WhatWorks
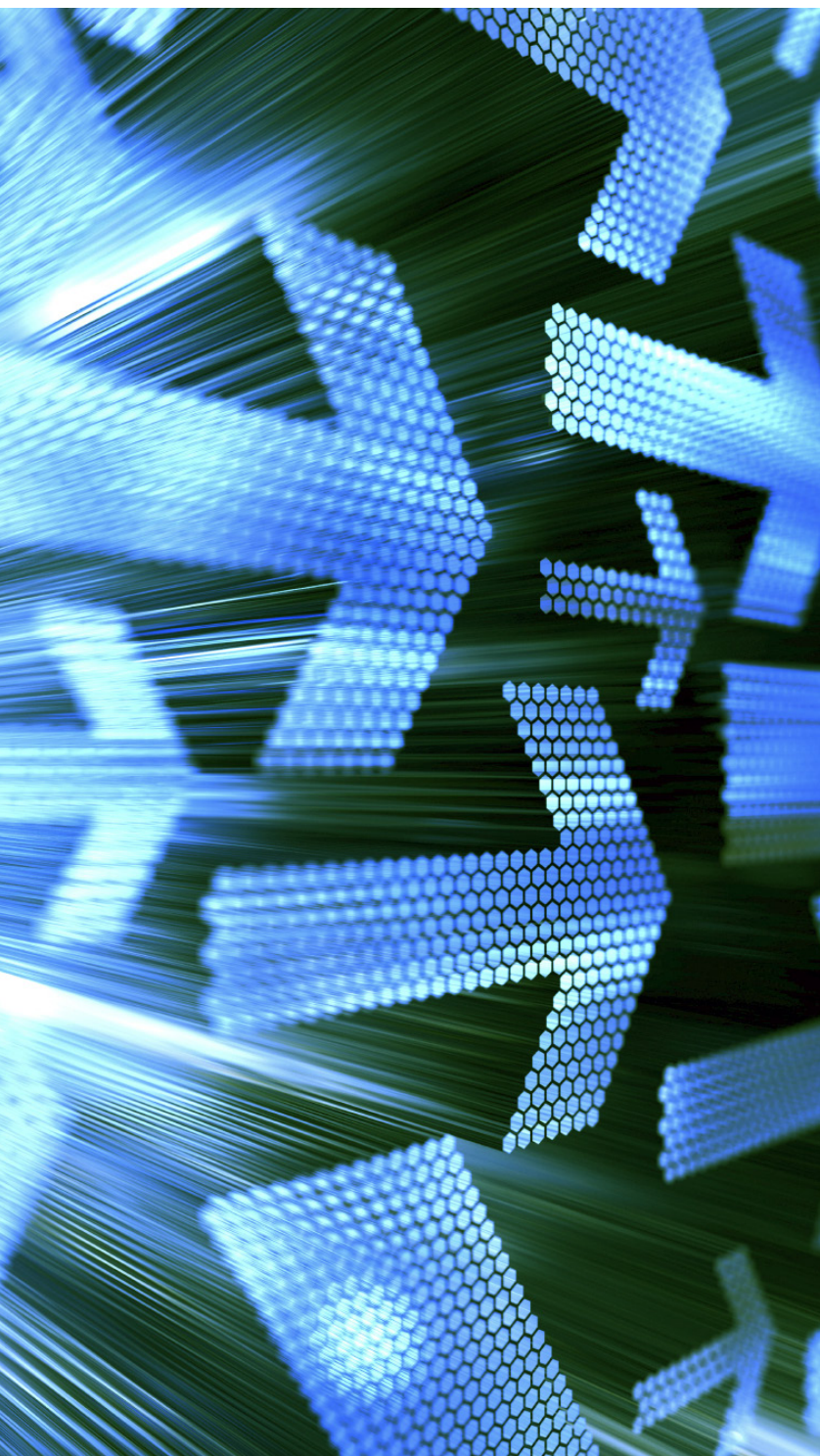
IN DATA INTEGRATION

**POWERFUL CASE STUDIES AND
LESSONS LEARNED FOCUSING ON:**

Data Governance

Data Integration

Data Quality

Master Data Management

Data Warehousing

FEATURE

Raising the Bar for Data Profiling

Philip Russom, TDWI Research
Learn about the four practice areas of data profiling and
10 best practices indicative of a mature, feature-rich
implementation.

TDWI RESEARCH EXCERPTS

Introduction to Next Generation Data
Warehouse Platforms

Introduction to Unified Data Management

tdwi
THE DATA WAREHOUSING INSTITUTE

Letter from the Editorial Director

This new edition of *What Works in Data Integration* offers a fresh, topically focused collection of customer success stories and expert perspectives. We're proud to offer this resource to enhance your understanding of the tools, technologies, and methods that are central to data integration today. We've arranged these case studies and lessons from the experts into specific categories to guide you through the articles: data governance, data integration, data quality, and master data management, as well as a special section on data warehousing.

Here's what you will find inside:

CASE STUDIES

What Works case studies are meant to present snapshots of the most innovative BI and DW implementations in the industry today. The case studies included in this volume demonstrate the power of data integration and data warehousing for industries ranging from marketing firms to automobile finance companies.

LESSONS FROM THE EXPERTS

Included in this issue of *What Works* are articles from leading experts in the services, software, and hardware vendor communities. These lessons provide perspectives about data integration and data warehousing best practices and trends.

FEATURE ARTICLES

In "Raising the Bar for Data Profiling," Philip Russom, senior manager of TDWI Research, details the four practice areas of data profiling and 10 best practices indicative of a mature, feature-rich implementation. Russom writes that technical users scrimp on data profiling by doing it rarely or shallowly, and this negligence of data profiling has engendered a body of poor practices.

TDWI RESEARCH

There's more from TDWI Research. *What Works* includes excerpts from TDWI's recent Best Practices Reports, *Next Generation Data Warehouse Platforms* and *Unified Data Management*, both by Philip Russom.

We hope you enjoy this collection of case studies, best practices, and expert insight focused on data integration and data warehousing. We look forward to your comments. If there is anything we can do to make this publication more valuable to you, please let us know. And please join me in thanking the companies that have shared their stories and successes, their technology insights, and the lessons they have learned.

Denelle Hanlon

Denelle Hanlon
Editorial Director, *What Works in Data Integration*
TDWI
dhanlon@tdwi.org

Table of Contents

Raising the Bar for Data Profiling

BY PHILIP RUSSOM
Senior Manager, TDWI Research

TDWI's position is that detailed profiling of the data in source and target systems is a prerequisite to successful data management projects for business intelligence (BI), data warehousing (DW), data integration (DI), data quality (DQ), master data management (MDM), customer data integration (CDI), and so on. The problem is that many technical users scrimp on data profiling by doing it rarely or shallowly, profiling only known systems or small pieces of them, and settling for profiles that are not very insightful or actionable.

These poor practices result in project overruns, the exclusion of important data, incomplete and inaccurate profiles, severe productivity losses, and low-quality or just-plain-wrong data in reports and applications. In other words, the negligence of data profiling—whether intentional or a product of minimal resources—has engendered a body of poor practices that hold data profiling back from making a more significant contribution to the many data management practices and related business initiatives of which it is a subset.

Hence, there's a real need to raise the bar on data profiling so that data management professionals and their business counterparts expect more and get more from it. For many organizations, raising the bar is about augmenting the current narrow definition of data profiling (generating statistics about data values in a column or table) to embrace more advanced functions for cross-system data discovery, regular data monitoring and remediation, and the collaborative improvement and sharing of profiles. To help organizations wishing to make such improvements, this article defines the four practice areas of data profiling, then lists 10 best practices indicative of a mature, feature-rich implementation.[1]

---

[1] This article is based on the TDWI Webinar "Best Practices in Data Profiling and Cross-System Data Discovery," available for replay on tdwi.org.

> ## Some folks rush through profiling, only to have to stop and do more later.
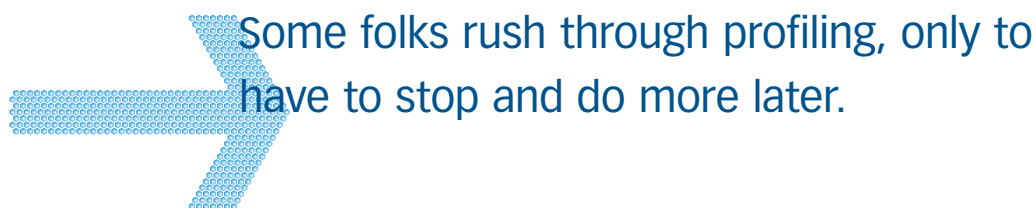
Defining Data Profiling and its Four Practice Areas

So we all start on the same page, here's a definition of data profiling adapted from Wikipedia:

> Data profiling is the process of examining the data available in an existing data source (e.g., a database or file) and collecting statistics and information about that data. The purpose of these statistics may be to find out whether existing data can easily be used for other purposes, provide metrics relevant to data quality and standards, assess the risk involved in integrating data for new applications, and assess whether metadata accurately describes the actual values in the source database.

It's convenient to have a single term and concept that everyone recognizes, like data profiling. But the single term—and the Wikipedia definition—suggest a smaller practice than is actually the case. In fact, data profiling consists of four practice areas, and for many users, raising the bar involves branching out into all four:

Data profiling. This is core functionality, focused on the development of a data inventory with actionable profiles. The greatest need for improvement here is for technical users to reach out and profile more data sources and targets, while ferreting out dependencies across them.

Data discovery. This helps you avoid bad habits, such as profiling only known sources. Through data discovery, you can find nuggets of un-mined data. And it helps you find the best source of data for a given purpose, not merely a source that works. Depending on how functions are organized within tools or solutions, data discovery may enable you to relate data structures distributed across multiple IT systems.

Data monitoring. Though profiling is done before developing a data management solution (for DI, DQ, MDM, etc.), monitoring provides a quick re-profile as data moves through the deployed solution. The point is to assess the current state of a source or target data set and keep a history of data's quality and structure (and sometimes its usage, too). From this information, a data steward or similar user can see how data's quality and content have changed since the solution last ran, as well as trends over time. Monitoring is critical to the continuous improvement expected of DQ and DW solutions. It can also spot changes to source data structures, quantify the rate of quality degradation, and classify data usage for data governance purposes.

Collaborative profiling. At the low end, data profiling is collaborative when multiple people on a technical team share and improve a common set of profiles. A better example, however, is when line-of-business people add value to the data inventory and its profiles by providing context about what the data represents and how departments use it. For example, business people can give meaning to columns of data that are poorly defined by metadata or documentation, and they are indispensable at prioritizing data sets and data structures for improvement based on business pain or need. Collaborative data profiling is a critical success factor with cross-functional and cross-system data management practices, such as BI, DW, DI, DQ, CDI, and MDM. As data governance programs deepen their policy-driven practices, they invariably seek automation for data governance from all four data profiling practices, but within the collaborative context of the data governance committee.

Ten Best Practices in Data Profiling

The following best practices raise the bar on data profiling but with realistic, achievable goals. Don't be surprised if you read about poor practices that you have perpetuated. Even seasoned data management professionals have some room for improvement as they raise the bar on data profiling and related practice areas for data monitoring, data discovery, and collaborative profiling.

Just do it! Technical users regularly voice their excuses for not profiling: "We know which sources to tap and we know their data so well that we needn't profile them." "We have a tight deadline, and data profiling is not a deliverable." Skipping data profiling puts data-intense projects at risk, because project workers don't know the data and its current state as well as they should. Inevitably, surprises pop up, such that the project concludes over time and over budget. So just do data profiling, as a prerequisite and critical success factor for projects in BI, DW, DI, DQ, MDM, data migrations, and so on.

Profile data thoroughly. Even when data management professionals conduct data profiling, they often do it superficially. A common but fallacious attitude is: "Running some ad hoc queries and SQL scripts tells us all we need to know." Some folks rush through profiling, only to have to stop and do more later. Be thorough from the start to avoid delays.

Produce more thorough data profiles. Whether hand coded or tool based, most profiling merely generates statistics about data values in one column in one table. This isn't enough. As you work, thoroughly document data sources, meanings, relationships, mappings, value types, and so on.

Discover and profile new data sources. Don't just profile known data sources. Look for those you may have forgotten, overlooked, or been unaware of. Assume that new sources are appearing and old ones are changing, else you'll miss opportunities.

Don't be content with just any source for a data point. Don't stop when a source is found, ignoring other potentially better sources. Distinguish similar data sources to ensure you've found the best source for a particular purpose (e.g., "customer" will be defined differently in different systems).

Rely on data discovery to find all the potential sources. Seek tool automation that assesses sources you're exploring and suggests data's meaning and relations.

Profile data across multiple IT systems. Don't profile only parts of tables or databases. Statistics describing the occurrence of values and data types in a table column are useful, but not broad enough. Instead, cross-system data discovery can operate on whole tables and databases, plus reach across multiple systems.

Discover cross-system keys, business rules, data mappings, data exceptions, and transformations. Documenting all these dependencies is essential for projects that are inherently cross-system, like MDM, CDI, and DQ. Cross-system data discovery is crucial with undocumented legacy systems, which are typical of database migration projects.

Map data as you discover and profile it. Don't just inventory data elements. Profiles should also catalog or develop mappings. After all, mappings are a big chunk of the work in developing DI and DQ solutions. Mapping is time-consuming and prone to error when manual. Avail yourself of any available software automation to make mappings faster and more accurate.

Re-profile data as it evolves. Data profiling is not a one-shot deal. Databases are alive, in that they grow, shrink, and evolve. For example, system owners change data structures and data definitions without telling you. And data sources come and go, as they appear, age, and retire.
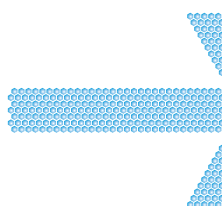
Thus, you must re-profile data sources periodically to assess their current state and discover changes in sources and targets that require adjustments to solutions for DQ, DI, MDM, BI, etc. This is true whether you are working with source databases you've handled before or you are building a target database that demands continuous improvement (as in CDI and DW projects).

Re-profile data daily via data monitoring. Data monitoring differs from profiling and discovery. Data profiling and discovery tend to be pre-development tasks that examine potential data sources and targets. Data monitoring re-profiles data touched daily as part of a deployed solution for DI, DQ, CDI, and MDM, to assure that these achieve predetermined metrics for data's quality, consistency, and validation. Monitoring also identifies structural changes in data sources and exceptions to be processed. While data profiling and discovery can be done with hand coding or mostly manual methods, data monitoring involves a high level of automation that is best executed via a vendor's tool.

Data monitoring reveals exceptions that need special handling the same way that deployed routines for DI, DQ, and MDM do. Therefore, data monitoring needs to integrate with other tools where a business or technical user processes exceptions.

Collaborate through data profiles. Poor practices limit collaboration via data profiles. "We've always just documented source-data profiles in Microsoft Offices files, which we e-mail to everyone." Documentation is an all-too-common product of manual profiling. Yet manual documentation is hard to update, share, and apply to multiple projects.

Tool automation enables collaborative profiling. Ideally, all team members (whether business or technical) should collaborate around a single view of source and target data. This requires a tool that can store and share data profiles and other valuable outputs of data discovery and monitoring via an easily accessed repository, data glossary, corporate portal, or an equivalent. You should also make profiles accessible to many users, so that both technical and business people can annotate them. This gives the profiles greater accuracy, richer metadata, and better prioritization. Reuse profiles across multiple projects for greater productivity and consistency.

Wean yourself off hand-coded and mostly manual methods in preference of tool-based data profiling.

Support many practices with data profiling, discovery, and monitoring. Let's be honest. Data management professionals make a living developing and maintaining primary solutions for BI, DW, DI, DQ, and MDM—not data profiling. Even so, primary solutions wouldn't be possible without secondary, supporting data management practices like data profiling. Equally indispensable supporting practices include metadata management, data modeling, and database administration. In fact, the quality, performance, and maintainability of primary solutions often depend on how well the supporting practices were performed—and that's probably the best reason for raising the bar on data profiling and other secondary practices.

Data management practices aside, data profiling also contributes to data-driven organizational practices, such as data stewardship and data governance. After all, these depend on developing a data glossary and quality metrics for data, which aren't possible without data profiling.[2]

Use a data profiling tool. Mostly manual methods are inhibiting data profiling. This includes ad hoc queries, SQL scripts, and visual perusals of data performed via a hodgepodge of utilities and hand coding. The result is documentation that's rarely kept up-to-date and nearly impossible to apply directly to primary data management solutions. Furthermore, a mostly manual approach to data profiling is feature poor, in that it lacks functions for discovery, monitoring, and collaboration.

Avoid these limitations by using the data profiling, discovery, and monitoring functions built into tools for DQ, DI, CDI, and MDM. Or augment these tools with a dedicated data profiling tool. Ideally, you want data profiles, data definitions, and mappings to be shareable across the multiple tools you use to create primary data management solutions. Data discovery, profiling, and monitoring need significant software automation to be sustainable and applicable for multiple users, tools, and types of data management projects. So wean yourself off hand-coded and mostly manual methods in preference of tool-based data profiling.

Recommendations

- **Move up to the next level of data profiling.** Do it more often and thoroughly, and share the results. Extend profiling to embrace related practice areas in data discovery, monitoring, and collaboration.

- **Realize that profiling is not a one-shot deal.** Data monitoring keeps profiles current and discovers changes. Remediate exceptions, don't just find and log them.

- **Eschew myopia!** Discover new data sources, and learn new things about old data.

- **Link the four data profiling practices to other practices.** After all, data profiling is a secondary practice that supports primary data management practices such as BI, DW, DI, DQ, CDI, and MDM. And profiling supports organizational practices, especially data stewardship and governance.

- **Collaborate over data profiles.** Solicit input from various technical and business people. Share profiles broadly for consistent data use and reuse.

- **Prefer tool functions over mostly manual methods.** This way, profiles are directly applicable to project deliverables. And tool automation enables you to raise the bar by profiling often, thoroughly, and collaboratively.

Philip Russom is the senior manager of research and services at The Data Warehousing Institute (TDWI), where he oversees many of TDWI's research-oriented publications, services, and events. Prior to joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research, Giga Information Group, and Hurwitz Group, as well as a contributing editor with *Intelligent Enterprise* and *DM Review* magazines. You can reach him at prussom@tdwi.org.

[2] For more information about relationships among data management practices, see the TDWI Best Practices Report *Unified Data Management*, available at tdwi.org.

Data Integration and Data Warehousing Defined

To help you make your way through the many powerful case studies and "lessons from the experts" articles in *What Works in Data Integration*, we have arranged them into specific categories: data governance, data integration, data quality, master data management, and data warehousing. What do these terms mean, and how do they apply to your organization?

DATA GOVERNANCE
page 7

Data governance is usually manifested as an executive-level data governance board, committee, or other organizational structure that creates and enforces policies and procedures for the business use and technical management of data across the organization. Common goals of data governance are to improve data's quality; remediate its inconsistencies; share it broadly; leverage its aggregate for competitive advantage; manage change relative to data usage; and comply with internal and external regulations and standards for data usage. In a nutshell, data governance is an organizational structure that oversees the broad use and usability of data as an enterprise asset.

DATA INTEGRATION
pages 8–17

Data integration (DI) is a family of techniques and best practices that repurpose data by transforming it as it's moved. ETL (extract, transform, and load) is the most common form of DI found in data warehousing. There are other techniques, including data federation, database replication, data synchronization, and so on. Solutions based on these techniques may be hand coded, based on a vendor's tool, or a mix of both. DI breaks into two broad practice areas. Analytic DI supports business intelligence (BI) and data warehousing (DW), and operational DI is applied outside BI/DW to the migration, consolidation, and synchronization of operational databases, as well as in exchanging data in a business-to-business context.

DATA QUALITY
pages 18–21

Data quality is a complex concept that encompasses many data management techniques and business quality practices, applied repeatedly over time as the state of quality evolves, to achieve levels of quality that vary per data type and seldom aspire to perfection. The most common technique is name-and-address cleansing, whereas the least common is the internationalization of data for quality purposes. Between these two extremes are numerous data quality techniques, including data standardization, verification, profiling, monitoring, matching, merging, householding, geocoding, postal standards, enrichment, and so on.

MASTER DATA MANAGEMENT
pages 22–27

Master data management is the practice of defining and maintaining consistent definitions of business entities, then sharing them via integration techniques across multiple IT systems within an enterprise and sometimes beyond to partnering companies or customers. Many technical users consider MDM to be an integration practice, enabled by integration tools and techniques for ETL, EAI, EII, and replication. When the system of record is a hub that connects many diverse systems, multiple integration technologies may be required, including newer ones like Web services and service-oriented architecture (SOA). More simply put: MDM is the practice of acquiring, improving, and sharing master data.

DATA WAREHOUSING
pages 28–31

At the highest level, designing a data warehouse involves creating, manipulating, and mapping models. These models are conceptual, logical, and physical (data) representations of the business and end-user information needs. Some models already exist in source systems and must be reverse engineered. Other models, such as those defining the data warehouse, are created from scratch. Creating a data warehouse requires designers to map data between source and target models, capturing the details of the transformation in a metadata repository. Tools that support these various modeling, mapping, and documentation activities are known as data warehouse design tools.

Data Integration, Data Quality, and Data Stewardship: Finding Common Ground Between Business and IT

By Daniel Teachey
Senior Director of Marketing, DataFlux

Businesses are discovering that their success is increasingly tied to the quality and reliability of their information. As companies collect more and more information about their customers, products, suppliers, inventory, and finances, it becomes more difficult to accurately maintain that information in a usable, logical framework.

The data management challenges facing today's businesses stem from the way IT systems have evolved. Enterprise data is frequently held in disparate applications across multiple departments and geographies. The confusion caused by this disjointed network of applications leads to poor customer service, redundant marketing campaigns, inaccurate product shipments and, ultimately, a higher cost of doing business. Add to that a sense of unrest and poor communication about data quality between departments and you're at a standstill.

This impending data disaster leads companies to wonder how to reconcile data that affects, to varying degrees, multiple departments and business units within a corporation. Are all data issues the sole domain of the infrastructure/technology management wizards in IT? What can business/finance managers and analysts do to help ensure that corporate information is high quality?

These questions point to a central theme—is data quality an IT problem or a business problem? This debate has raged for years. Proponents of IT data management argue that only technology-savvy overseers are capable of handling customer, product, and other data; only they have the knowledge and expertise to govern data round-the-clock.

This control, however, comes at a cost. IT executives often say they don't have time to manage an entire company's data and shouldn't have to clean up the business's messes.

Business executives are often eager to take control because data is essential to their day-to-day operations. Line-of-business staff members can feel frustrated that another department is pulling the strings on how quickly data can be integrated, cleansed, and amalgamated to provide reporting and analysis.

Nonetheless, the business side has a problem, too. They can't single-handedly manage enterprisewide data—that can lead to a fragmented, confused view of the organization. The systems are too complex, disconnected, or unassailable, and without IT's help, business users can't effectively manage data resources in most of their enterprise applications.

The answer is simple: both IT and business are responsible for ensuring data quality. But this answer leads to another level of complexity. If IT and business haven't been working together in the past, the best option is to bring in a referee. A referee, in this case, is known as a "data steward."

Data stewards can help a company make significant gains in managing its data. Simply put, data stewards are people trained to exclusively handle the middle ground between IT and business; they are technically savvy individuals who understand the corporate goals for the department, division, or enterprise.

The role of a data steward is to govern customer, product, supplier, and other types of data across various silos, including back-end and front-end systems across any department. They're a solution to the IT-versus-business debate because they're neutral and can see both sides of the equation.

Even savvy data stewards know the best technology in the world isn't enough to ensure smooth data management. In the end, the integrity of a company's data hinges on the integrity of its practices for collecting and managing data. Enabling an environment conducive to respect and cooperation can go a long way in effecting quality data that drives quality business decisions.

For a free white paper on this topic from DataFlux, click here and choose the title "Implementing a Data Management Methodology: A Do-It-Yourself Guide for High-Value Data Across the Enterprise."

Atlantic Detroit Diesel-Allison Enables Real-Time BI to Double Service Revenues Using CDC and Operational Data Replication

Commentary by Will Ortiz
IT Director, Atlantic Detroit Diesel-Allison

Looking at revenue growth opportunities, Atlantic Detroit Diesel-Allison, a leading Eastern U.S. distributor and service provider of engines, automatic transmissions, and parts, put a corporate initiative in place to empower its service department to process more orders, drive more revenue, and increase customer satisfaction.

Real-Time Information Is Key

To achieve these objectives, Atlantic DDA had to make real-time information available to service representatives so they could improve efficiency and relay the most up-to-date information to customers (e.g., when customers could expect to receive the equipment they have in for repair). In addition, having access to real-time information would enable up-to-date operational reporting and support current information in Web applications.

The Challenge

Making real-time information available proved to be a long-time challenge for Atlantic DDA. It would have required integrating and synchronizing heterogeneous data between the company's iSeries system and Oracle-based BI and Portal applications.

The Right Criteria

Atlantic DDA evaluated different replication solutions, looking for one that could meet its demanding time-to-market and budget requirements. Specifically, the replication solution needed to have low impact on the source iSeries system and support the het-erogeneous environment. The change data capture (CDC) technology also had to enable processing data changes in real time and on a continuous basis, which would put an end to the daily full bulk load integration the company had previously been tackling. In turn, the technology must help alleviate this time-consuming process and the strain on financial and labor resources. Finally, the solution would need to pass a "stress test," which included throughput measurements handling thousands of tables and more than a terabyte of data.

> "In addition to doubling our revenues with the ODR solution, we've also been pleased to improve our reporting to support business operations."
>
> **Will Ortiz**
> IT Director, Atlantic Detroit Diesel-Allison

Solution: ODR with CDC Technology for DB2/400

Atlantic DDA selected Attunity's Operational Data Replication (ODR) solution, which employs log-based CDC technology that monitors their DB2/400 database journals and captures only the changes made to the required tables. By using log-based CDC technology and offloading the changes as soon as they are captured, the ODR solution accomplishes minimal impact while simulta-neously scaling to meet the high throughput of changes in near real time. Atlantic DDA determined that the results garnered with the solution, combined with its price performance, made it the right choice.

Bottom-Line Benefits

The real-time information made available by the ODR solution delivered quick and tangible benefits. Atlantic DDA is now able to realize significant savings in cost and time, and reap the results from overall increases in process efficiencies.

The specific bottom-line benefits that please the company most include:

- Enabling the service department to double its order processing capabilities

- Doubling the revenue associated with increased repair orders

- Improved reporting and business intelligence for optimal operational efficiencies

- Increased customer satisfaction through improved responsiveness

For a free white paper on this topic from Attunity, click here and choose the title "Efficient and Real-Time Data Integration with Change Data Capture."

LESSON FROM THE EXPERTS

Just CDC It! Five Use Cases for Change Data Capture

By Itamar Ankorion
Director of Marketing, Attunity

Over the last few years, we've observed two clear and present trends: data volumes are growing rapidly and data latencies are quickly shrinking as users expect fresh and accurate data to be available in near real time, where and when needed.

As these trends evolve, many companies are realizing that their traditional bulk processing approach is doomed to fail. The solution of choice is to change the processing paradigm and only work with data that actually changed, which in many cases is a fraction of the source data (e.g., 5 percent). This paradigm is based on change data capture (CDC), a technology that reduces costs and enables improvement of data timeliness, quality, and consistency. The following applications can benefit from CDC, typically used in conjunction with ETL tools.

ETL for Data Warehousing

The most common case for CDC is in loading data warehouses, where processing changes can dramatically reduce load time, required resources (e.g., CPUs, memory), and associated costs (e.g., software licenses). In many cases, daily changes represent a fraction of the total data volume, so CDC has a big impact on efficiency and provides a solution for the continued and accelerating growth in data volumes.

CDC also enables processing small batches of data at higher frequencies (e.g., every hour, every minute, continuously), thus supporting lower delivery latencies for real-time data warehousing. Finally, CDC makes downtime and batch windows shrink or disappear, mitigating the risk of failure in long-running ETL jobs.

Slowly Changing Dimensions

Any data warehouse team needs to address slowly changing dimensions (SCD), which requires identifying the records and attributes that changed. For large dimension tables, this is a demanding and inefficient process, typically done by joining staging and production tables.

CDC delivers the changes to records and attributes right out of the box, reducing SCD processing time and enabling it to run more often, thus improving timeliness and accuracy.

> CDC reduces costs and improves data timeliness, quality, and consistency.

Data Replication for BI

As reporting and BI become more pervasive in supporting daily operations, more users require access to timely information from their production systems. A common solution is to offload production data to a secondary database that is then used by operational reporting applications.

CDC enables the replication of changes made to production tables with low latency and low impact on the source databases, and can be used with existing ETL tools to avoid the need to purchase expensive replication software.

Master Data Management

Key objectives of any MDM initiative are to improve and ensure the quality and consistency of master data, whether stored in a single repository or distributed across many. This requires timely responses to master data changes.

CDC makes it possible to capture and process master data changes efficiently and quickly, so quality and consistency can be ensured.

Data Quality

Improving data quality in source systems has become a common requirement, typically implemented by periodically scanning the data. By capturing and processing only changes, CDC enables ETL jobs (used to clean data) to run more efficiently and frequently. As a result, errors can be corrected faster, improving decisions and operations.

Conclusion

These use cases, and others that CDC enables, demonstrate the strategic nature of this technology. Mature solutions exist today, including independent CDC software that is not limited to a specific vendor and can support many tools, enabling one-time capture and use anywhere. For all these reasons and more, it's time to "just CDC it!"

For a free white paper on this topic from Attunity, click underline here and choose the title "Efficient and Real-Time Data Integration with Change Data Capture."

Epsilon Improves Client-Facing Data Services

Talend Integration Suite Builds a Demographic, Compiled List File from Multiple Formats

Commentary by Aaron Dix
Senior Technical Manager, Data Engineering, Epsilon

Leading Marketing Services Firm

Founded in 1969, Epsilon is the industry's leading marketing services firm. Ad Age ranks it the number one U.S. marketing services firm and the number one U.S. direct marketing agency. Epsilon combines the power of the world's most extensive collections of consumer and business data with world-class creative and proven techniques to maximize marketing success for clients worldwide.

Epsilon's problem was typical for a company that aggregates data for its clients. Data arrived in many different formats and required tedious hand coding. The company needed to streamline the integration process.

Interfacing with Legacy Systems

Epsilon had prior experience with proprietary tools and didn't want to experience the restrictions of a closed solution for its projects. "We wanted a solution that was more formally focused on data integration," said Aaron Dix, senior technical manager of the data engineering group.

Though cost was a consideration in looking for a new solution, Epsilon found that of the products they tested, Talend was one of the easiest to use. Because previous programs were written in Java or Perl, the developers were already comfortable with the technology. "Talend outperformed some of the other products we were testing," said Dix. "However, the determining factor was that the project involved legacy code that we needed to integrate into our build solutions. With its ease of use, Talend allows us to easily interface with external processes. Basically, we've overlaid a lot of legacy technology with Talend and it's much easier to maintain."

> "Our project involved legacy code that we needed to integrate into our build solutions. With its ease of use, Talend allows us to easily interface with external processes."
>
> **Aaron Dix**
> Senior Technical Manager, Data Engineering, Epsilon

Building a Demographic Database

Epsilon's project entailed building a compiled list containing geographic and demographic data, including age, ethnicity, occupation, and income—appending more than 800 attributes. This data is sourced from more than a dozen different formats and loaded into the database after processing through Talend. "It's a very large database," explained Dix, "containing around 430 million records.

"Initially the data is imported in flat files that are usually fixed or delimited—ASCII, DOS, or UNIX," he continued. "We actually use the Alterian Integrated Marketing Platform on the back end. Talend handles the integration aspect in between. Typically, Talend also performs file retrieval; it's got a very efficient set of components that integrates well into what we're doing.

"The code generation approach is also a plus. We can verify a job by looking at the generated code and then adapt it to our needs. We can also write code that is then invoked from within our Talend processes. By adding user routines, we create components or import them from Talend's extensive library. And we can reuse roughly 80 percent of it just by calling the routine we developed."

From Talend Open Studio to Talend Integration Suite

Epsilon invested in Talend Integration Suite. "It was well worth the investment," Dix said. Beyond value-added features for larger projects, the Talend Integration Suite subscription also includes technical support and IP indemnification.

"If you work on many different systems, even for testing, the product is very efficient," said Dix. "Instead of manually exporting your code over to many different systems, Talend Integration Suite lets you launch that code and test it on different systems from a single repository. It also facilitates reusability and makes teamwork pretty seamless."

Talend Integration Suite is also in use elsewhere in the company. "We have a sandbox for ad hoc projects," explained Dix. "Talend lets us take external data formats and quickly create a database where we can parse and analyze data."

For a free white paper on this topic from Talend, click here and choose the title "The Top 10 Reasons for Choosing Open Source Data Integration."

LESSON FROM THE EXPERTS

Dispelling Open Source Myths and Prejudice

By Yves de Montcheuil
VP of Marketing, Talend

Open source software, no matter its category, is still fighting prejudice in the business environment, despite an influx of successful adopters who support its worthiness. What is more surprising is that these misconceptions contradict everything that open source software stands for, and its true benefits are yet to be recognized.

Many companies are considering open source solutions as they hand over large amounts of money, year after year, on traditional software solutions. Open source is clearly making its way into the mainstream, offering the same power and reliability as expensive, proprietary software at dramatically reduced costs.

Lessons have shown that customers tend to use open source data integration software for smaller projects at first, quickly become convinced that open source is the way to go, and then roll out additional projects. Though IT still fights misconceptions that open source software is insecure and carries many hidden costs, these myths are still unproven and unfounded.

Licensing and security. The old myths that tout open source as software developed by anonymous communities that cannot be trusted is obsolete. In reality, the bulk of the software is developed by a team that leverages open source as their platform for development and is engaged to check the quality of any piece of software before it gets integrated. This practice ensures that all development complies with security and licensing requirements and the company's objective, and that the vendor is accountable for the software.

Flexibility. There is also apprehension that open source software isn't flexible and scalable enough to support established standards. This could not be further from the truth. Open source software proves to be generally more flexible than proprietary software. The community acts as a giant beta test and each developer has a different system configuration. If a function or interface is missing, experienced programmers can add it, thanks to the open architecture of the community. Hundreds of additional inputs verify that open source is a valuable method for advancing technology.

Hidden costs. There is a misconception that open source software is not really free for businesses due to hidden costs, including hidden administration fees and downtime caused by a lack of support.

Of course, nothing is entirely free. While no licensing costs will apply, the deployment and maintenance of any software must be performed by IT staff, and the hardware must be provisioned. However, these costs are still significantly lower than proprietary solutions. Even with the commercial version of open source data integration software, 80 percent savings are not uncommon. The licensing costs of existing solutions can also be reduced substantially through the complementary use of software, so that overall IT costs can be reduced on an ad hoc basis.

One thing is clear: The myths that have evolved are due to a lack of understanding and the fact that open source software is still new to a number of industries. Despite this, there is validation to support the statement that open source is now on par with proprietary software.

To remain competitive, it is paramount that IT managers consider open source software when reviewing new technologies. There are many positive case studies on open source implementations that highlight the benefits companies are realizing, and it is helpful to review and comprehend them. Breaking down the barriers to open source adoption will play a vital role in the development and future of the technology.

For a free white paper on this topic from Talend, click here and choose the title "The Top 10 Reasons for Choosing Open Source Data Integration."

CASE STUDY

Clickstream Analytics in a Competitive World

By Pete Benesh
Product Marketing Manager, Syncsort

Clickstream analytics is a powerful tool for businesses with an online presence. For online retailers, however, it's more than a tool—it's a requirement. In this $60 billion market, one giant has combined extreme data integration with clickstream analytics, further securing its position at the top.

The Need

This leading online retailer (LOR) makes more than $1 billion in revenue annually through its online channels. To meet its revenue goals, the company must obtain new business partners interested in selling inventory through its site. LOR must also quickly transform clickstream data into analytics that support strategies for planning, budgeting, determining product selections and offerings, and developing campaigns aimed at increasing Web site traffic. Key to this is maintaining a rich data inventory of customer-centric information, transactions, and Web site activity.

Using Omniture's SiteCatalyst to perform Web analytics, LOR can effectively report on Web site visits, page views, and conversions. But as with all SaaS Web analytics products, this page-centric reporting cannot segment Web site activity based on individual user sessions. This type of segmentation would allow LOR to connect multiple sessions over time to individual users, and link all of a user's online activities with activities from various operational systems, including the customer support and provisioning systems.

Responding to this need, the company developed its own clickstream data warehouse to augment the reporting available with Omniture's SiteCatalyst. Each day, Omniture delivers the previous day's complete page-tag data to the customer. These daily Omniture files range between 18 and 30 GB and require cleansing and processing before they can be loaded into the company's Teradata enterprise data warehouse (EDW).

The Challenge

Initially, LOR was using lengthy, complex, and custom shell scripts, AWK scripts, Perl scripts and UNIX utilities to parse, cleanse, and structure the Omniture data prior to loading into the EDW. Execution of these scripts caused performance problems.

In addition, the scripts were tedious and cumbersome to develop, modify, and maintain. The inflexible scripts prevented the company from enabling new business partners with unique business rules or processes in a timely manner. And because a single developer created all of the scripts—and was the only employee who could maintain or extend them—an unacceptable level of risk was created.

> "With DMExpress, our business managers reliably receive their deliverables on a timely basis, and we are confident they'll continue to do so well into the future regardless of data growth."
>
> IT Supervisor, Leading Online Retailer (LOR)

With the added pressure of rising data growth, LOR recognized the need to improve its file-based processing methodologies. "We were processing 30 to 60 GB of raw clickstream data each day," said an IT supervisor at LOR. "But our systems and processes were reaching their resource limits. Since our management was projecting significant year-over-year growth well into the next decade, we needed to shore up our ability to scale and grow."

The Solution

LOR sought a product to rapidly transform raw clickstream data and prepare it for EDW deployment. The company participated in a Syncsort proof of concept (POC), which was hosted within LOR's infrastructure and simulated actual production environments and data sets. Shortly after, LOR selected Syncsort's DMExpress Clickstream Data Integration Solution to pull the Web analytics files, perform complex transformations, and quickly pass the data off to the Teradata load utility. The solution allowed LOR to replace its custom shell scripts and successfully decrease processing time by 65 percent. The business eliminated its reliance on a single developer for maintaining and extending the code, and improved its ability to incorporate new business partners.

The company affirms that its new development environment is more agile, and the clickstream processes now complete faster and more efficiently than before the implementation. More important, as the IT supervisor at LOR remarked, "With DMExpress, our business managers reliably receive their deliverables on a timely basis, and we are confident they'll continue to do so well into the future regardless of data growth."

For a free white paper on this topic from Syncsort, click here and choose the title "Addressing the Destructive Business Impact of Data Performance Problems: Nondisruptive Strategies for Eliminating Performance Problems in Existing Data Integration Environments."

Weathering the Perfect Storm

By Pete Benesh
Product Marketing Manager, Syncsort

The Perfect Storm

Growing data volumes, shrinking operational windows, and the demand for low-latency data have created the perfect storm for data integration (DI) and data warehousing (DW) environments. In the face of this storm, IT departments fight a constant battle to provide a single, consistent, and current version of the truth readily accessible throughout the enterprise. This organizational pursuit supports one goal: maintaining a competitive advantage.

The battle often begins with the implementation of an enterprise DI strategy, which can include data warehouses, data marts, DI technologies, metadata management capabilities, a metadata repository, data quality, and profiling capabilities. Organizations commit significant capital and resources to build an environment customized for their specific data formats, sources, business relationships, processes, and service-level agreements. But even environments implemented using best practices and modern technologies feel the pressure of the perfect storm. When this pressure is compounded with the fact that most DI software is not sufficiently optimized for performance, it causes bottlenecks within the infrastructure that can have severe consequences to the business—lost revenue opportunities, increased costs, impaired decision making, and customer attrition.

Although DI bottlenecks manifest in different ways and in different architectural environments, they are all characterized by an inability to transform, integrate, or move data in a resource-efficient and scalable way within required timeframes. Some common scenarios that breed bottlenecks include:

- Building and maintaining large enterprise data warehouses

- Integrating data from many disparate sources into a single semantic format

- Daily data loads that support mission-critical enterprise business applications

- Application modernization/migration initiatives

- Clickstream data warehousing initiatives for Web analytics

- Various other scenarios that involve processing large data volumes in short operational windows

Rethink Data Integration

It is crucial to minimize or eliminate DI performance bottlenecks before they impact the business. The key is finding a solution that is optimized to function within existing DI environments, complementing an existing environment while solving data performance problems.

> Most important, to effectively eliminate bottlenecks caused by the perfect storm, the solution must also provide extreme DI performance.

Ideally, the solution should be implemented in such a way that is resource efficient, does not require significant investment in additional software and hardware, and is not disruptive to the company's existing DI environment. For this reason, traditional approaches such as acquiring additional software licenses, investing in extra hardware, or developing custom-coded tools are not the answer. Similarly, ripping out and replacing an expensive enterprisewide system is also not viable.

The solution should replace existing custom code and require short end-to-end implementation times. Most important, to effectively eliminate bottlenecks caused by the perfect storm, the solution must also provide extreme DI performance, with the ability to:

- Process massive data sets in the shortest elapsed time

- Leverage a minimum resource footprint on commodity hardware

- Dynamically optimize processing based on runtime data and system resource availability

An extreme performance DI solution that complements an existing DI environment will transform decision making in increasingly complex business situations and provide immense measurable value by:

- Enabling new and increased revenue streams

- Eliminating revenue losses caused by processing bottlenecks

- Reducing operating costs

- Delivering rapid time-to-value

- Extracting more value from existing infrastructure investments

- Eliminating costs associated with custom-coded solutions

Conclusion

Data performance problems occur within even the most carefully planned and well-executed DI environments. Highly efficient software optimized for performance within existing environments can eliminate these bottlenecks in a resource-efficient and scalable manner, without requiring costly rip-and-replace endeavors, large investments in additional hardware and software, or custom programming techniques. By easing the cost and effort of delivering rapid information to decision makers, the right solution will enable customers to effectively pursue strategic initiatives around revenue growth, customer acquisition and retention, cost reduction, and operational efficiency.

For a free white paper on this topic from Syncsort, click here and choose the title "Addressing the Destructive Business Impact of Data Performance Problems: Nondisruptive Strategies for Eliminating Performance Problems in Existing Data Integration Environments."

CASE STUDY

Pharmaceutical Company Finds the Remedy to Boost BI Process Performance

Automated End-to-End Scheduling Provides "Shot in the Arm" for Better ETL

By Thomas Hooker
Manager, Product Marketing, Cisco

THE SOLUTION AT A GLANCE

Industry	Pharmaceuticals
Company Size	Midsize (fewer than 2,000 employees)
Business Challenge	Supporting growth in number of products
IT Challenge	Increasing database sizes and workloads; free staff from manual activities to focus on strategic initiatives to support growth; scale BI to support growth without scaling up manual activities
Solution	Standardized scheduling automation of BI processes to manage coordination between disparate systems; provide a single pane of glass to view all jobs; and reduce dependency on IT staff to monitor and manage the process

Business Challenge

With business growth, IT is challenged to support dramatic increases in data processing. New and expanded product lines create more databases; increases in sales generate more data to process; and more transactions require increased individual intervention. In addition to workload increases, new strategic business initiatives need to be addressed by IT staff, and this requires freeing up valuable staff time. Doing this while successfully managing a larger workload involves sophisticated and intelligent automation in the data center.

A midsize, U.S.-based pharmaceutical company provides a good example. It sells a wide variety of both branded and specialty generic products, and the business is expanding. The company needed to increase its financial data warehousing to monitor direct sales and financial performance, and to closely monitor customer performance against contracts. Weekly reports are sent to a field organization of more than 800 representatives.

The key challenge was to ensure that IT could support the data center requirements for a growing business with a rising number of products. The company had to free staff resources to address the many new activities and processes necessary to support sustained growth.

Barriers to Success

IT depended too heavily on manually executing BI processes to support planned growth rates. Staff spent too much time manually conducting data loading, acquisition, and processing. These activities were not the ideal use of the staff's time, so the business had to eliminate this intensive manual intervention. Success depended on standardizing a scheduling automation solution that coordinated all processes from a single point of control.

Although some automated scheduling was in place, it was not freeing up staff time. The scheduling automation was a hodgepodge of application-specific schedulers, as well as a number of application-specific analytic applications. There was no interaction, however, so these disparate schedulers were riddled with production problems that required constant staff intervention.

Scheduled jobs were time dependent, and downstream schedulers were unaware of upstream problems. Also, different applications would not work together in a single job stream. IT staff members were always busy doing things manually, or individually monitoring other activities, and could not take on additional tasks. A standardized scheduling solution with a single view and single point of control over all the running processes would dramatically alleviate this problem.

Automating the Process

Central automation was the key to organizing and controlling these disparate processes in one coherent system. Not only would it provide automated solutions to many of the problems that required staff intervention, but it could also significantly shorten the time it took many of the processes to run. A centralized automation solution would eliminate all the "wait states," allowing processing of more event-driven activities and delivering vital information faster. When problems did arise, time to resolution was dramatically reduced, keeping critical processes on track.

The business quickly became comfortable with Tidal Enterprise Scheduler from Cisco. Referenced users told the company that although support was exceptional, the solution's ease of use was so strong that they seldom needed support.

Business Results

The company predicts it will soon shave 20 percent off the time it takes to run jobs that include multiple jumps. It expects to free up six to eight hours of performing manual activities. However, the main benefit is enhanced visibility into operational stability as more processes are built into the system. Ultimately, by standardizing on an enterprise scheduling solution that covers a broad range of systems, applications, and other technology categories, it will save the pharmaceutical company time and help it shift resources to more strategic endeavors.

For a free white paper on this topic from Cisco, click here and choose the title "BI and ETL Process Management Pain Points: A Look at the Most Pressing Pain Points, and Strategies for Addressing Them."

ETL and the Batch Window Challenge

A Practical Approach to BI Process Efficiency

By Derek Evan
Solutions Architect, Cisco

Despite the rise in real-time operational business intelligence (BI), the majority of BI implementations still depend heavily on batch processes. Creating an efficient process is key in meeting the challenge of timely information delivery within ever-shrinking batch windows.

With potentially hundreds of data sources, large volumes of data, and complex dependencies in the ETL process, the number of points where delays or errors can occur is many. Even when delays and errors are not factors, large variations in job running times cause a breakdown of processes based on traditional time-based scheduling. These variations lead to the inclusion of longer than needed "wait" times between steps, and an inability to complete processing in the available window.

In addition, significant manual effort is required to "babysit" processes, monitoring every step to ensure it completes in the manner expected, managing process step dependencies, and rescheduling or manually triggering steps when the process doesn't follow previously scheduled timelines. This is true even when using a packaged ETL tool; although these tools do a good job of managing the steps within the ETL process, they are unable to coordinate the ETL process with predecessor jobs in other applications, or with reporting solutions on the tail end.

Centralized and standardized event-based job scheduling eliminates several of these issues. A job scheduling solution that connects to different parts of the

BI and ETL environment as well as other applications and data sources outside of the environment can be used to build an automated process that eliminates wait states, manages complex dependencies, and sends alerts when things don't happen as expected.

For instance, consider the arrival of a file on an FTP server that needs to trigger a job in the ERP system. On completion, the ERP job needs to trigger an ETL workflow that has other conditions that must be met. The data loaded by the ETL is then used to build a cube, upon which reporting and interactive analysis is performed. This entire process can be completely automated by an enterprise distributed job scheduler that connects to all of the required business process components: FTP, ERP, ETL, and BI. No manual intervention is required to move from one part of the process to the next.

In addition, centralized job scheduling also helps manage resources. For instance, it can prevent two CPU-intensive jobs (with different dependencies) from running concurrently and creating resource contention, which could lead to job failure or running too long.

Standardizing on a cross-platform, cross-application job scheduler to manage the ETL and BI processes not only makes the process quicker and less error prone, but it also provides a single console from which to view and monitor the progress of the process, and frees up BI operational staff to focus on strategic tasks to help support business needs and increase the value of BI for the business.

Look for a job scheduling solution that is easy to use and has wide platform coverage and built-in integration for your chosen BI tools, and a vendor that has experience setting up BI and ETL integration to get the maximum value in the least amount of time.

For a free white paper on this topic from Cisco, click here and choose the title "BI and ETL Process Management Pain Points: A Look at the Most Pressing Pain Points, and Strategies for Addressing Them."

DIRECTV Improves Customer Satisfaction and Reduces Churn with Real-Time Data Integration Solution

Commentary by Jack Gustafson
Director of Data Warehouse, DIRECTV

DIRECTV is a leading worldwide provider of satellite-based television services. With annual revenue of $21.5 billion, the company serves approximately 18.5 million U.S. customers and more than 6.5 million Latin American customers. In the U.S., DIRECTV ranked higher in customer satisfaction than cable nine years running (based on data from the 2001–2009 American Customer Satisfaction Index).

DIRECTV's Challenges
DIRECTV needed to maintain high-quality customer care for its millions of satellite television customers to support growth and prevent customer churn in a highly competitive market. This included improving overall customer service levels by optimizing field technician routes for new installations and service calls.

DIRECTV also wanted to reduce fraud by proactively alerting field service teams to avoid truck rolls to new customers who are potentially fraudulent. In addition, the company wanted to attract new customers through improved execution of targeted marketing campaigns.

DIRECTV was challenged to satisfy the changing needs of its customer service agents averaging more than 600,000 customer calls per day, and had to handle large data volumes (10-plus terabytes).

Solution
DIRECTV implemented Oracle GoldenGate to consolidate disparate data marts into a central warehouse, which improved the timeliness, granularity, and accuracy of customer and service data available from its operational Oracle Siebel CRM system.

This solution enabled managers and more than 15,000 call center agents to conduct real-time data queries and analysis throughout the day—using dashboards, e-mail delivery, and end-user reporting tools—eliminating the reliance on outdated weekly or monthly reports.

> "By consolidating disparate data marts into a central, real-time data warehouse, we have been able to improve the quality of actionable data available to all decision makers."
>
> **Jack Gustafson**
> Director of Data Warehouse, DIRECTV

The primary benefit of the availability of real-time data using Oracle GoldenGate was decreased customer churn, which enabled agents to identify and rectify customer-related issues (for example, customers planning to switch to competitors). The solution opened access to critical data, such as reports on satellite service, and installation routes to 11 service and installation partners.

"In the fiercely competitive television service industry, customer retention is vital. Having the right information at the right time can make the difference between keeping a customer or letting him or her churn," said Jack Gustafson, director of data warehousing, DIRECTV. "By consolidating disparate data marts into a central, real-time data warehouse, we have been able to improve the quality of actionable data available to all decision makers—from executives to call center staff."

DIRECTV also increased the self-service functionality of its business intelligence (BI) system, cutting response times and reducing ongoing BI team requests from 100 to 30 or 40.

The Oracle GoldenGate product solution allows DIRECTV to address its changing requirements and future business opportunities with a scalable data integration platform.

Conclusion
Today, more than 5,000 customers are utilizing Oracle data integration technology products, including Oracle GoldenGate, Oracle Data Integrator Enterprise Edition, and Oracle Data Quality. These products provide a complete, open, and integrated foundation for real-time data warehousing, operational reporting, and BI solutions. Customers include DIRECTV, Overstock.com, J.Crew, Hyundai, Nestle, Ross, Verizon, Sabre Holdings, RIM, Verisign, New York Independent System Operator, and more.

For a free white paper on this topic from Oracle, click here and choose the title "Real-Time Data Integration for Data Warehousing and Operational Business Intelligence."

Accelerating Your Business with Real-Time Data Integration

By Dain Hansen and Irem Radzik
Directors of Product Marketing, Data Integration, Oracle

Batch windows are disappearing amidst 24/7 business operations and fast data growth—this poses a major challenge that can be solved with real-time data integration.

Today's most innovative companies recognize that integrating information effectively is key to accelerating the business. Those organizations seek opportunities to capture and respond to business events faster and more rigorously. Real-time data integration is ideal for maximizing value from an organization's information assets.

The good news is that real-time data integration solutions are readily available, proven in their benefit, and present a lower barrier to entry than ever before. To fast-track the adoption and advantages obtained from real-time data integration, organizations must determine not just where it will bring business value, but also the criteria for choosing the best technology option.
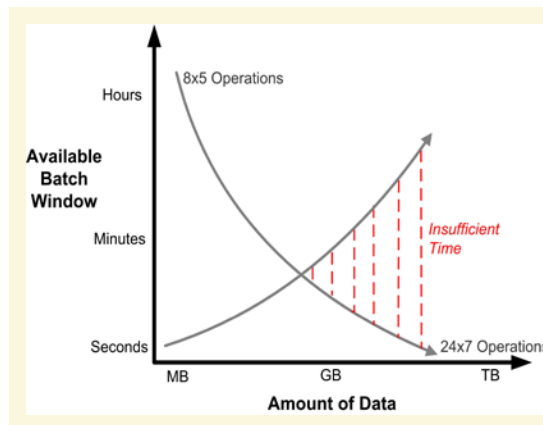
Value Drivers

Real-time data integration presents significant opportunities to help the business grow and excel, but it is not without architectural considerations when applied to data warehousing, business intelligence (BI), and service-oriented architectures (SOA).

Next-generation data warehousing and operational BI: More companies are empowering front-line employees with improved operational decision support through the use of fresh, relevant data. Here, day-old data is not acceptable. The data must be complete, accurate, and trusted—and now more than ever, timely. Operational BI is best supported by real-time data integration, typically fed to an operational data store or a data warehouse.

In today's 24/7 enterprise, traditional data warehousing solutions face a major challenge: moving very large and growing data volumes within shrinking batch windows. Real-time data integration supports a new era for data utilization, by streaming operational data to the warehouse without batch windows while enabling timely data.

Support for SOA: Real-time data integration can contribute to more successful SOA initiatives, providing the heavy lifting of data movement without burdening the business process execution engines. This helps lower costs (by consolidating legacy systems), reduce the risks of outdated data, and shorten the time to deliver new service offerings.

Selecting the Right Technology

Many real-time data integration solution offerings may sound similar on the surface. To make the best choice, there are five important aspects that should be closely evaluated:

1. **Performance.** The optimal data integration solution is capable of keeping up with the volume of changed data at the determined latency. This could mean moving thousands of transaction operations per second. Spikes during peak times should be handled with little to no additional latency.

2. **Impact.** Continuous, real-time changed data feeds should not impact the performance of source systems. Look for technologies that move only the necessary data (for example, only committed transactions) and offer features such as filtering

and compression to further minimize bandwidth requirements.

3. **Flexibility.** Beyond support for heterogeneous systems, the technology should be easy to implement, manage, and scale as needs change, and should support myriad topologies. For example, adding new data sources and/or targets should be straightforward and not require major overhauls.

4. **Data integrity and recoverability.** The real-time data integration solution should be durable and easily recover from human errors and unexpected interruptions, such as network issues, without losing or corrupting transactional data. During the evaluation period, test recoverability given several failure scenarios.

5. **Integrated and complete.** Look for solutions that combine the core elements of data integration—real-time and bulk-data movement, data synchronization, data quality, and data services—to ensure your investment serves the whole enterprise with timely and trusted information.

Conclusion

Real-time data integration helps to make organizations more profitable and competitive. Technologies that support a comprehensive and flexible approach to data integration deliver timely, accessible, and high-quality data. Companies should focus on solutions with high performance, low impact, and reliability to support today's 24/7 business.

For a free white paper on this topic from Oracle, click here and choose the title "Real-Time Data Integration for Data Warehousing and Operational Business Intelligence."

SMARTech Taps Total Data Quality Platform to Consolidate and Build Consistent Data Across Data Sets

Commentary by Jeff Averbeck
CEO, SMARTech

SMARTech is a leading provider of complex hosting, network services, and advanced Internet applications for enterprises that outsource their IT infrastructure. The company, whose parent firm is AirNet, provides a wide array of services, including Internet connectivity, disaster and data recovery services, audio and video streaming, VOIP services, Web hosting, co-location, and broadcast e-mail services.

For years, SMARTech/AirNet relied on applications developed by external vendors for their data hygiene, enhancement, and matching of data.

"Our previous vendor of data quality tools was a large software provider for whom their data quality tools were just another product off the shelf," said Jeff Averbeck, president and CEO of SMARTech/AirNet. "Integration within our proprietary applications—and even use of their product—was only done by select consultants … there were limited resources and a shallow knowledge base for the tools. Further compounding the problem was the instability within the ownership of the tools, the product having been purchased by three different companies (with different visions for the product) over the course of a few years."

The company's reliance on external vendors proved to be a challenge when implementing its smartCommunicator® product. smartCommunicator is the award-winning communication portal that integrates disparate data sources into a data-centric hub that customers can use for consumer and constituent engagement. The portal features several modules, including:

- **smartMailer** allows users to manage and send millions of e-mail messages per hour with a 95 percent or better deliverability rate

- **smartAnalyser** provides real-time analytics to determine ROI for marketing tactics

- **smartData** automatically cleans, standardizes, enhances, matches, and provides change-of-address updates to client data through the upload process

The main challenge was integrating disparate data sources into a data-centric hub for the purposes of consumer and constituent engagement, Averbeck said.

It was essential to utilize a data quality solution to help SMARTech/AirNet gain a single, accurate, and organized view of their customers by consolidating data across different data sources.

SMARTech chose to implement Melissa Data's Total Data Quality Integration Toolkit (TDQ-IT). TDQ-IT is a complete enterprise platform that works within the SQL Server's Integration Services (SSIS) data flow to deliver a wide range of data integration, transformation, cleansing, and enrichment functionality. TDQ-IT resides at the hub of the import tool for the smartData module within the smartCommunicator application.

"Name, address, geocoder, phone, and e-mail verification components, or transforms, are all put into action as each individual data upload is scrubbed and merged into a

single, uniform data set, creating a single anchor record per individual from which our customers operate multimillion-record mass communications strategies," Averbeck said. "The toolkit's ease of integration has also expanded our ability to provide customer-specific, data-based applications, ensuring that all data entering the applications is uniform, standardized, and cleansed."

Integrating TDQ-IT into the organization's workflow was seamless. "Rather than spending countless sums of money on training and specialized staff, it has been a breeze getting staffers who are already familiar with SQL's tools up and running with the new toolkit," he said.

Utilizing TDQ-IT proved to be a successful venture for SMARTech/AirNet. "Processing data within existing SSIS jobs, rather than relying on unstable Web service/command prompt calls, has increased our productivity and greatly stabilized the application as a whole … TDQ-IT's existence within the SQL Server Integration Services suite has enabled the company to build robust, stable ETL tools—in many cases, customizing data transforms to the specific needs of our clients," Averbeck said.

For a free white paper on this topic from Melissa Data, click here and choose the title "Six Steps to Managing Data Quality with SQL Server Integration Services."

Reversing the High Costs of Poor Quality Information

By Larry P. English
Information Impact International

What is the best way to get out of the financial crisis most organizations are in today? Lay off staff? File for bankruptcy? The answer is neither.

The best course of action is to eliminate the costs and wastes of process failure and information scrap and rework caused by broken information processes. World-class organizations are eliminating these costs and wastes as a regular part of their routine. Some organizations do not know how to measure and quantify the costs of poor quality information. But now there is help.

Measuring the Costs of Poor Quality Information

To develop the business case for investing in information—not data—quality management, you must be able to accurately measure and demonstrate the costs of poor quality information to your executive leadership team (ELT) in a way that cannot be rejected.

Chapter 6 in my new book, *Information Quality Applied: Best Practices for Business Information, Processes, and Systems,* describes an easy-to-follow, step-by-step approach on how to measure the costs of process failure and information scrap and rework (IS&R) that will get the attention of your ELT.

The short list of steps is as follows:

1. Identify key business performance indicators that your executives watch to measure enterprise success. You must understand what they deem important to success. These include mission accomplishment, increased profits, reduced costs, reduced cycle time, increased customer and employee satisfaction, and increased productivity.

Beware that some key indicators, such as excess bonuses or speed of delivery over quality, cause defective information.

2. Document critical business processes that have high failure rates and costs of recovery. Measure the costs of various processes and estimate the percent of all total processes.

3. Document the *official* costs of people time and other resources that are wasted in process failure and IS&R developed by the financial department.

4. Measure and calculate the direct costs of poor quality as a result of this failure. Use a time-and-materials cost worksheet and have employees measure the amount of time in recovering from process failure, workarounds, and IS&R.

5. Measure opportunity costs (missed and lost opportunity) if you know your current customer lifetime value (CLTV). Define CLTV if not. Analyze customer complaints by percent of complaints with IQ problems. Measure the attrition of customers with IQ problems. Calculate the lost customer lifetime value over the official lifetime of relationships with customers.

6. Once you have values for direct and opportunity costs, calculate the overall costs of poor information quality.

Note: Always use the official costing figures from finance and any procedures for estimating figures of waste not directly measured.

Chapter 6 in my book documents tips and techniques to give you an accurate measurement of the true costs of poor quality information, and chapter 7 describes how you can eliminate the costs of poor quality by identifying the root causes of broken information processes to discover improvements to prevent recurrence of the defects.

Measuring the ROI of Information Process Improvements

Once you have done this, follow process P3.7 in chapter 6 to measure the return on investment (ROI) of the information process improvements. If you follow these simple step-by-step guidelines, you will get the attention of your ELT to begin the culture transformation to a high-IQ enterprise that will become world class!

Melissa Data is proud to sponsor this informative article.

For a free white paper on this topic from Melissa Data, click here and choose the title "Six Steps to Managing Data Quality with SQL Server Integration Services."

Larry P. English is president and principal of Information Impact International, Inc. He is a highly respected authority in applying proven quality management principles, processes, and techniques to information processes. Featured as one of the "21 Voices for the 21st Century" in the ASQ's January 2000 issue of *Quality Progress*, he has pioneered information quality practices in his TIQM™ Quality System for Total Information Quality Management. English was recognized for his two decades of contributions in information quality management by MIT's Information Quality Management Program in July 2009. He has provided consulting and education services in more than 40 countries. The Heartbeat of America, hosted by William Shatner, awarded English the "Keeping America Strong Award" in a nationally televised news magazine program in July 2009 and can be viewed at www.infoimpact.com.

Sabre Takes off with Improved Data Quality

Global Travel Solutions Provider Improves Reliability and Integrity of Its Airport Data Intelligence Solution

Commentary by Darren Rickey
Vice President of Solutions Management, AirVision Marketing and Planning, Sabre Airline Solutions

Sabre Holdings Corporation is a Texas-based company with approximately 9,000 employees worldwide. As a world leader in the travel marketplace, Sabre Holdings offers travel products, distribution, and technology solutions for the travel industry. For example, hundreds of customers depend on Sabre Airport Data Intelligence as an accurate online source of demand and schedule information. Airports and other travel-related entities discover new revenue opportunities by analyzing its advanced booking information.

"We created Sabre Airport Data Intelligence to help clients in the travel industry to confidently study trends over a seven-year period," said Darren Rickey, vice president of solutions management, AirVision Marketing and Planning, Sabre Airline Solutions. "For example, if an airline wants to open a new route, it will first study the origin and destination (O&D) data to understand the opportunities. Airports use it to perform the same type of new route analysis an airline will, as well as analyze traffic patterns, predict flight capacities, and plan connections."

Sabre populates Airport Data Intelligence with a global database of O&D data that includes millions of rows of passenger information every month. The value of this information to Sabre, its clients, and the entire travel industry is inestimable. Ultimately, the information in this database impacts the routes, schedules, and even the fares that travelers depend on when they make a trip.

"We run seven to eight million rows per month through the database and commonly do internal audit trending analysis with 25 months worth of data," explained Rickey.

Sabre devotes a tremendous amount of time and effort to ensuring data quality. Its data quality efforts consisted primarily of focused, targeted analyses based on individual client needs. To automate its data quality efforts and deliver greater accuracy and convenience to customers, Sabre purchased Information Builders' iWay Data Quality Center (DQC). This unique software solution is designed not only to evaluate, monitor, and manage data quality in many types of information systems, but also to prevent incorrect data from entering these systems in the first place. It includes the Data Profiler, which conducts comprehensive analysis on the data to ensure its integrity and accuracy.

The iWay DQC solution uses bundled administration applications and does not require any external tools or third-party applications. The product ships with ready-to-use data quality modules for creating profiles, reports, and business tasks. Due to the breadth and connectivity of the iWay adapter family, it is easy to integrate iWay DQC with any existing IT infrastructure.

"iWay allows us to take a proactive approach to data quality," Rickey said. "The objective is to identify and correct discrepancies as the data is loaded so it has been scrubbed, audited, and certified as accurate before customers ever see it. This way, when clients contact us, it's not to tell us about problems, but to recommend enhancements. iWay Data Quality Center can cleanse many months of data in a fraction of the time of manually issuing SQL queries. One analysis that formerly took 10 hours was completed by iWay in one minute. iWay is smart enough and flexible enough to either find and fix or find and alert us to

In the airline industry, data quality is paramount, which is why Sabre Holdings is addressing this issue at the source, preventing invalid data from ever entering the system.

issues, and it can self-correct problems so it doesn't hold up a load procedure."

Because iWay DQC automatically audits its quality control processes, Sabre can share these reports with clients to boost their confidence in the reliability of the Airport Intelligence database. The software helps Sabre create audits and business rules to resolve data discrepancies such as incorrect airport codes, missing revenue figures, and inconsistent or incorrect flight capacity numbers.

"Clients often ask us to gauge the accuracy of the data, and now we can share iWay audit reports with them," Rickey concluded. "The security that stems from good data quality pays for itself. Our investment in iWay demonstrates our ongoing commitment in this area."

For a free white paper on this topic from iWay Software, click here and choose the title "Optimizing Data Quality in the Enterprise: How to Tackle Your Bad Information."

LESSON FROM THE EXPERTS

The Need for Proactive Data Quality Management

By John "JT" Taylor
Chief Technology Officer, iWay Software

Today's business transactions are highly complex, generating information in multiple ways and causing it to flow into and throughout an organization at a rapid pace. While this has optimized efficiency, it has also created major data quality problems.

Organizations are finally taking data quality seriously, implementing policies and tools to correct the invalid information that plagues their enterprise systems. However, most data quality initiatives take a reactive approach, when in fact only proactive quality control will fully ensure the integrity of all data, at all times.

Where It Starts

There's no single cause of data quality problems. Although manual data entry used to be a key contributor, new inbound information channels such as automated business-to-business (B2B) interactions and Web portals play a major role today. These next-generation sources deliver information that is more sophisticated, yet harder to harness, and only a real-time "data quality firewall" can mitigate the risks.

There Is No Such Thing as a Small Data Quality Problem

It takes just one piece of corrupt information to create monumental issues. Bad data multiplies at an astonishing rate, polluting not only the system in which it originates, but also the many other information sources it touches as it moves across a business. Therefore, the longer a company waits to detect and correct a bad record, the more damage it can do.

This is why taking a reactive approach to data quality, instead of a proactive one, can be an expensive decision. In a recent study, independent analyst firm SiriusDecisions notes what it calls the "1-10-100 rule," which demonstrates the benefits of proactive data quality. The rule states that it costs only $1 to verify a record upon entry and $10 to cleanse and dedupe it after it has been entered, but $100 in potential lost productivity or revenue if nothing is done.[1]

What You Need in a Data Quality Management Solution

When choosing a solution to support your data quality strategy, there are two key features to look for. First is the ability to protect agnostic sources. Many companies approach data quality only from the perspective of their own internal systems. However, a lot of data comes from outside corporate walls; it is collected from applications maintained by partners, or aggregated from various Web sites. Failing to consider these sources can leave huge gaps and create an environment fraught with risk.

The second capability—and perhaps, the most critical one—is real-time quality control. Yes, it's important to identify and correct any bad data that already exists. But quality must also be controlled as data flows instream, moving from system to system during the course of dynamic processes, or as it moves downstream as users access it for reporting and analysis. The key to truly keeping enterprise data "clean" is the ability to stop corrupt information as it heads upstream and enters the environment via various methods and formats, such as e-mail, manual data entry, B2B exchanges, and Web and self-service portals.

Only the most advanced tools on the market today can proactively control quality across all upstream, instream, and downstream data, using predefined rules and localized dictionaries to catch corrupt information as it is generated, stopping it before it reaches a database.

Summary

Data quality issues impact businesses of all types. Regardless of their cause, these problems cost billions of dollars each year. The longer they go undetected, the more damage they can do. Companies must leverage real-time quality control to not only correct existing records of subpar quality, but to also stop bad data from entering the environment in the first place.

For a free white paper on this topic from iWay Software, click here and choose the title "Optimizing Data Quality in the Enterprise: How to Tackle Your Bad Information."

---

[1] "The Impact of Bad Data on Demand Creation," by Jonathan Block, senior director of research, Sirius-Decisions.

How to Overcome the Limitations of Customer Data Integration

By Ravi Shankar
Senior Director of Product Marketing, Informatica
rashankar@informatica.com

Five Key Steps to Maximizing Business Value

The pioneers of customer data integration (CDI) had one goal: improve customer-facing operations such as marketing, sales, service delivery, and billing by providing a single view of the customer. To that end, many organizations selected a CDI technology to reconcile inaccurate, incomplete, and inconsistent customer data stored in different formats in multiple systems across the enterprise. Other critical business data, such as accounts, products, contracts, channel partners, and employees, was considered unnecessary and out of scope.

Although some benefits were gained with the single customer view, many companies realized in hindsight that their CDI projects significantly constrained business value. Business problems and opportunities rarely constrain themselves to a single dimension; they are, in fact, multidimensional, and hence require a multi-domain master data management (MDM) solution.

The Limitations of CDI

Take the hypothetical example of Traditional Bank, a financial services firm providing a range of banking, investment, and trust services to individuals and small and mid-size businesses. Its CDI project successfully delivered a single customer view, helping the customer-facing teams know if they were interacting with an individual, such as John Quincy Jones, or a small business, such as Jones Consulting.

However, the great limitation of the CDI system was that it did not provide visibility into the relationships between a customer, his or her organizations (such as an employer or an affiliated company), and his or her accounts. For example, the financial advisors and service delivery personnel could not see that corporate customer John Jones, the owner of Jones Consulting, was the same as individual investments client John Jones. Therefore, they were unaware that he had a corporate checking account as well as an individual brokerage account. This resulted in two problems:

- The bank jeopardized this customer's loyalty because the service delivery personnel were unable to accurately measure his value and delivered the wrong level of service to him.

- The bank lost revenue because its financial advisors were unable to make relevant cross-sell and up-sell offers to increase this customer's share of wallet.

Using Multi-domain MDM to Solve Business Problems and Capitalize on Opportunities

While providing a single customer view is a good first step, it creates only incremental business value. Maximizing business value requires more work. The next step is to enable the 360-degree customer view, associating the single customer view with all of his or her accounts or products across the company. The final and most lucrative step is to enable the extended customer view, revealing all of a customer's valuable family, business, and bank employee relationships.

Take another hypothetical example of Progressive Bank, a competitor to Traditional Bank with the same business model. Pro-

gressive Bank's team recognized that to maximize business value, its solution must extend beyond mere identification and correction of individual and corporate customers to a 360-degree view and, ultimately, the extended customer view.

The IT team was keenly aware that Progressive Bank's business problems and opportunities were multidimensional and involved information beyond customer data. Hence, the team used a multi-domain MDM platform, which handles multiple data types.

Although its goal was the same as that of Traditional Bank—improving customer-facing operations—Progressive Bank's multi-domain MDM solution delivered significantly more business value. Customer-facing teams gained visibility into relationships within and outside the household—they could see the relationships between a customer and his or her spouse, children, and parents, as well as all the businesses and organizations that were associated with the customer. In addition, they could view all the individual and shared products and identify which bank employees, such as financial advisors or stock brokers, worked most closely with the customer. As a result, Progressive Bank's team maximized business value threefold:

- **Improved profitability and customer retention.** Service delivery personnel could accurately measure each customer's value and align service levels to this value.

- **Increased revenue.** Financial advisors were able to make relevant cross-sell

and up-sell offers to increase each customer's share of wallet.

- **Increased business agility and speed.** The IT team could flexibly scale and adapt to changing business needs and support strategic imperatives.

The financial services example above is only one of many. Companies in other industries, such as manufacturing and high tech, consumer packaged goods, media, energy, healthcare, pharmaceuticals, and government have successfully transitioned from CDI to MDM to generate significant business value to their organizations.

One example is a pharmaceutical giant with an award-winning deployment that successfully extended its CDI initiative beyond customer data to include contracts, organizations, delivery channels, and pricing. It created a highly efficient system that utilizes relationships across different data types to:

- More accurately process payments, rebates, and disputes

- Produce reliable reports and performance metrics

- Better comply with regulations, such as Sarbanes-Oxley, which requires controls over data

- Increase customer service levels through error prevention, accurate orders, timely payments, and improved end-to-end experiences

- Adapt more quickly to changing market conditions

- Introduce new products more quickly into the sales channel

- Improve the sourcing of new customers

- Resolve potential problems faster

- Manage data and fix errors in contracting, finance, and the supply chain, using fewer resources
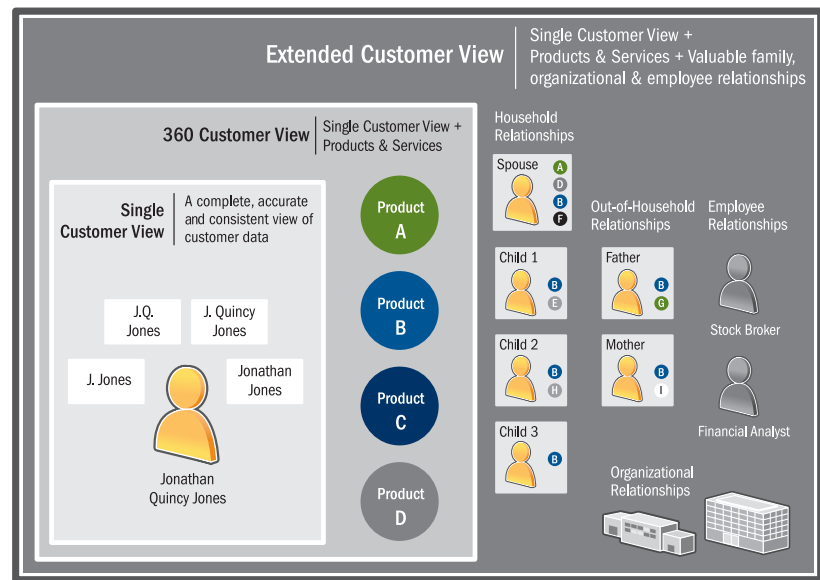
Evolution from single customer view to 360-degree customer view and extended customer view.

- Drive down the rate of write-offs, unresolved disputes, and deductions

Beginning the Journey from CDI to MDM

Multi-domain MDM can deliver on the undelivered promise of CDI by enabling the progression beyond a single customer view. These five simple steps start the journey from CDI to MDM:

1. **Identify the complete business problem.** Unreliable data typically impacts multiple areas of the business. Fixing only customer data will not solve the business problem. Identify the components of the complete business problem.

2. **Include the business users.** Even though IT implements a solution to solve a data problem, the business teams use the data. Include them in the discovery process to determine which types of data—and the relationships among and between data—need to be included to fully solve the problem.

3. **Create a single customer view.** Reconcile incomplete, inconsistent, and duplicate data in different formats in multiple systems across the enterprise.

4. **Deliver a 360-degree customer view.** Build on the single customer view by adding associated accounts across the enterprise.

5. **Provide an extended customer view.** Extend the 360-degree customer view by adding valuable but previously untapped family, business, and bank employee relationships.

Some companies that selected a CDI-only solution could not progress beyond the third step and hence limited their business benefits. However, others made an investment in multi-domain MDM, which enabled them to progress to the lucrative fourth and fifth steps. These visionaries have overcome the limitations of their CDI initiatives and delivered significant business value to their companies with increased profits, revenues, agility, and speed. By following these visionaries' footsteps, you will be well on your way to reaping the benefits from the larger multi-domain MDM initiative as well.

For a free white paper on this topic from Informatica, click here and choose the title "Is Your Approach to Multi-domain Master Data Management Upside-Down?"

Master Data Management Comes of Age

Commentary by David Smith
Senior Director, Enterprise Architecture, Citrix Systems

Organizations are increasingly turning to master data management (MDM) as a key enabler in improving the quality, timeliness, and reliability of business intelligence (BI), with the ultimate goal of improving business performance. And—as Citrix Systems recently learned—when it comes time to choose an MDM vendor, it's invaluable to have an experienced, strategic partner to facilitate the process. An MDM business consultant informs a more deliberate approach to implementation, including requirements gathering, business case development, and vendor selection.

Citrix is the leading provider of virtualization, networking, and software-as-a-service technologies for more than 230,000 organizations worldwide. Founded in 1989, its annual revenue in 2008 was $1.6 billion.

Citrix's goal is to continually improve its operational efficiencies and extend its predictive analytic capabilities for planning purposes. To achieve these goals, Citrix knew that laying a solid MDM foundation was key. The company chose the partner data domain as the nucleus to establish that foundation. The success of this project is predicated on the joint ownership between business and IT, represented by the partner operations group and the enterprise architecture (EA) group. The initial scope of the MDM platform was to support the partner network and eventually grow to other business units and data domains. Therefore, the MDM platform had to be extensible enough technically, and flexible enough from a business perspective, to handle business and data models from other departments.

EA and the partner operations group had a solid understanding of what end results they required. What they did not know was which MDM vendor to partner with, what resources would be required for the implementation and ongoing maintenance, and whether it would meet their business needs. EA had developed a thorough evaluation approach that depended on business participation and a strong partner in the MDM space to help achieve quality results in an aggressive timescale.

MDM Experts Hit the Scene

Citrix contacted Baseline Consulting, a business consultancy and technical implementation services company that helps companies use their data to improve business performance. The project was challenging: help Citrix evaluate four MDM platform vendors (a process that typically takes three to five months) and make a purchasing recommendation to the project steering committee, including the CIO, in six weeks.

> "Because of the thorough prep work, well thought-out process, and collaborative nature of the partnership, all team members were on the same page developing the vendor recommendation."
>
> **David Smith**
> Senior Director, Enterprise Architecture, Citrix Systems

A Baseline team of three hit the ground running. They partnered with a team of four Citrix professionals, including managers and directors from IT and the partner operations group. The Citrix team had done their homework: they developed high-level business and technical requirements, diligently researched MDM vendors, sent RFIs to a list of vendors, and visited potential vendors at industry trade shows.

Baseline's strategy was to limit each of the four MDM vendor candidates to a three-day

proof of concept (POC), during which Citrix employees would load and configure the data model and run through a dozen use cases with direction from the vendor. This forced the vendors to focus their presentations as the POC process typically takes weeks or months. In addition, it gave Citrix IT personnel a better grasp of how the technology would (or would not) work with existing systems.

The collaborative working environment at Citrix helped Baseline drive the vendor interfacing, schedule the POCs, and create vendor documentation, project status reports, and so on. Multiple scorecards were created and used to judge each MDM vendor on various technical aspects. Following the POCs, the team circled back around and spent several days interviewing candidate vendors and documenting the actual business requirements from the partner operations group.

Collaboration, Experience, Results

Baseline gathered all the documentation resulting from the POCs and reviewed and analyzed it with the Citrix team. Because of the thorough prep work, well thought-out process, and collaborative nature of the partnership, all team members—Citrix and Baseline—were on the same page developing the vendor recommendation. Better yet, the partner operations group and EA were able to present the recommendation to their executive sponsors one day before the deadline.

At the end of the day, joint project ownership between business and IT, the collaborative environment, and an experienced team of consultants shortened a process that typically takes months down into a six-week timeframe.

For a free white paper on this topic from Baseline Consulting, click here and choose the title "The Baseline on MDM: Five Levels of Maturity for Master Data Management."
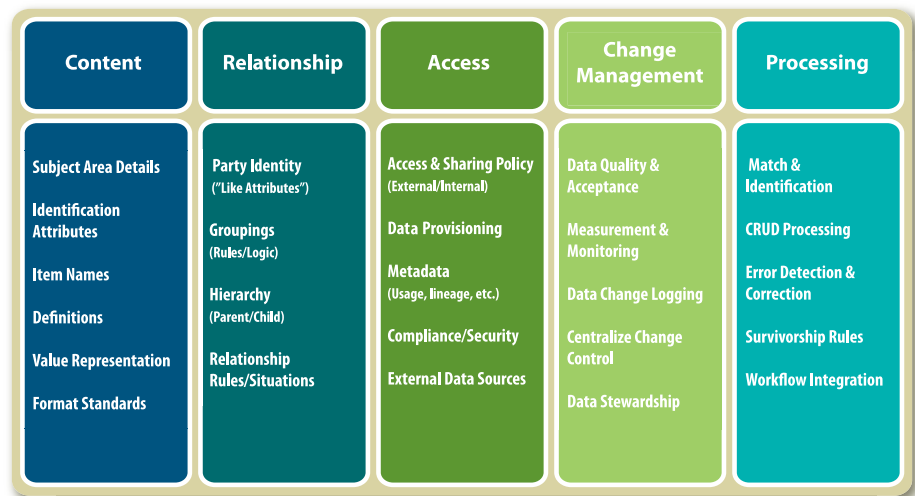
A Deliberate Approach to MDM Vendor Selection

By Evan Levy
Partner, Baseline Consulting

The phrase "MDM vendor selection" can be misleading. Most companies with master data management (MDM) hubs haven't performed "selection," per se. Instead, a vendor introduces its MDM solution to an interested IT manager or business owner. Naturally, the vendor focuses on the strengths of its MDM solution, rather than what the customer needs, and MDM selection becomes sales driven, instead of deliberate and requirements driven.

The company ends up with an MDM tool that doesn't meet its needs and eventually demands the creation of custom code. Time and resources are spent conforming business activities to the specifics of the application. And all too often, companies have to augment existing software products with "bolt-on" functionality to deliver MDM-like functions, only to discover that the resulting custom solution can't support operational data integration.

In all likelihood, the business hasn't spent time assessing its master data needs. Business priorities for master data are vague, so it's impossible to prioritize certain functions over others. Without a rigorous understanding of the desired state for MDM, vendor selection becomes a generalized function/ feature comparison, and not a mapping of capabilities to business requirements.

With any strategic technology, requirements should be ironclad before making an investment. This level of due diligence does not mean you should avoid MDM vendors, but keep in mind that vendors are there to sell products, and thus won't emphasize the capabilities their solutions can't support. Also, remember that you only need to master data if that data is shared between systems and the reference data in each of those systems is represented differently.

Content	Relationship	Access	Change Management	Processing
Subject Area Details	Party Identity ("Like Attributes")	Access & Sharing Policy (External/Internal)	Data Quality & Acceptance	Match & Identification
Identification Attributes	Groupings (Rules/Logic)	Data Provisioning	Measurement & Monitoring	CRUD Processing
Item Names	Hierarchy (Parent/Child)	Metadata (Usage, lineage, etc.)	Data Change Logging	Error Detection & Correction
Definitions	Relationship Rules/Situations	Compliance/Security	Centralize Change Control	Survivorship Rules
Value Representation		External Data Sources	Data Stewardship	Workflow Integration
Format Standards				

So how does a company define its MDM requirements? There are five functions associated with data management, as described below:

- **Content.** This function represents the necessary subject-area constraints for the data that will be mastered. It includes the terminology, definitions, and formatting of the data. For instance, if a car company needs to master "product color," we'd need to understand the valid color choices of exterior paint for each of its models.

- **Relationship.** This identifies whether there is a relationship between individual entries within the subject area, and the groupings, hierarchies, and interrelationships between those items.

- **Access.** Access infers privileges on data. A manager might be allowed to update an employee's compensation (based on the relationship function), but not be allowed to modify his own compensation.

- **Change management.** This function focuses on the need to track individual changes to the master data. For instance, if the automobile company wants to add a new product color, that color needs to be reviewed and approved by the product review team.

- **Processing.** Processing is the definition of rules for identifying, adding, or removing master data content. For example, when red cars need to be identified, colors like "maroon" and "candy apple" might also be considered red in the product catalog. When we need to manipulate master data, the rules associated with that data can determine how individual processes are executed.

When evaluating the functional requirements across these five areas, a company can determine fairly quickly which vendor products conform to their needs.

So what's the next step in determining your MDM requirements? First, understand where you've had challenges with your master data. Where does it vary across systems? Focus on the operational systems that originate the data. Document scenarios for when and how data is created, read, updated, and deleted. Then you should be able to distill and map your MDM requirements, and select a subset of vendors who can meet those needs. You might want to benchmark those companies using a subset of your own data—but that's a whole different set of best practices!

For a free white paper on this topic from Baseline Consulting, click here and choose the title "The Baseline on MDM: Five Levels of Maturity for Master Data Management."

B2B Data Management

By Richard Clements
*Senior Director, Strategy and Product Marketing,
Initiate, an IBM Company*

As Global 2000 companies face challenging times, procurement personnel, operations, marketing and sales executives, and risk officers look for new solutions to reduce operational costs, improve customer satisfaction and retention, and reduce customer risk. At the root of these challenges is a need for a deeper understanding of the data they have about their customers. The business problems that B2B data management initiatives are intended to solve are many and complex, but here is a brief summary of problems that organizations encounter:

- **Risk management:** Cannot avoid extending credit to high-risk customers due to a lack of legal hierarchies or the inability to understand exposure to customers who go bankrupt

- **Reporting:** Cannot report on activity at top accounts, conduct general roll-up reporting, or accurately recognize revenues

- **Relationships:** Unable to understand the relationship between individuals and organizations to find all relevant transactions and provide a relevant level of service

- **Holistic view of the customer:** Cannot assess the growth of the business by understanding customers, and associated contacts within the organization, across all lines of business

- **Multiple views:** Unable to understand customers through multiple views (e.g., legal, sales, pricing)

- **Sales territory management:** Unable to properly assign field sales force across territories, resulting in mis-parented sales accounts

MDM for B2B Is Different
Management of B2B customer data imposes market requirements for management frameworks, business rules, and process controls that aren't widely shared by consumer MDM. Why would this be the case?

Largely, it comes down to two factors: scale and value. Businesses that sell to other businesses typically have smaller customer bases, meaning B2B data tends to come in smaller volumes than consumer or B2C data. The smaller scale typically encountered in the B2B data set makes manual stewardship over business-impacting events in that data more feasible than it would be in a much larger B2C data set. And though the B2B data set is smaller, each record represents an account that corresponds to a larger share of the business's revenue base. Therefore, the organization has greater incentive to maintain oversight and controls over changes that could adversely affect that account.

Today's Data Governance Tools Don't Solve the Problem
Organizations that have B2B customers have a class of data management challenge that is not easily met by today's tools and technologies. These challenges exist because organizations are bombarded with information from internal and external sources. Incorrect, out-of-date, and out-of-synch information from one system can contaminate an entire enterprise like a virus, polluting the data ecosystem. Data governance directives that inoculate IT systems against bad data at worst do not exist, and at best, exist but cannot be implemented based on existing IT tools and processes. Data governance tooling on the market today does not solve these problems. Data stewards:

- Are unable to enforce data governance policies specific to initial data capture and ongoing data maintenance

- Do not have the capability to review potential issues before they impact existing data processes

- Are unable to track whether high priority issues are being resolved within required time frames

- Are unable to manage organizational relationships in the data, such as comparing hierarchies, creating different hierarchical views (e.g., legal, sales), viewing past representations of relationships, or creating tasks

- Are unable to automatically enforce logical workflow processes

Make no mistake, organizations have made significant IT investments to leverage their customer data assets; however, these data assets are still kept in silos and in views that cannot be readily exploited. Furthermore, the customer data is not presented in a hierarchical form that allows marketers and account managers to understand who their customers really are in the context of their overall organization. For marketing and sales organizations struggling with this challenge, a change is needed—these organizations must be able to reap the benefits of their data assets.

Active Data Governance for the Part-Time Data Steward
The ideal solution is a data stewardship/data management application purpose-built for MDM B2B use cases that typically involve lightweight, out-of-the-box data governance workflows and enable business-oriented users to collaborate and implement active data governance policies to ensure low-quality data is not transmitted throughout the broader organization.

LESSON FROM THE EXPERTS
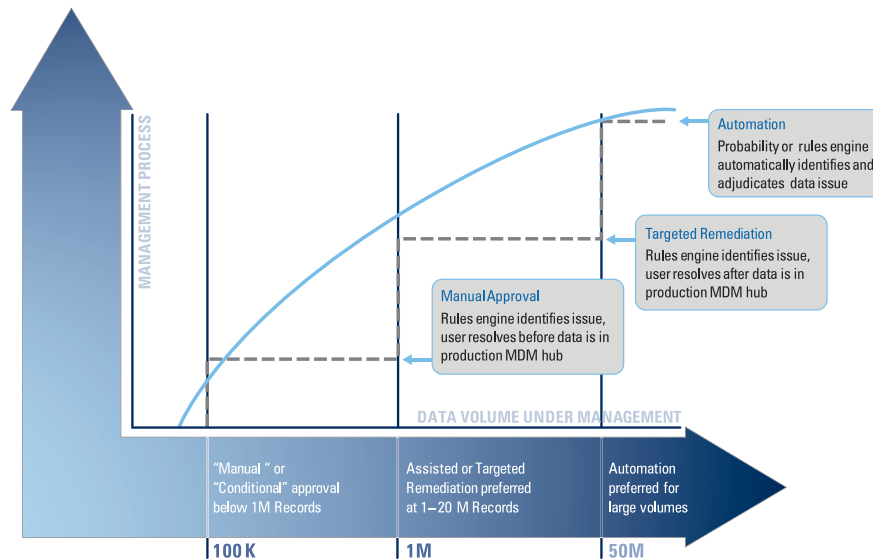
The data stewardship/data management application should allow collaboration and enable organizations to efficiently implement data governance policies that necessitate multiple stakeholders working together to improve data quality. It should also include native workflow-based data management capabilities to support processes and business rules the organization defines based on business priorities. The application should be intuitive and have an easy-to-use user interface that provides data stewards the ability to enforce data governance policies and examine a 360-degree view of all relevant business data and issues. The user interface should allow data stewards to manage multiple complex relationship types (often hierarchical in nature) that are the most important to their business. And, finally, it should be customizable and configurable, supporting individual user preferences and settings.

Benefits Can Be Enormous

With the appropriate MDM-based B2B data governance application, organizations and business executives can truly tap the power of knowledge and realize the following benefits:

- **Reduced operational costs.** Sales management can avoid paying multiple sales commissions for the same sales transaction because they can now link an individual account manager to a single account at a specific time. Also, duplicate entries can be eliminated, preventing multiple mailings or calls made to the same account. All of this reduces operational costs and generates fiscal control, both internally and externally.

- **Improved customer retention.** Marketers can now better analyze which profitable customers to keep and which customers are no longer profitable. As an example, account A may not seem very important when viewed separately. However, if you can see that account A is a subsidiary of a larger, more profitable organization, then the importance of account A is revealed. The ability to understand each account in its context allows marketers to create efficient retention programs.

- **Increased revenue via up-selling and cross-selling.** Because enterprisewide product-related information is now available as part of the holistic customer view, account managers can understand which product was sold to which company and when. As a result, they can leverage the trust generated with their existing customers by selling them additional products.

The potential gain is substantial; however, the effectiveness of B2B data management initiatives is only as good as your understanding of the underlying data that fuels the results. Active data governance applications targeted at B2B data management can deliver the critical understanding you need to meet today's challenges and anticipate tomorrow's needs.

For a free white paper on this topic from Initiate Systems, an IBM company, click here and choose the title "Beyond the Golden Record: Serving Many Masters Through Hierarchy Management."

AmeriCredit Creates a Single Source of Reference for Key Information Assets

Commentary by Nikitas Gogos and Aiman Gurji
Data Architects, AmeriCredit

As a leading independent automobile finance company, AmeriCredit Corp. relies on information technology to support its decision-making capabilities for credit risk management. Information is core to AmeriCredit's business, and it is critical that data is consistent, consolidated, and easily available to all employees for accurate and efficient decision making.

To create a single, consistent source of information for key information assets such as credit, payments, and collections, the data architecture team created a data warehouse that consolidated the dispersed data sources into a single source for reporting. This integration posed several challenges, according to data architects Nikitas Gogos and Aiman Gurji.

> "Our biggest business benefit has been the documenting and publishing of metadata to our end users."
>
> **Nikitas Gogos**
> Data Architect, Americredit

The first challenge was technical integration across multiple database platforms. In addition to platform migrations, data types, and formatting, the data needed to be consistent in the target warehouse. With hundreds of database tables, this was no simple task.

The second challenge was organizational integration and information sharing. Metadata had to be documented and standardized across the organization. Once these definitions were created, a variety of teams needed access to the information—including business intelligence analysts, database administrators (DBAs), and data architects. Publicizing information across the organization and creating consistency was key to the success of the project.

Consolidating Information

AmeriCredit created a single view of information using graphical data models. Through reverse engineering, the company created an inventory of the source and target systems of the warehouse to have a comprehensive view of the information landscape. These models were then stored in a central model repository so that information could easily be shared, standardized, and reused.

Once an inventory of the data was created, additional business information such as definitions and code values were added. To show the right amount of information to the right audience, business information was stored in the logical data model, and technical details were managed in the physical layer. To further manage the volume of information, subject areas were used to organize information into smaller, more understandable groupings such as applications, account services, and funding.

Getting the Word Out

After creating this valuable store of information, it was important to share it with the extended team. With a data warehouse team of more than 300 members, and business intelligence analysts using the information to build reports, it was necessary to develop a publishing mechanism that was both intuitive and easy to implement. HTML reports were published on the Web so that business defi-nitions, code values, and column details were easily accessible by end users.

Faster Access to More Reliable Information

Making information consistent and accessible saved AmeriCredit time and money as more and more users used the Web-based reports to find the information they needed. "Our biggest business benefit has been the documenting and publishing of metadata to our end users," said Gogos.

With information published and easy to find, the data architects spent less time answering questions and were able to focus on their primary tasks. Questions that did come up were more detailed, and as a result, the data architects had more productive conversations with team members. Training of new employees was also streamlined, since project information was consolidated in a central place. The models and reports were able to provide new team members with the information they needed to get up and running quickly.

For a free white paper on this topic from CA ERwin Modeling, click here and choose the title "The Benefits of Data Modeling in Business Intelligence."

LESSON FROM THE EXPERTS

Successful Business Intelligence Implementations Start with Good Data Model Design

By Donna Burbank
Director of Product Marketing, CA ERwin Modeling

In today's information-driven economy, organizations rely on business intelligence (BI) applications to make strategic business decisions. BI applications allow business users to create their own reports through an intuitive user interface, removing reliance on IT and putting the power of "self-service" into the hands of business users. The self-service approach is a driver in the increased popularity of BI applications, and a key justification for the return on investment (ROI) achieved.

A BI report is only as good as the database on which it is built; however, many business users are not involved in the design or population of the databases that drive their BI applications. Instead, database design and implementation are often performed by data professionals with specialized knowledge. This specialization is understandable, given the highly technical nature of database systems, but databases must be built to implement the needs of the business.

Start with a Good Design and Involve the Business

A common communication medium between IT and the business is a data model, which is typically built at various levels to suit the intended audience and purpose. For example, a logical or conceptual data model generally defines core business terms and rules, and a physical data model is used to create the database structures that are optimized to store and retrieve information. Usually, information about the business is gathered into the logical model by data modelers and architects through a series of interviews with business people.

This business-level analysis should not be overlooked—though the physical database design is important, the business rules and definitions are critical to the success of the business intelligence implementation. If a business user creates a report of product sales, the definition of "product" must be clear—does it mean finished products, or does it include raw materials? These core business terms need to be defined in the business-level data model.

Create an Inventory of Information Assets

In addition to understanding business requirements, the physical analysis of the database infrastructure is equally critical. With the massive volume of data in most organizations today, it is a challenge to understand what information exists and where it is stored. Data models provide a graphical road map of information assets by reverse engineering databases to analyze their structure and interrelationships.

Product information, for example, may be stored in multiple databases, on diverse platforms, and in disparate geographical regions. Data might be stored in various formats and use different terminology. The process of harmonizing information into a single, consistent design is facilitated by the graphical nature of a data model. Like a graphical road map that helps guide you through city streets, a graphical data model provides a map of your information assets and helps enforce standards so that information can be easily consolidated.

Get the Word Out

Once an inventory is built and business definitions are created, make sure to publish the information in a format that is easily accessible to multiple roles across the organization. For example, since business users may not have access to a data modeling tool, publish the models onto the Web for this audience. Even though graphic data models are intuitive, consider publishing model definitions in other formats such as spreadsheets, wikis, or even Word documents.

Although the analysis should be done in a robust modeling tool, the publication should be flexible to reach all audiences. Remember, the success of many BI applications was in their user-friendly nature. Make the data models that define the information just as accessible and user-friendly.

Conclusion

The growth in BI reporting has increased the need for robust analysis on the back end to ensure reports are built on correct information. To support the front-end reports, the back-end analysis requires both a careful inventory of database assets and an interactive discussion with the business to make sure data is defined correctly. Once this analysis has been done using data models, it is important that information is shared with the end consumers through intuitive, "self-service" publications.

For a free white paper on this topic from CA ERwin Modeling, click here and choose the title "The Benefits of Data Modeling in Business Intelligence."

Assuring 24/7 Analytic Availability to More Than 2,000 Major Online Sites

Commentary by Shawn Farshchi
Chief Operating Officer, Coremetrics

Optimizing Performance for Innovative Analytic Applications

As one of the industry's leaders in online marketing and business optimization solutions, Coremetrics guarantees its clients access to a high-performance SaaS platform that delivers valuable data-driven content for many of the world's major brands.

Coremetrics serves 2,000 companies that process more than $20 billion in transactions per year, and received top rankings for strategy and customer satisfaction from independent analyst firm Forrester Research, Inc. Coremetrics is well known for its service-level agreements (SLAs) that allow clients to fully leverage online customer interactions in real time, 24/7.

However, when Coremetrics planned to introduce more advanced analytic applications, soaring data volumes suggested that the company reevaluate its database infrastructure. Next-generation analytic tools would face formidable optimization limits to deliver the high performance that new applications would demand.

Maintain a Competitive Advantage in Both Cost and Performance

Because Coremetrics enjoys such an impressive reputation for meeting and exceeding its SLAs, it was essential to deploy a highly scalable, massively parallel database solution for current applications, and support future applications to maintain the company's competitive advantage.

Overall, the strategic business objectives required a fault-tolerant environment to maintain high availability, backup and restore functionality to assure reliability, and live server administration capabilities to eliminate downtime. But controlling software and hardware costs was also a key objective.

According to chief operating officer Shawn Farshchi, "Our goal was to deploy a mature database platform that provided superior scalability on commodity hardware—to give us both a tremendous cost and performance advantage. Aster Data was the only company we could find that gave us this capability."

Massively Scalable, Massively Parallel Data Warehouse

Aster Data *n*Cluster is the industry's first massively parallel data warehouse architecture that allows applications to be fully embedded within the database engine, which enables ultrafast, deep analysis of massive data sets.

> "Aster Data delivers superior scalability on commodity hardware—giving us a tremendous cost and performance advantage."
>
> **Shawn Farshchi**
> Chief Operating Officer, Coremetrics

It's a significant improvement over traditional data warehouses, which weren't designed to keep up with the data loads of today's big data applications. Aster Data's massively parallel data application server effectively uses Aster's patent-pending SQL-MapReduce with data and application parallelization to address the big data challenge.

Aster Data deployed its *n*Cluster data-application server across Coremetrics's commodity hardware cluster with multiple terabytes of capacity.

Support for New Innovative Applications, Sharply Lower Capital Expenditure Costs

Once the Aster Data *n*Cluster database was deployed for several Coremetrics applications, it was obvious the platform gave the company more control in managing its customers' data and could easily cope with soaring growth and large traffic spikes.

In fact, Coremetrics now supports a 20-fold increase in server calls—while maintaining 24/7 availability and resiliency on a highly reliable fault-tolerant platform. And because the solution runs on inexpensive commodity hardware, Coremetrics has had substantially reduced capital expenditures for its data infrastructure.

Most important, the company is no longer bound by old infrastructure limitations. Now customers have access to new and innovative applications that are deployed on top of the new and more robust data infrastructure.

These applications include Coremetrics AdTarget™ and Coremetrics Intelligent Offer™, which automatically generate personalized content and behavior-based product recommendations and content in real time to fully leverage customer data.

According to Farshchi, "Aster Data helped us meet or exceed all of our goals. We've improved performance and reduced our costs. We now have enormous freedom to introduce even more powerful data-driven marketing and analytics solutions that help some of the world's largest companies drive business results."

For a free white paper on this topic from Aster Data, click here and choose the title "Aster Data nCluster: A New Architecture for Large-Scale Data Analytics."

LESSON FROM THE EXPERTS

Emerging Technologies for Big Data Analytics

By Jon Bock
Director of Product Marketing, Aster Data

With more than 60 percent data growth per year in many enterprise applications, and over 100 percent in most Internet applications, addressing the challenges of "big data" has become a top priority. Organizations that want to gain an edge over the competition need to cost-effectively load and store terabytes to petabytes of data, perform rich analysis on these massive data sets, and do it at ultrafast speeds. Driven by uses such as fraud detection, customer behavior analysis, trending and forecasting, scenario modeling, and deep clickstream analysis, faster and deeper analysis of massive data sets has become a mainstream requirement.

Enabling this new generation of big data analytics requires overcoming the performance, scalability, and manageability limitations of traditional approaches. In the past, these limitations forced data analysts to accept limited and lower-quality analytics, but emerging technology in three key areas is enabling a new generation of analysis.

High Performance, Scalable Data Platforms

Traditional architectures struggle to provide fast results to queries on large data sets because they rely on monolithic architectures designed in an era of smaller data sets and slower data growth. They are typically expensive and cumbersome to scale as data grows due to inflexible systems, requirements that force "super-size" upgrades, required downtime, and error-prone upgrade processes.

Scaling data and analytics to manage terabytes or petabytes of data requires pervasive parallelism. Organizations are turning to massively parallel database architectures that parallelize all functions of the system, from loading to query processing. They are also turning to frameworks designed for large-scale data processing. Google, one of the first to face the challenge of analyzing petabyte-scale data, pioneered a software framework called MapReduce for fast processing of large amounts of data that is increasingly being leveraged for big data analytics.

Embedded Applications

Traditional architectures for data management and analytics are simply not built to move terabytes to petabytes of data through the data pipeline to the analytic application for processing. The larger the data volume, the larger the time and effort needed to move it from one location to another. The resulting performance and latency problems are so severe that application developers and analysts commonly compromise the quality of their analysis by avoiding big data computations.

> "It's really innovative, and I don't use those terms lightly. Moving application logic into the data warehousing environment is a logical next step."
>
> **James Kobielus**
> Forrester Research

New technology allows analytic applications to be pushed into the database so they are co-located with the data. This new approach allows businesses to embed existing application logic associated with data analysis into the database where data resides, eliminating data movement and delivering major improvements in speed of data analysis, depth of data analysis, and ability to analyze very large data sets rather than relying on data sampling.

Tools for Rapid Development

Traditional analytic application development within organizations is commonly plagued by complexity and inefficiency throughout the application lifecycle, from coding and integration to testing and deployment. These challenges are caused by the number of complex manual steps required, limited resources available for testing, and the need to involve multiple people and groups throughout the process.

Addressing these challenges requires simplifying and automating the process of creating and deploying analytic applications. Integrated development environments, desktop testing environments, and reusable building blocks for analytic applications are becoming available to address this critical pain point.

Conclusion

The need to provide fast and rich analytics on big data is forcing enterprises to rapidly evolve their data management and analytic architectures. Adopting technology that addresses the performance, scalability, and complexity challenges of big data analytics is critical to deliver the insights necessary to leverage the value in organizations' data.

For a free white paper on this topic from Aster Data, click here *and choose the title "Aster Data nCluster: A New Architecture for Large-Scale Data Analytics."*

INTRODUCTION TO NEXT GENERATION DATA WAREHOUSE PLATFORMS

BY PHILIP RUSSOM

If you're a data warehouse professional—or you work closely with one—you've probably noticed the many new options for data warehouse platforms that have appeared this decade.

We've seen the emergence of new categories of data warehouse (DW) platforms, such as data warehouse appliances and software appliances. A new interest in columnar databases has led to several new vendor products and renewed interest in older ones. Open source Linux is now common in data warehousing, and open source databases, data integration tools, and reporting platforms have come out of nowhere to establish a firm foothold. In the hardware realm, 64-bit computing has enabled larger in-memory data caches, and more vendors now offer MPP architectures. Leading database vendors have added more features and products conducive to data warehousing.

Those are mostly features within the data warehouse platform, especially its database. There are also growing practices that are demanding support from the platform, including real-time integration between the data warehouse platform and operational applications, various types of advanced analytics, and reusable interfaces exposed through Web services or service-oriented architecture (SOA). Furthermore, a number of data warehouse platforms and other business intelligence platforms are now readily available through software-as-a-service (SaaS) and cloud computing.

The good news is that the options for data warehouse platforms have recently become far more numerous. The bad news is that it's difficult for data warehouse professionals and their business sponsors to keep track of these advancements and select the ones that are appropriate for their needs.

To help organizations understand the many new options available to them, this report catalogs the new data warehouse platform products, features, and techniques that have appeared this decade, plus notable advances in more established data warehouse platforms. As examples, the report mentions many vendors and their products. From the survey data cited here, you'll see that many organizations are planning the next generation of their data warehouse, and this report provides information that can be instrumental for such planning. The focus is on technology, but this report also explains how technology's adoption in next generation data warehouse platforms is driven by real-world business and organizational needs and requirements.

Definitions of Terms and Concepts

DATA WAREHOUSE PLATFORM

For the purposes of this report, a data warehouse platform consists of one or more hardware servers, an operating system, a database management system (DBMS), and data storage. These communicate via a LAN or WAN, although a multi-node data warehouse platform may have its own specialized network. Note that a data warehouse platform manages a data warehouse, defined as a collection of metadata, data model, and data content, designed for the purposes of reporting, analyzing information, and making decisions. But the data warehouse is not part of the platform per se. (See Figure 1.) All these components and more have seen generational advances in recent years.

GENERATIONS OF DATA WAREHOUSES

TDWI's position is that certain relatively new technologies, techniques, and business practices are driving the majority of data warehouses and their platforms toward a redesign, major retrofit, or even replacement that we can recognize as a generation. TDWI takes the term literally, meaning that the current generation of a data warehouse will beget the next generation. In many cases, generational change is an evolutionary process that adapts the resulting data warehouse to changing business and technology requirements. In fact, generational change is often driven by these requirements, as is explained in detail in the next section of this report. In other cases, generational change is more of a maturation process that steps a data warehouse through multiple stages of a lifecycle.

NEXT GENERATION DATA WAREHOUSE PLATFORMS

What's next for a given organization's data warehouse platform can vary tremendously. For example, a next generation data warehouse platform may tap into leading-edge features, such as appliances, open source, and cloud computing. It may simply get you caught up with somewhat more established practices for real-time operation, advanced analytics, and services. Sometimes, the next generation addresses administrative issues, such as hardware upgrades
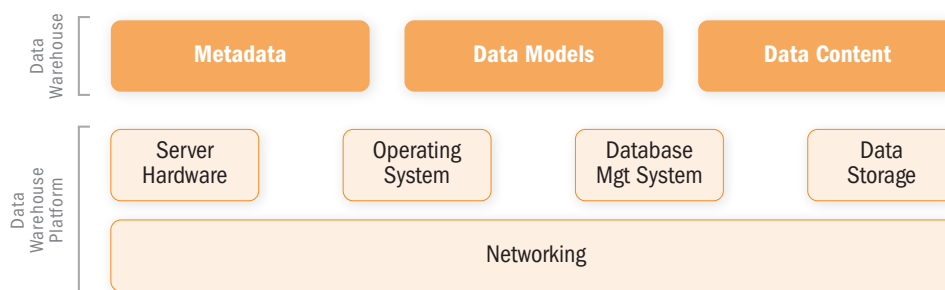
Figure 1. A data warehouse platform manages a data warehouse, but the two are separate.

(from 32-bit to 64-bit), data migrations (from one DBMS to another) or architectural changes (from SMP to MPP). So, let's keep in mind that a next generation data warehouse platform is a relative concept, because it depends on where you're starting, what new requirements you must address, and how many resources you have.

WHY CARE ABOUT DATA WAREHOUSE PLATFORMS NOW?

- **Businesses face change more often than ever before.** Recent history has seen businesses repeatedly adjusting to boom-and-bust economies, a recession, financial crises, and shifts in global dynamics or competitive pressures. Increasingly, businesses rely on the data warehouse and related business intelligence infrastructure to understand change and react appropriately.

- **DW platforms need updating to support changing business requirements.** In fact, many of the technologies associated with the next generation DW relate to change in some way, such as advanced analytics, scalable architectures, virtualization methods, reusable services, real-time integration with operational applications, and so on.

- **Successful DWs mature through multiple lifecycle stages.** This usually provokes changes in the underlying DW platform and elsewhere in the business intelligence (BI) infrastructure.

- **There's probably a new generation in your near future.** TDWI survey data shows that almost half of respondents are planning a data warehouse platform replacement in 2009–2012. Many others anticipate keeping their current platforms, but updating them significantly.

USER STORY MANAGEMENT REQUIREMENTS OFTEN DICTATE THE DESIGN OF A NEXT GENERATION DW AND ITS PLATFORM.

"We pulled together our current data warehouse a couple of years ago," said Karl Mikula, the data and BI manager at Hagerty Insurance Agency, America's leading provider of products and services for collectors of classic cars and boats. "Now that the company sees the value, we're building our next generation data warehouse and BI solution atop a platform that'll do what the company needs. In a nutshell, upper management wants to adapt a performance management methodology with scorecards. And they want self-service BI, where they can search a repository and pull data into reports or spreadsheets of their own design, presented through a corporate portal. To support this, we're designing a data warehouse that stores metrics and KPIs in a searchable repository. For the next generation platform, we have a database management system, a data integration tool, a reporting tool, a search engine, and an enterprise portal. All these come from Microsoft, and they're all tightly integrated out of the box."

Philip Russom is the senior manager of TDWI Research at The Data Warehousing Institute, where he oversees many of TDWI's research-oriented publications, services, and events. He can be reached at prussom@tdwi.org.

INTRODUCTION TO UNIFIED DATA MANAGEMENT

BY PHILIP RUSSOM

BACKGROUND

In most organizations today, data and other information are managed in isolated silos by independent teams using various data management tools for data quality, data integration, data governance and stewardship, metadata and master data management, B2B data exchange, content management, database administration and architecture, information lifecycle management, and so on. In response to this situation, some organizations are adopting what TDWI calls unified data management (UDM), a practice that holistically coordinates teams and integrates tools. Other common names for this practice include enterprise data management and enterprise information management. Regardless of what you call it, the "big picture" that results from bringing diverse data disciplines together yields several benefits, such as cross-system data standards, cross-tool architectures, cross-team design and development synergies, leveraging data as an organizational asset, and assuring data's integrity and lineage as it travels across multiple organizations and technology platforms.

But unified data management isn't purely an exercise in technology. Once it paves the way to managing data as an organizational asset, the ultimate goal of UDM becomes to achieve strategic, data-driven business objectives, such as fully informed operational excellence and business intelligence, plus related goals in governance, compliance, business transformation, and business integration. In fact, the challenge of UDM is to balance its two important goals—uniting multiple data management practices and aligning these with business goals that depend on data for success.

The purpose of this report is to help organizations plan and execute effective UDM efforts. Many need the help, because UDM is a relatively new shift in best practices for data management. Toward that end, the report drills into the business initiatives that need UDM, the data management practices and tools that support it, and the organizational structures that enable the cross-functional collaboration that's critical to UDM success.

Definitions of UDM

With all the above in mind, here's a nutshell definition of UDM:

> TDWI Research defines unified data management as a best practice for coordinating diverse data management disciplines, so that data is managed according to enterprisewide goals that promote technical efficiencies and support strategic, data-oriented business goals.

The term UDM itself seems focused on data management, which suggests that it's purely a technical affair. But this is misleading, because UDM—when performed to its full potential—is actually a unification of both technology practices and business management. For UDM to be considered successful, it should satisfy and balance both of the following requirements:

- **UDM must coordinate diverse data management disciplines.** This is mostly about coordinating the development efforts of data management teams and enabling greater interoperability among their servers. UDM may also involve the sharing or unifying of technical infrastructure and data architecture components that are relevant to data management. There are different ways to describe the resulting practice, and users who have achieved UDM call it a holistic, coordinated, collaborative, integrated, or unified practice. Regardless of the adjective you prefer, the point is that UDM practices must be inherently holistic if you are to improve and leverage data on a broad enterprise scale.

- **UDM must support strategic business objectives.** For this to happen, business managers must first know their business goals, then communicate data-oriented requirements to data management professionals and their management. Ideally, the corporate business plan should include requirements and milestones for data management. Hence, although UDM is initially about coordinating data management functions, it should eventually lead to better alignment between data management work and information-driven business goals of the enterprise. When

UDM supports strategic business goals, UDM itself becomes strategic.

Let's expand TDWI's terse definition of UDM by drilling into more specific details and issues.

UDM is largely about best practices from a technical user's viewpoint. Most UDM work involves collaboration among data management professionals of varying specialties (such as data integration, quality, master data, etc.). The collaboration fosters cross-solution data and development standards, interoperability of multiple data management solutions, and a grander concept of data and data management architectures.

UDM isn't a single type of vendor-supplied tool. Even so, a few leading software vendors (including all the vendor companies sponsoring this report) are shaping their offerings into UDM platforms. Such a platform consists of a portfolio of multiple tools for multiple data management disciplines, the most common being BI, data quality, data integration, master data management, and data governance. For the platform to be truly unified, all tools in the portfolio should share a common graphical user interface (GUI) for development and administration, servers should interoperate in deployment, and all tools should share key development artifacts (such as metadata, master data, data profiles, data models, etc.). Having all these conditions is ideal, but not an absolute requirement. As one interviewee put it: "The tools' servers have to interoperate or—at the end of the day—the solution isn't unified. So that's a 'must have,' as is shared metadata. If there are multiple development GUIs, I can live with that."

UDM often starts with pairs of practices. UDM is a matter of degree. In other words, it's unlikely that any organization would want or need to coordinate 100% of its data management work via UDM or anything similar. Instead, organizations opportunistically select combinations of data management practices whose coordination and collaboration will yield appreciable benefits. The most common combinations are pairs, as with data integration and data quality or data governance and master data management. Over time, an organization may extend the reach of UDM by coalescing these pairs and adding in secondary, supporting data management disciplines, such as metadata management, data modeling, and data profiling. Hence, the scope of UDM tends to broaden over time into a more comprehensive enterprise practice. And the scope can get rather broad, as a user interviewed for this report explained: "Enterprise-scale data management is like most things: it's a mix of people, process, and technology. The range of each is diverse, so there's potentially a place for just about anything."

A variety of organizational structures can support UDM. It can be a standalone program or a subset of larger programs for IT centralization and consolidation, IT-to-business alignment, data as an enterprise asset, and various types of business integrations and business transformations. UDM can be overseen by a competency center, a data governance committee, a data stewardship program, or some other data-oriented organizational structure. UDM is often executed purely within the scope of a program for BI and data warehousing (DW), but it may also reach into some or all operational data management disciplines (such as database administration, operational data integration, and enterprise data architecture).

UDM unifies many things. As its name suggests, it unifies disparate data disciplines and their technical solutions. On an organizational level, it also unifies the teams that design and deploy such solutions. The unification may simply involve greater collaboration among technical teams, or it may involve the consolidation of teams, perhaps into a data management competency center. In terms of deployed solutions, unification means a certain amount of interoperability among servers, and possibly integration of developer tool GUIs. Technology aside, UDM also forces a certain amount of unification among business people, as they come together to better define strategic business goals and their data requirements. When all goes well, a mature UDM effort unifies both technical and business teams through IT-to-business alignment germane to data management.
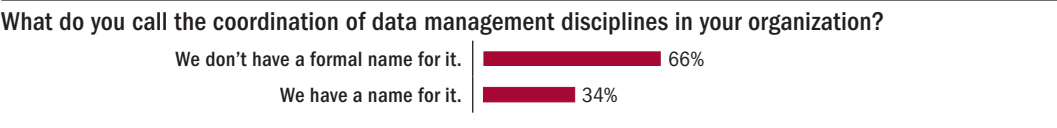
What do you call the coordination of data management disciplines in your organization?

We don't have a formal name for it. | 66%
We have a name for it. | 34%

Figure 1. Based on 179 respondents.

Related Terms and Concepts

Most likely, you've never heard the term "unified data management." After all, most organizations coordinating diverse data management disciplines do so without giving their actions a name. For example, the survey for this report asked: "What do you call the coordination of data management disciplines in your organization?" Two-thirds (66%) of survey respondents answered: "We don't have a formal name for it." (See Figure 1.) Likewise, in the user interviews conducted for this report, users and consultants alike described how they coordinate data management work and align it with stated business goals for data—but few had a name for it. Even stranger, most software vendors that offer a portfolio of multiple data management tools have no term for the coordinated use of the portfolio!

However, a third of survey respondents (34% in Figure 1) have given coordinated data management a name, and they typed that name into the online survey. The names they report using reveal much about how users think about such coordination. (See Figure 2.)

Generic terms for UDM. A lot of users keep it simple by referring to their coordinated efforts as simply data management (16%) or information management (11%). In fact, most users interviewed for this report stated that

they just assume that good data management involves technical people of diverse disciplines learning from each other, complying with data and development standards, considering cross-discipline architectures, and the other best practices this report associates with UDM.

Generic terms, but with enterprise aspirations. If you put the word "enterprise" in front of common terms like data management and information management (thus denoting a broad enterprise scope), then you get two of the most popular terms entered into this report's survey: enterprise data management (EDM, 15%)[1] and enterprise information management (EIM, 11%)[2]. By coincidence, these acronyms are strongly associated with vendors that have promoted them, namely EIM with SAP and Business Objects and EDM with DataFlux and SAS.

[1] See the 2009 TDWI Monograph *Enterprise Information Management: In Support of Operational, Analytic, and Governance Initiatives*, online at tdwi.org/research/monographs.

[2] See the 2009 TDWI Checklist Report *Enterprise Data Management*, online at tdwi.org/research/checklists.

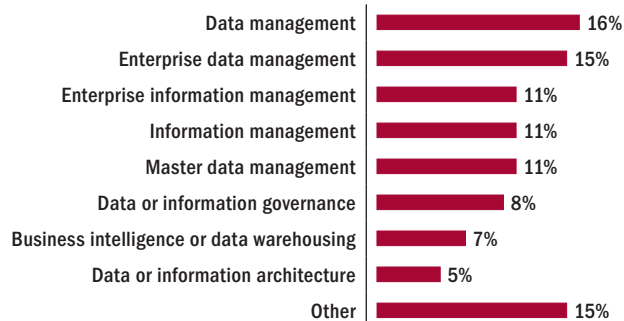**Common terms for UDM, according to survey respondents**

Data management	16%
Enterprise data management	15%
Enterprise information management	11%
Information management	11%
Master data management	11%
Data or information governance	8%
Business intelligence or data warehousing	7%
Data or information architecture	5%
Other	15%

Figure 2. Based on 61 respondents.

UDM as a subset of other programs. A number of survey respondents called their UDM-like activities by the names of other data management practices, in particular master data management (MDM, 11%), data or information governance (8%), BI or data warehousing (7%), and data or information architecture (5%). Each is a rather broad program, and each typically involves multiple data management practices. So it's possible that users responding to the survey were thinking of UDM as a subset that helps unify the data management solutions created and maintained by these programs.

Why Care about UDM Now?

There are many reasons why organizations need to step up their efforts with UDM:

Technology drivers. From a technology viewpoint, the lack of coordination among data management disciplines leads to redundant staffing and limited developer productivity. Even worse, competing data management solutions can inhibit data's quality, integrity, consistency, standards, scalability, architecture, and so on. On the upside, UDM fosters greater developer productivity, cross-system data standards, cross-tool architectures, cross-team design and development synergies, and assuring data's integrity and

lineage as it travels across multiple organizations and technology platforms.

Business drivers. From a business viewpoint, data-driven business initiatives (including BI, CRM, regulatory compliance, and business operations) suffer due to low data quality and incomplete information, inconsistent data definitions, noncompliant data, and uncontrolled data usage. UDM helps avoid these problems, plus it enables "big picture" data-driven business methods such as data governance, data security and privacy, operational excellence, better decision making, and leveraging data as an organizational asset.

Philip Russom is the senior manager of TDWI Research at The Data Warehousing Institute, where he oversees many of TDWI's research-oriented publications, services, and events. He can be reached at prussom@tdwi.org.

Solution Providers

The following solution providers have shared their data integration and data warehousing stories and successes, technology insights, and the lessons they have learned for *What Works in Data Integration*.

Aster Data Systems, Inc.

959 Skyway Road, Suite 100
San Carlos, CA 94070

+1 650.232.4400
Fax: +1 650.232.4405

info@asterdata.com
www.asterdata.com

Aster Data is a proven leader in big data management and big data analysis for data-driven applications. Aster Data's *n*Cluster is the first MPP data warehouse architecture that allows applications to be fully embedded within the database engine to enable ultrafast, deep analysis of massive data sets.

Aster Data's unique "applications-within™" approach allows application logic to exist and execute with the data itself. Termed a "massively parallel data-application server," Aster Data's solution effectively uses Aster's patent-pending SQL-MapReduce with parallelized data processing and applications to address the big data challenge.

Companies using Aster Data include Coremetrics, MySpace, Akamai, ShareThis, and Full Tilt Poker.

Attunity

70 Blanchard Road
Burlington, MA 01803

+1 781.328.0960
Fax: +1 877.896.2760

sales@attunity.com
www.attunity.com

Attunity, the leading independent provider of best-of-breed CDC technology, offers real-time data integration and event capture solutions. Using a log-based CDC model, Attunity's award-winning software provides a complete, easy-to-use, and affordable way to replicate and integrate source data efficiently and in real time. Capturing and delivering only the changes made to enterprise data sources, our solutions enable efficient data integration projects, including MDM, data warehousing, synchronization, replication, business intelligence/reporting, and legacy modernization/SOA.

Attunity's software has been deployed at thousands of organizations worldwide and is the choice of partners at leading companies, including Microsoft, Oracle, IBM, and HP. For more information, visit www.attunity.com.

Baseline Consulting Group

15300 Ventura Boulevard, Suite 523
Sherman Oaks, CA 91403

818.906.7638
Fax: 818.907.6012

inquiries@baseline-consulting.com
www.baseline-consulting.com

Baseline Consulting, an acknowledged leader in information design and deployment, helps companies enhance the value of their enterprise data, improve business performance, and achieve self-sufficiency in managing data as a corporate asset. Baseline provides business consulting and technical implementation services in four practice areas: data warehousing, data integration, business analytics, and data governance. Baseline brings certainty to results by blending its deep technical expertise in systems and data with its understanding of business processes and organizations. Founded in 1991 and headquartered in Los Angeles, California, Baseline changes how companies leverage information. To learn more, visit Baseline's Web site at www.baseline-consulting.com.

CA, Inc.

One CA Plaza
Islandia, NY 11749

800.225.5224
Fax: +1 631.342.6800

www.erwin.com

The CA ERwin® Modeling family has been the name associated with data modeling for years. Today, we offer more than just data modeling, and we've expanded our offerings to include data profiling, model validation, process modeling, ERP integration, metadata management, and model management. CA (Nasdaq: CA), the world's leading independent information technology (IT) management software company, provides software to unify and simplify the management of complex computing environments. Today, we are working with the majority of the Forbes Global 2000, as well as government organizations and thousands of companies in diverse industries worldwide, to transform IT management.

Cisco

170 W. Tasman Drive
San Jose, CA 95134

info@tidalsoftware.com
www.tidalsoftware.com
www.cisco.com

Business intelligence requires complex scheduling and processing for data to be accessed in the proper order, and processing to be completed within the operating window. Tidal Enterprise Scheduler's advanced BI integrations enable you to build and automate complex data flows across systems and applications to ensure the ETL processes that feed your BI systems are completed correctly and in the appropriate sequence each and every time, and that the reports are delivered on time, every time. It goes beyond managing ETL complexity to providing integration, recovery, and auditability—far more than is available with basic tools or scripting solutions.

DataFlux Corporation

940 NW Cary Parkway, Suite 201
Cary, NC, 27513

877.846.3589
Fax: 1.919.447.3100

info.us@dataflux.com
www.dataflux.com

DataFlux enables business agility and IT efficiency by providing innovative data management technology and services that transform data into a strategic asset. A wholly owned subsidiary of SAS (www.sas.com), DataFlux helps organizations manage critical aspects of data through unified technologies and expertise that provides the benefits of data quality, data integration, and master data management (MDM). To learn more, visit www.dataflux.com.

Informatica

100 Cardinal Way
Redwood City, CA 94063

650.385.5000
Toll Free: 1.800.653.387
Fax: 650.385.5500

http://vip.informatica.com/?elqPURLPage=483
www.informatica.com

About Informatica MDM
Informatica MDM empowers business users to improve their operations with reliable views of critical master data distributed across data sources. The award-winning solution provides comprehensive, unified, open, and economical master data management (MDM) on a single platform. It enables customers to manage multiple data domains and architectural styles and unifies all MDM requirements—data integration, profiling, quality, and master data management—on the same platform. Informatica MDM provides open data integration to all heterogeneous applications and data sources and delivers faster time to value, lower TCO, and superior ROI because it can be rapidly implemented and is easily configured to quickly accommodate ever-changing business needs.

Initiate, an IBM company

200 West Madison, Suite 2300
Chicago, IL 60606

800.992.0300
Fax: 312.759.5030

info@initiate.com
www.initiate.com

Initiate, an IBM Company, enables organizations to confidently share critical data assets. Multinational corporations, healthcare organizations, and government agencies rely on Initiate® software and services to deliver complete, accurate, and real-time views of data spread across multiple systems, domains, or databases. Initiate's multi-domain master data management (MDM) solutions facilitate interoperability, information sharing, entity resolution, and data quality to increase revenues, reduce costs, and mitigate risks. Initiate was acquired by IBM in March 2010. For more information, visit www.initiate.com.

iWay Software

Two Penn Plaza
New York, NY 10121

212.330.1700
Fax: 212.564.1726

info@iwaysoftware.com
www.iwaysoftware.com

iWay Software's integration methodology is simple: it uses a single, integrated set of graphical design tools to assemble powerful prebuilt components for enterprise-class business-to-business (B2B), business process automation (BPA), or enterprise information management (EIM) integration scenarios—without the use of custom code. Integration configurations can be deployed in a standalone manner to any environment supporting a JVM—or to Web application servers from any vendor. iWay delivers the fastest, most cost-effective, and simplest way to integrate and streamline critical business processes. iWay Software's vendor-, platform-, and protocol-neutral solutions achieve the promise of SOA: true reusability.

Melissa Data

22382 Avenida Empresa
Rancho Santa Margarita, CA 92688

1.800.MELISSA
Fax: 949.589.5211

sales@melissadata.com
www.melissadata.com

Since 1985, Melissa Data has provided scalable data quality and enrichment solutions that help SMB to *Fortune* 500 organizations verify, update, standardize, dedupe, and enrich their CRM data assets across the enterprise. Our multiplatform Data Quality Suite and Web services are employed by more than 5,000 customers to keep data clean and up-to-date for immediate ROI and significant cost savings. The newest product offering is the Total Data Quality Integration Toolkit (TDQ-IT), an affordable, single platform solution for enterprise data quality and MDM initiatives. TDQ-IT utilizes the native integration of Microsoft's SQL Server Integration Services (SSIS) to deliver a full range of data transformation, cleansing, matching, and enrichment functionality. Melissa Data offers free trials, unlimited tech support, and a four-month guaranteed ROI program.

Oracle Corporation

500 Oracle Parkway
Redwood Shores, CA 94065

1-800-ORACLE-1

oraclesales_us@oracle.com
www.oracle.com

Oracle data integration provides a fully unified set of products for building, deploying, and managing real-time data-centric architectures for operational and analytical data integration across the enterprise. Oracle's products combine to provide all the elements of data integration—real-time data movement, transformation, synchronization, data quality, data management, and data services—to ensure that information is timely, accurate, and consistent across complex systems.

Oracle (Nasdaq: ORCL) is the world's most complete, open, and integrated business software and hardware systems company. For more information about Oracle, visit www.oracle.com.

Syncsort

50 Tice Boulevard
Woodcliff Lake, NJ 07677

201.930.9700
Fax: 201.930.8290

info@syncsort.com
www.syncsort.com

Syncsort is a global software company that helps the world's most successful organizations rethink the economics of their data. Syncsort provides extreme performance for data integration and data protection solutions that transform decision making and deliver more profitable results to more than 90 percent of the *Fortune* 100 and more than 6,000 customers worldwide. For more information, visit www.syncsort.com.

Talend

105 Fremont Avenue, Suite F
Los Altos, CA 94022

714.786.8140
Fax: 714.786.8139

info@talend.com
www.talend.com

Talend is the recognized market leader in open source data integration. After three years of intense research and development investment, and with solid financial backing from leading investment firms, Talend revolutionized the world of data integration when it released the first version of Talend Open Studio in 2006.

Talend's solution portfolio includes data integration (operational data integration and ETL for business intelligence), data quality, and master data management (MDM). Unlike proprietary, closed solutions, which can only be afforded by the largest and wealthiest organizations, Talend makes data integration solutions available to organizations of all sizes, and for all integration needs. Visit us at www.talend.com.

About TDWI

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence and data warehousing industry. TDWI is a comprehensive resource for industry information and professional development opportunities. TDWI sponsors and promotes quarterly World Conferences, regional seminars, onsite courses, a worldwide Membership program, business intelligence certification, resourceful publications, industry news, an in-depth research program, and a comprehensive Web site: tdwi.org

MEMBERSHIP

tdwi.org/membership

In a challenging and ever-changing business intelligence and data warehousing environment, TDWI Membership offers a cost-effective solution for maintaining your competitive edge. TDWI will provide you with a comprehensive and constantly growing selection of industry research, news and information, online resources, and peer networking opportunities developed exclusively for its Members. TDWI offers a cost-effective way to keep your entire team current on the latest trends and technologies. TDWI's Team Membership program provides significant discounts to organizations that register individuals as TDWI Team Members.

WORLD CONFERENCES

tdwi.org/conferences

TDWI World Conferences provide a unique opportunity to learn from world-class instructors, participate in one-on-one sessions with industry gurus, peruse hype-free exhibits, and network with peers. Each six-day conference features a wide range of content that can help business intelligence and data warehousing professionals deploy and harness business intelligence on an enterprisewide scale.

SEMINAR SERIES

tdwi.org/seminars

TDWI Seminars offer a broad range of courses focused on the skills and techniques at the heart of successful business intelligence and data warehousing implementations. The small class sizes and unique format of TDWI Seminars provide a high-impact learning experience with significant student-teacher interactivity. TDWI Seminars are offered at locations throughout the United States and Canada.

CHAPTERS

tdwi.org/chapters

TDWI sponsors chapters in regions throughout the world to foster education and networking at the local level among business intelligence and data warehousing professionals. Chapter meetings are open to any BI/DW professional. Please visit our Web site to find a local chapter in your area.

ONSITE EDUCATION

tdwi.org/onsite

TDWI Onsite Education brings TDWI courses to customer sites and offers training for all experience levels. Everyone involved gains a common knowledge base and learns in support of the same corporate objectives. Training can be tailored to meet specific business needs and can incorporate organization-specific information.

CERTIFIED BUSINESS INTELLIGENCE PROFESSIONAL (CBIP)

tdwi.org/cbip

Convey your experience, knowledge, and expertise with a credential respected by employers and colleagues alike. CBIP is an exam-based certification program that tests industry knowledge, skills, and experience within five areas of specialization—providing the most meaningful and credible certification available in the industry.

WEBINAR SERIES

tdwi.org/webinars

TDWI Webinars deliver unbiased information on pertinent issues in the business intelligence and data warehousing industry. Each live Webinar is roughly one hour in length and includes an interactive question-and-answer session following the presentation.

BI Training Solutions:
As Close as Your Conference Room

We know you can't always send people to training, especially in today's economy. So TDWI Onsite Education brings the training to you. The same great instructors, the same great BI/DW education as a TDWI event—brought to your own conference room at an affordable rate.

It's just that easy. Your location, our instructors, your team.

Contact Yvonne Baho at 978.582.7105 or ybaho@tdwi.org for more information.
www.tdwi.org/onsite

tdwi
ONSITE EDUCATION

TDWI Partner Members

These solution providers have joined TDWI as special Partner Members and share TDWI's strong commitment to quality and content in education and knowledge transfer for business intelligence and data warehousing.

tdwi PARTNER

baseline CONSULTING

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

DATAFLUX
A sas COMPANY

DELL Services

IBM

INFORMATICA
The Data Integration Company™

Microsoft®

MicroStrategy®

NETEZZA

SAP®

Ssas

TERADATA®