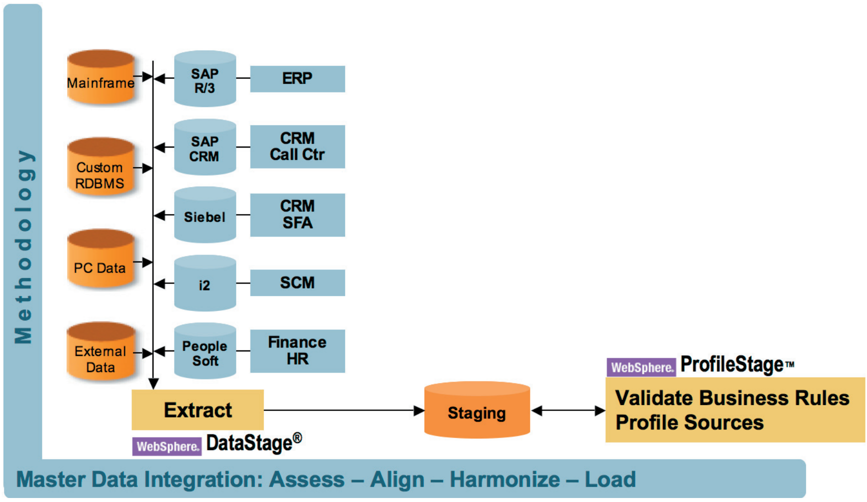


Figure 1.

2. Load the Customer Master Database

Typically all data integration projects have a requirement for extracting data from the source systems, integrating and consolidating the data, and applying transformation logic (calculations, algorithms, aggregations, lookups etc.) before loading the target system(s). As discussed in the previous section understanding source data before loading the customer master data repository is critical to a successful CDI MDM implementation. Organizations implementing WCC can either choose to undertake this work manually, or select tools such as IBM WebSphere DataStage to address these requirements. The IBM WebSphere Data Integration Suite enables organizations to implement this end-to-end data integration process with significant time, cost and risk savings.

WCC requires source data in the right format, clean and fit-for-purpose. Typically the initial data load will be undertaken from XML file(s) (WCC specific format) that satisfies the data requirements (integrated, clean, fit-for-purpose etc). The data loading is undertaken through the WCC services layer APIs, which are invoked through WebSphere DataStage, or the data can be loaded directly to the WCC schema as pictured below. The latter assumes that all the necessary data consolidation and harmonization has already occurred.

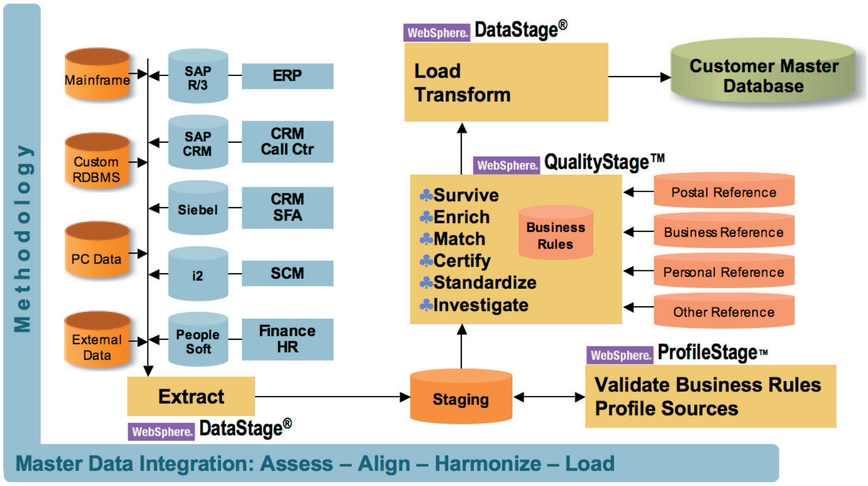


Figure 2.

Depending on the amount of data, time constraints, and size of production hardware, IBM may recommend that the initial data load into WCC bypass the WCC services layer and go directly to the WCC database. This is accomplished using the combination of IBM WebSphere DataStage and IBM WebSphere QualityStage to provide the necessary throughput and data quality required to maintain WCC's data integrity.

3. Synchronize

The WCC solution framework provides over 480 pre-built business transactions representing course and fine grained updates to the WCC data hub. Each of these transactions, which maps to the WCC data model, operates through the WCC services layer. This layer is accessible through pre and post-processing APIs using a variety of mechanisms, including a Web Service, Java code, COBOL, IBM WebSphere MQ and IBM WebSphere Business Integrator. Additionally, customers who have an investment in IBM WebSphere DataStage TX can leverage WCC's services layer directly.

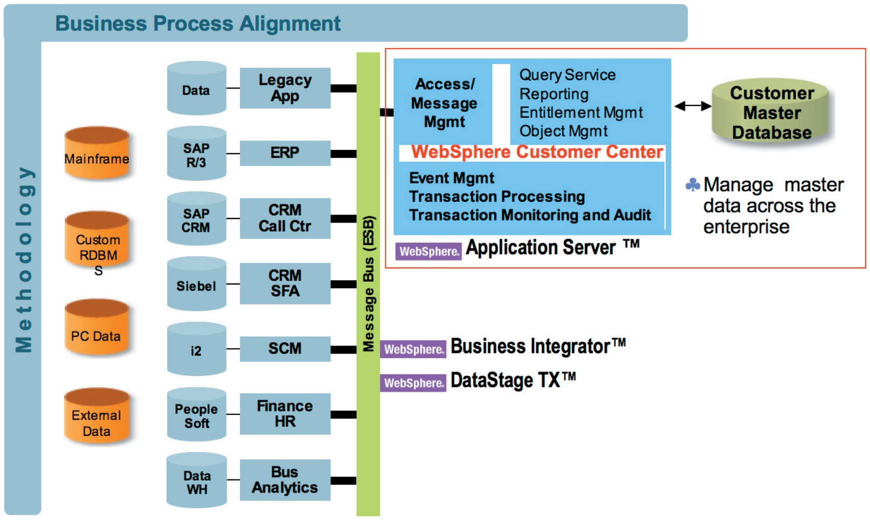


Figure 3.

4. Data quality

From an operational perspective, WCC transactions are triggered when business events occur such as the addition of a new customer record to the WCC data hub. It is imperative that before data is added to the WCC repository, the data undergoes stringent data quality processes to insure the data is of the highest quality.

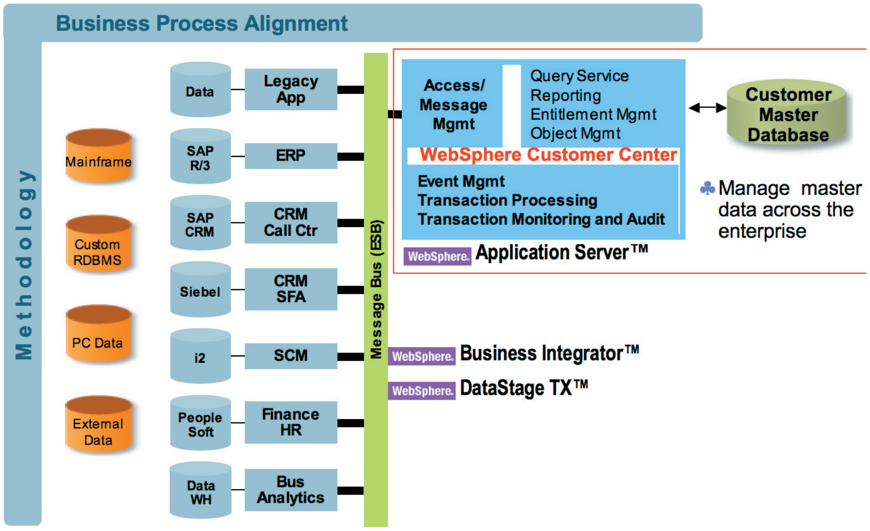


Figure 4.

Data preparation (often referred to as data cleansing) is critical to the success of any data integration project. Data quality functions typically include standardization, reconsolidation, validation, matching, de-duplication and data enrichment using third party sources of data. However, data cleansing often involves a more sophisticated analysis and processing of the data. For example, name and address cleansing requires multiple passes of the data to eliminate redundancy and impose consistency, not just record-by-record checking. In addition, today's enterprise requirements extend beyond name and address processing into personal and commercial "householding," where multiple contacts are tracked at the same location and across locations. Data cleansing beyond name and address processing – extending into customer attributes or even materials, parts and products adds considerable additional complexity. This may include matching WCC data with external data provided by third party vendors.

This processing generally breaks down into two functional layers;
1) standardization, and 2) validation:

1. *Standardization strongly types the meaning of the piece parts of an address and puts the address components into specific fixed fields within the WCC data model.*

2. *Validation determines if the location actually exists by comparing the location to a third party reference file. If the address can't be located on the reference file then it likely does not exist.*

For customers using WCC, IBM WebSphere QualityStage provides a complete standardization, validation and certification solution for global location data.

5. Party matching

One core data quality component of a CDI solution is the extent to which parties are corrected identified and linked (matched) to records already managed in the WCC repository, that are truly related. "Quality" in this context means that when a single new customer record is processed it should not be linked to another customer record incorrectly (if it's not the same person in the real world). However, if that customer is already in the repository than we should automatically identify that record as an existing customer. This critical function is handled by WCC and the use of specialized matching and record linkage technology such as IBM WebSphere QualityStage.

WCC manages the overall party matching process for operational customer data. For example, WCC will manage when to match, when to resolve, rules of survivorship for all customer data, and what to do based on certain types of matches. Within this process, WCC will rely on WebSphere QualityStage for the standardization, hygiene, and matching for the relevant customer identity data.

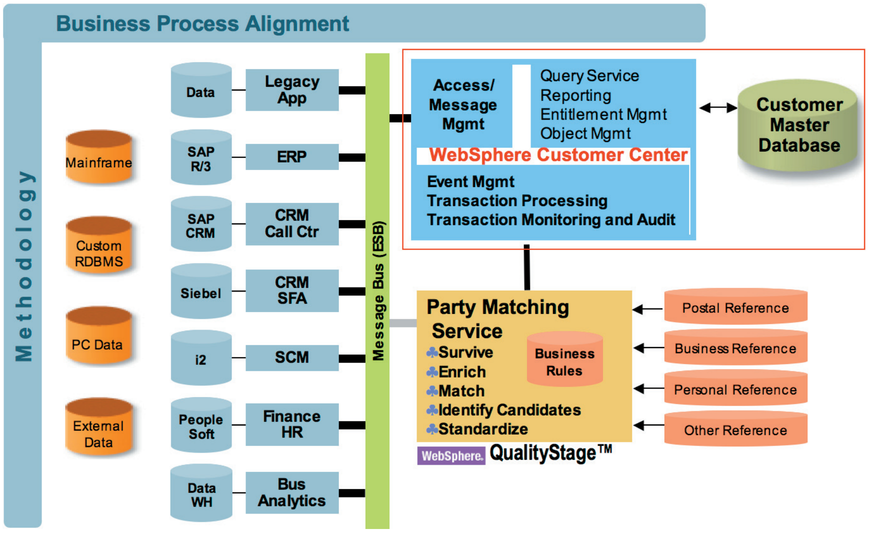


Figure 5.

6. Federated query including unstructured content

In many situations, the deployment of an MDM system motivates multiple follow-up projects. With an established reliable source of reference data the MDM deployment can be used as a basis to serve a host of new applications. For example, in the case of products, an organization might decide to extend the value of the MDM solution by supporting an e-commerce site and the creation of a print catalog. It's easy to see why having a centralized repository of product information would be the logical starting point. However in both scenarios, these solutions require a range of other on-demand information, such as product photos, real-time availability information, and affinity data like "customers also purchased these items." Access to this information outside the MDM system can be coordinated and presented simultaneously via federation. Federation provides a mechanism to query multiple sources of information simultaneously returning related information across applications, mainframe databases such as VSAM, IMS™ and DB2® Universal Database,™ and content sources such as IBM DB2 Content Manager, FileNet and Documentum. For example, pulling Robert Rich's contract history up from the data warehouse, querying for call details made to the support center and retrieving his photo from the HR system about contractors. This capability is part of the broader IBM WebSphere Information Integration portfolio.

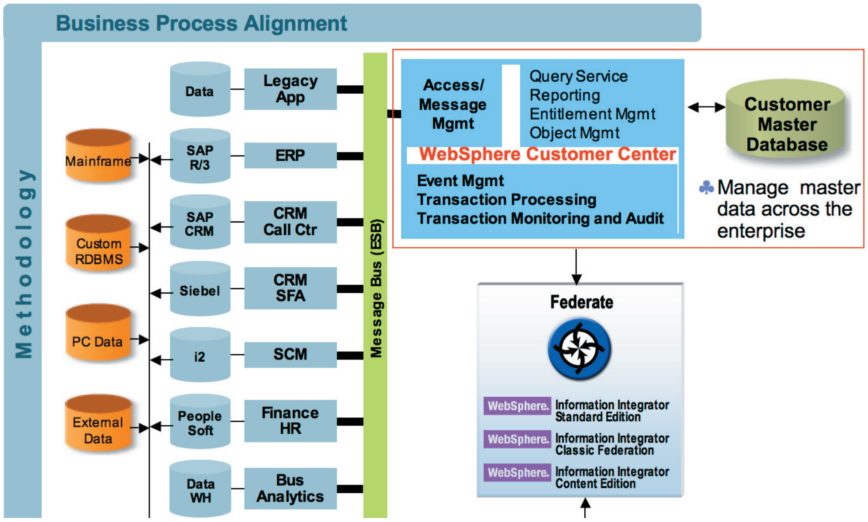


Figure 6.

7. New sources & WCC "targets"

As the task of incorporating new data sources and target data repositories are needed within the enterprise, the core capabilities of WebSphere Data Integration Suite provide the necessary tools for quickly and efficiently completing the process. This is accomplished by reducing the development time by using the reusable components, inherent capabilities, and business rules within ProfileStage, DataStage, and QualityStage. Relevant examples for WCC would be if a Retail Bank WCC operational instance needs to add a new line of business for WCC or has the need to import the WCC party view into a Basel II compliance solution.

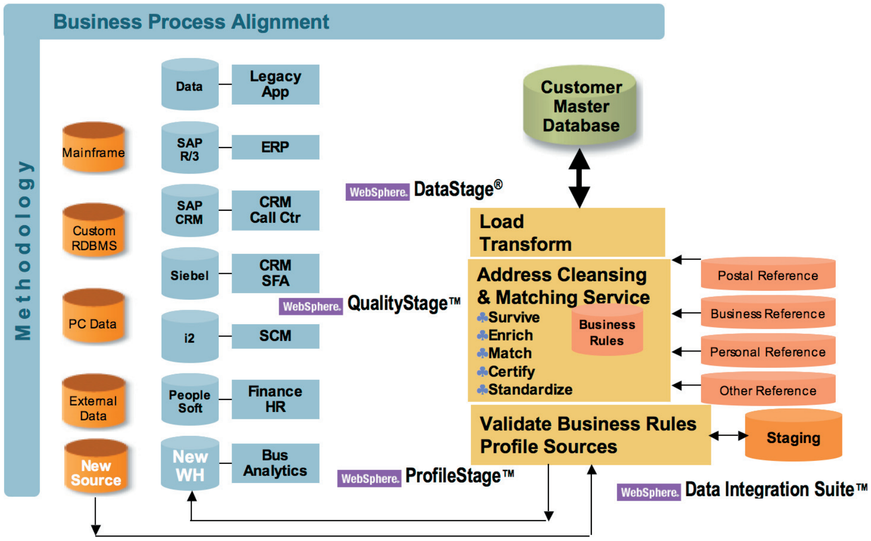


Figure 7.

WebSphere Customer Center overview

IBM WebSphere Customer Center (WCC) is an enterprise customer master data hub that provides a unified customer view and update environment to multiple channels. As an MDM solution, it aligns these front office systems with multiple back office systems in real time, providing a single source of customer truth across the enterprise. WCC's key technological strength is its "hub" or service oriented architecture. Unlike UI-driven CRM systems or product-centric back office systems that create islands of customer information, WCC is designed as a "headless" application that contains hundreds of packaged business services that may be integrated with front and back office business applications. This makes it faster and easier to integrate existing and new systems and solves the key issues impeding CRM implementations - multi-channel integration, incomplete customer knowledge and scalability. Leading industry analysts recognize DWL Customer™ (WCC) as the leading customer data integration (CDI) solution.

Here's why:

- *WCC is a real-time, transactional application developed on J2EE, EJB technology that contains more than 480 Java business services 'out of the box'.*
- *WCC operationalizes customer insight and knowledge. It maintains marketing department customer insight and injects that insight into operational processes.*
- *WCC manages 'net new' enterprise customer data and business processes. This includes privacy profiles, cross-channel interaction history, customer relationships and groupings (households), customer values, duplicate suspect processing, event notifications, among others.*
- *WCC is a fully service oriented application.*
- *WCC is a proven leader for massive performance levels and scalability.*
- *WCC derives real-time customer insight within the context of customer transactions.*
- *WCC detects customer events and provides insight on the context of that event based on the customer's profile via real-time event notifications.*
- *WCC contains sophisticated integration functionality and is neutral to all front-end CRM and back office systems.*
- *Unlike proprietary application-suite customer databases, WCC is a process-neutral CDI component that is built on open technology.*

IBM WebSphere Information Integration platform overview

The IBM WebSphere Information Integration platform enables organizations to apply a consistent and repeatable process to solve enterprise-class data integration problems across analytical, operational and transactional environments irrespective of data volume, complexity, or latency. Each of the core integration products operate as an integral part of the WebSphere Information Integration platform or as a standalone product.

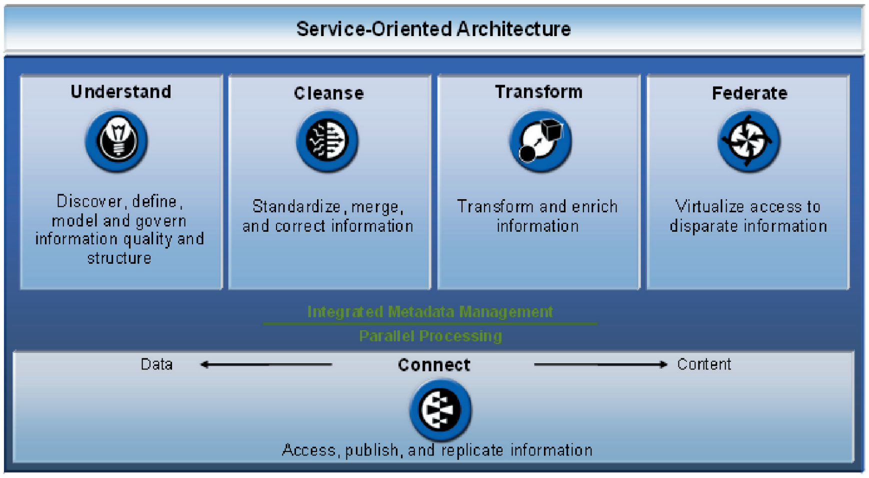

Service-Oriented Architecture
Understand
Discover, define, model and govern information quality and structure
Cleanse
Standardize, merge, and correct information
Transform
Transform and enrich information
Federate
Virtualize access to disparate information
Integrated Metadata Management
Parallel Processing
Data
Connect
Content
Access, publish, and replicate information

Figure 8. IBM WebSphere Information Integration platform.

The IBM WebSphere Information Integration platform enables businesses to perform five integration functions:

- **Connect** *to any data or content, wherever it resides*
- **Understand** *and analyze that information, including its meanings, relationships, and lineage*
- **Cleanse** *it to assure its quality and consistency*
- **Transform** *it to provide enriched and tailored information*
- **Federate** *it to make it accessible to people, processes, and applications*

Underlying these functions is a common metadata and parallel processing infrastructure that provides reuse, scalability and automation across the platform.

Each product in the platform also provides connections to many data and content sources, and the ability to deliver information through a variety of mechanisms. Additionally, these functions can be leveraged in a service oriented architecture through easily published shared services.

The breadth and flexibility of the platform enable it to address a wide range of business issues and related technology projects. The breadth of the offering optimizes the opportunities for reuse, leading to faster project cycles, better information consistency, and stronger information governance.

The IBM WebSphere Data Integration Suite
Part of the IBM WebSphere Information Integration platform, the IBM WebSphere Data Integration Suite, is the one solution that supports integration of data in all environments – transactional as well as operational and analytical – providing a solid basis for expediting transactions, streamlining operations, making optimal decisions and supporting customers. The Suite comprises the following components:

IBM WebSphere ProfileStage
Automates data profiling and source system analysis, reducing the time it takes to understand what data is available from months to weeks or even days.

IBM WebSphere QualityStage
Ensures that strategic systems deliver accurate, complete information to business users seeking a single view of customers, suppliers, and products from across their enterprise.

IBM WebSphere DataStage
Enables you to quickly and easily integrate enterprise information, regardless of the number of sources, complexity, volume, or latency of the data.

IBM WebSphere DataStage TX
Provides sophisticated EDI and data transformation and delivery for most application such as SAP, PeopleSoft and industry protocols such as HIPAA, SWIFT.

IBM WebSphere MetaStage
Provides metadata capture, sharing, management and reporting for all of the metadata in your Business Intelligence tool sets including data modelling, data profiling, ETL, data cleansing, and BI reporting. Processes defined within our framework or via our plug-ins will provide you with the data lineage and impact analysis never before achieved from any vendor!

Parallel execution
Provides unprecedented performance for all Enterprise Editions of DataStage, QualityStage, and ProfileStage tasks through massively parallel processing and distributed data base loading on SMP, clustered SMP, MPP and GRID environments.

Matching: WebSphere QualityStage vs. other approaches

Drivers for record matching and linkage
Record matching (and it's close cousin record linkage), 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 on a reference file, 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 there are three major functional pieces that 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 not feasible to compare each input record to every record on a reference file. Most record linkage applications use a "hash key" or a "match key" that pulls 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.*

For (1) above, IBM pioneered and continues to deliver the most flexible solution to break all manner of business information into discrete pieces and determine the business meaning of all the attributes associated with an organization's master data including customer, vendor, location, materials and parts. The balance of this paper discusses IBM's WebSphere QualityStage solution for record matching and linkage describing why probabilistic matching produces a result consistently superior to other "deterministic" approaches.

QualityStage's probabalistic approach to record matching

As mentioned above, detailed scoring of each input record to each record on a master file is not feasible, although the benefits of parallel processing allow more work to be done in short timeframes. Before the record-to-record scoring comparison begins, the blocking phase creates sets of "candidates" to compare in detail to each input record. This is done to limit the number of record pairs being examined and increases computational efficiency. QualityStage performs this task by first considering only records that agree on a "blocking key" composed of portions of one or more variables. For instance, to match individual at location, a blocking key may take the first 3 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 additional 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. QualityStage incorporates “optimal” blocking keys in matching templates, but also allows for end user modification of 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 ensure that no potentially matched pairs are omitted from the overall match process.

In the scoring phase, the 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.

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, then organizations risk dropping record pairs that are, in fact, matched. Matching criteria that are too loose 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% chance of random agreement. However, a Tax ID is highly discriminating due to 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 field values are much more useful for matching than common values. 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 the match is true. Relative scoring calculations draw heavily on the mathematics of information theory. WebSphere QualityStage's 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 insulated from any need to know 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.

Probabilistic matching differentiates QualityStage

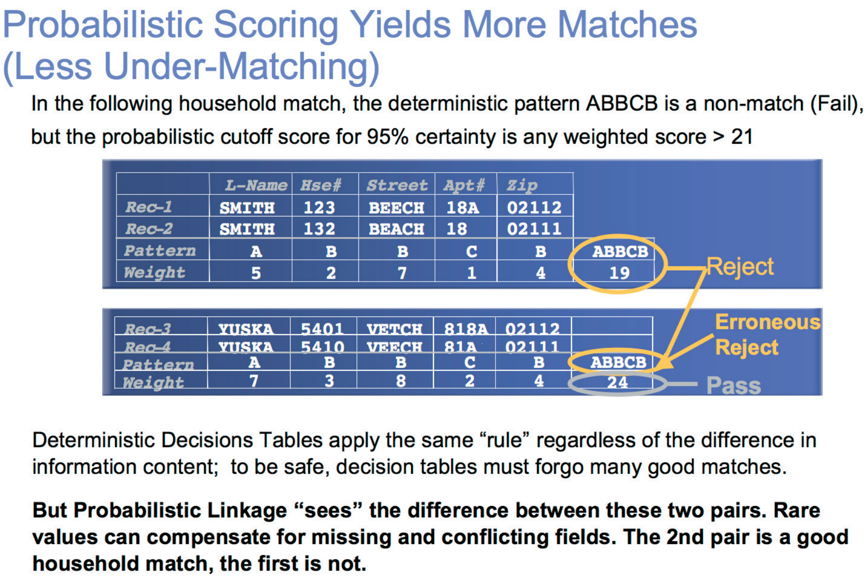
What truly differentiates WebSphere QualityStage from other competitive offerings are three characteristics of its probabilistic matching process.

- **Frequency analysis.** *Frequency analysis 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 as 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, 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 true matched pair.*
- **Significance.** *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 that have had emergency room treatment, then street address has very little significance in that match. A higher match value is assigned to fields that are of 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 due to the growing business requirement for extracting high-quality information from noisy and incomplete data. However, the more significant difference between 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 historical "best practices."

Decision tables don't work as well as setting scoring thresholds
The typical strategy is to rank or classify each field comparison by assigning a code (e.g., 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. WebSphere QualityStage 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 some 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, so 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.



Probabilistic Scoring Yields More Matches (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

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

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 2nd pair is a good household match, the first is not.

Figure 9.

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 "average" 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" don't have enough strength to justify matching. Using probabilistic scoring, QualityStage is able to extract more good matches from otherwise noisy or incomplete record groups.

Dynamic weighting strategies utilized in a probabilistic implementation result in far fewer clerical review cases than traditional decision table matching software. Where 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 an additional 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 straight forward 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 "wild-cards" 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.

Probabalistic matching consistently delivers more accurate results
Probabilistic matching suffers from neither of these constraints. With QualityStage, additional fields can easily be added to a match process. The more supplementary matching fields considered, the greater the statistical confidence. Decision tables can't 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 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, QualityStage will produce a 3% to 6% 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's 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 major positive benefits.

For more information

To learn more about Master Data Management, Master Data Integration and IBM's WebSphere Information Integration Solutions, contact your IBM marketing representative or IBM Business Partner, or visit:

IBM WebSphere Information Integration platform
http://ibm.ascential.com/products

IBM Center for Data Integration Excellence
http://ibm.ascential.com/services/cedi.html

IBM Master Data Integration
http://ibm.ascential.com/solutions/master_data_management.html

IBM Enterprise Master Data Solutions, including IBM WebSphere Customer Center and IBM WebSphere Product Center, please visit:

http://www-306.ibm.com/software/data/masterdata/launch.htm

IBM WebSphere Customer Center

IBM WebSphere Customer Center (WCC) is a Master Data Management offering for managing party information, including customer, also known as Customer Data Integration Solution (CDI). CDI enables organizations to achieve their goal of becoming a more customer-centric corporation by enabling the synchronization of their key party data across multiple applications and line of business channels. With WebSphere Customer Center, IBM markets the industry's leading, most robust and mature customer master data hub. It's unique set of more than 480 pre-built 'business services' enables Fortune 1000 companies to build a solid foundation of 'actionable' customer data for consumption by front and back-office systems, in real-time or in batch. The party-centric data model provides robust support for managing customer information for both individuals and organizations.

IBM WebSphere Information Integration

WebSphere Information Integration solutions provide a broad integration platform that integrates and transforms any data and content to deliver information you can trust for your critical business initiatives. The WebSphere Information Integration platform provides breakthrough productivity, flexibility and performance, so you, your customers and partners have the right information for running and growing your businesses. It helps you understand, cleanse and enhance information, while governing its quality to ultimately provide authoritative information. Integrated across the extended enterprise and delivered when you need it, this consistent, timely and complete information can *enrich business* processes, enable key contextual insights and inspire confident business decision-making.

IBM Software Group
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
01-06