



Tuesday, November 7, 2000

Business Intelligence: Enabling E-Business



Stay Informed: DM Review Article

► Search

[Advanced Search](#)

► Stay Informed

[Home Page](#)

[DM Review](#)

[Magazine](#)

[DM Direct Newsletter](#)

[Archive](#)

[Industry News](#)

[Online Columnists](#)

[Industry Events](#)

[Content Alert System](#)

[Career Center](#)

► Resource Portal

[Analytic Apps/DSS](#)

[CRM](#)

[Data Quality](#)

[Data Visualization](#)

[DW Design,](#)

[Methodology](#)

[DW Storage](#)

[DW Survival Kit](#)

[E-Commerce / E-Business](#)

[E-Intelligence](#)

[ERP](#)

[Supply Chain e-BI](#)

[XML](#)

[Business Intelligence](#)

[Data Marts](#)

[Data Mining](#)

[DW Engines](#)

[Data Transformation](#)

[DW Servers](#)

[\(Hardware\)](#)

[Knowledge Mgmt.](#)

Know Your Data: Data Profiling Solutions for Today's Hot Projects

By [Craig Olson](#)

*Published in DM Review
in March 2000*

Global 2000 companies have deferred strategic business system initiatives for years while they grappled with massive Y2K upgrades. With the witching hour here and gone, these companies are now turning their attention to high-priority projects such as customer relationship management (CRM), e-commerce and data warehousing.

Many innovative enterprises are discovering data profiling and mapping solutions that help solve the enormous data migration, integration and consolidation challenges these projects entail. By gaining a thorough knowledge of their data, these companies achieve faster project completion, improved data quality, reduced risk and rapid return on investment with data profiling and mapping.

CRM The New Priority

You can scarcely pick up a copy of *DM Review* or other IT periodicals without reading about CRM. But what exactly is it? CRM can be a set of applications or an application suite that supports the entire customer-facing life cycle of marketing, sales and customer support. It can also be a set of business intelligence applications supported by customer information in a data warehouse or data mart. In either case, CRM relies on accurate, complete and easily accessed data to provide a comprehensive view of an individual customer.



Recommend It!

If you found this editorial to be helpful recommend it to others.

[Cast your vote!](#)

*Currently, there are
488 vote(s) for this
editorial.*



E-mail It!

Know someone who should read this editorial?

[Send an E-mail!](#)



Print It!

A printer friendly version of this editorial for your private use.

[Print from Here!](#)

[Meta Data](#)[OLAP](#)[Query and Reporting](#)[DW in Healthcare](#)[DW in Telecom](#)**► General Resources**[Winner's Circle](#)[Vendor Categories](#)[Hall of Fame](#)[Classic Articles](#)[Product Reviews](#)[Software Demos](#)[White Paper Library](#)[Online Posters](#)[Bookstore](#)[Author Index](#)[Glossary](#)[Internet Resources](#)**► Get In Touch**[Discussion Forum](#)[Associations and Organizations](#)[Ask the Experts](#)**► Corporate Info**[Subscriptions](#)[Industry Awards](#)[Press Releases](#)[Editorial Calendar](#)[2000 Media Kit](#)[Reprints](#)[Customer Service](#)[Contact Information](#)

The META Group reports that CRM has been driven to the top of many Global 2000 agendas by a combination of the following factors:

- CRM is required for the successful deployment of supply chain management, e-commerce and other strategic applications.
- Traditional practices such as product innovation and discounting are losing their ability to sustain competitive advantage, making the personalized service supported by CRM a critical differentiator.
- The high cost of acquiring new customers increases the importance of implementing CRM practices and systems in order to retain them.
- Globalization and deregulation have increased competitors' market reach, while the Internet has increased customers' ability to do comparative analysis. As a result, it can be harder than ever to retain existing customers.

CRM projects can be enormously valuable but difficult to implement successfully. Using data from multiple sources, with diverse data formats and definitions, is a tremendous challenge. Unidentified or unresolved data quality problems will compromise the effectiveness of a CRM system and, if severe, render it unusable.

E-Commerce Integration

Reduced administrative costs, fast response time and a flexible, personalized buying experience draw buyers and sellers to Web-based e-commerce. Many early e-commerce success stories, such as Amazon.com, involved business-to-consumer (B2C) interactions; but an increasing number of Global 2000 companies are putting their business-to-business (B2B) transactions on the Web.

You may not realize it when you plan your e-commerce initiatives, but you are about to embark on a major systems integration project. A second-generation e-commerce application will exchange data with systems throughout your organization.

- Web-page content, format and links may be personalized based on customer data drawn from existing systems (previous purchases, frequent buyer programs, negotiated discounts, etc.).
- Detailed product selection pages should include data from inventory and/or ERP systems (product options and prices, stock on hand, substitute products, etc.).
- When a visitor makes a purchase, the e-commerce application may exchange information with credit card, credit approval, order entry, invoicing, sales analysis and other systems.

At each point where an e-commerce application integrates with another system, there are at least two views of the data. When the

two views have inconsistent fields, formats or definitions, the differences must be resolved before the systems can be integrated.

Most organizations rely on manual analysis to resolve data quality problems and differences between systems. This difficult and time-consuming task generally results in lengthy delays in deployment of the e-commerce system, reduction and simplification of the integration points which means that desirable features such as personalization are either eliminated or reduced in effectiveness, and integration errors and major data quality problems which anger customers and lose business.

As Ken Rudin noted in the September 1999 issue of *DM Review*, the recent interest in CRM solutions has once again put data warehouses in the limelight. CRM requires complete knowledge about your customer, but this knowledge is stored on disparate operational systems throughout your organization.

Before the advent of the data warehouse, extensive manual data compilation and analysis were necessary to bring together related information housed in different systems. It took a long time to gather and present the information, and the manual processes often introduced errors. Many potentially valuable reports and analyses could not be done because they were simply too expensive.

In recent years, companies of all sizes have developed data warehouses as a comprehensive and integrated source of business intelligence. A large data warehouse may bring together data from diverse in-house systems, business partners and other sources. The data warehouse is intended to be an easily accessible, centralized source of integrated data. Data warehouses have been built to support all sorts of business activities and analysis, including customer support, market analysis, cross-selling and virtually every aspect of business.

Data-Driven Project Challenges

Efforts to migrate, integrate or consolidate existing data for CRM systems, e-commerce applications and data warehouses often fail due to data quality and integration issues. According to a report by The Standish Group, 83 percent of data migration projects overrun their budgets or fail entirely.¹ One of the primary reasons cited was a lack of a clear understanding of the source data at the beginning of these projects.

CRM, e-commerce and data warehouse projects all face several common data problems:

- Finding the data Similar data often has different labels in different systems. The customer name, for instance, may be

called "CUSTNAME" in one system and "CUSTOMER" in another, while a third might break the name into three fields called "FNAME," "INIT" and "LNAME."

- Understanding the data Data formats and content in two different systems may be entirely different. A customer's date of birth may be formatted as "12-07-1952" in one system and "December 7, 1952" in another.
- Eliminating redundancy and quality problems in the data The Automobile Association of the U.K. recently reported that they found 12 different spellings of the word "beige" in their database. A human reading the data might know that "bage," "biege" and "beige" all mean a light brown car. But a data warehouse sorting and selecting data on color would treat these diverse spellings differently.

To identify and solve problems like these in the systems you are integrating, it's not enough to examine meta data (detailed definitions of field format and content created by the development staff). Undocumented changes and inconsistent data entry practices often render meta data obsolete or inaccurate. To identify potential problems and fully understand your data, you must examine the actual data itself. When you have compiled an accurate profile of the content, structure, quality and integrity of your existing data, you can then create the extraction, cleansing, transformation and load programs you need to migrate the data to your new database or integrate the data with your new applications.

The Solution: Data Profiling and Mapping

Unfortunately, conventional methods for analyzing real data take a great deal of time, involve only small samples of the data and fail to deliver a complete understanding of the source data. Manual or semi-automated processing techniques cannot possibly compare the thousands of attributes and millions of values necessary to uncover the relationships, dependencies, redundancies and quality problems you need to understand.

The answer is a new category of software called data profiling and mapping which offers a fast, accurate and automated way to understand your data. It enables a small, focused team of technical and business users to quickly perform the highly complex tasks necessary to achieve a thorough understanding of source data. This level of understanding cannot be achieved through conventional approaches. Colin White, the founder of DataBase Associates International, highlights this point by saying, "When significant quality problems exist in source data, data profiling should first be used to analyze and identify problem areas."

Data profiling and mapping software enables data migration and integration projects to be completed successfully the first time, eliminating extensive design rework and late-stage project

cancellations. It can even warn IT management when the business objectives of the project are not supported by the data, dramatically lowering project risk and enabling valuable resources to be re-directed to other, more fruitful projects. Finally, data profiling and mapping delivers higher data quality.

How Does It Work?

Data profiling employs sophisticated algorithms to examine the content, structure and internal relationships of legacy data sources so they can be understood in detail. It consists of three steps, with each step building on the information produced in the previous steps.

Column Profiling analyzes the values in each column or field of source data, inferring detailed characteristics for each column, including data type and size, range of values, frequency and distribution of values, null rule and uniqueness characteristics.

Dependency Profiling analyzes data across rows and infers all dependency relationships that exist between attributes within each table. It identifies primary keys and "gray-area dependencies" dependencies that are true most of the time, but not all the time usually indicating a data quality problem.

Redundancy Profiling compares data between tables, determining which columns contain overlapping or identical sets of values. It identifies attributes containing the same information but with different names (synonyms) and attributes that have the same name but different business meaning (homonyms). It also helps determine which columns are redundant and can be eliminated.

Once the data profiling process is completed, the data mapping process begins. Using the profile results, a fully normalized relational model is built based on the consolidation of all the data sources. Users can modify this model by adding structures to support new requirements or by adding indexes and denormalizing the structures to enhance performance. Source-to-target transformation maps are then created, including attribute-to-attribute flows and scrubbing and transformation requirements. These maps provide essential information to the programmers creating routines to move data from the source to the target applications.

As Global 2000 organizations turn their attention to CRM, e-commerce, data warehouses and other major new initiatives, data profiling and mapping is becoming an increasingly important solution for integration and data quality problems. This approach results in a thorough understanding of the data and a set of transformation maps that are 100 percent supported by the data. Understanding the source data means that these strategic projects can be completed on time, on budget, with less risk of failure.

¹ Includes projects with budgets of \$1 million or more. *Migrate Headaches*. January 1999.

Craig Olson, vice president of marketing for Evoke Software, has 16 years of senior marketing management experience from several successful high technology companies including a high-growth Internet company and a leading provider of customer relationship management software. Olson can be reached at colson@evokesoft.com or (415) 537-9127.

[^ top](#)



© 2000 DM Review. Thomson Financial Media [Privacy Statement](#)
COMPAQ Powered by Microsoft SQL Server 7.0 **Microsoft**