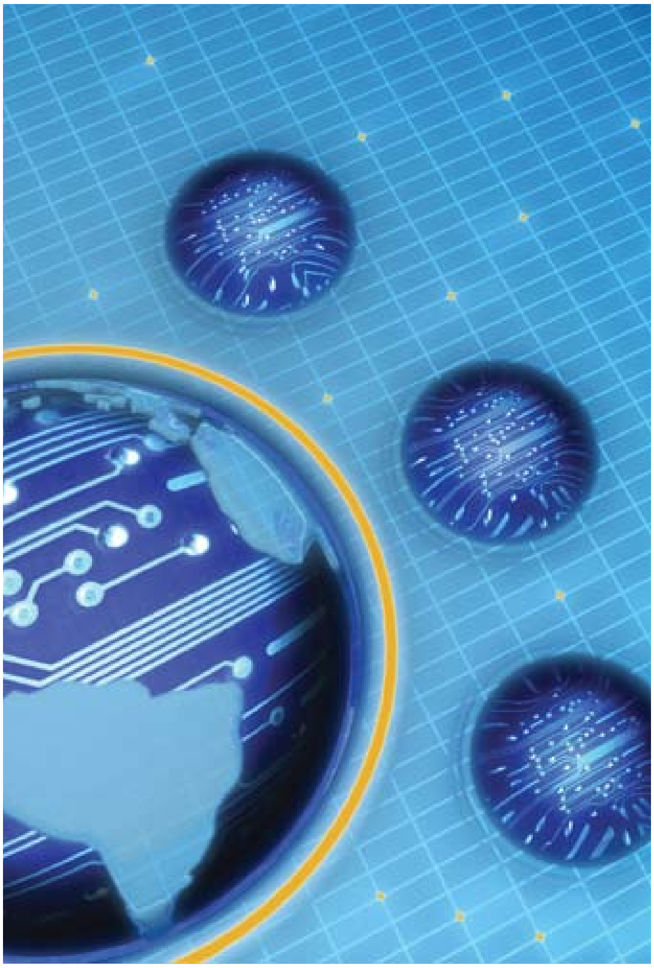
EXCLUSIVELY FOR TDWI MEMBERS

Volume 13 • Number 4 • 4th Quarter 2008

Business Intelligence Journal

THE LEADING PUBLICATION FOR BUSINESS INTELLIGENCE AND DATA WAREHOUSING PROFESSIONALS



tdwi
THE DATA WAREHOUSING INSTITUTE

VOLUME 13 • NUMBER 4

Business Intelligence *Journal*

Business Intelligence
JOURNAL

VOLUME 13 • NUMBER 4

www.tdwi.org

Advertising Sales: Scott Geissler, sgeissler@tdwi.org, 248.658.6365.

List Rentals: 1105 Media, Inc., offers numerous e-mail, postal, and telemarketing lists targeting business intelligence and data warehousing professionals, as well as other high-tech markets. For more information, please contact our list manager, Merit Direct, at 914.368.1000 or www.meritdirect.com.

Reprints: For single article reprints (in minimum quantities of 250–500), e-prints, plaques and posters contact: PARS International, Phone: 212.221.9595, E-mail: 1105reprints@parsintl.com, www.magreprints.com/QuickQuote.asp

tdwi

General Manager	Rich Zbylut
Art Director	Deirdre Hoffman
Graphic Designer	Bill Grimmer

1105 MEDIA INC

President & Chief Executive Officer	Neal Vitale
Senior Vice President & Chief Financial Officer	Richard Vitale
Executive Vice President	Michael J. Valenti
President, Events	Dick Blouin
Vice President, Finance & Administration	Christopher M. Coates
Vice President, Information Technology & Web Operations	Erik A. Lindgren
Vice President, Attendee Marketing	Carmel McDonagh
Vice President, Print & Online Production	Mary Ann Paniccia
Chairman of the Board	Jeffrey S. Klein

Reaching the staff
Staff may be reached via e-mail, telephone, fax, or mail.

E-mail: To e-mail any member of the staff, please use the following form: FirstinitialLastname@1105media.com

Renton office (weekdays, 8:30 a.m.–5:00 p.m. PT)
Telephone 425.277.9126; Fax 425.687.2842
1201 Monster Road SW, Suite 250, Renton, WA 98057

Corporate office (weekdays, 8:30 a.m. – 5:30 p.m. PT)
Telephone 818.734.1520; Fax 818.734.1528
9121 Oakdale Avenue, Suite 101, Chatsworth, CA 91311

Business Intelligence Journal
(article submission inquiries)
Jennifer Agee
425.277.9239
Fax: 425.687.2842
E-mail: jagee@tdwi.org
www.tdwi.org/journalsubmissions

TDWI Membership
(inquiries & changes of address)
Margaret Ikeda
425.226.3053
Fax: 425.687.2842
E-mail: membership@tdwi.org
www.tdwi.org/membership

From the Editor

The more things change, the more BI practitioners adjust. It isn't easy sometimes, and often we find ourselves kicking and screaming in protest. In this issue we look at a variety of changes in store for anyone supporting or using BI technology.

Take data, for example. We're all used to the speed (and, to me, surprising accuracy) of Google searches. Although such user-friendly searches are the Next Big Thing in BI applications, David Besemer argues that when it comes to today's enormous data volumes, users may feel they have left the familiar world of easy queries behind. Besemer explains what change is needed. Change is also on Lis Strenger's mind; she looks at how enterprises are adjusting so "big data" doesn't become a big problem.

Many BI professionals are used to alerts when something goes wrong. Nari Kannan takes that idea further, explaining that we can longer be satisfied with traditional BI that looks at the past. Instead, we need to change and start looking *ahead* so we can *prevent* problems. This isn't just real-time BI—this is forward-thinking BI.

Speaking of real-time, Bill Jacobs looks at the tricky business of database replication and how you can turn a large database from a data jailhouse to a data powerhouse. Replication isn't just for disaster recovery anymore.

Data quality programs have been set up in most enterprises, but Dan Sandler advocates you grow your data quality program, starting by establishing a framework that promotes data quality from the top down. He describes the four elements that ensure a sturdy platform that will support data quality initiatives throughout the enterprise.

Change is also the subject of our case studies: we look at how a police department improved how staff retrieves information and how a leading pharmaceutical company tapped a model-driven data warehousing infrastructure to revamp and revive an overworked BI system. Kirby Lunger challenges our thinking by debunking three myths of pervasive BI. Senior manager at TDWI Research Philip Russom looks at how enterprises are focusing ever more carefully on customer data integration.

Tim Wormus argues that traditional BI is becoming obsolete, and that we need to prepare for hybrid systems that will be embedded into business processes, and senior editor Hugh J. Watson looks at the changes business schools need to make to remain relevant to future BI practitioners.

We're always interested in your comments about our publication and specific articles you've enjoyed. Please send me your comments and feedback: jpowell@1105media.com.

James E. Powell

Business Schools Need to Change What They Teach

Hugh J. Watson

Hugh J. Watson holds a C. Herman and Mary Virginia Terry Chair of Business Administration in the Terry College of Business at the University of Georgia. He is a Fellow of TDWI and the senior editor of the *Business Intelligence Journal*.

hwatson@uga.edu

Several years ago I helped write a case study about a bank that had implemented an innovative customer intimacy strategy (Cooper, et al, 2000). To support this strategy, the bank had integrated customer transaction, demographic, and behavioral data; analyzed the data to identify market segments; determined the profitability of every customer and product; and optimized its distribution channels. The bank had moved from intuitive to fact-based decision making using analytics.

It was interesting to learn from the CEO that this new way of running the bank was accompanied by a large turnover in personnel. Before the change, the marketing department had 12 marketing analysts, and afterward there were still a dozen, but none from the original team—all the workers had changed positions or left the bank. As the CEO described it, “The original group thought of marketing as giving out ‘balloons and suckers’ and running focus groups and were unwilling or unable to adapt to the new demands of their jobs.”

In this and other organizations, the “good jobs” are changing. As BI becomes more important and pervasive in companies, people must have a different skill set. They need to understand how data is stored and be able to access and analyze it using a variety of tools.

Universities are behind the curve in recognizing and responding to this change. This is true even of business schools, where you would think they would be fast to respond. As I talk to my colleagues around the country, I find that most B-schools require proficiency only in

Microsoft Office applications, and even then there may be little or limited work with Access. Typically, unless students are majoring in management information systems (MIS), they learn little about relational databases and the wide variety of ways to access and analyze data.[1]

Typical B-School Curriculum

Unless you are a B-school graduate, you are probably unfamiliar with the curriculum. It is similar at most schools and begins with general education classes (e.g., English, mathematics, and history) during the first two years. Students planning to major in business also take courses in economics, accounting, statistics, and information systems. In their junior year, they take the common body of knowledge in business, which includes courses in management, marketing, and finance.

The MIS major typically requires about seven courses, which are taken during a student's junior and senior years. For example, at Georgia, MIS majors are required to take Java programming, data management, systems analysis and design, advanced application development, and project management. To complete their major, students can choose from courses in business intelligence, business process management, enterprise resource planning, accounting controls and security, managing the IS resource, and global IS.

All MIS majors work with relational databases and several data access tools and applications. Consequently, they are well prepared for the "good jobs." This is not the case, however, with most non-MIS business graduates. They learn only what is in the required information systems course and what is taught in their major (such as finance or marketing), and this is often not enough.

The information systems course that is required of all B-school students typically covers a smorgasbord of topics—the information-based organization, data as a corporate resource, network computing, e-business, systems analysis and design, transaction processing systems, decision support systems, enterprisewide systems, and using IT for competitive advantage. In addition, there is usually a lab where students develop their hands-on technical skills. Unfortunately, this seldom goes much further than using Microsoft Office.

The additional skills students learn depend on their major. For example, finance majors spend dozens or even hundreds of hours working with Excel. Marketing and accounting majors may elect to take database courses. Meanwhile, management majors seldom take any courses that expand their hands-on skills. Overall, the non-MIS students are not well trained in how data is stored and organized and how it can be accessed and analyzed though a variety of data access and analysis tools.

> All B-school students need to understand how data is organized in relational databases, because that is how they are going to find data when they join the working world.

What Students Really Need to Know

All B-school students need to understand how data is organized in relational databases, because that is how they are going to find data when they join the working world. They should understand that the relational model stores data in tables made up of rows and attributes; that tables are joined using primary-foreign key combinations to access specific attributes that are needed; what normalization is and why it is used; and so on. From this understanding they will gain insight into what data is available, where it is located, how it is organized,

[1] Some B-school computing programs are called "information systems" (IS) or "computer information systems" (CIS).

and how to bring it to bear to solve problems and make better decisions.

In addition to understanding tables based on entity-relationship (E-R) data models, students should be taught the star schema data model. My experience is that this is trickier than it seems at first. I'm always surprised by how difficult it is for students first trained on E-R data models to understand star schemas, even though ease of understanding is a touted feature. My guess is that students trained first on star schemas would find it at least as difficult to then understand E-R data models. Regardless, to fully understand how data can be stored in databases, students need to be familiar with both approaches.

> In addition to understanding tables based on entity-relationship (E-R) data models, students should be taught the star schema data model. This is trickier than it seems at first.

K. Liddell Avery and I conducted a study several years ago in which we asked TDWI Best Practices Award winners about their BI training programs (Avery and Watson, 2004). In the study, we asked the companies about the biggest shortcoming in their programs. At the top of the list was end users' lack of understanding about what data is available to them and how it is organized. The companies' solution was to provide training programs in which end users were given assistance with projects they brought to class. With this approach, the end users met an actual need, and in the process became more familiar with warehouse data and how it could be accessed and analyzed.

When I teach my students how to access and analyze data, I start out by telling them that there are many approaches; several tools can be used; and there is a simplicity/flexibility trade-off. The simplest tools are the least flexible, while the most flexible tools are more complex, forming a continuum of options.

The most flexible (but most difficult) approach is writing a SQL query. If the data is available, you (or someone else) can probably write a query to access and analyze it. However, it may not be easy, depending on the query. Although we should teach MIS majors to write complex SQL queries, it is not realistic to have the same expectations of non-majors. Sub-queries, for example, are challenging to understand, especially for people with little technical aptitude or inclination. However, I do think that it is realistic and important that all B-school graduates be able to write a simple query that selects the attributes to be output, identifies the tables to be used, specifies the necessary joins, and puts conditions on the results.

Students must also understand managed query environments, exemplified by products such as MicroStrategy, Cognos, Business Objects, and Hyperion. For example, I use MicroStrategy in my classes. With these products, students learn about dimensions, measures, drill-down, roll-up, and how to "point and click" the user interface to get what they need. My experience is that it takes about two to four hours for a good student to become proficient using one of these products. Because many students will use one or more of these products in the workplace, they should understand how the tools access data, and be able to use the tools to generate needed information.

Knowledge of dashboards and scorecards is also vital. Because they are designed to be easy to use, there is little training involved in these visual tools. People quickly pick up on the various ways that data is presented (e.g., chart or gauge), the traffic light metaphor (i.e., red/yellow/green) used for highlighting exceptional conditions, and how to drill down to underlying data. To keep things interesting, I'll normally have students react to some of Stephen Few's (2005) innovative ideas for designing dashboards.

Finally, students should be familiar with executive information systems (EIS), the epitome of no-training, point-and-click access to information. While these systems have morphed (to some extent) into dashboards and scorecards, and although most EISes have advanced functionality such as drill-down and roll-up, they are useful for identifying the upper end point on the flexibility/simplicity continuum.

Conclusion

Though these hands-on skills (e.g., writing basic SQL, using a dashboard) can be taught in either the required information systems course or courses in a student's major, it is likely that MIS faculty will have to lead the way. Only a few non-MIS faculty have the perspective, skills, and inclination to work this material into their courses without some help. It is up to MIS faculty to make sure that the next generation of B-school graduates has a better understanding of data and the many ways that it can be accessed and analyzed.

This change will not be easy to achieve. It may require taking topics out of the required IS course to make room for the new material. Adding a hands-on course to the curriculum is one option, but the B-school curriculum is typically a "zero-sum game" and whenever a course is added, another course must be removed. There will be staunch supporters for the courses currently in place.

Another alternative is to get non-MIS faculty to work the materials into their courses. However, this requires getting the silos (i.e., academic departments) to work collaboratively, and this is at least as challenging in business school as it is in business organizations. ■

References

Avery, K.L., and Hugh J. Watson [2004]. "Training Data Warehouse End Users," *Business Intelligence Journal*, Vol. 9, No. 4, pp. 40–51. http://www.tdwi.org/research/display.aspx?ID=7326

Cooper, Brian L., Hugh J. Watson, Barbara H. Wixom, and Dale L. Goodhue [2000]. "Data Warehousing Supports Corporate Strategy at First American Corporation," *MIS Quarterly*, December, pp. 547–567. http://www.misq.org/archivist/vol/no24/issue4/cooper.html

Few, Stephen [2005]. "Dashboard Design: Beyond Meters, Gauges, and Traffic Lights," *Business Intelligence Journal*, Vol. 10, No. 1, pp. 18–24. http://www.tdwi.org/research/display.aspx?ID=7487

Where Structured Data Search Fails

Why Relationship Discovery is Critical When Working with Structured Data

David Besemer

David Besemer is CTO of Composite Software, Inc. Previously, Besemer was CTO of eStyle; headed software product marketing at NeXT Computer; built program trading systems on Wall Street; and researched natural language processing systems at GE's corporate R&D center.
dbesemer@compositesw.com

Why Structured Data Search is Critical

"Toto, I have a feeling we're not in Kansas anymore," Dorothy famously told her dog as she realized they had entered the strange new land called Oz. Similarly, when it comes to today's enormous data volumes, enterprises and government agencies may feel they have left the familiar world behind. Consider these facts:

- Data is growing by a factor of 10 every five years, at a compound annual growth rate of nearly 60 percent (Mearian, 2008)
- At a large enterprise such as Chevron, this growth rate means two terabytes of new data are created daily (Anthes, 2006)
- As a result, "typical" managers spend a couple of hours a day looking for data, yet half of the time they cannot find the information they need, although the data exists (McGee, 2007)

Data volume is not the only challenge. System complexity is accelerating along with rapid business change. Unanticipated requirements arise from new business opportunities. Unique combinations of data never before modeled or reported result from the work of savvy business analysts and a range of analytical professionals such as engineers and researchers. One-off requirements come and go in the course of day-to-day business. Sometimes there is an immediate need to combine data from a newly acquired company before the independent systems have been formally integrated into an enterprisewide transaction system and data warehouse.

The business impact of the growth of data and complexity is significant. Revenue growth, cost control, and governmental compliance are achieved using an information foundation. Traditional approaches that require large IT investments in data access, schema modeling, and BI reporting, while providing an excellent foundation and many core requirements, simply do not evolve fast enough to keep pace in this new environment. Therefore, business cannot rely on IT alone. Just as Dorothy and Toto called upon the Scarecrow, the Tin Man, and the Cowardly Lion to help them successfully navigate Oz, business is now looking to new tools to help address the explosion in data volume and complexity.

Search is one of the tools enterprises are using to help navigate the vast amounts of data. However, not all search methods are alike. From its roots as a tool for searching documents within the enterprise (and later to search HTML pages across the Internet), enterprise search has expanded to address a broad range of unstructured, semi-structured, and structured data found both inside and outside the enterprise. Enterprise users are gaining significant benefits in the efficiency and effectiveness of data gathering, and they are putting this information to good use as they make decisions and solve problems. To meet this growing need for more self-service solutions, many vendors are moving quickly to provide enabling technology.

When it comes to structured data searches, however, many of the initial technology solutions—typically based on unstructured search technology—fail to hit the mark. The reason is simple. When it comes to structured data, the structure is just as critical as the data. Search solutions must consider schema, metadata, syntax, security, and tabular format.

How Unstructured and Structured Data Search Differ

With most search interfaces, users specify one or more items of interest in a search window, and the search engine finds matching items. The items meeting the criteria specified in the query are typically sorted or ranked in a search results page that provides links to pages displaying the details. To provide matches quickly, a search engine typically collects metadata about the group of items under consideration beforehand using a process referred to as indexing. Search engines store only the indexed information and not the full content of each item, which remains in place at the source.

Most search engines support a range of unstructured, semi-structured, and structured data. However, at this point of technical maturity, no single search engine can fully support the deep requirements for each data type; each data type provides distinct challenges in terms of what is being searched, why search is being used, and how search needs to work to solve the business problem at hand. Table 1 summarizes the major differences between structured and unstructured search.

Unstructured search engines are the most mature offering and the most often used today. Ubiquitously

	STRUCTURED	UNSTRUCTURED
Data repository	Databases and file systems	Web sites and document repositories
Metadata focus	Column attributes	Keyword tags
Relationship discovery	Schema based	Keyword based
Access standards	ODBC, JDBC, SOAP	HTML, file system
Retrieval approach	High-performance federated query to source data	URL link to document or Web page
Data security	Source, application, column, row	Source and document
Detailed results display	Tabular via Web browser or Excel	Text via Web browser or word processor
Final use	Analysis and reports	Documents and e-mail

Table 1: Structured versus unstructured search

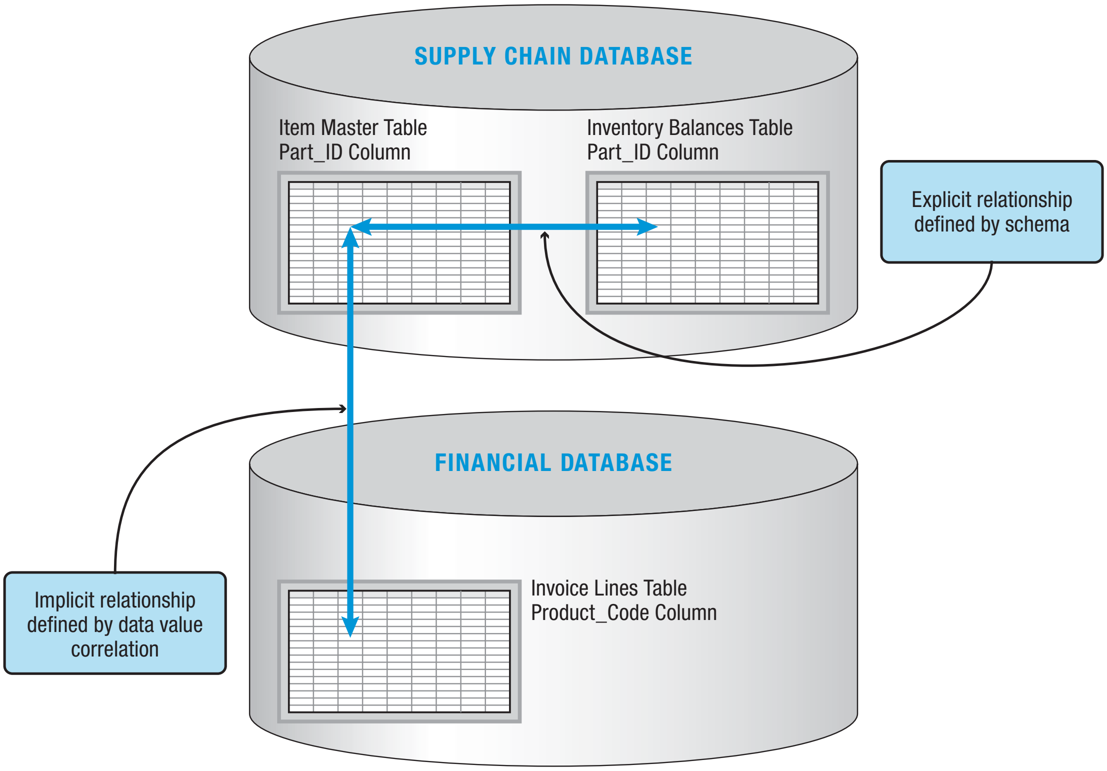


Figure 1: Explicit and implicit foreign key relationships

offered by Google, Yahoo!, and Microsoft, along with specialized offerings from other providers, these tools are optimized for text that is typically stored in documents or HTML files. Specialized capabilities focus on global reach, relevance ranking, natural language support, and other text-oriented requirements. Metadata is usually in the form of keyword tags. To index and display actual documents, unstructured search engines leverage standard HTML and text document format standards to greatly simplify the access required. Security outside the enterprise is typically controlled by the searched Web sites themselves. Inside the enterprise, file system privileges dominate, and directories such as LDAP (lightweight directory access protocol) and Active Directory typically control repository-level and, to some degree, document-level security.

Display is the easy part, because final results have already been formatted as HTML pages or actual documents. Any browser or word processor is more than adequate for displaying the results, and therefore, these viewing tools are not directly integrated in the search solution. If reuse of the data in a separate document is required, such as in an e-mail or a status report, users can simply cut and paste either the URL link or the actual text from the found source into the new document.

Structured data search engines have come to market in the past two years. These are optimized for structured data (typically stored in databases) and often used as a complement to BI applications such as reporting and ad hoc query. With structured data search, metadata about the item is as important as the item itself. Metadata in the form of a column name helps distinguish 2317 Elm St. as a shipping address, 2,317 as a number of units shipped, 23:17 p.m. as a shipment transaction time, and 2317 as a shipped item's unique ID.

Further, as any data modeler or report developer knows, when it comes to structured data, the relationships between items—such as customers, invoices, shipments, and returns—are also critical. In fact, BI has been created to fulfill businesses' need to better understand the relationships between disparate data elements. Structured data search tools also provide a means of discovering and leveraging foreign key relationships in existing schema. Indeed, next-generation structured data tools now being introduced will also discover schema from the data values themselves in multiple-database scenarios where foreign key relationships are not explicit. Figure 1 is an example of these explicit and implicit relationships.

Accessing the actual data, either when building the index or performing the live query, requires deep understanding of diverse data file formats. Oracle tables, ISAM (indexed sequential access method) files, SAP structures, and Teradata slices are the tip of the iceberg in a typical enterprise. ODBC and JDBC standards are useful, but the power of structured data search solutions can be inadvertently constrained by the breadth of sources available.

Although the performance of both the index build and the live data query are critical, unstructured and structured data searches differ in their approaches by the type of data being accessed. When accessing data from multiple systems—for example, an Oracle data warehouse and a Microsoft SQL Server data mart—structured search relies on high-performance federated query algorithms to optimize the SELECT statements generated by the search engine.

Data security is another key consideration when searching structured data. For example, customer, employee, financial, sales, marketing, and supply chain data are governed by a wide variety of individual security and compliance policies. Powerful search tools cannot ignore these requirements, nor should they introduce or require another security infrastructure. The right approach is to leverage existing source-, column-, and row-level security rules as implemented in corporate directories or packaged applications (such as Siebel, PeopleSoft, and SAP).

Finally, displaying structured data is less straightforward than displaying unstructured data. With unstructured data, the Web page or document is the end display from the search result. With structured data, users want to see and refine tabular result sets showing the many rows of data with appropriate column headers. This means providing a spreadsheet-style workspace where users can easily add, move, and delete rows and columns as they build reports. In other words, users want to produce a report that looks more like Microsoft Excel than Microsoft Word.

> Finding reports is often a challenge because most large organizations have hundreds, if not thousands, of them, and enforcing report naming conventions is difficult.

Although structured and unstructured search each has specialized tools, semi-structured data (typically in the form of XML documents and RSS feeds) may be supported by either approach.

How Structured Data Search Complements BI and Reporting

Analyst firms estimate that between 15 and 20 percent of users successfully utilize BI tools (Pettey, 2008). Structured data search complements BI in two ways, both of which should drive greater adoption of these analytic solutions, increasing business benefits.

The first way is to use search to find existing BI reports. Finding reports is often a challenge because most large organizations have hundreds, if not thousands, of them, and enforcing report naming conventions is difficult. Reports can be difficult to find even in organizations that use a hierarchical folder navigation structure or a set of predefined parameters; replacing these with search greatly simplifies the process. Of course, this assumes that the user is looking for data that is available in an existing report.

More often than not, no report exists that provides the information required. The information may be a new business requirement, a unique combination of data, or perhaps a one-off or short-term need. Prior to the availability of structured data search, the typical scenario would involve a few users—those skilled enough to effectively use ad hoc query tools—building the reports they needed. However, for the majority of users, the solution was to ask IT for help.

Structured data search provides a complete solution that includes modern search paradigms, intelligent leveraging of metadata and schema, proper data security, appropriate tabular display and refinement tools, and good integration with Microsoft Excel.

In the second approach, structured data search fills this information gap by providing data from structured sources even when a report doesn't already exist—and in a simple, self-service way that works for all business users, not just the 15 to 20 percent who are technology experts. This approach to structured data search provides an end result that is similar to ad hoc query analysis and reporting. However, it is packaged differently, leveraging new search and relationship discovery technologies that make it easier to use than earlier ad hoc query solutions.

Structured data search provides a complete solution that includes modern search paradigms, intelligent leveraging of metadata and schema, proper data security, appropriate tabular display and refinement tools, and good integration with Microsoft Excel. Together, these capabilities provide an easy and fast method for end users to retrieve and explore data with little or no IT assistance.

Why Users Need to Understand Data Relationships

The importance of data relationships is evident in any enterprise BI report. For the purposes of this article, we'll use a gross margin report as an example. Such a report combines a variety of data from several sources: revenue data from financial systems, shipment data from order management systems, cost data from supply chain systems, and so on. These varied inputs need to be correctly correlated to ensure accurate gross margin calculations.

Beyond enabling standard reports for well-known business requirements, relationships can be a critical success factor in ad hoc situations. What happens when users need to find the answers to new questions, solve unique problems, or make unanticipated decisions? The old adage, "Sometimes you don't know what you're looking for until you find it," is especially applicable.

For example, a customer service manager trying to resolve why a key customer didn't receive a shipment might need to relate order quantity to available inventory to see if the problem was caused by an out-of-stock condition. The manager might also try relating order value to credit limits to see if the problem involved a financial hold. If neither of these provides the necessary insight, the manager might relate shipments and returns to see if the problem was due to an incorrect ship-to address.

In the process of uncovering the data required, the manager would have effectively created a complex join of seven tables (customer master, open orders, order lines, on-hand inventory, credit limits, shipments, and returns) from at least three different systems (order management, supply chain, and finance).

The traditional, IT-based approach to reporting solution development has been optimized to leverage data relationships to create more complete views of data from the wide range of underlying sources. Using this approach, developers work with end users to determine

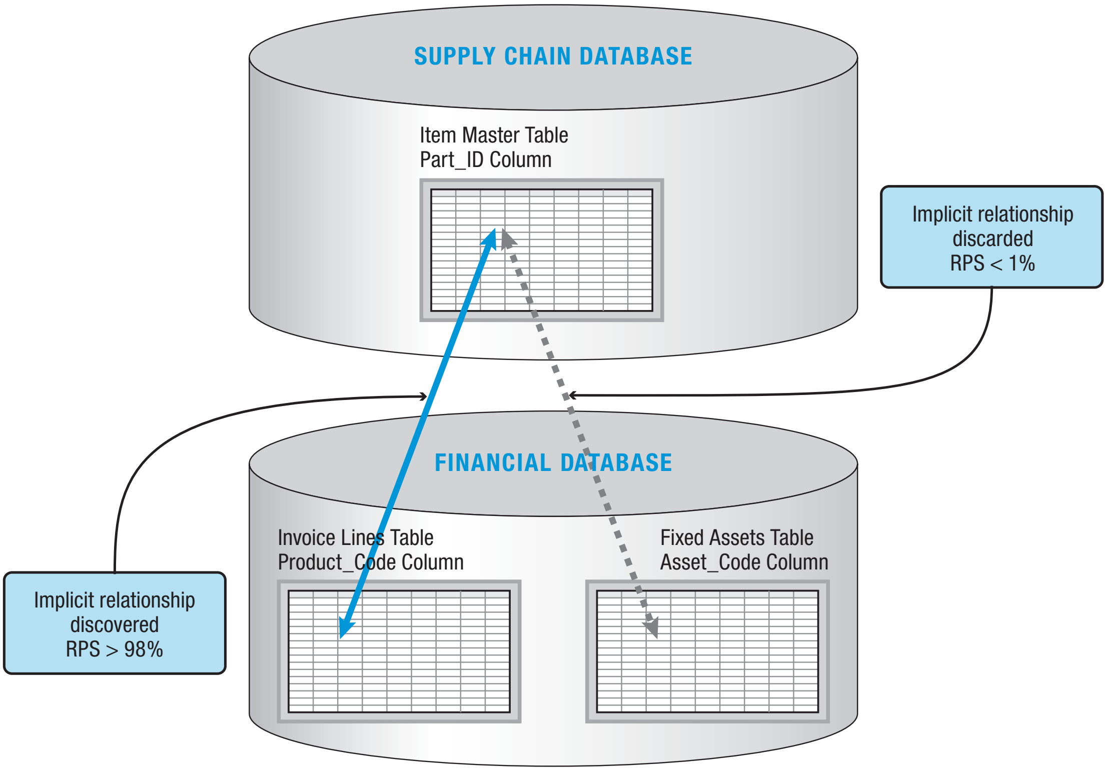


Figure 2: Relationships discovered through data value correlation

their requirements based on an agreed-upon design and development process. Developers work comfortably with raw data held in disparate sources and diverse formats. When faced with data they don't understand, they can leverage database administrators and data architects to fill in the gaps.

Developers have been trained to consider foreign key relationships between entities as well as the standard set of SQL operands used to define and manipulate them. Developers can use powerful graphical studios to design new tables, views, cubes, and even SOA-standard data services that integrate related data in a way better suited to reporting. Alternatively, they can simply code the same result in SQL, Java, or their language of choice. From there, developers can use their organization's standard analytical and reporting tools to build the end solution.

End users typically lack the training and tools enjoyed by developers. They are often uncomfortable with source data, schemas, metadata, relationships, or any of the details that underlie enterprise data. Few, if any, can perform a seven-way join. However, although they lack a developer's mindset and resources, most users are very comfortable with self-service data analysis and reporting—as long as Excel is the analysis and reporting tool.

To be effective as a complement to BI solutions that provide the structured information business professionals require, structured data search tools must make it easy for end users to work with data relationships. The best structured data search tools use automated relationship discovery to achieve this goal.

How Structured Data Search Solutions Discover Relationships

Discovering relationships is one of the biggest challenges in structured data search because two distinct approaches are required: schema-based and data-value-based. Discovering relationships within a single database is typically easy because the relationships within that database are already defined by its schema. When building its index, the search tool interprets the foreign key relationships captured by the data definition language (DDL) statements used to build the database. In essence, the search tool simply reuses the work performed by the original database designers.

> Discovering relationships within a single database is typically easy because the relationships within that database are already defined by its schema.

What happens when relationships across disparate databases are important? A common example is when supply chain management and financial data need to be searched to analyze cost overages. There are no explicit foreign key relationships already defined or easily leveraged to address this challenge. The structured search solution needs to discover the implicit relationships that already exist. Such relationship discovery uses advanced correlation algorithms that compare the actual data values.

For example, how can a relationship be discovered between part numbers in the "Part_ID" column from the item master table within the supply chain system and part numbers in the "Product_Code" column from the invoice lines table within the financial system? Probabilistic matching techniques can evaluate the data correlation and calculate a relationship probability score (RPS). The RPS between the data in the "Part_ID" column and the "Product_Code" column would be high.

In contrast, the RPS between the data in the "Part_ID" column and the "Asset_Code" column from the fixed assets table within the financial system would be extremely low. Figure 2 shows these relationships along with their RPS. Administrators determine what level of RPS will be considered a meaningful relationship. Above this bar, a foreign key relationship is automatically assumed.

Calculating the RPS between every dual column permutation across databases is a significant undertaking. However, by using a combination of intelligent algorithms and parallel processing, this work can be accomplished efficiently and effectively, even in the largest enterprises.

Evaluation Criteria for Structured Search Offerings

The goal of structured data search is to help users get the structured data they need to answer questions, solve problems, and make decisions. A complete structured data search offering needs to efficiently and effectively support the entire data gathering and analysis process from end to end.

Minimum requirements include:

- Search box entry screen to capture keywords
- High-performance search engine optimized for indexing structured data
- Access to structured data via standards such as JDBC and ODBC
- Adherence to enterprise data security and governance policies
- Summarized search results on display screens that display "hits"
- Tabular, detailed results workspace with tools for selecting, combining, sorting, moving, and removing rows and columns
- Simple export to Excel or other formats for additional analysis, formatting, etc.

- Ease of use at each step

Additional functions include:

- Relationship discovery that automatically identifies both explicit and implicit relationships
- Advanced search result display that shows relationships in addition to hits
- High-performance federated query that supports a full range of SQL operands
- Specialized data access for packaged applications such as SAP, Oracle, Salesforce.com, Siebel, etc.
- Saved searches (sometimes called "recipes") that enable reuse of the entire process from initial search to ultimate display
- Excel plug-in that enables saved searches to be run live from Excel
- Enterprise 2.0 enablement to encourage users to enrich enterprise metadata, thereby accelerating broader adoption and increasing productivity
- Rapid implementation and low total cost of ownership enabled via an appliance or software-as-a-service delivery model

Current Marketplace Offerings

At this early stage, the budding market landscape consists of vendors from adjacent market categories, including unstructured data search, BI, data integration, and data connectivity. What each of these vendors delivers varies significantly—especially when it comes to delivering data relationship discovery and display—and is correlated to how they leverage their core strengths. It is prudent to keep an open mind during product evaluation.

- Traditional unstructured search solution providers—most notably Autonomy Corporation; Endeca Technologies; FAST Search and Transfer (now a Microsoft subsidiary); and Google via its OneBox for Enterprise product—are extending their existing solutions in an effort to properly support structured data.
- Several BI vendors provide solutions that extend their strengths in structured data reporting and analysis, including Business Objects (an SAP company) with its BusinessObjects PoleStar and Information Builders with its WebFOCUS Magnify.
- Progress Software has moved beyond data connectivity with its Progress EasyAsk, which utilizes natural language processing when searching structured data.
- The Composite Discovery appliance builds on Composite Software's strengths in structured data modeling and high-performance virtual data federation (query), thereby satisfying nearly all the functional requirements listed here.

Calculating the relationship probability score between every dual column permutation across databases is a significant undertaking. However, by using a combination of intelligent algorithms and parallel processing, this work can be accomplished efficiently and effectively, even in the largest enterprises.

Best-Use Cases for Implementing Structured Data Search

Structured data search can be the fastest route from question to answer when data cannot be found in an existing report. By providing a search interface to enterprise data, users are empowered to retrieve and work with structured data, including both original source data and consolidated stores such as the warehouses and marts specifically designed to support more traditional BI and reporting.

Through its ease of use and automation (neither SQL nor report-writer skills are required), structured data search solutions open the door to a much wider set of users and use cases than earlier ad hoc reporting tools or other forms of structured data analysis. Certainly every analyst, nearly every analytical business professional (researcher, scientist, engineer, etc.), and most managers will find structured data search a valuable addition to their personal information tool kits.

However, the value to both business and IT goes beyond users getting their questions answered faster. The self-service approach creates more independence from IT. This allows IT to act as an enabler, providing data governance policies and supporting infrastructure. When its report-writing load is reduced, the IT team can focus scarce business funding and resources on other high-value projects. Finally, in cases where these ad hoc analyses prove repeatable and are used frequently, users' recipes provide an excellent report specification list that clarifies exact reporting requirements and specifies data sources, selects, and joins (among other details) to make the IT team's development efforts easier.

Conclusion

Faced with exponentially growing data volumes and complex information systems, enterprises might feel that they've left the familiar and entered a strange land called Oz. To leverage their information systems, enterprise IT teams employ BI and search solutions to help business professionals answer business questions. Current versions of BI and enterprise search tools fall short of fully empowering these business analysts and related professionals when it comes to self-service discovery of structured data and data relationships across disparate data sources.

Structured data search leverages search, relationship discovery, high-performance query, and a simple, self-service approach that contributes to bottom-line profitability. It provides business users a faster, easier way to access, work with, and benefit from the structured data they need every day, and frees critical IT resources to address other business and IT opportunities. ■

References

Anthes, Gary [2006]. "Chevron: Where Size is Opportunity," *Computerworld*, October 30. http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=265831&pageNumber=3

McGee, Marianne K. [2007]. "Managers Have Too Much Information, Do Too Little Sharing, Says Study," *InformationWeek*, January 8. http://www.informationweek.com/news/showArticle.jhtml?articleID=196801833

Mearian, Lucas [2008]. "Study: Digital Universe and its Impact Bigger than We Thought," *Computerworld*, March 11. http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9067639

Pettey, Christy [2008]. "Gartner Says Emerging Technologies will Marginalize IT's Role in Business Intelligence," Gartner press release, March 18. http://www.gartner.com/it/page.jsp?id=625810

Exception Intelligence

The Key to Timely, Specific, Actionable Business Intelligence

Nari Kannan

Nari Kannan is co-founder and CEO of Ajira, a company that designs and develops real-time business intelligence software tools.
nkannan@ajira.com

Abstract

Tracking business events that happen as expected, and those that don't happen as expected, is of immense help in keeping business processes humming along. Traditional business intelligence looks at the past; exception intelligence can provide event-based intelligence in a number of verticals and business processes. In time-bound processes such as home loan processing, events such as income verification and appraisals must happen within planned time frames so that the loan-closing deadlines are met. If any event deadline is missed, those responsible need to know so they can proactively make course corrections.

Timely, actionable exception intelligence can help people set wayward business processes back on track and move an organization toward a real-time enterprise. It can make more efficient and effective use of resources and help continuously evaluate and improve business processes.

Introduction

Increased adoption and use of business intelligence (BI) tools points the way to an "intelligence explosion," just as the collection and use of information once led to an "information explosion." With all of the BI products producing volumes of data and reports, new "exception intelligence" tools are needed to create specific, timely, actionable business intelligence and send it to only the most appropriate people. This differs from today's business intelligence tools, which generate static, postmortem reports that are distributed to all relevant people.

Anyone familiar with monitoring systems and networks in a data center environment is also familiar with the differences between network/system logs and receiving exception alerts on a pager or phone. The logs record everything of note, while exception alerts notify you only when defined exceptions occur. Apply the same principles to business in general, and you can see the differences between business intelligence and real-time, actionable exception intelligence. Business intelligence turns data into information, while real-time exception intelligence uses business rules to report business exceptions.

> Business intelligence provides aggregations of data into information. However, exception information includes things that happened as well as things that were expected but did not happen.

Webster's Revised Unabridged Dictionary defines an exception as "that which is excepted or taken out from others; a person, thing, or case, specified as distinct, or not included; as, almost every general rule has its exceptions." This captures the essence of exception intelligence and what it will achieve for organizations. Business intelligence provides aggregations of data into information. However, as the rest of this article shows, exception intelligence provides information at the right time to the right people. The exception information provided includes things that happened as well as things that were expected but did not happen. The time is right for real-time enterprises to look at exception intelligence to be provided by customizable, easy-to-deploy tools.

Here are some examples of how exception intelligence can be helpful in streamlining business processes—in some cases mandated by law, and in others for the sake of optimal internal processes:

- **Home mortgage loan processing.**
  Loan processing contains many distinct steps that all need to happen within a certain period of time—loan application completion, rates locked, appraisal received, credit check completed, employment verification completed, etc. Some of these events depend upon other events being completed, while other events may occur asynchronously at various points in time. Exception intelligence can be used to highlight business events, both when they happen and when they do *not* happen. For example, if you expected the home appraisal to be completed within 10 days of the application completion, exception intelligence can notify the appropriate person if the appraisal was or was not received.

- **Automobile accident claims processing.**
  If the original auto repair estimate has been received but the accompanying digital photos have not been uploaded within an allowable period of time, the person responsible (often a third party to the insurance company) may need an e-mail reminder that is copied to the claims processor responsible within the insurance company.

- **Debt collection processes.**
  The Fair Debt Collection Act mandates that specific steps be followed with associated time intervals for governing the debt collection process. Notices may have to be sent out on anniversary dates and if certain events happen (or do not happen) as these processes change.

- **Supply chain optimization.**
  In complex supply chain interactions, downstream transactions and their timeliness have an impact on upstream processes. Exception intelligence can notify the right people who can take action when potential bottlenecks are identified on the supply chain critical path.

- **Collection optimization.**
  Accounts receivables aging statements are useful for following up on payments from customers, but in practice, companies may want to monitor payments from specific customers more closely (for example, if they seem likely to declare bankruptcy). In such

cases, exception intelligence provides one more level of actionable, timely intelligence that is more useful than routine, static business intelligence.

- **Logistics.**
  Logistics is all about getting the right things to the right place at the right time. If key logistical processes fail, the right people need actionable exception intelligence at the right time to prevent or recover from the process failure.

Need for Exception Intelligence

Let us briefly look at why exception intelligence may be needed above and beyond business intelligence.

- **Realize the real-time enterprise.**
  When businesses are wondering what a real-time enterprise really means and what it can do for their business, exception intelligence seems to be the practical first set of steps toward adopting a real-time enterprise strategy. The logical first step is recognizing business exceptions in real time and responding to them in a timely manner.

- **Increasing workloads, limited time.**
  Company budgets and resources have been (or are being) cut. When workloads are increased for people in operations or sales, they have less time to read voluminous business intelligence reports and take appropriate action. A more practical solution is to send just the right intelligence at just the right time—no more, no less.

- **Priority setting.**
  Priority setting has been the key to effective management when it comes to network management or security management. Network and system administrators have used exception alerting and escalation effectively. Exception intelligence can highlight daily, actionable items that a person should have on their list of priorities. Exception intelligence extends this practice to business processes.

- **Timely communication to the right people.**
  There is considerable truth in the cliché, “when everybody is responsible, nobody really is.” Exception intelligence will help assign actionable business intelligence to the right people at the right time, so you don’t have to watch this cliché come true in your organization.

- **Continuous evaluation of business performance.**
  We see an emerging trend: increased use of business performance measurements and analytics software. Many Hyperion customers have started using alerts built on top of Hyperion reports. The more voluminous and detailed the reports become, the more the critical information inside those reports becomes blurred and sometimes difficult to decipher. However, the reports become more useful once exceptions are defined for the business performance metrics and alerts are issued.

> When businesses are wondering what a real-time enterprise really means and what it can do for their business, exception intelligence seems to be the practical first set of steps toward adopting a real-time enterprise strategy.

Implementing Exception Intelligence

One of the best ways to get started with implementing exception intelligence is to start with business processes. Events within an organization always happen in the context of business processes. Alternatively, they could be useful in the context of measures (key performance indicators) that may turn critical. The following is a road map for exception intelligence implementation:

- **Identifying the right business processes or measures.**
  Not all business processes can benefit enough from

exception intelligence to justify the costs of implementation. An order-to-cash business process may be a key area where exception intelligence can highlight bottlenecks. When these bottlenecks are addressed, the process can keep flowing. Since the process is directly connected to customer satisfaction, cash, and collections, this may be a good business process for exception intelligence implementation. On the other hand, an office supplies ordering process may not justify the expense of an implementation. In the case of measures, the daily order flow may be a key measure for highlighting exceptions.

> One of the major pitfalls in exception intelligence reporting is using it when routine reporting would have sufficed! If a weekly report can highlight the same information, exception intelligence may be superfluous.

- **Identifying the right exceptions.** The right exceptions to highlight in a business process should be ones that can result in some action. For example, if a credit verification step is overdue in a mortgage application process, sending a message to the concerned person inside or outside the organization may result in follow-up and completion of the process. Exceptions that are outside the control of the organization, on the other hand, may not be worth highlighting. For example, if an exception refers to an overdue regulatory approval, the organization may not be able to do anything about this except wait. With business processes or measures, the cost of implementation of exception intelligence should be balanced with the *actionable value* gained.

- **Adjusting the detail and periodicity of the exceptions.** Exception intelligence, when overdone, can lose its utility quickly. If you get a few e-mail messages or exception intelligence reports a day, you are likely to pay attention to them individually. If you get a few messages every hour, you are less likely to pay attention to any one of them—much less use them for targeted corrective actions. The level of detail in these exception intelligence reports also affects their usefulness. Timely, precise, and summarized brief reports or e-mail messages stand a better chance of being read and acted upon compared to voluminous reports.

- **Exception reporting should be only for exceptions.** One of the major pitfalls in exception intelligence reporting is using it when routine reporting would have sufficed! If a weekly report can highlight the same information, exception intelligence may be superfluous. Exception intelligence should be used for time-sensitive events or non-events that must be followed up with timely actions.

Implementation Best Practices

Here are some best practices for the implementation of exception intelligence:

- **Combine heterogeneous sources of information.** Exception intelligence has the maximum level of impact when it combines information from multiple back-end software systems to highlight exceptions. Individual software systems can produce periodic reports that can be the basis for actions. *When exceptions can only be highlighted with information drawn and combined from multiple back-end systems, they have much more utility and impact.* For example, in an order-to-cash business process, an exception intelligence system that combines the flow of orders by extracting information from multiple systems such as order management, financial management, and warehousing systems provides a view of the flow of orders that any single individual system would be unable to provide.

- **Leverage graphs, dial charts, and dashboards.** Exception intelligence in many cases is best communicated with graphs, dial charts, and dashboards. Text-based reporting may be suitable for business process exceptions, but exceptions related to performance measures can leverage visual tools to communicate information optimally.

- **Escalation mechanisms.** In many business processes, escalation mechanisms may need to be employed so that exceptions are addressed within allowed time periods. If the exceptions are not addressed and removed, the system should escalate them to the next level of responsibility. This can ensure that the actionable part of the intelligence is acted upon.

Implementation Technology Available

Exception intelligence can be implemented with a number of technology options:

- **Service-level management solutions.** These are the most comprehensive solutions for providing the widest variety of exception intelligence reporting. Ajira, DigitalFuel, and Oblicore are some of the companies that provide these software product solutions.

- **Real-time business intelligence solutions.** These solutions are built on top of database management systems like TIBCO or Cognos' Celequest Appliance. These are particularly useful for highly time-sensitive exception intelligence applications like rapid changes in stock prices that trigger buy or sell orders.

- **Business process orchestration solutions.** These solutions tie together many disparate software systems that are executing parts of an overall workflow. They have limited capabilities to provide exception intelligence solutions.

Summary

Hockey player Wayne Gretzky is known for saying, "Skate to where the puck is going, not where it is now." The business intelligence "puck" seems to be moving toward establishing rules on how to recognize business exceptions when they arise and quickly delivering that information to the right people in an actionable format. Exception intelligence seems to be the answer. After all, exception intelligence is nothing new to people who manage mission-critical IT infrastructure. The same concepts could be applied in any business, paving the way to a true real-time enterprise. ■

Four Elements of Successful Data Quality Programs

Building a Strategic Framework to Improve Information From the Ground Up

Dan Sandler

Dan Sandler is a principal consultant for Collaborative Consulting.
dsandler@collaborativeconsulting.com

Abstract

Most organizations that have addressed application and data integration have also launched data quality campaigns to meet the challenges head-on. These campaigns can include any combination of four elements to achieve data quality improvements: processes, technology, governance, and people.

In building and growing data quality programs, the first order of business is to establish a framework that promotes data quality from the top levels of the organization. It is important to administer this framework centrally because the four elements are integrated yet independent. The integration between the four elements ensures that a data quality safety net is cast across multiple projects, providing continuity as well as alignment with strategic business objectives. The independence promotes data improvements through ground-level activities within each element.

Establishing these four elements ensures a sturdy platform that will support data quality initiatives throughout the enterprise. This article provides practical examples that illustrate how to independently grow these elements and integrate them into a complete data quality framework.

The Four Elements: A Natural Fit

It is difficult to create a lasting data quality program without the support of processes, technology, governance, and people. Those who have faced data quality issues realize the root causes can be systemic and the solutions are typically not quick fixes. Each step toward a long-term data quality program requires ongoing effort and support.

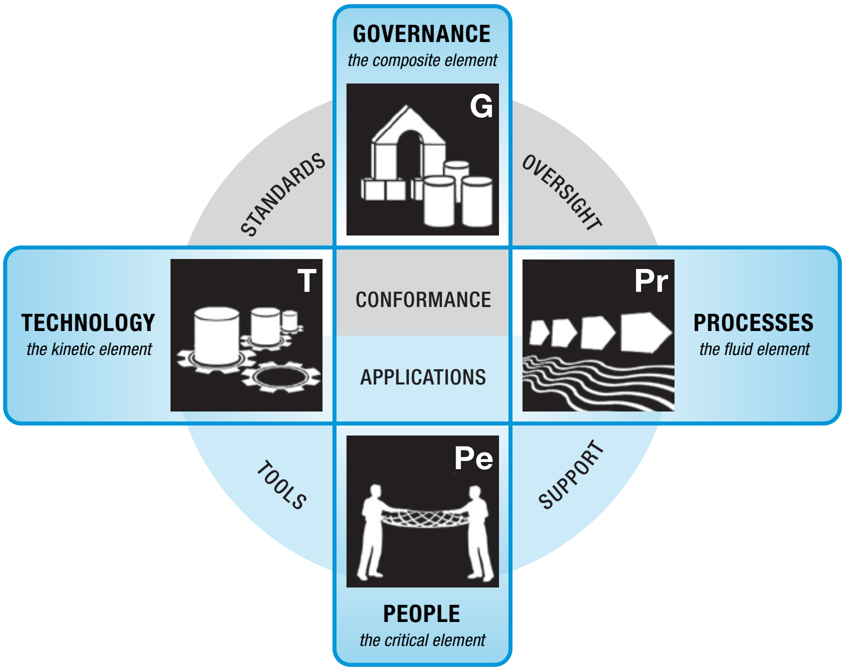


Figure 1: Process, technology, governance, and people are the four elements that support data quality initiatives throughout the enterprise

The first step toward building a data quality program is identifying the processes that require support. In some scenarios, tuning an existing business process can provide immediate data quality benefits. Where greater attention is required, the data quality program will need its own set of processes to identify and prevent issues at the source.

Once the underlying business processes have been scrutinized, the role of technology in the data quality lifecycle can be examined. Data quality software or master data management applications can help remedy data quality issues at the point of origin. Aging legacy systems may also cause data quality defects. Whether it is the cure or the disease, technology is a critical success factor in a data quality program.

Governance provides oversight and standards to the data quality program. The data standards and data definition may already exist within the organization, but a successful data quality program requires formalized governance that centralizes these standards from all areas of the organization, including business, legal, and operational levels. Governance involves more than oversight; it's also about process management that supports the people at all levels who fix the data.

Finally, people are the core element to any data quality program or initiative. Putting a data quality program into action requires ongoing effort and support. Long-term improvements will not materialize unless the data quality program is fully staffed.

These four elements are interrelated. In simple terms, governance provides data quality standards supported by technology; governance also provides oversight to business processes to ensure standards and regulations are not violated, and ensures data quality activities comply with these standards. Technology provides applications to support business processes and data quality tools that allow routine tasks to be automated. Finally, people support business processes, filling the gaps where necessary and addressing data exceptions as they occur.

Figure 1 illustrates the interplay and connectedness of the data quality program elements.

Processes: The Fluid Element

Conceptually, horizontal processes convey data quality best practices that are adapted to support specific (vertical) business processes and their functional requirements. In a data quality white paper from Collaborative Consulting, this horizontal process is represented as a data quality continuum that should seem familiar to most data practitioners (see References).

> Poor data quality results in suboptimal levels of service to prospects, leads, and customers during critical touch points in the operational marketing cycle.

The data quality continuum is an iterative lifecycle that identifies, analyzes, improves, and continuously measures quality. As Figure 2 illustrates, the data quality process begins by identifying the business drivers, then identifying relevant data. The data is then profiled, standards are established, and the data is further evaluated against the standards. Next, corrective action is determined and implemented, ideally resulting in enhanced and consolidated data. Finally, data quality improvements are measured. Based on the progress reflected in the measurements, a new iteration is generated.

Once we understand these horizontal processes, we can explore the many vertical business processes that require data quality support. As an example, Figure 2 illustrates the connection between the horizontal data quality processes and vertical business processes common to operational marketing.

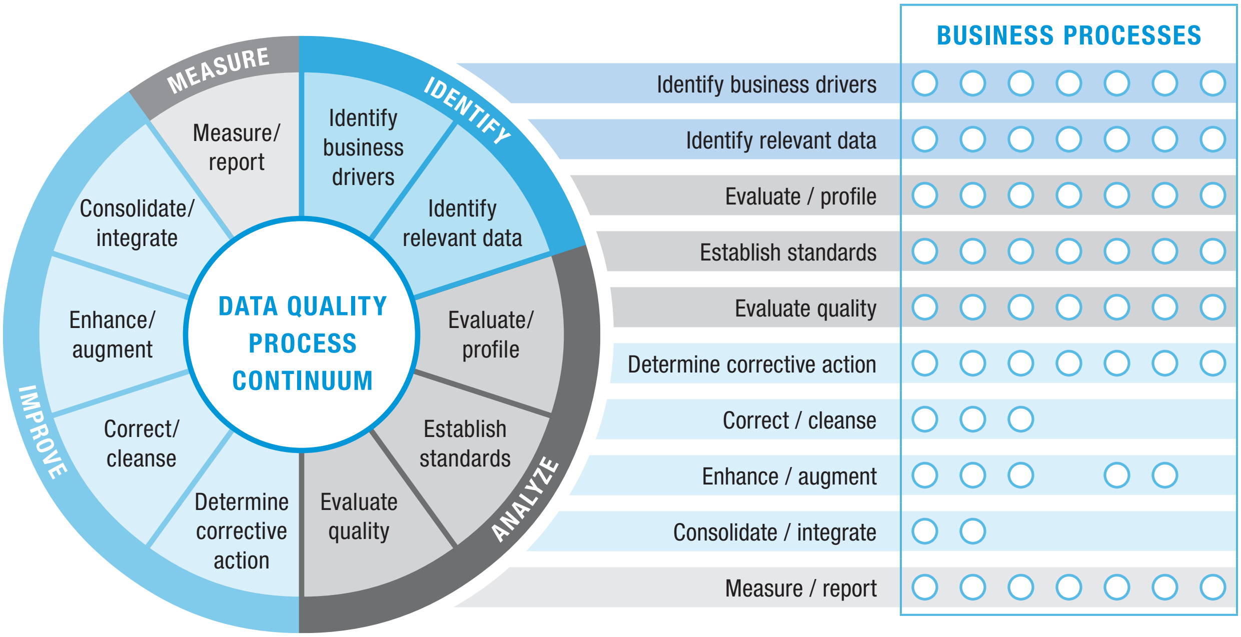
Briefly, the operational marketing processes begin with campaign design, execution, and monitoring (also known as channel, contact, and response management). Once designed (based on customer profiles and past behavioral patterns), campaigns generate marketing and lead-relevant activities, which are subsequently managed. For lead-relevant activities, there is a handoff from the campaign to a lead qualification process. From the leads come potential marketing and sales opportunities, which must be assessed for response potential. From there, the opportunities are primed and enabled for marketing and sales.

If we take this vertical slice of marketing-related business processes and align the steps with the data quality horizontal processes, there is an overlap. This overlap is not a full intersection, but nevertheless, the core data quality elements make an optimized process possible. One best practice is to examine the horizontal streams and identify business drivers, analyze data quality, and determine corrective action across all marketing processes.

Beyond these data quality processes, the overlap is sparser. The need to cleanse, match, and consolidate data is greatest early in the operational marketing lifecycle, when master data for organizations, contacts, and corporate subsidiaries is collected. The cost of merging and unifying customer data increases as time passes and more child company transactions must be tied to now-duplicated customer records. More important, poor data quality results in suboptimal levels of service to prospects, leads, and customers during critical touch points in the operational marketing cycle.

For example, during lead qualification, customer data may be enhanced with data from external sources, such as competitor data, customer profiles, product/service gap assessment, company data (head count, annual revenue, and facts about the parent, child, and sibling companies to exploit global opportunities), and opportunity funding. Such relevant information sets the stage for the next business process; an enterprise need not play catch-up during the sales cycle.

A larger point should be recognized. There is a partial intersection between the data quality horizontal components and the marketing vertical processes. *This intersection is purely a function of the vertical business process being engineered.* Figure 2 illustrates this overlap by indicating where the data quality process supports the business.



	BUSINESS PROCESSES						
Identify business drivers	○	○	○	○	○	○	○
Identify relevant data	○	○	○	○	○	○	○
Evaluate / profile	○	○	○	○	○	○	○
Establish standards	○	○	○	○	○	○	○
Evaluate quality	○	○	○	○	○	○	○
Determine corrective action	○	○	○	○	○	○	○
Correct / cleanse	○	○	○				
Enhance / augment	○	○	○		○	○	
Consolidate / integrate	○	○					
Measure / report	○	○	○	○	○	○	○

Figure 2: The partial intersection between a data quality program's horizontal components and the verticals of business process

It is important to identify where the X and Y axes intersect when engineering the vertical process—prior to architecting the solution. The vertical process itself has a profound influence on the technology in terms of applications and tools. As the next section will illustrate, the degree of matching and consolidation differs between the vertical business processes. The solution in this particular process (and most processes) is ultimately achieved through a blend of technology, governance, and people. The appropriate mix of these elements varies across business processes.

Technology: The Kinetic Element

The data quality technology stack is not limited to software vendors that specialize in data cleansing. Extract, transform, and load (ETL), business intelligence, and database vendors provide support to identify, analyze, improve, and measure data quality. In particular, ETL and database vendors typically support phonetic functions out of the box. While these functions provide matching capabilities, vendor support is not necessarily a substitute for enhanced matching capabilities (particularly probabilistic matching or householding). A coarse-grained matching solution may sustain a business process such as merging lists to create mass mailings for consumer retailers, but more demanding business processes, such as matching patient data from a claim form or verifying that airline passengers are not on a "no-fly" list, require more precision (see the sidebar "Phonetically Incorrect," next page).

In supporting the data quality horizontal process, vendors such as Trillium Software, Business Objects (FirstLogic), IBM Ascential QualityStage, and DataFlux offer end-to-end functionality, particularly when consolidating customer records for customer data integration (CDI) initiatives. By rationalizing their technology with the data quality horizontal, the software may be adapted to the demands of each vertical business process. Using the operational marketing process described earlier as an example, matching records from the customer relationship management (CRM) to online profile database can be done in batch or real time. Even if a batch horizontal process is implemented, the batch rules can be leveraged in real time when unsolicited events result in first-time user registration. No matter which stage in a customer lifecycle the customer record enters, the data practitioners can employ standard batch routines or Web services for real-time data quality. Moreover, as the sidebar illustrates, the support for enhanced matching can help avoid the pitfalls of relying on coarse-grained matching techniques.

Beyond tools, there are applications that can be employed to support the horizontal and vertical processes inherent to a data quality program. Business applications (CRM, enterprise resource planning, human resources, CDI, product information management [PIM], and financials) are modeled and configured to support specific vertical processes. Although these applications inherently provide thin support for data quality, most applications have certified interfaces into leading data quality tools that help ensure the appropriate level of data quality support is achieved.

Perhaps the biggest convergence in the technology stack that supports the horizontal and vertical processes is the advent of master data management (MDM) applications. Whether it is a PIM, CDI-MDM, or a generic MDM package, these applications attempt to provide flexible workflow and business rules that directly reflect the core business events, thereby enforcing data quality rules at the point of entry. Moreover, as the undisputed source of truth for each subject area, these applications rely on tight data controls so that data is integrated, consolidated, enhanced, and published in a manner that upholds data hygiene and consistency.

Governance: The Composite Element

This element is not a stand-alone practice but rather a composite of other subject areas. In this sense, data governance also resembles data quality since it is a horizontal and vertical discipline; governance involves, but is not exclusive to, data quality (and vice versa).

At any level, data governance involves data architecture. As all data practitioners recognize, data architecture heavily influences data quality.

Data normalization was introduced to remove data redundancies that could introduce update anomalies. Data architecture will also specify the logical data constraints that establish guardrails, which will be

Phonetically Incorrect

The use of phonetic devices to systematically match data has served data practitioners for more than 100 years. For those who have tackled tricky data matching scenarios, the use of phonetics can result in a surprising number of false positive matches, particularly when supporting international data. This example demonstrates how technical solutions do not always support the business process or even the data itself.

Anecdotal Evidence

Li and Lee are common surnames in the Asia-Pacific region. Using the phonetic device NYSIIS, Li evaluates to “L” and Lee evaluates to “LY.” Thus, NYSIIS views Li and Lee as phonetically distinct values, even though to the human ear they are phonetically equivalent. Based on phonetics, NYSIIS cannot support this APA surname for matching purposes (specifically blocking).

Conversely, SOUNDEX will resolve “Li” to “L000” and equates “Lee” to “L000”. While SOUNDEX views these surnames as phonetic equivalents, what is gained? You would not base a yes/no match result on this information, even if Li and Lee evaluate as the same phonetic value. They are fundamentally different, and should be handled as such when comparing surnames.

Most phonetic functions impede accurate and precise matching. If you are sending out mass mailings as part of retail operations (and are willing to send your target group duplicate communications), perhaps you can live with this low level of precision. In less forgiving business processes, however, a data practitioner must proceed with caution.

Technically, pseudo-phonetic functions such as DIFFERENCE (available in SQL Server 2005) cannot handle this example. When comparing “Lee” to “Li,” the DIFFERENCE function returns a 4, which indicates a best possible match in terms of characters and SOUNDEX.

Historical Context

As this example shows, the use of phonetics to support matching international data is like performing surgery with a chainsaw—the instrument is too blunt for matching customer data and constructing a reliable version of the truth within the context of demanding business processes.

If we think about the origins of phonetic functions, it is not surprising that they cannot support non-Western data sets.

SOUNDEX was patented in 1918 and conceived to index U.S. census forms. NYSIIS was invented in 1970 to support a New York State identification and intelligence system. Thus, both functions have a distinctly Westernized view and are not intended to support non-Western names or dialects. It's debatable which one is more accurate. However, neither was designed to handle the nuances of local dialects (both character sets and accents), nor has either established its ability to handle local flavors introduced by non-Western dialects.

Still, these functions have technical merit in the context of the appropriate business processes. Phonetic functions have been used by telephone companies to automate phone menu prompts through voice activation. Here, the level of precision is normalized with the demand of the business process. Prompts must react correctly to the phonetics of the customer's vocal responses to provide adequate customer service.

Beyond simple business processes, phonetics have a poor track record of false positive matches. Faulty phonetic-based matching has wreaked havoc on innocent travelers who have been incorrectly identified with individuals on a "no-fly" list (Kehaulani Goo, 2004). This is only one modern illustration of the significant missteps of phonetic-based matching.

Bottom Line

Database and data integration tool vendors package and preach NYSIIS and SOUNDEX. It is part of their training and their manuals. In positioning phonetics as a key component to matching business-critical master data from different sources, vendors sometimes do a disservice to data practitioners. These westernized and dated tools cannot always meet the demands of changing data demographics and challenging business processes. There are technical shortcomings as well. The phonetic functions support Latin-1 character sets but not double-byte character sets such as Mandarin.

Can phonetics be used for matching? The answer is yes, as long as a small contributor to the overall match "score" *or* the phonetic value is used to block data, *and* the data is westernized. The last caveat is no longer a binary yes or no, because socioeconomic trends have blurred the line between West and East, North and South.

Perhaps newer phonetic devices will improve reliability (e.g. Metaphone, Caverphone), but on the surface, these appear westernized as well. Even as improvements surface, phonetics should not be seen as a shortcut to true data standardization. Data standardization is the key to improving the reliability of match results, and standardization improves the overall data quality—whereas phonetics has no such added value. In any event, data practitioners should aware of the origins of phonetic devices, as well as their failure to keep pace with trends such as changing demographics.

Data quality support for customer matching and consolidation is a natural fit for a CDI-MDM or generic MDM package, but PIM applications can also capitalize on data quality functionality. Challenges with creating a product master such as standardizing manufacturer names, supporting multiple passes of matching (e.g., pass one on GTIN, pass two on UPC, or passes specific to different levels in the MDM packaging hierarchy) in a manageable fashion, and finally survivorship of core item fields (dimensions, various descriptions) can reliably construct the best internal view of the packaging hierarchy.

Note: You would want to tighten the degree of fuzzy matching when constructing the packaging hierarchy—standardizing manufacturer names or descriptions is one thing, but the matching on dimensions, pack size, and textual fields such as manufacturer must be precise within an allowable variance, e.g., a 10 percent difference.

Just as MDM vendors have certified adapters for data quality software, inevitably, data quality technology will be embedded within the MDM applications as the next generation of master data applications emerges.

physically enforced within some layer of the application architecture (e.g., presentation layer, business logic layer, or database layer). This includes allowable domain values, data type specification (i.e., precision), relationship identification, and cardinality. Without these constraints, the data could violate fundamental laws of the business. Finally, sound data architecture provides unambiguous meaning to objects and verbs, preventing unwanted data violations such as overloaded attributes.

Data governance vertically supports a data quality program as a platform for sound data architecture and data policies. Beyond data quality, governance provides direction and guidance to tricky data scenarios that occur in business processes.

In addition to providing the blueprint for data quality within an application, governance plays a key role in defining the data policies that help sustain a data quality program. The data governor will typically help define survivorship rules that determine how to construct the single source of truth record following a match. Moreover, data edits and rules (beyond data architecture) are specified by data governance. Governance also helps determine the severity of exceptions to these business rules, and defines ways resolve data errors. Finally, data governance formulates these policies in a cohesive manner that takes cross-project or cross-system dependencies into account.

Data governance vertically supports a data quality program as a platform for sound data architecture and data policies. Beyond data quality, governance provides direction and guidance to tricky data scenarios that occur in business processes. For instance, the online registration process for a customer requires tight governance if a preregistered profile needs support. In particular, true governance would recommend that passwords for registered and preregistered profiles be used as well as secure sessions (https) with timed logouts after periods of inactivity. In addition, before displaying the preregistered profile, the e-marketing application would authenticate the user with a unique, system-generated ID and password. The combination of the unique ID and the e-mail address should be used to link from the URL contained in the invitation e-mail to the landing page.

Returning to data quality, governance also provides the priorities and resources to fix problems. Fixing data has to be a corporate objective. For instance, users own data, and data owners need to repair data exceptions. However, users have other priorities that compete with their data cleanup tasks. An authoritative and empowered source such as a data governor should play a role in convincing data owners that fixing data is more important than getting orders out the door.

People: The Critical Element

Since there is no silver bullet or magic wand that cures all data quality issues, a successful data quality program includes a human component. After all, technology can fail us, the process may have holes, or governance can appear high-minded or heavy-handed from the outside looking in. As a safety net, a well-formed data quality program should be staffed to handle data exceptions and requests. We will refer to these resources as data custodians.

Knowledgeable about the business processes, trained in the technology, and fluent in data governance policies, data custodians make data quality improvements possible. We have saved the most significant element for last; data custodians are the heart of any data quality program or initiative. The data custodians are at the ground level, making sure data quality improves one record at a time. Thus, people are the foundation of the data quality program.

Any data cleansing project involves some manual intervention. Certain conditions require the insight and expertise of data custodians to handle outliers. Data custodians also help stabilize the launch of new data quality processes and technology, providing cleanup in the event that unexpected results corrupt the data.

Even the most forward-looking processes overlook some unexpected data scenarios. Data custodians can help bridge this gap by servicing user requests. Data custodians in a data quality program field common requests such as adding more values recognizable by data standardization routines or "unmerging" the results of false positive matches.

An authoritative and empowered source such as a data governor should play a role in convincing data owners that fixing data is more important than getting orders out the door.

The people element is often overlooked and might be the most difficult element in the program to stabilize. Ramping up data custodians takes considerable training, and turnover is expected. To help cement this pillar of a data quality program, the processes should be engineered with the data quality staff in mind. In particular, the processes must address when to escalate data exceptions to data custodians. The technology element should support the data custodians by providing user-friendly, high-quality software to resolve data issues and requests quickly.

The tools may also provide mass update or data quality reporting capabilities. Finally, data governance should treat data custodians as a key constituency group when formulating policies. To this end, data governance should ensure that this component is properly funded from the top of the organization. Without this key element, data quality initiatives are likely to struggle and eventually fail.

Conclusion

Data quality is achieved from the ground up, one record at a time. A successful data quality program is designed to operate at the ground level, aligning day-to-day tasks with strategic business objectives.

At the core of this program are four elements that provide the building blocks for a successful data quality program. Much like elements in nature, each one depends on the others to create and sustain a viable ecosystem. Ultimately these elements work together to ensure that a cohesive and integrated data quality program is established that provides coverage throughout the enterprise. ■

References

Kehaulani Goo, Sara [2004]. "Faulty 'No-Fly' System Detailed," *The Washington Post,* October 9. http://www.washingtonpost.com/ac2/wp-dyn/A18735-2004Oct8

Williams, John. "What a Data Quality Initiative Can and Cannot Do for Your Organization: Measuring and Improving Data Quality to Provide Business Value," Collaborative Consulting white paper, p. 9. http://www.collaborativeconsulting.com/thought-leadership/white-papers/view/7/

BI Case Study

Policing Data for Better Information

Linda L. Briggs

As law enforcement officials will admit, police departments are very good at collecting data but often not so skilled at retrieving that data and making good use of it.

The Erlanger, Kentucky police department wanted to change that with a new business intelligence system that included integrated search capabilities for easy, intuitive access to a rich repository of crime data. By integrating a variety of crime-related databases from 19 local and state government agencies, and updating the data every 15 minutes, the new system lets officers quickly find and link to information about suspects that previously languished in assorted databases.

The system has been deployed to Erlanger police leaders and appropriate administrators, and is being gradually rolled out to officers, according to Marc Fields, Erlanger chief of police. Fields says that a need for better information sharing between agencies the department works with drove the project. Although it's too soon to pinpoint actual cost savings or crime reduction, Fields says that once the system is fully deployed, the department will gain better insights into criminal activity. That can help reduce operating costs through more efficient crime-solving and prevention, and help the department assign its officers more effectively.

The BI technology Erlanger PD is using comes from Information Builders, whose professional services group used WebFOCUS Magnify to help the department build the real-time search index. WebFOCUS Magnify is a search navigation tool intended to help bridge the gap between BI and search capabilities. It scans indexed content, including both structured and unstructured data, and presents Google-like results with links to a range of document types and formats.

Minimal Training Needed

Key to the early success of Erlanger's new system is a basic, Google-like entry page to the custom search engine. The interface is so intuitive, according to the department's Public Safety Communication Center Manager Steve Castor, that "every officer can pretty much just look at it and know what to do." The

Public Safety Communication Center, which falls within the police department, serves 19 law enforcement and fire departments and nearly a dozen government agencies, along with some 75,000 citizens. With the new system, 10 cities throughout northern Kentucky can share crime records and incident reports going back more than five years.

Because the interface so closely emulates Google, whose simplistic entry page has become the gold standard of search engines, Castor says that officers are using it with virtually no upfront education. Castor says he is ready—but still waiting—for an officer to ask for help with the interface. "People understand, just looking at [our] simple box and simple button, how to put information in there and get things back."

In addition to the simple entry page, WebFOCUS Magnify presents a simple search results page in which each entry has a headline and short summary. A list in the left margin includes a breakdown by police-specific categories of exactly where the search term turned up. From there, officers can drill further down into the information. That breakdown is an important clue for police, since the categories a term appears in can give an officer valuable background. "[If] I see that my search term appeared four times in Incident Disposition," Castor explains, "I'm going to click there. I can then go deep [into the database], filter the data, and find my information quickly."

While officers in the field access the system in their patrol cars through cellular-powered displays, back at headquarters, dispatchers and supervisors use several BI dashboards. The dashboards, built using WebFOCUS and Arc/IMS from ESRI for geographic data, allow them to view the same crime data as field officers, but dissected in different ways. An interactive portal, for example, displays real-time views of incidents, arrests, emergency calls, and other events.

Linking Small Bits of Data

Like many police departments, Erlanger PD already had mechanisms in place to collect and store crime data. A traditional police accident report, for example, already contained reams of data for insurance purposes. What was missing was an easy method of accessing that data once it was collected and entered into the system. "We gather tons of information," Castor says. "We've stored it, but we had no [easy] way to get it out."

With the Information Builders system, the search engine can link seemingly tiny bits of data from recent incidents. For example, an officer who stops a speeding driver in one city could access the database on the spot from the patrol car and discover a link to a hit-and-run accident in a neighboring city earlier in the day. The search can pick up potential links even if only the car model and a portion of the license

Law Enforcement Turns to Analytics

Reducing crime by using BI analytic software to see patterns is beginning to draw interest from law enforcement. In a noted case profiled at the August TDWI World Conference in San Diego, the police department in Richmond, Virginia, once one of the five most violent cities in the U.S., reported that it had reduced crime dramatically using data integration and analysis software from Information Builders and SPSS.

The award-winning project uses data integration capabilities from Information Builders' WebFOCUS, predictive analysis from SPSS, and geographic information system mapping from ESRI to predict the likelihood of crime in city sectors.

Using what it calls Law Enforcement Analytics, Richmond combines data from multiple databases, including crime reports and GPS information about the city, along with known crime triggers such as holidays, paydays, and city events. The department also includes what it calls "interdependent factors" such as weather and phases of the moon. Combined and analyzed, the data has enabled them to understand crime patterns and deploy officers more effectively, and may eventually help to predict crime at a granular level before it occurs.

plate number were captured from the earlier incident.

"I can [link] two crimes that I probably would never have related before," Castor says, "and I can do it from the [patrol] car."

In the past, Castor says, all the information would have been captured for both incidents, then stored somewhere in the system, where all too often "it was lost forevermore." Although there were methods to retrieve incident reports and other data, they involved asking a system administrator to perform a text search. Even then, the results would not be in a very accessible format, he says.

Contrast that to the new system, in which information is updated every 15 minutes. That can make a tremendous difference to police on the street. Officers investigating a burglary, for example, can return to the patrol car and use the search engine to check for seemingly unrelated crime incidents—simple criminal mischief, for example—occurring on that street or in that area. With a search of street names, police can draw patterns and elicit clues about the crime, all without returning to the station.

Ambitious Plans Ahead

As powerful as the new system is, Castor sees plenty more potential going forward, mostly for accessing additional data. "I think that we've only hit the tip of the iceberg," he says enthusiastically. "A lot of what the system does today—and even how it looks—at this time next year will probably be dramatically different." The search engine itself will retain the simplistic entry page that is working well, but the value of what can be pulled from the system, Castor says, will expand dramatically.

Because the interface so closely emulates Google, whose simplistic entry page has become the gold standard of search engines, officers are using it with virtually no upfront education.

One planned addition is a crime-mapping ability available to the public that will use geographic information to show the location, on a map accessible through the Internet, of each crime logged. The department already has such a system for officers as part of the WebFOCUS system, which logs calls onto a map that officers can access through a Web page. The page includes key performance indicators (KPIs) such as current crime statistics broken down by category and compared to a year earlier. At a glance, the page allows administrators and officers to see areas that need more resources.

Other KPIs are set to show the location of every call received over the past 24 hours. A cluster of pinpoints on the map, Castor says, tells an officer immediately where to focus his or her attention for the day.

The crime-mapping system will be extended for public access later this year. That will help address regular phone calls to the department, Castor says, from citizens asking about the crime rate in a particular area of the city. Eventually, the department will simply point callers to a Web page that maps exactly what has been reported in each city sector over a given period. The public will also be able to look up recent crime incidents—a burglary or car accident in their neighborhood, for example—without tying up police phone lines with questions. ■

Based in San Diego, Linda L. Briggs writes about technology in corporate, education, and government markets. lbriggs@lindabriggs.com

Database Replication: Solving Real-Time Movement of OLTP Data

Bill Jacobs

Bill Jacobs is Technology Evangelist for Sybase.
wjacobs@sybase.com

Introduction

Large transactional databases, running at increasingly high speeds, emerged in the 1980s to serve core functions in industries such as financial services and air travel. Today, bigger, faster versions of these same high-performance OLTP databases are common in all large industries, and providing ever-higher transaction rates remains a major focus of the database industry.

However, these same databases—now the backbone of many highly successful organizations—have also become the data jailhouse. Real-time reporting and dashboards, an increasingly important component to many businesses, place large query workloads on these engines, slowing operational applications and increasing the risk of missed service-level agreements. Anyone who has stood in a long line at an airport check-in counter, waiting because the computer is responding slowly, understands this problem. The days when reporting could be run at night are long gone in today's 24x7 business environment, as today's OLTP databases know no idle hours. Increasingly, as businesses demand real-time operational reports and dashboards, reports run at off-peak hours are one to eight hours late, which is no longer acceptable in many industries.

Starting in the 1990s, companies began extracting transaction data from OLTP databases to load into data warehouses, operational data stores (ODSs), data marts, and report servers so that data could be rapidly analyzed and reported without impacting OLTP system performance. However, the batch-process scripts and ETL (extract, transform, and load) tools used to extract this data could not capture and copy transaction changes immediately after they occurred, which meant that latency was introduced into the report servers and operational data stores. In many highly competitive industries, this data latency can cause losses or reduce profits. Companies need to report data changes in real time—not days, hours, or even minutes after they have occurred. This requires solutions that can copy OLTP data into reporting repositories in real time without impacting production system performance.

> Existing techniques such as batch scripts and ETL tools have become inadequate in today's economic and competitive climate, where real-time reporting is no longer a "nice-to-have" option.

Consider the following real-world scenarios:

- In Asia, production lines move so fast that if a quality issue develops—such as a failing production machine approaching its wear limits—and is not detected for an hour, the resulting scrap cost can rise into the millions of dollars. To prevent such a problem, a solution must be in place to copy and report quality data in real time to predict impending problems and initiate intervention before product quality dips below acceptable levels and scrap costs rise.

- In the financial services industry, financial institutions lose tens of millions of dollars a year to credit card theft. By analyzing the charges on individual cards for specific fraud types, they can identify and prevent fraud, cutting losses dramatically. This analysis must be done with real-time data, since stolen or copied credit cards are typically used quickly in multiple locations and discarded within hours. Historically, this meant that the financial damage was done before the firm was even aware of the fraud.

Replication Requirements

Existing techniques such as batch scripts and ETL tools have become inadequate in today's economic and competitive climate, where real-time reporting is no longer a "nice-to-have" option. Many companies today need a faster way to replicate information to support reports, dashboards, and decision support systems that:

1. **Provide real-time, continuous replication.** Changes to data can be moved quickly and efficiently into a duplicate repository, decreasing latency toward sub-second response. Cutting latency improves the accuracy of real-time analysis, dashboards, and up-to-the-minute reporting. A database replication solution should capture transaction changes in the log files rather than directly from individual tables to prevent any performance degradation in the production systems. It should also preserve the transactional integrity of the data being moved.

2. **Work in a heterogeneous environment.** By replicating data from Oracle, Microsoft, IBM, and Sybase ASE databases into a similar set of heterogeneous targets, a replication solution breaks down silos and information barriers across systems and organizations. The ability to replicate IBM to IBM or Oracle to Oracle, for example, is insufficient. The solution must be able to replicate any-to-any in any combination that the organization needs, now and in the future. To do this, the solution must be able to easily and rapidly translate differences in data schemas between these multi-platform databases.

3. **Replicate data from multiple sources simultaneously.** This enables consolidation of data into a single repository that can be shared among different departments and groups. By supporting a common repository, a replication solution can cut the number of distinct data stores, which in some large enterprises has exceeded thousand of systems.

4. **Distribute data geographically and support local autonomy.**
   The administrators of each target site must be free to decide which sets of data the site will receive and how it will capture, store, view, and modify the data. Thus, the solution must support multiple vendors, operating both as data sources and replication targets.

5. **Make efficient use of network resources.**
   Replicating large amounts of data puts major loads on the network. The replication solution must be designed to use the available network bandwidth in the most efficient way, thus minimizing the impact of data replication on other network users.

6. **Offer selective replication from source databases.**
   The solution must be able to replicate entire databases or subsets of databases, such as specific tables, columns, rows within a table, or event types, enabling selective optimization.

7. **Provide central administration across the enterprise.**
   A powerful systems management tool with an easy-to-use graphical user interface is critical to managing such a distributed environment. This systems management tool allows administrators to manage and monitor all distributed components of the enterprise client/server replication environment from a single site, reducing the labor required to manage real-time reporting systems.

8. **Include rich data modeling capabilities.**
   Rich data modeling offers the ability to create and capture the metadata used to describe the replication topology, and to make changes to that topology rapidly and flexibly. It enables DBAs to automatically generate many of the scripts needed to create the replication logic definitions.

Gaining a Business Advantage

Database replication can be used in many ways to achieve multiple business advantages, depending on the needs of the enterprise. We describe three of them here.

Data Distribution: One Source, Multiple Targets

Database replication provides an efficient way to distribute information from a central site to many replicate sites where the information will be read only and will not be modified. Unlike snapshots, the distribution mechanism maintains the transactional integrity of the data.

> A database replication solution should capture transaction changes in the log files rather than directly from individual tables to prevent any performance degradation in the production systems. It should also preserve the transactional integrity of the data being moved.

In one scenario, a company uses a central OLTP database in San Francisco to process incoming orders. Order entry applications are connected directly (as terminals or clients) to the central OLTP system, which is controlled by corporate IT staff. A large number of decision makers in other parts of the organization (e.g., finance in San Francisco, manufacturing in Dallas, and sales in New York) would like to view order information to make timely decisions in their respective organizations. However, due to the taxing nature of their ad hoc queries or customized batch reports, or simply due to their large number, they are currently not allowed to access the OLTP system. These decision makers have to look for the information they need in standardized reports generated by the IT staff.

Database replication can allow individual divisions to “subscribe” to a subset of the central site’s data and view that information locally. With a local copy of the data needed for decision support, finance, sales, and manufacturing can carry out their operations without having to log into the central IT site in San Francisco. Their decision support work is consequently isolated from network or remote system outages. Furthermore, local users are likely to receive the results of their queries faster, since they are accessing local (replicated) rather than remote data. Thus, replication can provide a simple way to increase the scalability of a system by providing data access to more corporate decision makers.

> Database replication can allow individual divisions to "subscribe" to a subset of the central site's data and view that information locally. With a local copy of the data needed for decision support, departments can carry out their operations without having to log into the central IT site.

Data Consolidation: Multiple Sources, One Target

In addition to distributing information from a central location, true database replication provides a simple mechanism to bring together corporate information from several locations and consolidate it at one site. One example is a corporation that gathers information from remote production sites for central reporting at corporate headquarters; another is a multinational company with several facilities distributed worldwide—say, production sites in Seoul, Frankfurt, and Mexico City, and corporate headquarters in Boston—that needs to have an up-to-date, read-only view of production status worldwide. Consequently, information from the three production sites is consolidated in Boston. While the data in Boston may be a few minutes old, it is near real time and current enough for the corporate applications. Note that the database schemas in Seoul, Frankfurt, Mexico City, and Boston may be different.

Reporting consolidation is also proving to be vital for developing a single view of other elements of a business—such as the customer or a product—particularly in large enterprises where multiple divisions interact with the same customer. Customer data integration (CDI) is a particular area of interest in many leading companies today. Such consolidated customer or product views are impossible to create when the data lies in a multitude of data stores scattered around the enterprise, and therefore need to be consolidated in real time into one reporting system.

Data Sharing: Bidirectional (Peer-to-Peer) Replication

The two previous examples described cases where replication occurred in a single direction: in the first case from one site to many, and in the second example from many sites to one. Replication, however, does not necessarily need to involve data transmission in a single direction. Corporate data sharing provides one illustration of this.

Under this configuration scenario, several distributed systems are set up so that they include primary as well as replicated data. For example, a company with several sites across the U.S. (Atlanta, Chicago, and Seattle) wants to consolidate its employee data. Each employee has only one home office. Each regional office manages local employee information. Database replication provides a mechanism for the corporation to share this distributed set of employee data.

Since each employee has only one home office, the example illustrates the case where distinct data items are combined from each site and where the distributed application does not involve update conflicts. While each site includes primary as well as replicate data, no two sites are allowed to modify information on the same employee. Thus, the Atlanta office will only modify information on Atlanta employees; the Chicago office will edit information only on Chicago employees, and so on. The result is that all sites can view near-real-time employee information for the entire corporation.

Database Migrations and Upgrades

Upgrading a production system to a new version of a data server is a difficult and risky task. New features or changes in functionality can cause unexpected downtime. Applications must be fully tested with the new version before they are rolled out into the production environment. Database replication is a fast and effective way to test a new version of a database management server before it is put into production.

> Reporting consolidation is also proving to be vital for developing a single view of other elements of a business—such as the customer or a product—particularly in large enterprises where multiple divisions interact with the same customer.

Suppose IT is responsible for upgrading a database server that houses data for a business-critical inventory system. The DBA wants to upgrade the server to take advantage of new performance features. The DBA installs the new version of the database server on a test system and creates a replication connection to the existing production system. All production transactions then begin to be replicated to the new database server version in the test environment.

The new version of the server is then tested in parallel with the production application. Because the test application is fed live data, at ordinary (production) speed, the DBA can test the new performance features as well as conduct basic regression testing. Because the test and production versions remain in sync using replication, the DBA can be more confident that the new database is accurate and ready for production. Transition risks are greatly minimized, as are risks of disruption to source OLTP systems.

Summary

Historically, database replication has been used primarily as a disaster recovery solution by creating warm standby database versions at remote sites. However, as the need for real-time data movement and reporting in modern enterprises grows, database replication technology is the ideal solution to meet a large variety of other business challenges.

These solutions provide log-based, efficient, real-time data movement with minimal impact on production systems. They provide the ability to handle replication of transactions across the enterprise in multiple directions under efficient central administration. Best of all, they are mature, stable, scalable products that sustain enterprises growing from modest to multinationals, from a few nodes to thousands. ■

Debunking Three Myths of Pervasive Business Intelligence

How to Create a Truly Democratic BI Environment

Kirby Lunger

Kirby Lunger is SVP, Corporate Development
at Attivio, Inc. in Newton, MA.
kirby@attivio.com

Abstract

As your new BI application grows in popularity—expanding into multiple data warehouses—you increase hardware capacity to meet demand. You may even install a warehouse appliance to handle the workload. But have you truly realized pervasive BI just because your application is being used more? Are you fully delivering on the promise of BI? We explore three myths about pervasive BI and show you how to move your organization closer to the goal of democratic performance management.

Introduction

Suppose you built a departmental BI application with an operational data store or data mart to support your business unit. Then the departmental executive got promoted and started pitching the concept of a company-wide dashboard. When you realized the dashboard supported only a small sliver of your company's transactional data, you built a data warehouse to support additional data sources.

Somehow that evolved into multiple data warehouses. You developed an enterprise information management group to regulate information hierarchies and master data between applications. Meanwhile, you increased hardware capacity, until finally someone invented data warehouse appliances that made analysis faster. Software vendors devised better licensing models and easier implementation capabilities so more people could actually see the reports and analytics formerly seen by only a very small group within your company.

Now, after all this work, why don't you feel like your analytical environment is actually delivering the promise of business intelligence capabilities to the majority of people using the platform?

"Pervasive BI" is a sometimes overused term that can refer to several characteristics of a BI solution:

- How many people use the solution within your company
- How many people in your enterprise ecosystem use the solution (e.g., partners and customers)
- The kinds of data types the solution accesses
- The kinds of analysis the system provides

Many software and hardware vendors claim to have solutions to "democratize BI for the masses." The simple truth is that although more people are accessing these products—and although these products are accessing more data sources—the quality of the analysis hasn't truly improved because these systems are still limited by the types of data they access and how the data is presented. By taking steps that address the three biggest myths of pervasive BI, you will be able to move your organization much closer to the goal of democratic performance management.

Myth #1
Making your BI application "pervasive" by growing your audience increases the power of your BI environment.

Mae West is widely quoted as saying, "Too much of a good thing can be wonderful." This thinking sums up the growth in BI audiences in companies over the past several years. Most BI solutions were designed for a very small audience—usually senior executives or a group inside an operating line or function. If the BI application is successful with this small group and useful for upper management, the conventional thinking goes, it should be great for the rest of the company, too.

Most BI tools excel at providing two types of analysis: static reports and "guided" analysis. The challenge with this functionality for a larger audience is that your executives will look at high-level data and ask additional questions that do not work within the guided analysis implemented with the majority of BI tools. Rather than asking, "What number?" or "What quarter?" they may want fuzzier information, such as "Why are the numbers forecasted to be low in the second quarter next year?"

Generally, the employee who creates the activity leading to whether the quarter is off is also asked to perform this type of ad hoc research and provide that needle-in-the-haystack answer back to the executives. (Some call this "verbal commentary" to go along with the transactional information typically displayed in a BI application.)

> Rather than asking, "What number?" or "What quarter?" BI users may want fuzzier information, such as "Why are the numbers forecasted to be low in the second quarter next year?"

Providing access to your structured, hierarchical BI environment will allow more people in your company to access some part of the data and perform some kind of analysis, but more pervasive access does not necessarily create more pervasive BI.

Myth #2
Your current BI system contains most of the data you need for pervasive BI.

Let's get real: No analytical platform will ever provide a way to analyze 100 percent of the content within a company, if for no other reason than some of your information is still stored in e-mail messages and paper files! Even with that expectation clearly set, BI tools alone cannot always provide access to enough data sources to allow for true performance management within your company.

Along with the enterprisewide dashboard you have implemented in the quest for democratic BI, you have probably implemented some form of performance management system, like a balanced scorecard. In the process of rolling out whatever methodology you liked best, you may have discovered that although your existing data warehouse(s), data mart(s), and operational data store(s) contain considerable transactional content, they are missing the most crucial information of all. This brings us back to the ad hoc querying issue mentioned in Myth #1: to find the "needle in the haystack," you need to be able to search for information that is not highly structured and is not a good fit for the traditional, "structured" environment.

Without evolution (and some might say revolution) in how you organize your architecture to support BI and performance management, you really cannot access the data you need for "pervasive" or preventive BI.

The goal of performance management is to move away from lagging indicators (things that have already happened, like your company's financial results) and instead manage performance by looking at leading indicators (such as how well your employees are performing their jobs so they can influence financial outcomes). Your company's transactional systems are, by their very nature, full of lagging indicators, because they primarily house financial data. The majority of predictive information is sitting in content formats and systems that your BI solution does not touch, such as e-mail messages, PowerPoint and Excel files, call notes in your CRM or SFA systems, or even sources outside your organization where customers might be posting comments on your products in the blogosphere.

Without evolution (and some might say revolution) in how you organize your architecture to support BI and performance management, you really cannot access the data you need for "pervasive" or preventive BI.

Myth #3

BI solutions today contain adequate analysis capabilities to enable pervasive performance management inside your organization.

BI solutions are optimized to allow for very fast analysis of hierarchical financial data. They were not built to address the following representative types of analysis or use cases.

Basic Search

If you want to search for a piece of information that is not part of the predefined hierarchy you use to find information in your BI tool, the closest you will get is using a BI platform that offers search plug-ins from Internet search players such as Google. The challenge is that most of these search engines were built to search the Web, where companies and people try hard to organize the metadata about their information in a searchable format.

This is not the case inside your company—or your BI application. The people creating information in your company usually have no motivation to tag content to make it easy to access or search. More important, most of this information isn't even accessible in your analytical environment.

Exploratory (Ad Hoc) Analysis or Search

For this kind of analysis, you must provide a mechanism to correlate items not obviously associated, or see patterns in data sets that are not readily apparent. For example, there are specific techniques an analyst can use to address the ad hoc question introduced earlier, "Why are the numbers forecasted to be low in the second quarter next year?" Imagine if this user could compare the financial forecast against actuals and drill through or back to a specific region or salesperson's CRM or e-mail text records to understand if there was a low volume of communications with certain customers. The person conducting the analysis might also evaluate the sentiment of communications with a certain group of people to determine if there are problems with your company's products or services.

THE SOLUTION

Establish a Foundation for Pervasive BI

During this discussion, we assumed the goal of pervasive BI is to give most of the people in your organization access to the majority of information appropriate to their roles with a breadth of analysis techniques. This allows your employees to access the right information at the right time to make the most informed performance management decisions possible.

> The people creating information in your company usually have no motivation to tag content to make it easy to access or search. More important, most of this information isn't even accessible in your analytical environment.

Some of the issues with the three big myths of pervasive BI are structural and cannot be fixed anytime soon, such as the fact that some percentage of your company's information lives on paper and in people's brains. There are, however, some short-term steps you can take toward truly providing pervasive BI.

First, limit your audience. Until you can provide the tools and analysis your end users require, do not bother to buy licenses and implement solutions that give them access to information that does not meet their needs.

Second, understand the types of analysis people want to perform that they currently cannot do with the BI tools at their disposal. This analysis will most likely look at the drivers of financial and customer outcomes. This driver information is contained in content such as e-mail messages, documents, or PowerPoint files that usually does not integrate well into your current analytical architecture. Start thinking about what interim steps you can take to aggregate and display this information in a usable format, such as a relational table or search index.

Third, start investigating new technologies that let your users access all of your organization's information assets, whether it's your structured BI data or the fuzzier but critical unstructured information.

As Mark Twain said, "The secret of getting ahead is getting started. The secret of getting started is breaking your complex overwhelming tasks into small manageable tasks, and then starting on the first one." Pervasive, democratic BI is not a goal that is 100 percent possible with the tools available today, but if you can start to take small steps in this direction, you will start to see major, positive changes in your organization's ability to conduct better performance management. ■

Customer Data Integration

Philip Russom

Who *are* your customers? Which products and services are they buying across your enterprise? How much business have they transacted with your enterprise so far this year? Where do they conduct business besides your firm?

If you don't have an enterprisewide solution for customer data integration (CDI), it's unlikely you can answer any of these questions with a respectable level of accuracy. That's one of the assertions in a recent report from TDWI Research, *Customer Data Integration: Managing Customer Information as an Organizational Asset.*

Before we investigate some findings of that report, let's begin with a definition of CDI from the report:

> **Customer data integration (CDI) uses information technologies, business processes, and professional services to collect customer data from disparate enterprise and third-party sources and integrate the data into a singular 360-degree view of each customer that's complete, up to date, accurate, clean, and standard. TDWI's position is that customer data is an enterprise asset that should be integrated, shared, and leveraged broadly via CDI techniques.**

Philip Russom is a senior manager at TDWI Research and the author of TDWI's new report on CDI. We recently caught up with him to learn more about his findings.

Business Intelligence Journal: **Why is CDI important to so many BI practitioners—what benefits do they expect?**

Philip Russom: Many of the most common business questions that BI seeks to answer concern customers. This includes questions such as: Who are our most profitable customers? What's the demographic profile of our customers, sorted by profit, geography, product preferences, financial bracket, and so on? How many customers do we have? How do our customers relate to each other?

You can't answer these questions accurately with a BI solution unless BI is backed up by an analytic CDI solution.

What's driving interest in CDI at this particular time? After all, don't companies already have CDI projects?

I think there are several compelling reasons why business and technical people should revisit CDI, even if they feel they've "been there, done that."

First, improving CDI is one of the many things firms have to do as they continue the slow but profitable process of becoming customer centric. Second, U.S. firms deployed far too many CRM and CRM-ish applications in the 1990s (many of these with their own CDI solutions), and consolidating CRM applications (a common IT project today) usually requires consolidating CDI, too. Third, if we look at how CDI was deployed in the 1990s, it often provided a complete view of a customer, but only from a sales and marketing viewpoint, a financial viewpoint, a customer service viewpoint, and so on. Many organizations now need to consolidate or integrate these customer views to get a true 360-degree view.

Finally, many CDI solutions are now old enough—or were "low-end" solutions to begin with—that they need replacing or updating just to support new practices and technologies, such as master data management and service-oriented architecture.

What are some of the key features of a CDI solution?

At the risk of stating the obvious, CDI is a form of data integration, so successful CDI solutions are those with rich data integration functionality. With that in mind, organizations that lack a well-developed data integration infrastructure and skilled team are somewhat at risk when they attempt to develop a homegrown CDI solution in house. Likewise, multiple functions for data quality are key features of a CDI solution, because customer data evolves constantly as customers change their addresses, jobs, financial brackets, preferences, names, and so on. If an organization lacks experience in data quality, its ability to infuse CDI with data quality functions may be hamstrung from the beginning.

> There are several compelling reasons why business and technical people should revisit CDI, even if they feel they've "been there, done that."

CDI doesn't exist in a vacuum, of course. It's wrapped up in such things as master data management and data quality.

Absolutely. We've already talked about how data integration and data quality are the leading key features of a CDI solution. Master data management (MDM) is also a high priority for CDI because MDM gives CDI one of its most coveted goals—consensus-driven definitions of the customer applied consistently across multiple IT systems and departments. Despite the obvious benefit, I don't think enough users have started to practice MDM with CDI. For example, the CDI report's survey found that 26 percent of respondents are applying MDM functions to CDI today. This is higher than I was anticipating, but it's still not good enough.

Part of the problem is that organizations are doing a lot of CDI but in disconnected silos. As they connect the silos, consistent definitions of customers become a higher priority; hence MDM becomes a higher priority. Those silos are often legacies that predate the modern practice of MDM, so few of them support MDM. As companies integrate CDI silos and update legacy CDI solutions, they also retrofit CDI with MDM. We can see this in the report survey, which revealed that 55 percent of users plan to add MDM to CDI in the near future.

ROI is an important consideration in getting projects approved. How does an enterprise calculate the ROI of a CDI project?

The business case for CDI is both easy and hard to make. Note that most CDI scenarios include an implicit domino effect. For instance, up-to-date customer information leads to more efficient customer service, which yields higher customer satisfaction so that customers churn less. More complete customer data enables more accurate customer segmentation and one-to-one marketing, which leads to better targeted marketing campaigns, resulting in higher conversion rates.

Because CDI is the first link in each of these chains—and the effect on revenue is the last—it's hard to link a direct causal effect from CDI to revenue. We sometimes forget the soft benefits of the intervening dominos. Even so, conventional wisdom, based on the hindsight of many completed CDI initiatives, now says that CDI easily yields a return on the investment, whether raising revenue through sales and marketing or retaining customers through better service. Another way to put it is that CDI is an underlying data management infrastructure that contributes to business initiatives and their financial goals, such that CDI yields ROI, albeit indirectly.

According to your survey, a lot of people think CDI's impact is mediocre. Why isn't success greater?

According to our survey, half of respondents (50 percent) rated their organization's overall success with CDI as medium. A few rated their success as high (11 percent), and a considerable percentage (38 percent) rated their success as low. This shows that CDI solutions have plenty of room for improvement in most organizations, which is why so many CDI upgrades and replacements are happening today.

I personally suspect that the mediocre success of CDI relates to its silos, single application or department orientation, inability to push improved customer data back to operational applications, and lack of support for modern technologies such as MDM and SOA. I also talked with people who said that CDI was so hyped in their firm—similar to how CRM has been hyped—that it could never measure up to end users' unrealistic expectations.

Your survey pointed out that less than half (49 percent) of CDI solutions currently share data across the enterprise, and that many projects serve only smaller departments. What can be done to expand this scope?

Increasing the scope of CDI solutions is partly a technology problem but mostly a business problem. Customer data isn't shared as much as it should be (so the firm gets as much leverage as possible), because the silo nature of most CDI means that customer data owned by department heads, line-of-business managers, and other folks who funded CDI with their own budget.

In the report, I described "data as an enterprise asset program" in which a central body (typically corporate IT) takes ownership of data that has been isolated in departmental silos. That's pretty radical, so it takes a corporate culture that's very open to change to succeed. A milder and less risky approach is for a data governance committee to set up procedures for gaining access to departmentally owned data, as customer data usually is. Once these organizational barriers are breached, the technology piece of CDI is relatively straightforward.

What recommendations can you make based on the survey results?

We were just talking about the importance of transforming the business, such that departmentally developed assets such as customer data are opened up for use and repurposing by multiple business units. Once that foundation is in place, everything else goes much faster and has a deeper impact.

From a technology standpoint, the challenges in most organizations stem from numerous preexisting CDI solutions. These overlap and sometimes serve up contradictory data, especially when they lack MDM. Most are tied to a single application, whereas their valuable data should be accessible to many applications, possibly through services, so sorting out the sordid silos is a key first step.

A plan of action must describe which silos should be consolidated, which should be left in place but synchronized with others, and which need upgrades and enhancements. Despite the insidious silos, there will be gaps where important applications or business functions aren't yet served by CDI. Once all these possible actions are cataloged, prioritize them by business pain points and potential ROI, yet with an eye to your ultimate goal: sharing consistent, quality customer data broadly across many business units and applications. ■

Coping with "Big Data" Growing Pains

Lis Strenger

Lis Strenger is the director of product marketing at Dataupia.
lstrenger@dataupia.com

Abstract

Data volumes are growing exponentially. That's a given. How organizations respond to the growth, however, is not. A recent survey of data center managers shows that newer practices such as single-instance storage are gaining traction but not replacing older methods.

To stay on top of data growth, IT groups are combining approaches such as data deduplication, archiving, deletion, workload prioritization, hardware acquisition, hosted repositories, and database optimization. However, most are still making decisions about handling "big data" based on their previous experience with "small data," and risk making costly choices or missing opportunities to improve information management.

This article explains how "big data" changes the rules of the game. Strategies that worked well for gigabytes of data fall short in the terabyte world. We answer the top five questions data warehousing professionals have about working with terabytes. After looking at common industry approaches under the terabyte lens, readers will have a relevant framework for evaluating technology solutions and business process changes to support their growing data assets.

Introduction

I recently read a short report in *Time* about a Russian fast-food chain that has swiftly grown to become fourth largest in that nation (McGrane, 2008). The founder, Mikhail Goncharov, made a comment about managing growth that caught my attention: "If you're chopping 100 kg of mushrooms, you do it one way. If it's 200 kg of mushrooms, you do it a totally different way." That, in a nutshell, is the mindset to have when facing the need to manage large data volumes.

How do you find that different way to chop twice the amount of mushrooms? Do you pull out the colanders and knives you used yesterday and see how far they get you? Of course, you count on your prep-chef experience to tell you when and how to make adjustments along the way. Do you reach for the largest cleaver you have, following the principle of "the larger the volume, the larger the tool," or do you look at the tower of boxes and rethink the whole procedure? The latter choice isn't likely if you are running a real business with limited resources and concrete requirements—although you might start researching appliances designed specifically to chop fungi.

> Instead of counting terabytes or the number of data rows you store, determine whether the amount of data you need for operations exceeds the capacity of a component within your infrastructure by a few gigabytes.

Begin to think through how to handle large amounts of data by evaluating your current tools and deciding what you can scale (what can be extended) with supporting technology. Perhaps in parallel, research other available solutions, mindful that it is unlikely you will find the perfect fit. You will have to tailor a solution to meet your requirements. Whatever approach you emphasize when looking for solutions, there are two things you need to know when you develop your requirements and evaluate options:

- Define "big data" in your organization's or project's context. Technology vendors, industry analysts, and academics each have a different definition. However, as well-founded and valid as these definitions are, they will not apply directly to your situation. Understand where your data comes from, the business context of each data set, and where data volume is largest.

- Know the capacity and extensibility of the tools and methods you are investigating, especially with respect to how they perform at the top range of your forecast. *Predictions about how our information environments will expand are usually accurate.*

Taking Measure of Your Data

Data complexity is a result of working within memory, disk space, CPU, I/O, and network constraints. An operation executed against large data sets is split into subtransactions or subprocesses, and continuity must be guaranteed. That threshold is shifting as advances in technology ease some of the constraints, but there is still a size barrier. Like the sound barrier, the size barrier can be crossed, but some of the rules change on the other side.

There is a more pragmatic way to determine whether your data qualifies as "big." Instead of counting terabytes or the number of data rows you store, determine whether the amount of data you need for operations exceeds the capacity of a component within your infrastructure by a few gigabytes (GB). If so, you face many of the same scalability challenges as someone trying to add terabytes (TB) of data. For example, Microsoft SQL Server has its own internal size markers. Its COUNT function tops out at two billion records, after which COUNT BIG must be used. This is just one example of how, at the fundamental level of query processing, size changes the rules. COUNT BIG mimics COUNT as much as possible, but the differences can have a ripple effect. The database can operate against a big data set, but can your infrastructure support passing a big result set through the application stack?

Your data warehouse is swelling. You know it takes in 500 GB per day, but from how many feeds and at what rate? To understand the impact big data is having on your infrastructure, identify the points in your data management infrastructure that touch large volumes of data, either serially or all at once. These parameters are critical when architecting your solution. Managing two 200 GB batches every 12 hours is different from managing multiple lightweight sources fed hourly.

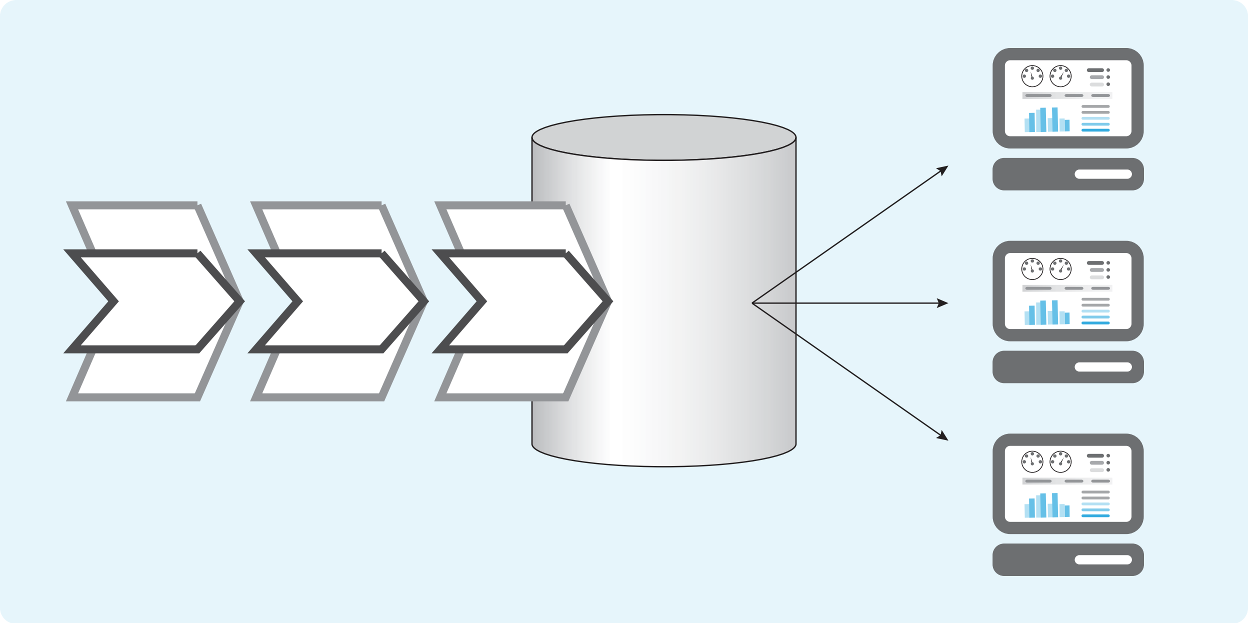


Figure 1: Some data warehouses must scale to adapt to the growing size of data sets. In this illustration, only the size of batch feeds grows, not the frequency or the degree of application access.

Figures 1 and 2 show examples of the factors that require a data warehouse to scale. Although these examples look simple, accommodating just one growth area can result in work at almost every layer of the data warehouse's infrastructure and surrounding networks and applications. The type of changes required by growth factors will vary, however. It is essential to map all the ways a data warehouse will grow at the outset of any expansion project. Consider whether it will need to accommodate larger data sets, more data sets, higher frequency of data loads, more users, more applications, and so on.

Assessing Impact Points and Approaches

Just as there are many ways a data warehouse might need to grow, there are many points at which its infrastructure must scale. Figure 3 (page 49) shows the many points where opportunities exist to make big data more manageable: at the point of collection, before sending data to the data warehouse, before archiving, and before off-site storage. Each is associated with performance or accessibility trade-offs. For example, the more that is done at the point of data collection, the more performance degrades from the users' perspective. Shifting all work to the end of the process puts pressure on maintenance procedures; backup and replication might collide with data collection and access.

Figure 3 shows a solution map you can use to identify your anticipated growth areas. This is the first step to selecting approaches that address the particular challenges your data warehouse faces.

The second step in the process is to look at the short- and long-term impact each approach has on both the IT and business users. To measure "impact," examine expense, labor, time, future scalability, and whether accessibility to data increases or decreases for the business user. Table 1 (page 50) compares the relative impact of each approach or technology.

The art in applying this information lies in striking a balance among all factors that your organization can live with. No single practice stands out as being a panacea. Instead, your organization's tolerance for limited accessibility to data, your IT group's appetite for taking on new technology, and resource constraints will determine a solution's viability.

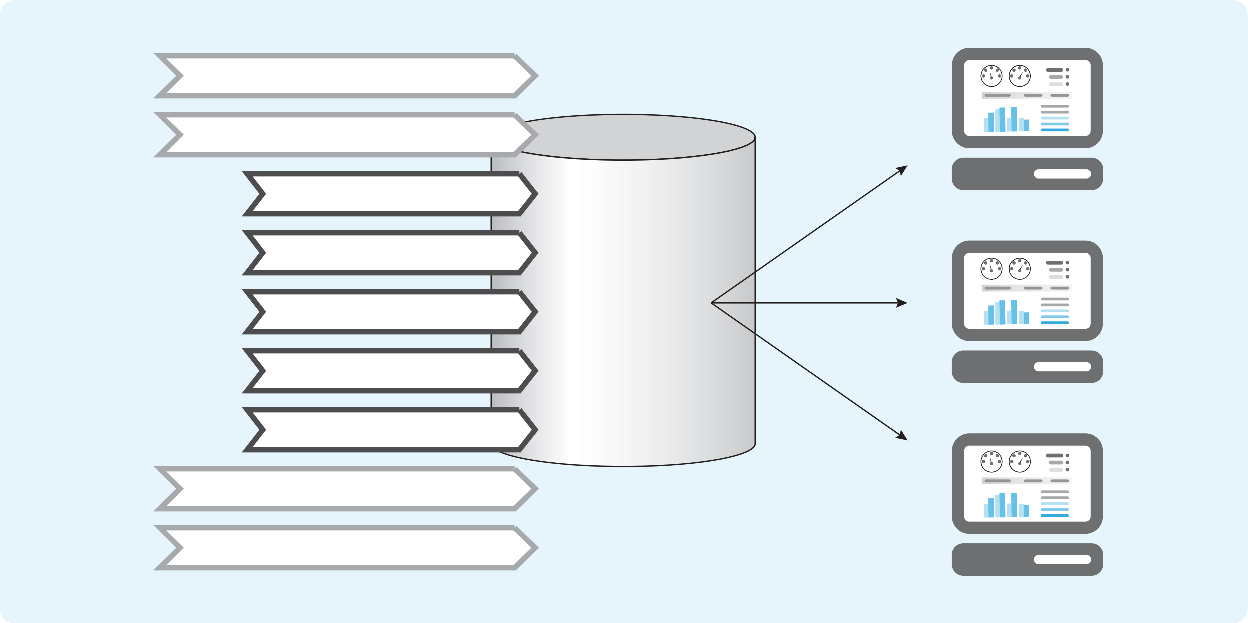


Figure 2: Adding new data sets, even small ones, can cause scalability challenges for a data warehouse even if all other parts of the environment stay the same.

Up to this point you've identified the sources of your growth and the stress points the added volume will cause in your data warehouse infrastructure. You've been introduced to the approaches that ease scalability problems, identified where in the architecture they operate, and determined what factors will play into your choices. That's the framework for evaluating solutions. You still need information to evaluate the solutions on their own merit.

Use Existing Capacity More Effectively

The clear first task is to ensure that existing resources are used appropriately and to their fullest extent. Techniques that help stretch existing resources include work prioritization, system optimization, and virtualization. However, these approaches can provide only a temporary fix, because your existing resources were selected to fulfill very different requirements (at least in terms of scale) than those you have now.

Using existing capacity to its fullest is a matter of applying basic resource allocation techniques, although now you are making rules based on a new set of conditions. You still have two options: to manage hardware or manage users. When classifying units of work for resource allocation or for query prioritization, size becomes critical. In particular, you are no longer concerned only with the number of CPU cycles a request consumes. You must also consider the time it takes to access secondary storage and the application impact of moving several terabytes of data across the network.

Assessing the impact of an operation requires an in-depth knowledge of the database, its table sizes, data density, and indexing. Scheduling with an eye to minimizing resource conflicts is a possible solution, though low-usage windows are scarce in multinational organizations. The result is that prioritization will increase your actual capacity but at the cost of additional complexity, expertise, and slower system performance.

Established practices of managing workload using query prioritization and job scheduling are still effective here, though you will have to revisit the algorithms you use and the business rules they embody. For example, standard reports, when run against massive data sets, can consume more system bandwidth than the business will tolerate. Larger data sets also present new opportunities

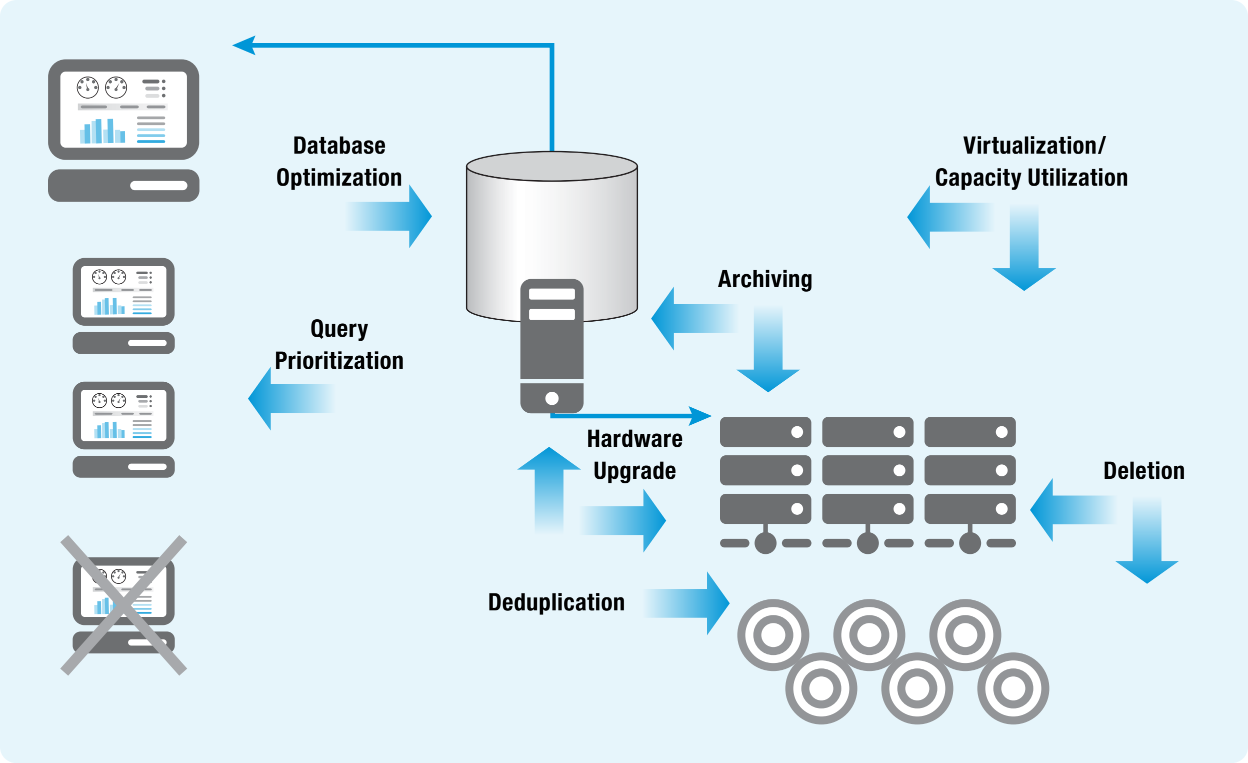


Figure 3: The approaches to handling a data warehouse's growth increase scalability at specific points in the architecture.

to gather information, especially through analysis. The standard reports once viewed as mission-critical may be less desirable if running them limits root-cause or trend analysis that can provide more meaningful information.

Optimize Databases

When you optimize your database for large volumes, remember that the database platform you are using was designed for OLTP, not for data warehousing. Although you might have built your data warehouse from the ground up, the technology you used had many built-in assumptions stemming from its OLTP roots. If you transformed an OLTP database into a data warehouse, you should review for transactional legacy even more rigorously. For example, are cursor numbers, buffer sizes, and limiting parameters set appropriately? Have you fully exploited your database platform's native optimization capabilities, such as partitioning, distribution, and query tuning?

Deploy Virtualization

A 2008 *McKinsey Quarterly* article describes a common problem businesses have with unused capacity:

> Well-managed companies use 80 percent or more of their available storage, but in others that figure hovers around 40 to 50 percent. One large IT organization used only 50 percent of its storage capabilities. Some of its individual storage systems were at just 10 to 20 percent of capacity, and one of its businesses utilized only 33 percent of the entire amount of storage it had requested.

Technologies such as virtualization have a strong impact on the ability to tap into a broad range of existing resources as needed. Virtualization alone cannot address the challenges of working with large volumes of data, but it can stretch your resources, especially where you have many small streams of data or a high rate of data operations. If your large data is made up of smaller data

	HARDWARE COST	LABOR	TIME	SCALABILITY	LIMITED ACCESSIBILITY
DATABASE OPTIMIZATION	Low	Moderate	Low	Moderate	High
HARDWARE UPGRADES	High	High	Moderate	High	None
VIRTUALIZATION	Moderate	Moderate	Low	Moderate	Low
DATA DEDUPLICATION	Moderate	Moderate	High	Moderate	Low
ARCHIVING	Low	High	Low	Low	High
DELETION	None	Low	Low	Low	High

Table 1: Impacts of each approach to, or technology for, handling growth

sets, you can farm out the units of work to available CPUs and distribute the data blocks to disk space spread through the virtual array. However, if you intend to run operations against the whole data set, neither virtualized servers nor virtualized storage will suffice. At some point, you will need enough CPU and disk capacity that can function as a single engine. Virtualization management tools are not yet advanced enough to distribute a single unit of work and guarantee its integrity.

Acquire or Upgrade Hardware

Virtualization will extend some of the capabilities of your physical infrastructure until data growth exceeds your hardware's capacity. More often than not, adding or upgrading hardware is the first solution data center managers turn to when confronted with growing data. The first signs of problems show up as hardware performance issues signaled by high CPU activity, hanging processes, and running out of memory or disk. Adding more hardware without altering the physical architecture will not address the root cause, which is the need to move too much data through the network.

For physical infrastructure to support working with big data, three components must be scalable: CPUs, disk I/O, and network connectivity. The latter is important because unless you have a mainframe, your big data cannot be processed and reside on the same server. A typical architecture for big data has at its core a robust, multi-CPU server hosting the database and network-attached storage. There might be additional servers for auditing services, backup facilities, ETL, data cleansing, data staging, etc. You can add more CPUs at the front end and you can add more storage, but you also need to address bandwidth by moving to more powerful backbones (Fibre Channel or 10 Gigabit Ethernet) and switches.

Databases that have tens or hundreds of terabytes require some form of parallel processing. Massively parallel processing (MPP) is becoming the best-in-class architecture for very large databases. MPP configurations are designed to recast operations as sets of subprocesses, distribute them for parallel execution on an array of CPUs and disks, and marshal the results. Both the hardware and the database have to be designed for MPP for this level of parallelism to occur. IBM's DB2 was an early MPP database, but most deployments are on servers that have minimal parallel processing. MPP systems are available as data warehouse appliances or as specially configured hardware/software bundles.

Adjust Information Lifecycle Management Policies

Why are so many organizations overwhelmed by exponential increases in data? One of the main drivers behind this growth is that organizations have explicitly decided to collect these volumes of data. More demand has led to more data—more data will lead to increased demand.

Putting draconian archiving or purging policies in place would stem the data tide, but the goal is to maintain access to data despite size challenges. Information lifecycle management (ILM) based solely on age, access frequency, or compliance conflicts with the purpose of amassing data, namely to maintain enough historical, detailed data to support strategic and tactical decision making. Even data that is collected and retained primarily to satisfy regulations can be mined or analyzed for trends and patterns. A recent study by Nemertes Research on security and information protection showed that 27 percent of participants kept compliance data “forever.” Leveraging that data for intelligence gathering would make up for some of the cost of housing it (Burke, 2008).

To the extent that big data is putting a strain on resources, you will want to refine the processes by which you rank projects, subject areas, and data sets. Age of data takes on a different meaning if the business has decided it wants to collect clickstream or shopping cart data for three years to discover seasonal customer behavior. Operational data like this would once have just been analyzed on the fly, then deleted. Now it becomes more significant and has a longer shelf life. Although it won’t be accessed often, analysts and management may want it on demand, as might customer service reps in companies embracing operational BI.

Even compliance to data’s ILM might have to change. Since the costs of compliance are so high, the business might want to recoup some of that investment by leveraging it for its BI value, which means that it needs to stay online longer instead of being archived to less expensive storage as quickly as possible. One solution is to treat read-only data differently from data that continues to be updated. Instead of two stages (online and archive), you would use three stages: online current, online read-only, and archive. Read-only data requires fewer resources. Creating and maintaining a separate repository for rolling off read-only data is not a trivial task, but reducing the pressure on resources might justify this additional step.

These changes in data usage have more to do with the reasons organizations are collecting so much more data than with the issue of big data. Negotiations on service-level agreements will have to balance the increased demand against resource availability. New ILM guidelines need to take business significance into account. ILM decisions have typically been made primarily in the data center in consultation with data governance and risk management groups who represented the business. Now, sponsors of the many input data streams will be needed to provide the business context for ILM and resource allocation decisions.

Research Deduplication—An Emerging Practice

A single terabyte of data needs 53 terabytes of storage over its lifetime because of the number of times it is instantiated across multiple applications or data marts, time series snapshots, backups, and replication (Darrow, 2008). Deduplication techniques reduce these many images to a single one, which promises to decrease storage needs by a factor of 20. As a result, 16 percent of data center managers surveyed by InfoPro Incorporated plan to adopt deduplication within a year. For the first time in several years, entrenched data center practices are losing ground to a relatively new technology.

> To the extent that big data is putting a strain on resources, you will want to refine the processes by which you rank projects, subject areas, and data sets. Age of data takes on a different meaning if the business has decided it wants to collect clickstream or shopping cart data for three years to discover seasonal customer behavior.

Business users reading this description might jump to the conclusion that deduplication will bring about the elusive “single version of the truth,” but the impact of deduplication is far removed from the business user. The single image refers to a single image of data on backup tapes or in third-tier storage. As such it is also removed from data integration and master data management initiatives. As the technology evolves, we might see it deployed as part of the data warehouse.

Although deduplication will not reduce the amount of data in a data warehouse, it will reduce the number of tapes or disks in your physical warehouse and lower end-of-life storage costs. Deduping technology is available via software and hardware solutions. It can be implemented off-line (that is, after data has been stored) or in-line (as the last step before data is archived). Even this technology, which seems tailor-made for big-data scenarios, has hard limitations when it comes to disk space. Most solutions require that you first store the data on a storage server, then run it through the deduping procedure. Today's solutions cannot accommodate more than 20 TB for off-line deduping.

A deduping implementation requires the same amount of storage up front—a 20 TB database needs 20 TB of storage space. The reduction comes once a history of deduping the same database is established. Applying the 1-to-53 rule, the initial TB will require 1 TB, but 53 subsequent snapshots will not require significantly more space than that initial TB. The advantage to such reduction is clear: backup and replication will be more efficient, hardware costs will be reduced, and energy and space requirements will be minimized.

Design Your Solution

Retrofitting your infrastructure to handle massive amounts of data is a complicated proposal. There is no single tool or methodology for scaling resources to keep pace with data growth. Neither is there a set of tools or a reference blueprint that lays out a clear path. Best practices are just now beginning to emerge as more organizations cross the size barrier. Even today, on community sites dedicated to SQL, database analysts post questions about how peers handle environments for databases larger than 100 GB.

Most of us are taking inventory of current resources—hardware, software, and expertise—and seeing how far they'll take us. At some point, however, we have to shift perspective, because the size of data changes many of our assumptions about collecting, using, and storing data. The methods that have made data centers more efficient and data more secure need careful review to uncover the weaknesses that appear only when working with large amounts of data. Even the business and usage policies have to be adjusted to scale along with the data and to accommodate new usage patterns.

Even if there is no prescribed approach to managing data growth, there are some guidelines for evaluating solutions and whether they fit your own environment and requirements. Any operation on large data will depend on CPU capacity, network bandwidth, and disk I/O. Many expansion initiatives don't show the expected results because of a failure to consider all three of these components. The burden on these components can be addressed through database and query optimization, workload prioritization, virtualization, hardware investments, or usage and ILM policies.

New data management approaches must address the sheer presence of so much data as well as anticipate how the users change what they expect from their big data and how they leverage it. In other words, don't let the storage dimension overshadow access considerations. The world of compliance offers up some insight here. Let's not forget that every data retention regulation has a "timely access" clause. ■

References

Burke, John [2008]. "Compliance-Related Costs are Rising," *NetworkWorld* Executive Guide: Storage Heats Up, white paper, page 18.

Darrow, Barbara [2008]."Is Fibre Channel Dead?" *NetworkWorld* Executive Guide: Storage Heats Up, white paper, page 5.

McGrane, Sally [2008]. "The Czar of Crepes," *Time*, June 5. http://www.time.com/time/magazine/article/0,9171,1812074,00.html

The McKinsey Quarterly [2008]. "Meeting the Demand for Data Storage," June. http://www.mckinseyquarterly.com/Information_Technology/Management/Meeting_the_demand_for_data_storage_2153?gp=1

Complex Event Processing

Analytics and Complex Event Processing: Adding Intelligence to the Event Chain

Tim Wormus

Tim Wormus is an analytics evangelist for the Spotfire division of TIBCO.
twormus@tibco.com

Abstract

Traditional business intelligence (BI) has largely fulfilled its purpose, and although it will continue to provide reports on structured data, BI is becoming obsolete. Operational decision-support systems will be born from combinations of technologies such as analytic systems (both predictive and ad hoc) and continuous/complex event processing (CEP). These hybrid systems will be embedded into business processes and address day-to-day operational issues. They will be the vehicles by which a greatly evolved BI will become central to the infrastructure of the modern enterprise.

Introduction

Classic business intelligence was created for a world in which a company's most important data was stored in well-structured databases and provided well-defined information to an easily described set of managers and executives. Although not all deployments have achieved the visionary goals described by industry analysts, traditional BI has provided considerable value. However, businesses now face a different world than they did when Gartner first popularized the term *business intelligence* in 1989.

Even as data warehousing technology has improved by leaps and bounds, much time-sensitive, operational data is stored in as diffuse a manner as ever. It may reside in local or departmental databases, but it may also exist in spreadsheets or in the heads of domain experts. In many organizations, it's all of the above. This is hardly the fault of data warehousing vendors or IT departments; it's the nature of operational data. Business challenges are highly dynamic, and the data required to address them is often a poor fit for highly structured data warehouses. However, this doesn't stop a multitude of key, front-line decisions from being made on the basis of such operational data.

Along with the volumes and types of data captured, the technology to process that data has changed enormously. Everything is cheaper, faster, and easier to use, and the technology that was designed to solve problems in yesterday's world is no longer appropriate. This isn't to say it should disappear, but the business problems it addresses best are essentially solved—or at least addressable by commodity technologies. Now, as more data is collected, and as people become more familiar with technology and better connected, we need new ways to address current challenges.

Although BI will continue in its traditional role of providing reports on structured data, operational decision-support systems will be born from combinations of new technologies such as analytic software, CEP systems, and business process modeling systems.

New technologies have appeared in response. Complex event processing allows sets of individual events to be integrated into a cohesive and meaningful whole, enabling sophisticated, rule-based processes to be largely automated. Analytics, both model-driven and interactive, enable analysts to draw vastly more insight and value from ever-growing volumes of corporate data.

Although BI will continue in its traditional role of providing reports on structured data, operational decision-support systems will be born from combinations of new technologies such as analytic software, CEP systems, and business process modeling systems. These hybrid systems will be embedded into business processes and address day-to-day operational issues.

Each of these maturing technologies has demonstrated significant promise. However, in the spirit of Enterprise 2.0, the greatest possibilities result from mashing them up. Combinations of these technologies will be the vehicles by which BI—though no longer bearing much resemblance to the BI of the past—will become pervasive.

Complex Event Processing—An Unfulfilled Promise

To understand the transformative potential of CEP, we will describe how normal, or simple, event processing works. A good example is a bank that sends an automated notification to a customer when his or her checking balance goes below a certain amount. This is a straightforward application of event processing, and it helps the bank provide better customer service. Although few would argue against improving customer service, it's also hard to imagine such notifications transforming the relationship between the bank and its customers.

Imagine that in the same week the same customer also applied for a short-term personal loan and transferred money from a long-term savings account to his checking account. None of these individual events is particularly unusual, but the combination in a relatively short time period could indicate that this customer is in an unusual situation.

A complex event processing system can correlate such isolated events and build a coherent picture of this customer that has meaning at the business, rather than the transactional, level: our hypothetical customer may be having financial difficulty. It's this ability to automatically detect—and potentially react to—subtle and complex cues that makes CEP so powerful.

However, this same capacity to collect and synthesize individual, transactional events into meaningful business events also poses one of the biggest challenges for CEP. Simple events can be combined in many ways, not all of which have business relevance. Even for those combinations that are important at the business level, it's not always possible to determine the best response in advance.

This means that current CEP implementations wisely target the highest-value events that are likely to occur and for which the appropriate response can be readily determined. This substantially under-delivers on the potential of CEP.

Analytic Support for Complex Event Processing

Adding analytics to the mix enables CEP to deliver its full potential. By intelligently processing information about the business, customers, and other relevant data, analytics—both automated and interactive—enables us to move from detection to insight and then on to decision and action, all within the same interface.

The difference between detection and decision is the critical one. In the case of our banking customer, it's impressive that we're able to automatically detect his potential financial difficulties, but it's important at the business level only if our inference is correct *and* we're able to take some action. This is an important point—processes and systems that don't lead to action, or that lead to incorrect action, not only fail to add value, but are also costly and distracting.

Some responses can be prespecified and therefore (in principle) automated. In many cases, however, the appropriate action may not be immediately clear without additional information or processing.

The actions that can be automated should be. If such responses require no additional intelligence, there is no reason to involve a human or another system in the process, except perhaps to receive a notification. (Such notifications should be tracked as events themselves, processed accordingly, and made available for subsequent data mining.)

When the appropriate action isn't calculable in advance, analytic systems make an enormous difference. They can either provide recommendations for action based on statistical analysis—which, depending on your level of trust in the statistical model and the consequences of being wrong, could be automated—or they can provide a context in which human intelligence can be brought to bear on the decision in a rapid and effective manner.

Business intelligence is especially relevant here. Statistical and predictive analytic systems are key elements of adding intelligence to an organization, but for the moment remain largely the province of quantitative specialists.

The ability of interactive analytics tools to enable efficient and effective decision making makes them an ideal fit with CEP. Classic business intelligence has always been about decision support, and by mashing up CEP with analytics and some of the data assimilation and aggregation capabilities of classic BI, new systems can be built that are much more operational in nature. These systems can directly address the needs of front-line users within the context of their business processes.

> Simple events can be combined in many ways, not all of which have business relevance. Even for those combinations that are important at the business level, it's not always possible to determine the best response in advance.

This ability to be embedded in users' business processes, and even enable new processes, separates next-generation BI from classic BI. Before we come back to *how* this can be done, let's briefly discuss why it *should* be done.

The Need to Operationalize BI

Top-down, command-and-control management isn't really how most businesses operate anymore, simply because it takes too long for information to flow up the hierarchy and for decisions to propagate back down. Major, strategic decisions are still handled at the top, but those are already well-supported by classic BI capabilities. It's the day-to-day analysts and business professionals who are making key business decisions without the tools they need.

Even for the brightest and most knowledgeable people, rigorous analytic decision-making beats intuition. Ian Ayres, writing in *Super Crunchers,* gives example after example where regression beats expert knowledge. From medical diagnoses to predicting the behavior of customers, from determining whether a given movie script will result in a blockbuster to finding neighborhoods where housing prices are likely to rise or fall, it's been repeatedly demonstrated that model- and data-driven estimates and predictions, on average, beat expertise and experience.

> Classic business intelligence has always been about decision support, and by mashing up complex event processing with analytics and some of the data assimilation and aggregation capabilities of classic BI, new systems can be built that are much more operational in nature.

Many of Ayres' examples apply statistically driven analytic systems to large data sets. The value of such systems is quickly becoming well known (see Thomas Davenport and Jeanne Harris's *Competing on Analytics* for further examples of organizations that derive tremendous value from such systems). However, they're still often positioned as strategic rather than operational. For one thing, they're incredibly powerful, and it makes sense to have them where they can provide the biggest business impact. In addition, they typically rely on large volumes of highly processed data and sophisticated models to generate their impressive results. These requirements have made deploying statistics-based analytics in an operational environment challenging.

Yet, it's in operational environments that analytics has the potential to be of the greatest benefit, especially by increasing the odds that users will make the right decisions. Under uncertain conditions, no person or system is going to make the correct call every time. However, if analytics improves the chances of getting the correct decision by, say, 10 percent, that can provide enormous lift to a business—*provided that it applies to many decisions.*

The more widely deployed analytics are, the greater the impact. To deploy them widely, however, businesses must determine who needs which applications and how to deliver them within the available bandwidth of IT organizations.

Unfortunately, it's essentially impossible to specify all of the applications people need in advance. Business processes evolve in response to changing conditions, whether due to regulations or a competitive marketplace (or some other factor), and it's unrealistic to expect that the appropriate tools can always be built in advance. Users will always need some flexibility to deal with new situations. The problem with the current situation is that classic BI technology—which effectively addresses reporting and the need to disseminate important information to important people—is being retooled to make it available to more people to do more or less the same thing.

Ultimately, this isn't going to be enough. The challenges that classic BI addresses well are roughly the same across many enterprises. For instance, most finance departments have similar goals. Though there are cultural and budgetary differences, CFOs in most companies are primarily concerned with the same big questions. This is true for most other C-level positions, which is why successful CEOs, CFOs, and CIOs can move from one industry to another. It's as you move further down in the organization that the dissimilarities from one organization to the next become apparent and critical. The business processes of the front-line, operational decision makers in a consumer packaged goods company are quite different from their counterparts in a telecommunications company.

There are certainly processes that are analogous—supply chain management, sales force management, and so on—but the details of day-to-day operations are different. It's not plausible that the same application will meet operations demands in every environment, at least not without so much customization that you may as well call it another application entirely.

To provide intelligence and decision support at an operational level, applications should either be tailored to fit the business process they support or be flexible and easy enough to use that front-line professionals don't need to prespecify their needs. Realistically, it will be a combination of the two. By developing applications in an environment that allows for the ready reuse and combination of pieces of key functionality—using a service-oriented architecture (SOA), for instance—tailored applications can be developed at a sufficiently high level that they don't dictate the process required to use them.

Flexibility is important. Deploying classic BI in an SOA environment won't magically make it operationally useful. The applications that provide operational BI actually need to be part of the business process. A key indicator that this is the case is when front-line professionals think of the application they're using as "the tool we use for process X," rather than "our BI system."

How CEP and Analytics Embed BI in Business Processes

Let's go back to our unfortunate customer who has made a series of transactions that seem to signal financial distress. What would be the next steps in an intelligent business process? The first step would be to validate whether the customer is in trouble; next, to evaluate what response (if any) is the most appropriate; and finally, to act.

A CEP system without analytical capabilities severely limits a user's ability to carry out this workflow. As far as validating the event, we are limited to sending a notification (which is easy), storing the information somewhere for subsequent reporting, or taking some automated action. None of these options is particularly attractive. A notification will at least make someone aware of the situation, but it doesn't provide the user with any additional information or guidance on what action to take. An automatic action might be helpful, but only if the appropriate response is known beforehand, and here it's not clear. Finally, logging the event is good, as eventually we'll want to mine the database that records such events, but it doesn't do us or our customer any good now. We want to know what immediate action to take.

The applications that provide operational BI actually need to be part of the business process. A key indicator that this is the case is when front-line professionals think of the application they're using as "the tool we use for process X," rather than "our BI system."

Integrating the CEP system with an analytics platform vastly improves our ability to determine the appropriate response. For instance, even though we've received an alert, we don't really know whether this particular customer is actually in financial pain. Without embedded tools, a bank employee will need to take several manual steps to validate the event. Perhaps he or she will need to pull data from a few different systems, transforming the data and probably dropping it into a spreadsheet to make a determination. This is generally slow and painful, and ultimately leads to suboptimal decisions.

This is where the combination of CEP and analytics excels: it optimizes the delivery of insight over the "last mile" of technology. Imagine that the CEP system has been configured to deliver a pre-packaged visual analytics application populated with data related to this customer. Included in the application is our particular customer's transaction history, data on his or her credit history or total level of indebtedness, or information about other accounts. We don't know in advance which pieces of this data will be useful, but leave it to the analyst to derive the key insights from the data.

By packaging all the relevant data and the analysis tools, such an application enables the analyst to quickly get a picture of the customer's financial state. It doesn't specify the correct action in advance, but lets the operational decision maker evaluate all the relevant information and come to an appropriate decision.

This is how human intelligence can best be leveraged in the context of event processing. It's not possible to automate human decision making, but it is possible to expedite it by removing the hassle of finding, loading, and aggregating data and enabling a user to manipulate the data as quickly as possible. By presenting all the data necessary to make a judgment in a visual context, we enable the analyst to spend his or her time gaining insight, rather than gathering and merging data.

Delivering such applications to users in response to events is real-time, event-driven BI! In our example, a CEP system has noticed that one of our customers has executed a set of transactions that we have reason to believe indicates some financial distress. One of our analysts has not only been alerted, but has also been automatically presented with an interactive application that lets her validate the event and choose the best course of action. This is an excellent blend of automation and human intelligence.

The marriage of CEP and analytics can be taken even further. It's great to supercharge analysts and let them spend their time dealing with the analytical questions for which they are trained, but in cases where we have the necessary historical data, a predictive analytics system can take this a step further. By comparing the current customer to other customers who have been in similar situations, predictive analytics can calculate the likelihood that our inference about the customer's situation is correct and indicate which actions have historically led to the best outcomes.

Here our analyst receives information about the current customer *and* sees a detailed comparison to similar customers and a projection of which actions are likely to have the greatest positive impact—improving the likelihood that he or she will make the correct decision.

Summary

Blending analytics with CEP systems allows us to leverage automated systems for the things they are best at (monitoring and notification), while enabling business professionals to focus on making key decisions. This makes high-value professionals more efficient and also eases the process of fact-based decision making and improves the decisions.

These targeted analytics applications, delivered to end users by CEP systems, don't have much in common with the static reports with which BI has traditionally been associated. However, by providing a data-driven context for decision making, they continue in the same tradition—though at an operational, rather than strategic, level.

The combination of these two technologies is only one of the many combinations we will see in the coming years. I fully expect to see analytic front-ends on BPM systems and event-driven process optimization enabled by mashing up BPM with real-time event architectures. This is to say nothing of the Web 2.0 collaboration technologies that will eventually be woven into the fabric of Enterprise 2.0. ■

References

Ayres, Ian [2007]. *Super Crunchers: Why Thinking-By-Numbers Is the New Way to Be Smart,* Bantam.

Davenport, Tom, and Jeanne Harris [2007]. *Competing on Analytics: The New Science of Winning*, Harvard Business School Press.

BI Case Study

The Very Model of Success

Stephen Swoyer

Abbott Laboratories tapped Kalido's model-driven data warehousing infrastructure to revamp—and revive—an overworked BI system.

Business change disrupts, undermines, subverts, and in many cases wrecks the best-laid plans or strategies of business and IT executives.

It doesn't have to be that way. Along with disruption, change also produces *opportunity*—to revisit, tinker with, improve upon, and in some cases radically alter a business or IT status quo. Change can be a positive force—even in the most risk-averse of organizations.

Take Abbott Laboratories, a multinational pharmaceutical giant with operations in more than 130 countries. Over the last half-decade, Abbott pursued an ambitious growth strategy in which it significantly augmented its international sales and marketing activities. Growth of this kind entails disruption, and as Abbott sought to integrate and consolidate sales data from 65 international sites, its business intelligence (BI) infrastructure started to show the strain.

Abbott officials saw the impetus for change as an opportunity—not just to address existing or foreseeable demands, but also as a means to revamp existing business processes, introduce new services, improve information delivery and reliability, and insulate their BI infrastructure against the disruptiveness of change.

Abbott tapped the Kalido Information Engine from data warehousing (DW) specialist Kalido to power its new BI infrastructure. Abbott's Kalido-based data warehouse lets the company have it both ways: its BI and DW infrastructure is largely shielded from the disruptive effects of change (e.g., merger and acquisition activity, changes in the business cycle, planned or unplanned growth) even as it gives Abbott more flexibility—and more reliability—than ever.

For example, says Peggy Mathias, manager of HQ IT applications with Abbott, company managers once used aging or incomplete information to supplement their decision making. With the new Kalido-powered system, however, they're able to view accurate, timely data in each of several different contexts—historical, current, or future. The new system also gives Abbott's sales and marketing executives—some 250 of them, in 65 different locations—better insight into how efficiently the company is allocating its sales and marketing investments, particularly in emerging markets. From an IT perspective, Kalido delivers as promised: Abbott can now reconfigure its DW and BI infrastructure in as little as two weeks to accommodate acquisitions, mergers, new services, or other disruptions. No, Mathias concedes, it isn't quite on the fly, but for a multinational company with 65,000 employees, it's breathtakingly responsive.

Best of all, she concludes, Kalido brings a business-first perspective to DW and BI, courtesy of its Business Information Modeler, a graphical, declarative modeling tool that encourages collaboration between DW architects and business stakeholders. "It really lets us manage the data warehouse from a business point of view, and not requiring technical skill to design database tables or figure out what a star schema is and build a star schema," she observes. "We now have a single version of the truth for sales and marketing information that we're using to populate the people who need that information as a verified source, and it [also] ties together with our general ledger system, which was the key, and now we're adding more data because of this success."

Change You Can Believe In

It wasn't always smooth sailing, of course. Two years ago, in fact, Mathias and her colleagues were ready to throw their hands up in frustration—with good reason.

Abbott has had a busy seven-year stretch. In 2001, it acquired Knoll, the pharmaceutical division of German conglomerate BASF. Since then, it has spun off an internal group (its hospital products division, which became Hospira), acquired several entities (including TheraSense, which Abbott merged with its MediSense division, and the vascular products division of life sciences manufacturer Guidant), and notched an agreement with General Electric to sell assets worth about $8 billion. (The latter deal ultimately collapsed—but not before both parties had committed time, money, and energy to transition planning.)

One constant was disruptive change. Mathias and her colleagues had built a functional, reliable BI system, based on BI and database software from Cognos and Oracle, respectively. Abbott's growth was beginning to pose problems, particularly as the company sought to augment its international sales and marketing efforts. "We've had difficulty with having multiple systems with the same sort of redundant information for sales and marketing in particular. We are responsible in [our] Chicago [operations center] for consolidating the results of operations across 65 countries. Once we got all that data here in Chicago, we were having too many different systems that were reporting the information at different levels with different calculations and getting different answers," Mathias explains. "The result was that everybody was spending most of their time doing maintenance with reconciliation and cleaning up the data."

Kalido and its Information Engine were not a silver bullet, Mathias stresses. Before Abbott could lay out its new BI infrastructure, it first had to do some housecleaning: chiefly studying its reporting processes, identifying bottlenecks and other potential problems, and—most important—improving the quality of the data it received from its international offices.

"We had about a year-long process to improve the process of reporting sales information from the 65 countries. We worked for a solid year on the financial reporting process and really cleaning up the data that was being fed into Chicago to make sure that it was consistent and had good quality. After that, we started with the data warehousing," she explains. "We ended up putting a lot of editing rules in place that weren't established previously. We actually put the onus back on the countries to make sure that they were doing the editing at their systems and were sending us data that was fully edited."

Time-to-Implementation You Can Believe In

Kalido was on Abbott's short list once Mathias and her colleagues had finished cleaning up (and improving) its sales and financial reporting processes. There was a reason for that, she acknowledges: at least two other units inside the company had already deployed the Kalido technology. Based partly on feedback from these units and on her own assessment, Mathias decided to tap Kalido.

"What we were intending to do was sending data marts out to each of our 65 affiliates so that they could have their own local copy of the data to analyze locally," Mathias says. "Kalido was already in use in two other divisions in the company, so they were one of the first [options] we considered," she continues. "They weren't the cheapest, but what was really of interest to our executives was the time-to-implementation. Like I said, they weren't the cheapest, but I doubt anyone else could have been implemented as quickly."

Rapid time-to-implementation, more than anything else, ultimately helped clinch the deal. "We did the proof of concept [where] Kalido came in, took all of our data, and in less than a week, had Cognos hooked up to it," she says. "It was amazing how quickly they had it [a proof concept] up and running. We just did the business case that if we had to do the development ourselves, it was going to take a lot more time. It was really the time that attracted us to them."

During the assessment phase, Mathias and Abbott got to see just how flexible—i.e., how quickly reconfigurable—Kalido really is. "When we did the initial demonstration for our CFO, for example, he wasn't happy with the hierarchies, so he asked if they could be changed. It turns out that the way we were reporting financial information wasn't really the way that [finance] liked [to consume] it," she explains. "Now that [reconfiguration] is a non-trivial task in most environments, but with Kalido, it was surprisingly simple. We started out with something like five views in Cognos and we ended up with 20."

Kalido's rapid configurability derives from its emphasis on business modeling, which uses an abstraction layer (basically, a conceptual layout of a business—represented by core business entities such as customers, products, or assets and bound together by pre-defined business rules) in place of hard-coded data models. One advantage of such an approach, Kalido officials maintain, is that it lets line-of-business stakeholders and BI architects view integration through a business-centric prism, understood in terms of declarative business concepts or terms. "Typically you develop [your data model] with hand-coded ETL tools and hand-coded BI configurations, and it becomes … rather expensive to deliver business change," argues Cliff Longman, Kalido's chief technology officer. "The Kalido Information Engine is driven by a business model … where if you make changes, you're making changes to the *abstraction* of the business model. That makes the whole infrastructure much more agile."

Kalido CEO Bill Hewitt frames the issue even more starkly: "It is an abstraction layer that [is designed to] insulate you against the effects of change," he explains. "[Y]ou can make some significant changes to the physical implementation [of a warehouse] that only involves really simple changes to the business model, and if you have the right [modeling tool], you can automate [those changes]."

That jibes with Mathias' experience, too. Once Abbott got down to brass tacks, overall implementation took about four months. Abbott tasked a six-person team—consisting of Mathias and five of her colleagues—with getting things done. As soon as they were finished, Abbott's new Kalido-powered system got its first real test, Mathias says—in the form of a huge potential disruption.

"[R]ight after implementation, we had a geographical reorganization. It was our first geographical reorganization in 30 years! If you ask any business intelligence professional about that, they'll tell you that it's going to be a headache [reconciling] all of that with the data warehouse," she says. "It was all pretty painless in Kalido. Two weeks after that reorganization, we had made the changes in the warehouse and everything was running smoothly."

Outcomes You Can Believe In

Mathias thinks Kalido's business model-driven approach has another advantage, too: it helps foster closer collaboration between line-of-business stakeholders and BI practitioners like herself. "It brings business units and IT closer together. It's designed so that the business and IT can collaborate together [to define the model]. It has an intuitive, graphical interface, and it's clearly designed so business users can have meaningful input into the process," she indicates.

One upshot of this, she says, is that business executives have quickly become hip to the capabilities—and the potential—of Abbott's new BI infrastructure.

"We're starting to add more data [sources] because of the success we've had. For example, our general ledger owner was very skeptical about using his data in the data warehouse, but we've been able to win his business, and we'll be doing more reporting of his results through the warehouse," she explains.

Abbott, like many adopters, didn't commission an official ROI study to assess the dollars-and-cents benefits of its Kalido-centric BI overhaul. Mathias believes such a study would have been both misleading and insufficient; there's a sense, after all, in which ROI studies are backward-looking propositions. To be sure, they *do* help organizations understand how much they're saving (or, in many cases, how much more they're spending) vis-à-vis the ante status quo, and they also identify income that stems from the introduction of new products or services, but they're less compelling when it comes to pinpointing ROI that derives from intangibles—such as potentially costly business disruptions. It's difficult to put a price on agility, Mathias argues, citing the geographical reorganization that Abbott completed in just two weeks after it went live on Kalido. An effort like that would have been extremely costly—chiefly in terms of person hours—using Abbott's old, loosely federated system.

Moreover, she maintains, much of the system's value accrues from as-yet-unrealized projects, services, and products, many of which Abbott plans to roll out over the coming months. To a surprising degree, business stakeholders—and not just IT visionaries—were quick to see the potential value in the system. "I had a really strong business sponsor who … did some prototyping … before we went ahead [with the implementation]," she says. "He was doing some really interesting things to build the business case for how valuable this [project] could be and working through the executives to show them how valuable this can be."

One forward-looking project involves the unprecedented (for Abbott) introduction of analytic capabilities into greenfield business processes. Abbott's CFO is particularly excited about an effort to augment the company's monthly sales reports with analytic insights, Mathias reports.

"We're working with the CFO now to do some more use cases with him, so we can start to embed the use of the warehouse into the business process. The main thing is the monthly sales flash process, where [analysts are] looking at the numbers and starting to give their initial feedback on why the numbers are what they are. Initially, we're looking to be able to do some analysis on the monthly sales flash—what's good, what's bad, what's going on."

You Have to Believe

Abbott completed its Kalido implementation in just four months, but—in many respects—its broader BI project remains unfinished.

At this point, Mathias says, managers can generate reports based on key performance indicators, product profitability, and finances. They have better insight into sales information, too, inasmuch they can now break down sales by location, currency, product, or even financial metrics. Mathias sees this as just the beginning: the tactical benefits, so to speak, of a Kalido-based BI strategy. "It just introduces this whole new model for delivering information to the business. This is just a foundation to build upon for other data types and for other applications. The flexibility is just tremendous, and we're very confident that we have a very solid strategy to build on going forward." ■

Stephen Swoyer is a technology writer based in Athens, GA. Contact him via e-mail at stephen.swoyer@spinkle.net.

Editorial Calendar and Instructions for Authors

The *Business Intelligence Journal* is a quarterly journal that focuses on all aspects of data warehousing and business intelligence. It serves the needs of researchers and practitioners in this important field by publishing surveys of current practices, opinion pieces, conceptual frameworks, case studies that describe innovative practices or provide important insights, tutorials, technology discussions, and annotated bibliographies. The *Journal* publishes educational articles that do not market, advertise, or promote one particular product or company.

Editorial Topics for 2009

Journal authors are encouraged to submit articles of interest to business intelligence and data warehousing professionals, including the following timely topics:

- Project management and planning
- Architecture and deployment
- Data design and integration
- Data management and infrastructure
- Data analysis and delivery
- Analytic applications
- Selling and justifying the data warehouse

Editorial Acceptance

- All articles are reviewed by the *Journal's* editors before they are accepted for publication.
- The publisher will copyedit the final manuscript to conform to its standards of grammar, style, format, and length.
- Articles must not have been published previously without the knowledge of the publisher. Submission of a manuscript implies the authors' assurance that the same work has not been, will not be, and is not currently submitted elsewhere.
- Authors will be required to sign a release form before the article is published; this agreement is available upon request (contact journal@tdwi.org).
- The *Journal* will not publish articles that market, advertise, or promote one particular product or company.

Submissions

www.tdwi.org/journalsubmissions

Materials should be submitted to:
Jennifer Agee, Managing Editor
E-mail: journal@tdwi.org

Upcoming Submissions Deadlines

Volume 14, Number 2
Submissions Deadline: February 27, 2009
Distribution: June 2009

Volume 14, Number 3
Submissions Deadline: May 22, 2009
Distribution: September 2009

Volume 14, Number 4
Submissions Deadline: September 4, 2009
Distribution: December 2009

BI StatShots

The Perceptions and Realities of Sharing Customer Data

We asked respondents to rate their organization's perceptions and efforts as high, medium, or low for issues related to sharing customer data. (See Figure 1.) Issues were grouped in pairs to reveal conflicts between perception and reality.

- **Most organizations think that sharing customer data is highly valuable (59%).** But sharing mostly reaches medium and low percentages of the enterprise (44% and 38%, respectively). Here, as with many CDI issues, the perception of potential benefit is way ahead of the action being taken to achieve the benefit.
- **Employee access to customer data is medium to high in most organizations.** Yet, the completeness of the data is mostly low (55%). The good news is that organizations are sharing customer data; the bad news is that the data being shared is sketchy.
- **All respondents rated very highly both benefits and problems.** Few respondents rated benefits or problems as low. Clearly, shared customer data yields perceptible benefits, just as the lack of it results in noticeable problems.
- **The perceived quality of customer data is mediocre, as is the effort put into improving it.** This isn't bad, given that customer data (due to its constantly changing nature) is more prone to quality problems than most data domains. This explains why the majority of data quality solutions focus on customer data, whereas other domains get little or no quality improvement.
- **Half of respondents consider their organization's CDI success mediocre (50%).** Few rate their success as high (11%), and a considerable percentage (38%) rate their success as low. This shows that CDI solutions have plenty of room for improvement in most organizations.

—Philip Russom

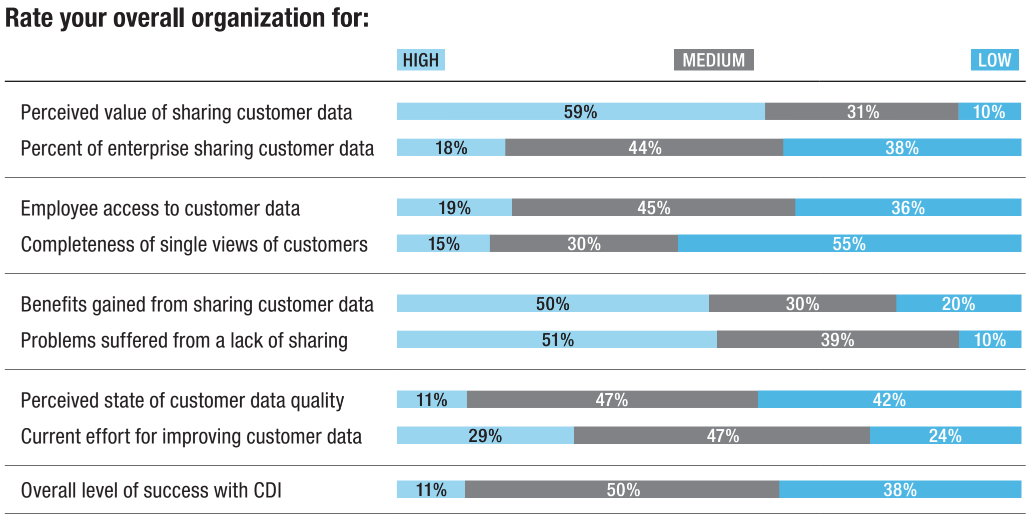


Figure 1: Based on 352–357 respondents per answer

Source: Customer Data Integration: Managing Customer Information as an Organizational Asset, *TDWI Best Practices Report, Q4 2008.*

Bogged down
in your BI project?
We can help.

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

Business Objects™ an SAP® company

COGNOS® AN IBM® COMPANY

CONNECT: The Knowledge Network

DATAFLUX® A sas COMPANY

DATAllegro DATA AT THE SPEED OF BUSINESS

DATAUPIA™ Free your data

DecisionPath CONSULTING

hp® invent

IBM®

INFORMATICA® The Data Integration Company™

Information Builders

kognitio Competitive advantage from data

Microsoft®

MicroStrategy®

NETEZZA Question Everything™

ORACLE®

Pitney Bowes GROUP 1 SOFTWARE

sas®

syncsort

TERADATA® Raising Intelligence