EXCLUSIVELY FOR
TDWI PREMIUM MEMBERS

VOLUME 18 • NUMBER 1

Business Intelligence Journal

THE LEADING PUBLICATION FOR BUSINESS INTELLIGENCE AND DATA WAREHOUSING PROFESSIONALS

tdwi
THE DATA WAREHOUSING INSTITUTE

TDWI ONSITE EDUCATION

BI Training Solutions:
As Close as Your Conference Room

TDWI Onsite Education brings our vendor-neutral BI and DW training to companies worldwide, tailored to meet the specific needs of your organization. From fundamental courses to advanced techniques, plus prep courses and exams for the Certified Business Intelligence Professional (CBIP) designation—we can bring the training you need directly to your team in your own conference room.

YOUR TEAM, OUR INSTRUCTORS, YOUR LOCATION.

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

tdwi.org/onsite

tdwi
ONSITE EDUCATION

VOLUME 18 • NUMBER 1

Business Intelligence JOURNAL

Business Intelligence Journal

VOLUME 18 • NUMBER 1

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

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 818.814.5200; Fax 818.734.1522
9201 Oakdale Avenue, Suite 101, Chatsworth, CA 91311

Business Intelligence Journal **(article submission inquiries)**
Jennifer Agee
E-mail: journal@tdwi.org
tdwi.org/journalsubmissions

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

From the Editor

It's never too late to learn something new. This issue of the *Business Intelligence Journal* is loaded with lessons learned and new perspectives on many of BI's hottest topics.

Senior editor Hugh J. Watson is joined by Barbara Wixom and Bruce Yen in reviewing the six lessons that GUESS, a designer clothing company, learned from its mobile BI efforts. As you'd expect, making applications speedy and providing support for bring-your-own-device are on the list, but there are some surprises as well.

As every schoolchild knows, you need the three Rs—reading, 'riting, and 'rithmetic—to get ahead in this world. To that I would add "reality"—as in learning from real-world experience. Kevin Lewis goes beyond a textbook's broad concepts to explain the nuances he learned from building his own data warehouse. Lewis's hands-on perspective will help you learn what makes a data warehouse project a success.

Speaking of keeping up with the times, Scott Walters understands how social media has changed the way people communicate and connect with each other. He explains the challenges of leveraging social media and lays out a six-step plan to help you get started.

One thing IT has learned is that users increasingly want to go the "do-it-yourself" route. That's one reason why self-service reporting has become such a hot topic. Myron Weber will teach you the right way to provide such reporting using a unified design, targeted outputs, sound development practices, and active governance.

There are many more things to learn in this issue. Troy Hiltbrand explains his study of human behavior as managers dealt with budgets and how that knowledge was used to build a successful predictive analytics system. Our experts tell a job seeker what he needs to know before accepting a BI director position. Our case study explains how Merkle learned to deal with big data to create personalized marketing campaigns, and David Teplow details why relational databases aren't suitable for handling big data. He describes the advantages Hadoop offers.

Katrina Read clarifies why "in-memory" technology isn't always the right choice, and Frank Buytendijk explains why philosophy is so important in our fast-paced world of innovation.

What have you been learning? We welcome your feedback and comments; please send them to jpowell@tdwi.org.

James E. Powell

Delivering Value through Mobile Business Intelligence

Hugh Watson, Barbara Wixom, and Bruce Yen

Hugh J. Watson is a professor of MIS and a C. Herman and Mary Virginia Terry Chair of Business Administration in the Terry College of Business at the University of Georgia.
hwatson@terry.uga.edu

Barbara H. Wixom is an associate professor of IT in the McIntire School of Commerce at the University of Virginia.
bwixom@virginia.edu

Bruce Yen is a BI director at GUESS Corporation.
bruceye@guess.com

Mobile BI is high on many BI directors' agendas. Mobile BI offers portability and easy access to BI, which can potentially drive pervasive BI use throughout an organization. Much as the movement from client/server to Web-based BI was transformational, so, too, is the shift to smartphones and tablets. Although mobile BI's adoption rate has been slower than was initially anticipated, surveys show that it is now taking hold and is widely considered to be very important or even critical to business success (Dresner, 2011).

BI directors have many decisions and choices to make when planning for mobile BI. Which devices should they support? Should they provide users with devices or have them bring their own? Should they deploy Web-based or "native" applications? Which vendor should they use? How should privacy and security issues be handled?

Over the past few years, many companies have moved into mobile BI, including GUESS, a $2.69 billion global company that designs and sells contemporary clothing. In 2011, GUESS won a TDWI Best Practices Award for its GMobile initiative, which delivers BI on iPads (Briggs, 2011). This article will describe GUESS's approach to mobile BI and lessons learned by GUESS and other companies that can help BI directors in their initial or ongoing mobile BI efforts.

About GUESS

Since 1981, GUESS has been one of the most widely recognized apparel brands in the world. GUESS designs contemporary clothing and accessories for women, men, and children and distributes its merchandise to stores in 87 countries. Its designers focus on creating fashionable product lines, and they closely monitor best sellers to understand trends. Merchandisers ensure that products

Figure 1: GMobile iPad application.

are placed in the right stores at the right times to meet the needs of GUESS's fashion-savvy customers. Designers, merchandisers, and executives travel extensively across the global GUESS network and are a highly mobile workforce.

BI on Mobile Devices

In early 2008, GUESS successfully rolled out a mobile BlackBerry application to executives and regional sales directors for sales reporting.

In 2011, GUESS delivered a BI app for the iPad after CIO Michael Relich and BI director Bruce Yen became convinced that the iPad could be a game changer in delivering graphical analytics to the company's highly visual and creative business users. The app provides interactive access to best-seller trends and current sales figures for products, arranged by styles, colors, and stores. The app meets the needs of a wide range of work styles and processes and was adopted by a diverse set of business users. GUESS refers to the iPad app as GMobile.

Executives use the app to understand the company's current sales, profitability, and performance trends over time. Merchants use GMobile for store visits. The app allows a merchant to understand how a store is performing and what mix of products accounts for the store's best sellers. Designers use the app to view and analyze their best sellers so they can understand the current sales and design trends. See Figure 1 for a screenshot of GMobile.

The BI team at GUESS found that deploying BI on mobile devices was different from their past BI efforts. Here are seven lessons they learned.

Lesson #1: Design with the device in mind

Not all mobile devices are created equal. In fact, each device—whether it is an iPad, a BlackBerry, or an Android phone—has unique capabilities to leverage and important constraints to consider.

Although smartphones are important devices for delivering certain kinds of information (such as alerts),

they have inherent limitations (such as screen size) that make them less than ideal for displaying some kinds of information. At GUESS, the BI team distributed reports via BlackBerrys. Because the devices have small screens, the team incorporated few bells and whistles, and they listed the most popular reports first to reduce the need to scroll through a long list to find information.

With the iPad, GUESS developers took a much different approach. They created a highly interactive and visually appealing app that takes advantage of the iPad's graphical capabilities. The team produced colorful charts and graphs, and combined data with maps to communicate performance by geographical regions down to the store level.

Instead of replicating dashboard displays that they created for their non-mobile, Web-based dashboard application, the team created GMobile with an investigative workflow that allows users to take a myriad of paths through the app and easily return to an earlier screen—or the home screen—at any point. The app incorporates common Apple-supported gestures such as swipes and taps. Because of the interactive nature of the interface, GMobile integrates content that is equivalent to 12 dashboards from the Web-based system.

Lesson #2: Ensure the device delivers BI quickly

Mobile BI veterans agree that speed is the number-one usability factor for mobile BI; most try to achieve a response time of three seconds or less for their mobile BI applications. Enabling fast response can be easier said than done, however, particularly when working with devices such as iPads, which have limited processing and memory capacities. To achieve performance improvements, development teams may need to break up long-running queries into multiple screens or limit background loads on the device when it is idle.

The GUESS BI team decided to help close the performance gap for the GMobile app by implementing a data warehouse appliance because the iPad simply could not perform quickly without it. The team implemented a columnar database appliance, and queries that took 20 to 30 minutes on their traditional data warehouse technology improved to 5 to 10 seconds, making product affinity and market basket analyses more feasible. Because of its speed, the appliance was jokingly called "the Maserati" after the race car, and the nickname stuck with users.

Lesson #3: Develop a "bring your own device" (BYOD) policy

When the iPad app was first introduced at GUESS, about half of the users brought their own iPads to work and wanted GMobile loaded onto them, either because there was limited availability of devices or because users didn't want to carry two iPads. The BI team was concerned that an iPad might be lost and that confidential company information could be compromised. To address this issue, users signed a waiver permitting software to be placed on their iPads that would enable the BI team to remove the GMobile app in case an iPad was reported lost or stolen. Today, more company iPads are available, and most employees opt for a GUESS iPad rather than using their own.

Every company needs to create a BYOD policy that states whether and how employees can use their own devices and the controls that users must accept if they do. A balance must be struck between the risks associated with the loss or theft of a mobile device and the impact that security controls have on ease of use. Appropriate mobile security software must be selected to safeguard information that is available through the device. Increasingly, mobile security software incorporates location awareness as an additional way to control where, when, and what data is available for viewing on devices.

Lesson #4: Leverage the excitement of the "it" device

Mobile devices such as the iPad have a "coolness" factor. Nearly everyone wants the trendy mobile device, including executives. At GUESS, some users initially wanted BI, in part, because they really wanted an iPad. This did not concern the team because over time, users with iPads ultimately became highly engaged BI users.

At GUESS, an interesting consequence of using a popular device for mobile BI is that the staff fields many questions about iPads that are unrelated to BI. Users approach the GUESS IT group for device help. Instead of explaining

that such questions are out of scope, the team uses the requests as an opportunity to satisfy their users, thereby strengthening the IT-user relationship.

Lesson #5: New roles are needed to best exploit the device

The development of mobile BI applications at GUESS required two new types of project team roles. The first was a developer who was adept at creating apps specifically for the iPad. This individual ensured that the app leveraged Apple widgets and was designed with Apple-savvy users in mind. The BI team realized that their GMobile app was not competing with other IT systems, but rather against other iPad apps that users interacted with regularly, such as the game *Angry Birds*.

Another new role at GUESS was a graphic artist who ensured that GMobile looked good. The designer helped implement a visually appealing app that incorporated a theme with related graphics and colors. The theme was a virtual storefront with a clean, modern, and futuristic look that incorporated best-seller product photos and campaign images.

Lesson #6: Expand the BI experience to include device features

Devices come with a wide variety of features to appeal to consumers, and some of these features (such as cameras, location awareness, and SMS) may serve new and helpful purposes for BI applications. At GUESS, users found their iPad cameras to be particularly helpful for decision making and began incorporating their photos into daily work processes. They took pictures using the iPad to capture store layouts, window designs, and even competitor marketing efforts. These photos were then e-mailed to headquarters or decision makers, or saved to the iPad to be referenced in upcoming meetings.

Some companies are investigating how to use device cameras to serve as bar code readers. This would enable users to scan product codes and use them to generate reports. Similarly, a device's location awareness can be incorporated into BI reporting and produce geographically based reports according to where a user is working.

Lesson #7: Expect a variety of benefits from a mobile workforce

In many companies, decisions are not made behind a desk. Instead, they are made by highly mobile employees who make decisions and take actions throughout the day and who require on-the-go, instant access to BI. Saving time is one benefit of mobile BI that can really add up. At GUESS, store merchants used to take hours preparing for store visits because they needed to gather reports and review them before entering the store. Now, merchants can sit on a bench outside a store with their iPads and get up to speed in minutes.

Similarly, a biotech company calculated that each of its sales representatives saves 30 to 90 minutes each day using mobile BI, which translates into a more than $4 million annual savings and productivity boost for the company.

Some benefits from mobile BI are less tangible, yet important. For example, some companies are reaping reputation benefits from having employees using devices that are perceived as leading edge. At GUESS, the iPad device and appealing GMobile app resonated well with the company's many global partners. In addition, GMobile's adoption and popularity fostered innovation internally at GUESS, prompting other iPad-related projects elsewhere in the company. ■

References

Briggs, Linda [2011]. "BI Case Study: Apparel Company App Melds Fashion, Mobile BI," *Business Intelligence Journal*, Volume 16, Number 4.

Dresner, Howard [2011]. "Mobile Business Intelligence Market Study," Dresner Advisory Services.

Four Lessons Learned Building a Data Warehouse in the Real World

Kevin Lewis is a professional services partner with Teradata specializing in enterprise data management strategy.
kevin.lewis2@teradata.com

Kevin Lewis

Introduction

It has taken me a long time to fully appreciate that the context within which a data warehouse program is established is at least as important as the architecture and design of the data warehouse itself.

The most popular books and articles on data warehousing do a good job of explaining broad concepts and even provide some detailed information on what a well-architected data warehouse should look like. Unfortunately, these books present what sometimes appear to be idealized versions of reality. They don't reflect the nuances that make a program successful in a real-life, complex organization.

In this article, I share what I've learned initiating, building, and institutionalizing my own data warehouse program so that others can learn how to make their data warehouse project a success. [Editor's note: The author changed some minor details to maintain company confidentiality.]

Background

Other than my sometimes painful struggle to balance theory and pragmatism, the common theme running through my career has been distaste for seeing the same work done over and over again. I have also discovered that if you complain about something enough, eventually the boss will tell you to go ahead and fix it.

Before I became a data architect at a large grocery retailer, I had noticed that several independent projects were collecting the same data from the same sources in slightly different forms—repeatedly. In addition to the projects that were proactively collecting data from these

sources—the point of sale (POS) and product management systems being the most popular—there were many ad hoc reactions to production issues, which resulted in a number of database copies being created to prevent people from directly querying transaction systems. These individual projects contributed to an already-complex environment of redundant and overlapping data that had been created over the years.

Initiating the Data Warehouse

After reading a couple of books about data warehousing and deciding that I'd found the solution to our disparate data problem—at least in the decision support environment—I began trying to sell the idea to management. In my first attempt, I assembled a presentation that showed our current systems on the left-hand side (a typical spaghetti chart) along with a nice, neat data warehouse "target state" picture on the right-hand side that depicted all the sources with a single feed to a central data warehouse. I explained that each of these integration points costs money, and that it would make sense to integrate each data source once rather than multiple times. I even provided a rough estimate of the time and cost of building and supporting the more significant interfaces. This did not generate any excitement.

What went wrong? I decided that the problem was that I hadn't involved business users and sponsors. Everything I'd read said the program should be "business driven," and I realized that up to this point I had spent the time I had available understanding and explaining the technical benefits. I identified and spoke with end users in several business areas, learned about their data challenges, and worked with a few to calculate the return on investment for the new analysis they would do and estimate the time they could save if only the data was integrated and easily accessible. I was sure that this more comprehensive presentation would get the attention (read: funding) that I was looking for. It did not.

Later, when I was looking at some papers on my desk, I had an idea. Like everyone in a leadership role, I had been given a list of the "top 10" projects for the coming year. This list was a set of initiatives in priority order. They were the most important projects that involved IT—not "data warehouse" projects, just projects. I looked down the list and noticed what many of the projects had in common. There were projects to optimize shelf space, improve pricing strategies, and select sites for new retail locations, among others.

I didn't need to know all the details to figure out that if left alone, these projects would once again collect sales data from the POS and product data from the product management system. In addition to the read-intensive, integrated data they would need to feed their proposed applications, it was also easy to predict that—whether they planned for it or not—a decision support system would be needed to monitor and manage the new business capabilities.

I didn't use PowerPoint for my next pitch. I simply brought the Top 10 list with a few of the projects circled. I explained how each of these projects would be doing similar work if we were to approach them independently, as planned, and that this work was going to cost significant time and money. I had plenty of evidence based on the current state of IT systems, alluding to much of what I'd explained in previous attempts to sell the program. I then proposed that we remove the sales and product data collection from these projects—adjusting schedules accordingly—and establish a project that I would lead to collect the data for them.

We would also need to introduce new work that wasn't planned in these projects—the analytical needs that would eventually have been built reactively. (I referenced many examples of this sort of reactive work that had happened in the past, such as the many times a "copy" of an operational database was created to meet the "query" needs of users.) With my proposed approach, the data collection, integration, and delivery work would be done once and the results would be shared. That was when the CIO asked what I needed to make it happen. Thus began data warehouse phase 1.

Lesson #1: Find business initiatives that are already important to the organization and explain how the data warehouse will enable those initiatives. Complement planned initiatives; don't compete with them.

Extending the Data Warehouse

Just before the end of phase 1—which we delivered a little bit late and over budget (but with exactly the results we had promised)—I began thinking about what we would need to work on next. Unfortunately, I hadn't completely grasped the importance of the lesson I had just learned. Instead of looking at the next batch of planned business initiatives and the common data they would need, I focused primarily on which data domain should be next—a subtle but crucial difference.

Yes, there were projects that could benefit from the inventory data I proposed for phase 2, and yes, this inventory data was important and widely used within my enterprise, but the connection from the data warehouse to the business initiatives and objectives was very loose. I did not establish a hard dependency. Without a hard link to business initiatives (real, already-approved business initiatives that increase sales or reduce costs), I didn't have the best quality business drivers and, more important, I had no way to drive out the detailed scope of the data warehouse phase.

Without clear business objectives driving data warehouse phase 2, any and all inventory data was in scope. For example, there was no good way to decide which data elements should be modeled and integrated. Instead of asking a question such as "Will either the vendor accountability or distribution network initiatives require these elements?", we would ask, "How important do we think the data will be? Is it 'core' data?" With this approach, any data quality issues discovered were equal candidates to be dealt with in the project. Not surprisingly, we significantly exceeded the budget and missed the deadline. It was only through force of personality and a few testimonials from the end users on the importance of the data that the project wasn't cancelled. In the end, thankfully, the data was used for real business purposes and was integrated at a detailed level with the data previously delivered in phase 1. In short, I was lucky.

Lesson #2: Scope each data warehouse project based on the data needed in near-term business initiatives and deliver *only* that data.

With subsequent phases, we began to triangulate back to more targeted purposes by linking to business initiatives that required the data being delivered in the data warehouse. One reason that the scope of previous phases had been too large was my own fear that we would miss something. I was worried that if we didn't keep the scope wide—at least within a given data domain such as sales or inventory—that we would regret it later and have to perform significant rework to meet new requirements. We learned that if we focused on near-term requirements *and* built the data warehouse to be *extensible* and *scalable*, then we could expand the data warehouse with each phase while controlling scope and completing the project in a reasonable time frame. We established principles to:

- Model data at the lowest level of detail—that is, even if transaction-level data wasn't called for in the immediate requirements, we would still model at this level, a carefully controlled exception to the "meet only known requirements" rule. This allowed us to extend and summarize the data in any way.

- Stage as much data as possible from a new source, but bring the data only as far as the staging area if it isn't immediately needed for near-term business initiatives. (This is the difference between a monstrosity of a project that never ends and a focused project that delivers value.) Only model, integrate, and manage the quality of data needed in the near term. Staging additional data elements made it easier to integrate and manage data later when it was needed, and provided some history for initial loads.

- Obtain data from the source closest to the original point of entry (the "system of record"). With this rule, new projects are able to obtain additional data that may not be available from intermediate sources. It also allows the timeliest data acquisition possible.

- Build "right-time" and adjustable integration processes. This allowed us to meet near-term data freshness goals (e.g., yesterday's sales data versus sales data from 10 minutes ago) and enabled future business objectives to increase data freshness as needed. (Near-real-time integration can have significant complexity and cost,

so this principle avoids investments until that level of data freshness is needed—another lesson I learned the hard way.)

- Model integrated data based on stable business entities, not source system structures. Following this rule provided us a scaffolding to integrate the data in a way that was meaningful for the business and also enabled us to modify or replace source systems without significant (or any) ripple effects for the data warehouse data model or dependent applications. (I made a very big deal about the work that never had to be done because of this approach. It saved millions of dollars.)

- Leverage scalable infrastructure. This minimized our investment in technology by deferring it until needed, and took advantage of the increase in infrastructure capability over time. We deployed infrastructure with as close to linear scalability as possible. Incidentally, the same applied to building the data warehouse organizations and processes—the rule was to implement only what was needed when it was needed.

From time to time, I had to re-argue the case for an enterprise, integrated approach to data warehousing, which is understandable. A common question was, "How are you going to build a data warehouse to meet requirements you don't even know yet? This is a nice idea, but can't work in reality." This is when I would answer—after acknowledging previous scope challenges and the reasons for them—that we could apply rules that would give us at least a high likelihood that we could scale in the future. I'm happy to say we repeatedly proved this to be true. There were many unforeseen applications that ended up using the same, shared data. Yes, there was additional work to extend the model and integrate additional data, but there wasn't significant rework. By focusing only on near-term needs, the size and scope of the projects I proposed (and we delivered) were from 50 to 75 percent smaller than data warehouse projects that had not followed this approach.

Lesson #3: Establish architecture and design principles of extensibility and scalability so that each data warehouse project contributes to an integrated data warehouse.

Institutionalizing the Program

As we were building out the data warehouse phase by phase, I was always on the lookout for projects that, if left untouched, would deliver redundant and overlapping data. I would like to claim that I was always able to leverage the data warehouse in these projects when appropriate, but the truth is that I won some and I lost some. Sometimes the projects were too far along in their life cycle to make any significant changes. Once dates are set, they are hard to change, so I kept looking for ways to find out about these projects as early as possible. I didn't want to wait for "requests" to come my way because I wanted to be sure that the data we were proposing for the data warehouse was the right data—that is, data that requires integration, is shared by multiple business areas, is high volume, read intensive, and, most important, required by business initiatives.

Fortunately, there were funneling mechanisms that all projects had to go through to secure approval, which I realized could give me the opportunity to review the project proposals and assert the role of the data warehouse. There was an annual process in which each business area, along with its assigned IT liaison, would propose projects for the coming year. I would review this list line by line to find common data needs. It was like playing *Concentration* (the card game—and later television game show—where you find cards that are alike).

I did not attempt to uncover any detailed requirements, as that was not possible with a one-line description of each project. However, there was enough information to understand the core data that would be needed and a very likely role for the data warehouse. For example, with only the phrase "improve targeted marketing," it didn't take much thought to conclude that any CRM application would require customer, sales, and promotion data, and that the solution would need analysis capabilities to monitor and improve the promotion effectiveness.

Over time, I found other processes to which I could attach the data warehouse program in order to make it an integral part of the organization. These included:

- **Strategic business and IT planning:** As the corporation as a whole and individual business areas began planning their major initiatives further in advance (roughly 18 months to 5 years out), I did everything I could to get involved so that I could plan the data warehouse deployment in support of these road maps. This would help to refine my own road map while ensuring linkage to pre-vetted business value.

- **Enterprise architecture:** Like most (if not all) organizations, data hasn't been the only victim of project-by-project thinking over the years. Enterprise architecture has emerged and reemerged as a way to begin sharing business processes, applications, infrastructure, *and* data across projects. By making usage and extension of the data warehouse part of the overall enterprise architecture program (under the enterprise data architecture umbrella), we were able to work as a team to find checkpoints before and during projects. (We tried to emphasize "service" over "enforcement.")

- **Program management:** The project management office (PMO) had established funding, tracking, and inter-project dependency management for all large programs, so I made sure to register the data warehouse program as just another program to be managed, and I had a full-time program manager assigned to the program to manage delivery across projects. I considered it a sign of success if another project outside of the program was dependent on a data warehouse project. By funding the deployment of shared data through a central funding mechanism, I didn't have to wrestle with application projects to apply some of the rules I mentioned earlier, even if they didn't directly benefit the project. Of course, I had to do this with great care to avoid delaying the dependent projects.

- **Solution development life cycle (SDLC):** We had an SDLC that was used relatively consistently, so we were able to embed data management practices directly into the methodology and establish a template work breakdown structure (WBS) for data warehouse projects. With a deliverable and milestone for solution architecture in all projects prior to design, we established a formal checkpoint (with enterprise architecture) to see if the data warehouse would be used appropriately within each project or if the project was about to deploy redundant data.

My goal was to make effective planning, delivery, and use of the data warehouse a natural part of the organization's processes. As I have already confessed, this was accomplished with fits and starts, but I do believe that we created something of real value and struck a healthy balance between theory and pragmatism.

Lesson #4: Link the data warehouse program to planning and execution processes within the organization to make the data warehouse a natural part of the organization's functions.

Taking Lessons on the Road

I've since moved on and have been a data warehouse consultant for the past five years. It's only after visiting dozens of companies and studying their situations that I have been able to carefully consider the lessons I learned in my own program. I have seen many patterns in these organizations that reflect the same mistakes I made and some of the same hard-earned course corrections.

There are certainly some success stories, but there are also many programs in serious trouble. Although there are multiple root causes, typically these data warehouses are at one of two extremes. They are either too closely tied to individual business projects (thus exacerbating their disparate data problem), or they are insufficiently connected to business initiatives and projects, and are therefore working hard at delivering data and buzzwords with no apparent business purpose driving the scope, and little—if any—anticipation from the business.

Few companies are institutionalizing their data warehouse programs by linking them to other processes in their organizations. Often, managers and practitioners within these programs realize something is wrong, but they can't quite figure out what it is. If this is your situation, I hope that you will find a few hints from my lessons learned as to what's happening and some ideas to get your data warehouse program on track. ■

Beyond Listening

Six Steps for Integrating and Acting on Social Media

Scott Walters is the global solution leader for social intelligence at HP Enterprise Services.
scott.walters@hp.com

Scott Walters

Introduction

The onset of social media technologies has fundamentally changed the way people communicate and network with other people. Because of these technologies, information flows instantaneously between friends, family, and businesses via numerous devices.

The popularity of these social media platforms is such that Facebook has 1 billion active users (as of October 2012)[1]; the average number of tweets people send per day is about 340 million (as of March 2012)[2]; and 72 hours of video are uploaded to YouTube every minute.[3]

Although much of this content relates to noncommercial topics, an explosion of business-to-consumer and consumer-to-business interactions via social channels has changed consumer behavior and their expectations about how they interact and transact with businesses.

Social networking channels contain vast amounts of consumer behavior information. Much of this social media activity can be tied back to individuals to create highly valuable customer profiles. Leveraging social media data to create more complete customer profiles is critical to effective marketing. Enterprises that understand their customers and engage with them on their terms—when and where they want—will be at a significant competitive advantage.

---

[1] Source: Facebook Key Facts (http://newsroom.fb.com/content/default.aspx?NewsAreaId=22)

[2] Source: Twitter Blog: "Twitter Turns Six" (http://blog.twitter.com/2012/03/twitter-turns-six.html)

[3] Source: YouTube Statistics (http://www.youtube.com/t/press_statistics/)

The Benefits

Recognizing the value of these sources of social data and leveraging them appropriately is a relatively new challenge for many organizations. Social intelligence is a discipline that combines social media data with traditional customer data sources to yield deep insights that drive better marketing and overall business decisions. The benefits of using social intelligence include:

- Enhancing visibility and understanding of customers for better insight and foresight to develop customer advocacy
- Proactively protecting and enhancing brand reputation by monitoring and managing market sentiments, perceptions, and trends
- Spurring product or service innovation by leveraging market insights gained by listening, analyzing, and engaging with customers through social media

The Challenge

Although many benefits come with leveraging social intelligence, social media data can also bring challenges:

- **It is big:** It is much larger in scale than traditional customer information
- **It is different:** It consists of unstructured data that doesn't fit into typical data warehouse structures
- **It is out of the enterprise's control:** The data is created by people who are not affiliated with the enterprise, and the data resides on external domains such as Facebook and Twitter

Most organizations already struggle to integrate all relevant data sources and use the results to create a consistent, customer-focused experience. Unfortunately, traditional business intelligence (BI) approaches may fail to provide the flexibility, timeliness, and mobility required to respond to the real-time marketing demands of this new environment. In addition, legacy BI solutions struggle to support analysis of social media data sources and the integration of these sources with existing customer information. As a result, vital insight remains unavailable to marketers and decision makers across the enterprise.

Legacy BI solutions struggle to support analysis of social media data sources and the integration of these sources with existing customer information.

In many cases, a social intelligence effort will require rethinking and redesigning existing information management ecosystems. New information management and analytical platforms, techniques, tools, and governance processes are needed to unlock customer insights and make them available in real time. New roles and skill sets that combine business acumen and analytical/technology savvy may be required.

Social and other unstructured customer data sources such as call center recordings, videos, chat sessions, and e-mails will be a significant challenge for IT staff who may already be overwhelmed with legacy data challenges. In many cases, business users have taken matters into their own hands and implemented point solutions in an attempt to exploit these valuable data sources. This creates an opportunity for IT staff to help the enterprise create structure and order. The IT department knows how to plan, manage, and govern complex information platforms, and by marrying that expertise with the business context, IT can bring value.

How to Get Started

Social intelligence is a relatively new area, but it requires many of the traditional information management activities—albeit with some changes in focus and scope. Enterprises should maintain focus on connecting the people, ideas, business processes, and technologies needed to make smarter decisions faster. Although there

are many new tools available to address "social business" opportunities, these tools alone will not necessarily drive the anticipated business value.

A key element of all successful social media initiatives is a carefully designed business plan that clearly defines the strategic and economic value of such programs. Listening to a customer and understanding the customer's opinion about your products and services in a strategic vacuum is worthless. You must first address these questions:

- Will your definition of a customer change to encompass the many online personas each has?
- How will value be measured in the future and how will it differ from today?
- How will you make social media insights actionable within your organization?
- Will you need to transform your technology-enabled business processes to maximize the opportunities of conversational marketing?

Here are a few steps to integrate social media into business intelligence plans.

Step 1: Take stock

Understand where you are by completing a thorough self-assessment of your current state. Consider how you are leveraging social media data and to what extent you are integrating it with other customer data. Are you able to leverage that integrated customer data at the point of interaction with customers? Finally, consider your current IT infrastructure. Is it struggling to adapt to the volume, variety, and velocity of data generated by this new social world?

Step 2: Identify your goals

Define your target state by outlining your objectives in leveraging social media data. Be sure the social strategy supports strategic enterprise goals. To get started, determine if you want to:

- Integrate social media data with your other customer data and analyze it
- Influence existing customers or gain new customers and increase sales/revenue
- Change public opinion of a brand/product/company
- Conduct customer research at a lower cost
- Provide improved customer service
- Drive product/service innovation using the voice of the customer
- Improve employee productivity and knowledge sharing

Social and other unstructured customer data sources such as call center recordings, videos, chat sessions, and e-mails will be a significant challenge for IT.

Consider a few industry-specific objectives to clarify the point. A consumer goods company might want to leverage social media data to understand how a product is perceived by different market segments or to track the impact of a particular marketing campaign. A pharmaceutical company might consider using social media to analyze brand perception after a new drug launch or to support patients via blogs and health forums. A telecommunications provider can leverage social network analysis to provide information on mobile number portability, accurately revealing customer migration and churn behaviors to better understand new customer decision paths.

Step 3: Create a social business plan

Your plan should include specific details of what you want to accomplish. There are thousands of things you might do, so you have to be as specific as you can about your focus. Include a highly detailed definition of the target state, which should be validated with industry-specific benchmarks. In addition, the plan should include a detailed definition of the actions required and an estimate of the financial impact of implementing your program. A proof of concept is needed to bring together structured and unstructured data and determine what is (and is not) valuable.

Step 4: Design a social business and IT transformation plan

Include a detailed requirements analysis in your plan. Incorporate designs for business processes—whether new or changed—and the IT transformation required to enable these new processes. Your plan should also include estimates of transformation time and cost and be incremental in nature. Like any complex transformation, the most successful outcomes tend to come from a phased approach focused on specific business capabilities or use cases with carefully defined metrics and business-unit ownership. Finally, the plan must address the new skills and roles that will be required to drive and exploit these new capabilities.

Step 5: Implement your social business plan

As you implement your social business plan, keep these characteristic activities in mind:

- Combine structured data with unstructured social data to drive actionable customer insights that promote growth
- Engage rather than communicate with customers
- Recognize influencers and high-value customers in new segmentation models
- Encourage participation, sharing, and co-creation internally first, then externally
- Develop real-time analytic capabilities
- Learn from interactions and respond—social learning is applied throughout the business, not just in marketing
- Understand how being social can increase profitability

Implementing a socially enabled business includes three main phases: (1) capture and analyze social media data; (2) integrate it with existing data; and (3) filter it back into business processes. In other words: listen, analyze, and engage.

The first phase of a social intelligence program is "listening" for consumer dialogue on sites and communities. Today, sophisticated software tools can automate and scale listening to huge quantities of social media data. Enterprises can use these tools to "crawl" the Internet looking for the voice of customers.

The next phase is "analyzing" the gathered information to extract actionable meaning. A large amount of information on the Web includes opinion. Traditional information processing tools were not designed to interpret opinion; however, new analytical tools help overcome this gap by assessing user-generated content to identify the opinion or emotional state of a writer and the most recurrent conversational values. This is a complex undertaking because:

- Feelings and emotions are subjective
- Sentiment is rarely an all-or-nothing expression but rather comprises a range of feelings and tones
- Gauging sentiment is strongly associated with context

Although listening to customers yields good information, not all of this data has inherent value. Integrating social media data with other data sources creates deeper insights that drive better decision making. For example, services such as Klout and PeerIndex assign a score to a digital persona to identify his or her influence. This may be based on the number of followers, the total engaged network, or the likelihood that a recommendation will be acted upon.

The final phase is "engaging" to deliver insights to customer management processes and enable organizations to act quickly, decisively, and appropriately. Social intelligence, combined with other customer intelligence, allows enterprises to manage real-time (or near-real-time) conversations with customers and deliver offers, information, and customer services via social media channels. Increasingly, these communications can be delivered through mobile devices that provide content exactly when people need it.

Step 6: Govern a socially enabled business

Companies must design a governance model and align it with the overall organization. For example, implement new governance processes that take into account the need for real-time analytical insights at the point of customer interaction.

When establishing key measures for your program, be as specific as you can about what success looks like, keeping in mind that the metrics will vary depending on the type of business or government entity. Some measures of success include:

- Metrics about the customer and sales conversion rates before and after implementing a social intelligence solution
- The average elapsed time to respond to customer inquiries
- The number of customers engaged with the social intelligence solution, benchmarked over set periods of time
- The number of customer complaints mitigated

Most important, learn through the whole process. Test plan elements to see how they impact key measures. Constantly evolve and apply what you learn in the context of the overall plan for a self-reinforcing loop.

For enterprises that need help with their social intelligence solution, consider a provider with a complete portfolio of integrated solutions to manage, govern, and analyze social data. The provider should also be able to integrate social media data with existing data to provide real-time insights.

Social intelligence, combined with other customer intelligence, allows enterprises to manage real-time (or near-real-time) conversations with customers.

Conclusion

A social intelligence initiative is a new priority for success in today's instantaneous market. Within an organization, any employee interacting with customers needs access to all customer information at all times. Armed with this information, employees can change a conversation, modify a tactic, or extend a new product offer—whatever it takes to ensure customer satisfaction. This new reality requires both real-time information and real-time analytic platforms.

Point solutions, such as individual social media listening applications, are only one component of a social intelligence initiative. In order to harness the true potential of social intelligence, companies must deploy a broad information management and analytics program that is tightly linked with the overall business strategy. Significant first-mover advantages exist for companies that execute an effective social intelligence strategy before their competitors do. ■

Keys to Sustainable Self-Service Business Intelligence

Myron Weber is the founder and a managing partner at Northwood Advisors (www.northwoodadvisors.com). info@northwoodadvisors.com

Myron Weber

Abstract

Self-service reporting allows business executives, managers, operational decision makers, analysts, and knowledge workers to access the data they need whenever and wherever they need it to support the decisions and actions critical to business success. Business intelligence (BI) software vendors and industry experts recognize self-service reporting as a key feature of BI because it eliminates obstacles to timely insight and decision making and lowers the costs of reporting, analysis, and metrics-driven management by putting data directly in the hands of those who need it.

This article begins by addressing what's broken:

- **The 1 percent have access but the 99 percent need access**
- **Analysts don't get to analyze**
- **BI developers are unhappy**

Readers will learn the steps to providing self-service reporting the right way with unified design, targeted outputs, sound development practices, and active governance. We will explain how to implement correct procedures via strategy, road maps, governance, and best practices and examine the costs of poorly run self-service reporting systems. Finally, we offer advice for identifying a catalyst or compelling event and starting small with the right mix of motivated users, a committed team, and targeted objectives.

Introduction

Self-service business intelligence (BI) enables business executives, managers, operational decision makers, analysts, and knowledge workers to access the information they need whenever and wherever they need it, providing key data to support the decisions and actions that are critical to business success.

The meaning of the term *self-service BI* is fairly intuitive, but for clarity's sake, consider the following three functional characteristics of self-service BI:

- The majority of decision makers who need to make data-driven business decisions can access BI reports or queries that provide the information to answer the most common business questions. These reports and queries are dynamic, allowing filtering, sorting, and drill-down, among other features.
- Analysts who must dig deeper to answer uncommon or complex business questions are able to conduct the majority of that analysis through BI tools.
- For exceptionally complex, one-off analyses that must be performed in offline tools (spreadsheets, etc.), most of the required data can be extracted from BI sources; users do not have to go digging through source systems to find it.

Companies that succeed at self-service BI immediately begin to recognize a number of tangible benefits—some obvious and others less so. First, self-service BI lowers the direct labor cost and time-wasting cost of reporting, analysis, and metrics-driven management by eliminating the middleman and putting data directly in the hands of those who need it.

Self-service BI also eliminates bottlenecks to timely insight and decision making. This provides several potential benefits, including:

- Competitive advantages that can be gained by more agile decision making
- Reduced frustration of data consumers
- Better forecasting based on more accurate, up-to-date information
- More consistent returns with data-driven decisions

Enterprises investing in business intelligence often take for granted that a BI solution will automatically deliver self-service capabilities. This expectation is reinforced as BI software vendors and industry experts promote self service as a key feature of BI solutions. Most enterprise BI platforms as well as niche BI tools provide good to excellent self-service capabilities.

BI stakeholders want self service for compelling reasons. The unfortunate and frustrating reality is that many companies, including many that have invested tremendous resources in BI, don't succeed in delivering *sustainable* self-service BI. In some cases, they fail right from the start; in other cases, they begin delivering self-service BI but are unable to sustain it over time.

Why don't they succeed? This article will examine the symptoms of a lack of self-service BI, four key reasons companies fail at self-service BI, and corresponding alternatives to avoid falling into those traps. It will also look at strategies for correcting course if a company has already started down the wrong path. In the process, both real-world failures and successes in self-service BI will be examined. Whether a company is just starting on its journey to self-service BI or has a current BI investment that is not delivering self-service functionality, we focus on *sustainable* self-service BI.

For more information about self-service BI, see *Self-Service Business Intelligence: Empowering Users to Generate Insights* by Claudia Imhoff and Colin White, a 2011 TDWI Best Practices Report available at tdwi.org/bpreports. Imhoff and White give an excellent introduction to what self-service BI is and what tools and approaches companies should consider to deliver self-service BI.

Signs That Your Organization Needs Self-Service BI

There are several hallmarks or symptoms that are common to companies that need a self-service BI solution. These

pain points indicate that effective, sustainable self-service BI is needed.

Symptom #1: The 1 percent has access but the 99 percent needs access

Today's enterprises are facing a clear distinction between the workers who can access data and those who cannot. Within organizations that lack self-service BI, the figurative 99 percent are data users who are dependent on the 1 percent of users who have the technical expertise, domain knowledge, and security access to be data independent.

Furthermore, use doesn't equal satisfaction. Common complaints from business decision makers in enterprises without self-service BI include the lack of available data and the difficulty they experience getting the data that is available.

> In an effort to give users what they *want*, IT sometimes errs on the side of giving users *everything*.

Symptom #2: Analysts don't get to analyze

When asked what percentage of time is spent trying to get needed data, data analysts will typically offer a range of 60 to 90 percent, with most right in the middle at 75 percent.

If analysts at your enterprise typically spend only 25 percent of their time analyzing because they have to work so hard to gather data, the need for self-service BI is clear. Workers who aren't using their skills also result in a sometimes unrecognized cost: increased turnover among these valuable knowledge workers as they become frustrated and leave.

Symptom #3: Unhappy BI developers

BI technical team members are often data independent, which seems like a position of benefit and privilege. However, these workers face continuous pressure to deliver more information from increasingly complex and brittle data systems. The BI team—despite its best intentions—becomes a bottleneck in the system. For most BI professionals, this is the opposite of what they envisioned for their careers, leading to job dissatisfaction. In addition, rather than being part of a cross-functional community of BI stakeholders in the company, the BI team becomes increasingly isolated as they try to insulate themselves against the valid criticisms of frustrated business users.

Where Do Things Go Wrong?

Many companies make significant investments in BI over the course of many years, but rather than maturing into well-functioning BI solutions, they exhibit all the symptoms of failed self-service BI. There are four specific and common problems associated with these struggling efforts.

Problem #1: Taking the wrong approach

In an effort to give users what they *want*, IT sometimes errs on the side of giving users *everything*. In the words of one BI team manager whose company had failed at self-service BI, "We thought: let's just give everyone all the data. Then they can do whatever they want." That's a sweeping and impossible goal that is typical of the mindset in struggling organizations.

A guiding concept of BI is "one version of the truth," but giving everyone all the data so they can do whatever they want can result in different approaches, different results, and ultimately, different versions of the truth.

This big-box-store approach is evident in some cases where self-service BI works for a time but eventually becomes unsustainable. Typically, the early phases of BI attack "low-hanging fruit," an approach that often succeeds because it deals with targeted issues. This early success gives way to disappointment when the scope and complexity of the "all-the-data" approach overwhelms the design, the technology, and the users.

Problem #2: Partitioning design expertise

Another common approach of a struggling BI team is designing the data warehouse in isolation from the rest of the BI solution. The data warehouse is thrown to the BI metadata and report designers with a hearty "Good luck!"

If the analytical requirements and business rules are relatively simple, it is possible to design the data warehouse in relative isolation and produce meaningful outputs by adding BI software. However, in any environment where the reporting needs are moderately or highly complex, sustainable self-service BI can be achieved only by designing the overall solution holistically.

Problem #3: Shoddy practices

Poor BI development practices will sabotage efforts to deliver self-service BI. This is apparent with companies that start out delivering some self-service BI, then find they can't sustain or scale it.

A comprehensive set of best practices is beyond the scope of this article, but the shoddy practices fall into a few categories:

Data warehouse design

A data warehouse design that doesn't support specific query requirements will make self-service BI virtually impossible. One BI team developed a star schema data warehouse that allowed straightforward SQL queries of 95 percent of their data. The 5 percent exception was a very complex hierarchy that required queries that could not be generated automatically by BI software. Unfortunately, even though this was a small part of the overall data, this hierarchy was required in nearly all the reports and queries demanded by the business! To achieve self service, the team had to find a way to simplify that hierarchy so that most of their business questions could be answered with SQL generated by the BI software and only a small remainder required programmer support for report authoring.

Metadata development shortcuts

Although every vendor's BI software is different, one common feature is the need to follow vendor recommendations for BI metadata development. Every member of the BI team must avoid the common temptation to take shortcuts—violating the best practices in the metadata layer of the software with the assumption that it can be fixed in reports where the queries are defined. Violating vendor-specific metadata best practices leads to short-term self service that will prove unsustainable.

Software development life cycle (SDLC) practices

SDLC practices that are either too lax or too rigid will undermine self-service BI. Some BI teams, in an attempt to be "agile," allow a Wild West approach to development, with every developer empowered to make changes and deploy them to production. This seemingly nimble approach grinds to a slow crawl over time as inconsistency, duplication, and poor development practices are allowed to creep into the system through lack of control.

On the other hand, some organizations take a completely different approach, applying rigid SDLC or IT standards that don't embrace the concept of users developing their own queries and reports in a live production environment. When everything within the BI solution is under strict control and review, users cannot serve themselves.

Sound SDLC for BI avoids both extremes and adopts a nuanced approach that empowers users within defined boundaries and provides review mechanisms for developers.

Undefined BI team roles

Failing to define clear roles and accountability within the BI team undermines self service by enabling the other shoddy practices. When everyone is responsible for everything, it's hard to make anyone responsible for anything. At one company, the most knowledgeable metadata developer gave up trying to enforce sustainable best practices because the management model imposed by the director allowed any member of the BI team to make changes to the metadata model.

When practices are shoddy, the BI solution becomes increasingly brittle and complex over time. As one BI team member said about a BI solution, "We have so much duct tape and bubble gum holding things together, it's a wonder anything works at all."

Problem #4: Ineffective governance

Organizations struggling with problems in BI programs can often trace the issues back to governance. Executives and senior management should consider themselves part of the BI effort, but instead of saying "we" and including themselves, they sometimes resort to a blaming "they" in describing the team's shortcomings. The BI technical

team reads the situation easily and sees that they are, as one put it, "being hung out to dry."

In the absence of engaged, empowered, cross-functional governance of the BI program, the isolated BI team must determine when to say "yes" and when to say "no" to the myriad of requests they get from stakeholders. When BI is treated as a cost center without cross-functional governance, it becomes like an open bar that closes early—everyone can ask for whatever they want regardless of cost, but no one is ever convinced they received enough value. In the end, major investment decisions are often made by technical staff in relative isolation.

Poor BI development practices will sabotage efforts to deliver self-service BI.

Doing It Right

There are specific strategies organizations can employ to avoid falling into the traps that lead to failure in self-service BI.

Create targeted output

Self-service BI requires a targeted approach. Start by asking which decisions need to be supported in a self-service fashion, for whom, at what times, and in what ways. Design the BI solution to support self service for those targeted areas. Prioritize those subject areas into a road map that outlines short-term, targeted objectives in the context of a long-term, holistic vision.

Unify your design

Don't forget to include the BI software experts in the entire design and architecture process. As obvious as that might sound, it's a common mistake to bring the business domain experts and the data warehouse designers together while assuming that the BI software experts can come in later. This works in a limited way if the analytical requirements are simple, but when they are complex, the data warehouse design must include asking, "How, exactly, will the BI software need to query the data warehouse to meet these complex business requirements?"

Use sound practices

Sustainable self-service BI requires strict adherence to sound practices. Without this focus, any success in self-service BI will be isolated or short-lived. Adherence to sound practices has to start with defining and documenting those practices, then establishing a governance process to assure compliance.

Sound practices for self service must include at least these key elements:

Optimize the data warehouse for self service. Designing and building a data warehouse for sustainable self service requires focusing on the queries (SQL, MDX, etc.) required to answer key business questions with the goal of optimizing querying for most of the business questions that self-service users will be asking.

Follow vendor-specific BI practices. Every BI vendor's software has its own best practices. Mastery of those practices by the BI team is essential to sustainable self service. The design must be driven by an understanding of exactly how queries will be formed and generated by the BI query and metadata layers for delivering the self-service outputs.

Set development standards for approval and deployment. As technical standards are established to support self-service BI, the way to make them sustainable in the real world is to establish a process for the technical experts on the BI team to review and approve content before it is deployed into the self-service production environment. For example, the domain experts on metadata development, report authoring, and business data definitions can maintain standards and lend their expertise to others without becoming bottlenecks who must do all the work within their specialty areas.

Effective Governance

A proven approach advocated by Northwood Advisors (the BI and decision systems advisory firm I founded and where I am a managing partner) is aligned governance, which ensures that there is effective governance across the

full spectrum of BI activities. A simplified view of this approach identifies four levels of governance:

- **C-level enterprise leadership.** Obviously, the C-level leadership is not strictly part of the BI organization. Rather, they are the executives who ultimately run the business and set the objectives and strategy for the organization. Aligned governance begins by understanding the organizational objectives and strategy set by C-level leadership.

- **BI executive governance.** Often labeled the *BI steering committee*, this group must include C-level leaders who are ultimately responsible for the BI program. This group also includes cross-functional executives and managers who have a stake in the outputs of BI, along with managers from the BI team who serve as the conduit to the operational governance board and BI team.

- **BI operational governance.** This group governs the operational, day-to-day functioning of BI. It includes cross-functional, operational, and technical representation; specifically, BI team representation, as well as representatives of business stakeholders and BI champions.

- **BI champions and power users.** The BI champions and power users group drives adoption of the BI solution by spreading and supporting the rollout of BI to business departments, as well as by identifying requirements, opportunities, and problems with BI that must be escalated to the operational governance board. The group should include a mix of dedicated BI team members and users who are embedded within other business departments.

BI governance that delivers sustainable self service must be aligned at all levels of the governance organization: mandated by C-level leadership, owned by the BI executive governance group, executed by the BI operational governance team, and embraced by BI champions and power users.

Making the Change

Companies that are young in their BI maturity and have the opportunity to approach self-service BI properly can avoid much heartache. Companies that have already made a significant investment in BI but aren't getting the desired results, as well as those exhibiting the hallmarks of a need for self-service BI as described here, can correct their course and get on track to sustainable self-service BI by employing the following core strategies.

A proven approach is aligned governance, which ensures that there is effective governance across the full spectrum of BI activities.

Strategy #1: Understand the cost

Stakeholders typically recognize two types of costs right away when contemplating the necessary course correction. The first is the sunk cost—the money that has already been spent. The second is the cost of making the change—what it will take to actually correct what has gone wrong and move forward on the right path.

There are additional costs that aren't so easily recognized, such as the long-term cost of ownership of an unsustainable BI solution. The mounting complexity and brittleness of the "duct tape and bubble gum" solution become staggeringly expensive to maintain over time.

The final cost category is the opportunity cost of not having the benefits of self-service BI. This includes the hidden costs of frustration, lost productivity, and turnover that result from the gap between the 1 percent and the 99 percent. It includes missing competitive advantages, efficiencies, and operational improvement that self-service BI offers.

With a clear view of the costs and benefits, it's easier to set a new path forward.

Strategy #2: Find a catalyst

Sometimes the best way to force a change and accelerate progress in the right direction is to find a compelling event or catalyst. For example:

- A major version upgrade or migration of BI software or database platform
- Acquiring, replacing, or retiring key business source systems that feed BI
- An organizational change or realignment that shifts power centers
- Business changes that alter the focus or intensity of demand for BI outputs

Strategy #3: Choose a compelling niche

As a company takes its first steps to correct its BI course, it's important to do so with the right niche—meaning a first self-service deliverable that will shift momentum from the entrenched solution toward sustainable self service.

The first defining characteristic of the right niche is a group of motivated, positive, and clear-thinking stakeholders. If the BI team is trying to drag reluctant, cynical users along on the first attempt at making a positive change, the effort may well be doomed. That makes it crucial to find a subject area and stakeholder group that "gets it" and is on board with the plan to do something new.

The second key to finding the right niche is to assign dedicated technical resources who understand and support the new approach. Although much of the BI team's energy will continue to be drained supporting the old "duct tape and bubble gum" solution, changing directions will require setting aside BI developers who can blaze the trail, establish new practices, and teach others.

The final part of building a compelling niche is defining a high-ROI subject area. This means a set of targeted outputs for which the relative risk from complexity, data quality, and unanticipated requirements is low, and the relative reward in terms of business value is high. In no case, however, should the lure of a high-ROI subject area blind the team to the importance of motivated, positive, and clear-thinking stakeholders. If forced to choose, the BI team should take the lower-value subject area with more positive stakeholders.

Conclusion

Self-service BI can be achieved and sustained with the right approach, and the benefits are well worth the effort. ■

Behavior-Based Budget Management Using Predictive Analytics

Troy Hiltbrand is an enterprise architect and IT strategy manager for Idaho National Laboratory.
troy.hiltbrand@inl.gov

Troy Hiltbrand

In times of financial austerity, budget management becomes critical for any organization. When trying to optimize operations and do as much as possible with a limited set of funds, the ability to forecast spending becomes an imperative for business and government organizations alike.

Historically, forecasters have primarily based their predictions upon two common factors: time and money. Although these are important aspects for determining future spending patterns, organizations represent a complex system of unique individuals with a myriad of associated behaviors, all of which affect how budgets are utilized.

Forecasted budgets often reflect a guessing game about how budget managers will behave under a given set of conditions. This becomes messy when human nature is introduced; different managers will react differently under similar circumstances. One manager may become ultra-conservative during periods of financial austerity, while another might be unfazed and continue the same spending habits. Both managers might revert into a state of budgetary protectionism, masking their activities in order to keep as much budget and influence as possible regardless of the greater good of the organization.

To more accurately predict future outcomes, models should consider time, money, and observed behavior patterns. Predictive analytics is poised to provide the tools and methodologies organizations need to capture and leverage behaviors of the past to predict the future.

At Idaho National Laboratory (INL), budget management is at the forefront of every management discussion,

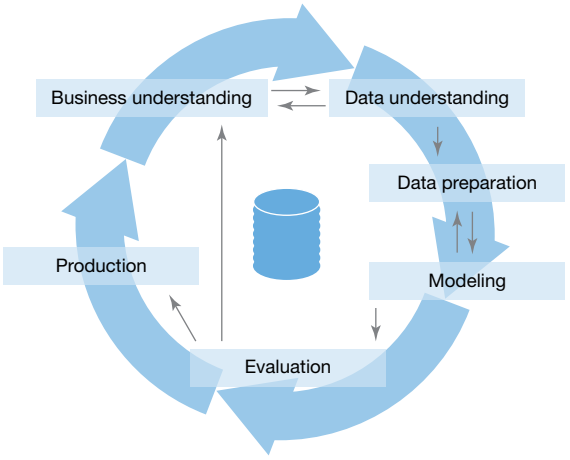

Figure 1: CRISP-DM methodology.

and this will continue for the foreseeable future. We needed a more accurate model and methodology for predicting future budgetary outcomes and supporting future business operations. Predictive analytics helped us build a behavior-based budget management model and apply it to support delivery of the right information for decision makers to optimally manage laboratory funds.

Our Methodology

Predictive analytics is the practice of using patterns in historical data to anticipate future outcomes. Because predictive analytics is based on a set of unknown attributes, it is as much art as science and requires much more effort than simply running a data set through a tool to get an answer. Predictive analytics requires massaging the data and looking at the information inputs and outputs to find the optimal solution from a body of potential solutions. Even though it is an art in the end, there are industry-accepted methods for guiding teams through the process so it progresses smoothly and the results meet anticipated business objectives.

Of these methodologies, the most widely accepted is the Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM was started in 1996 by a consortium of industry experts from DaimlerChrysler, SSPS, and NCR (Chapman, et al, 1999). This group set out to develop a method that would encapsulate the steps necessary to perform a data mining or predictive analytics project in an orderly and organized fashion.

CRISP-DM is the iteration of six fundamental activities (see Figure 1). Each process is supported by procedures and processes that further clarify needs as a solution set is built.

In this article, we will explain how INL used the CRISP-DM methodology to develop a predictive analytic model that empowered the laboratory to make critical budgetary decisions.

Activity #1: Business Understanding

The first step in the process is to understand the business challenge you are trying to solve and identify the relevant stakeholders. With budget management at INL, two primary groups have a significant stake in understanding how budgets are managed at the end of the year.

The first stakeholder is at the enterprise level. Idaho National Laboratory is owned by the Department of Energy (DOE) but is managed by Battelle Energy Alliance (BEA). Based on BEA's prime contract for managing operations at INL, which is used by DOE to provide oversight, budgets cannot be carried over from year to year. Unlike some enterprises in private industries, any money not expended to further the mission of the laboratory by the end of the fiscal year is forfeited and cannot be used to accomplish future work.

In addition, BEA is not permitted to run operations in a deficit. This puts BEA in a precarious situation if budgets don't come in precisely on target at the end of the fiscal year, and adds pressure at the enterprise level to closely monitor budgets throughout the year and make mid-year course corrections to complete the year as closely as possible to the target. The more clearly the trends are understood, the earlier decision makers can make course corrections, leading to a greater breadth of available options and a more effective use of funds in accomplishing the laboratory's mission. Without visibility until near the end of the fiscal year, investment options are limited because of resource constraints.

The other group with a vested stake in understanding where the budgets are predicted to stand at the end of the budget year consists of work managers. These managers are allocated a defined budget at the beginning of the fiscal year and are expected to complete a specific amount of work within this budget.

Each month, the cumulative budgets are matched against the cumulative actuals. If there is a significant variance between budgets and actuals, the work manager is required to justify what is being done to accomplish the work within the budget. If the work packages are over budget, mitigation is used to reduce the overages so that they do not inflate the laboratory's spending. If the actual amount spent is significantly under budget and the justification for future plans is not adequate, the money may be diverted to other projects in the queue. If work managers do not have the right level of visibility into the budget forecasts, they may be out of balance near the end of the fiscal year and have too little or too much money to accomplish their assigned tasks.

As part of the first step of CRISP-DM to gain an understanding of the business, your team must identify what is being measured and understand the project's criteria for success. In the case of INL, we identified the measure as the forecasted actuals by month for the remaining months of the fiscal year at a level that could be viewed by work package and expenditure type. Managers also needed to be able to roll up these figures to an overall view of the budget (planned and actual) at the enterprise level. This target measurement became the dependent variable in our budget forecast model.

The deliverable of any predictive analytics project is a model composed of a formula (or set of formulas) that can derive future values of the dependent variable. Other informational inputs act as independent variables that are known at the time of computation to generate a predicted dependent variable.

Activity #2: Data Understanding

Once the business objective is understood and the dependent variable is defined, your team must understand the data inputs into the model and how that data is currently stored and structured. This requires understanding the potential behaviors that play into the prediction of the dependent variable.

As your team starts to identify behaviors associated with decision-making activities attributed to budget management, it can easily define hundreds or thousands of different attributes. It is unlikely that all of these attributes will be used in the final model. However, it is important to identify a sufficient breadth of measurable variables so those with the highest correlation with the dependent variable can be readily identified. Sometimes obscure data elements have a high impact on the outcome, and unless they are identified as part of the data understanding step, they will be missed, weakening the potential of the end model.

At INL, we looked at many behaviors associated with determining future budget estimates. When dealing with people and their individual thought processes, we found that what is cogent and logical to one group in managing their budget may not be so to another group. Even within a specific group, there are often many different behaviors manifested among managers. It is even possible to see divergent behaviors from the same work manager when dealing with various specific budgets.

First, we looked at the variance reporting and justification process. Work managers do what they can to avoid reporting variances; they behave in a way that keeps their budgets within the tolerance of variance so that they don't trigger the justification process. To capture this, attributes such as whether the work manager was required to fill out a variance report in the previous month or how many months in succession the work manager has been in a state of unacceptable variance were identified and captured.

Second, we looked at the breadth of work packages the work manager was responsible for. At INL, the work package is the mechanism associated with budget allocation. This work package is the lowest component of a multi-level work breakdown structure. Every work package has an associated work breakdown structure that places it in the context of the laboratory's mission.

Work managers often own or have influence over a higher level of the work breakdown structure. If they are out of variance in one work package, they may move money from another work package under their purview to cover that variance, or they may shift charges between work packages to balance spending, masking the true nature of spending within the work package. With some work packages, this shifting of budgeted funds or actual spending is perfectly acceptable and is a viable mechanism for managing budget variances. In other work packages, it is strictly forbidden due to contractual limitations governing fund usage.

The goal was not to correct behavior but to understand its true nature so that it could be accurately modeled and forecast.

In cases where moving money between work packages is allowed, managers often view their budgets at a higher level than the work package level and forecasts must account for the behavior associated with this higher level of budget management.

Third, behavior over time is a factor in the decisions made by work managers. As the fiscal year comes to a close, work managers often become more vigilant about current expenditures. This creates peaks and valleys in spending patterns. If the variance has been negative, work managers will scramble in the last months of the fiscal year to bring budgets back into alignment. In the opposite case, if they are running a positive variance, they will often crash schedules or approve procurements that will benefit them in the future so they can bring their budget back to an acceptable level.

Such influential behaviors are not limited to the work manager. As work is done, there are behaviors related to how budgets are used at all levels of the organization. Labor is a huge portion of the budget, and each time an individual is approved to work overtime or take advantage of personal leave, their behavior alters the usage of these budgets. They do not think of their actions in terms of budget spending patterns—they are merely focused on accomplishing work—but this type of behavior can influence budget swings. Although personal leave does not change the enterprise view of the budget, it impacts the forecasts at the manager level because funding for personal leave comes out of a different work package from standard labor hours.

Many of these potential behaviors were identified by interviewing work managers across the organization in an honest and non-threatening manner. It was important that these interviews remain non-threatening or the work managers would be inclined to provide answers based on expected behaviors and mask their true behaviors to avoid being reprimanded. The goal of these exchanges was not to correct behavior but to understand its true nature so that it could be accurately modeled and forecast.

Once these behaviors were understood, the team looked at the data to identify what pieces could be consolidated to represent the behaviors as information inputs to the model. These inputs do not have to be limited to numeric data. Textual and categorical data can be used in predicting future values of the dependent variable, but it has to be synthesized down to discrete values that can be factored into the prediction calculation.

Key elements associated with data understanding include the timing and durability of the data used to train the model. As the predictive model is built, its optimization will be based on the quality of the input data. If either the timing or durability of the data gets out of alignment, the model will ultimately be ineffective.

If the dependent variable was calculated based on a particular attribute that was not available until *after* the forecast was made, the result will be heavily skewed and will be useless in a production environment. For example, when using variance, the calculation can use a *previous* month's budget variance to predict the outcome of future budget, but cannot use the *current* month's variance because it would not be known. This causes a problem

because the prediction of the forecasted actual would be calculated prior to the beginning of the month and the variance would not be available until after the month was complete. Depending on how far out the forecast will be calculated, certain attributes will be outside the scope of usability from a timing perspective. If the previous month's variance is an input to the model, it will be available for the next month, but not for successive months because the values for their previous-month's variance will still be in the future.

From a durability perspective, the attributes used in the formula must be consistent with the value at the time that the forecast would have happened. For example, if the demographics of the work manager are an important factor in the model, the training set will need to have the work manager who was responsible for that period's budget and not the current work manager. If the current work manager is superimposed on past budget decisions, the results will group past decisions with the current work manager. If this is a tangible predictor, the model will be trained using incorrect assumptions and behaviors of the past work manager's decision-making process will be attributed to the current work manager, leading to inaccuracy.

This second step in the process—understanding the data—becomes a foundation for moving forward. If the team does not understand what data is important and available, their ability to create an effective forecasting model will be limited.

Activity #3: Data Preparation

Once the pool of potential attributes is known, your team must determine what data is available and what data will need to be transformed or cleansed to be consumable. Some of the attributes will be relatively easy to determine and will come directly from operational systems; others will be significantly harder to access and will have to be derived or—even more challenging—your team will have to determine when processes must change to enable data collection. Each informational attribute will have a relative cost associated with its use in the model, which will factor into your team's final choices among candidates.

Clustering

When the domain of data is too vast, it is often difficult for a model to consume it directly. If it can be clustered together in groups based on similarity, the model is much more likely to be able to consume the derived group attribute. In budgeting, the exact size of the budget might be useless as a raw number of dollars, but if it can be classified into groups such as small, medium, large, and enormous, then other behaviors associated with how that budget is expended start to take on greater meaning.

Key elements associated with data understanding include the timing and durability of the data used to train the model.

It is also possible that both the exact size and the size category can be used in different ways in the model. With categorical data, the category will not be a direct attribute in the formula, but it will allow for the creation of multiple formulas categorizing the data into multiple buckets, each with a different formula to produce a predictive forecast of the dependent variable.

Filling in Data Gaps

If attribute values are missing on a subset of records, your results may be skewed. This is especially common with null values in the data set. Nulls can easily be replaced with a default value, but the existence of the null could have predictive potential in and of itself. One way to capture this is to replace the null value with a default and create an independent attribute that acts as a flag indicating that the value was originally null. Both of these values can be tested to identify their correlation with the dependent attribute.

Other attributes can be derived from a combination of attributes. For example, a postal code can be derived from the address. Due to the categorical nature of postal codes, they might be more highly correlated to the dependent

variable than addresses alone, so populating missing postal codes can increase the effectiveness of the model.

Eliminating Outliers

Predictive analytics is not as fundamentally based on the central tendency theory as is classic statistics, but the concepts of mean and standard deviation can be used to find significant outliers within the data. Removing these outliers may improve the model's predictive capability, but you must be careful to ensure that such removal does not mask behaviors that need to be captured.

For example, if your organization had a large, one-time, unbudgeted fine for unacceptable practices and there is no indication that this will happen again, you may remove the outlier to reduce how it skews the model and thus maintain the model's integrity. On the other hand, if a manager is spending significantly more or less than their budget, these expenditures might be outside of the normal distribution of spending yet are important for predicting the spending patterns of this manager or the organization as a whole.

Event-Response Data over Time

Unlike a record in a transactional system (where each event lives in an individual record), it is often important to flatten historical data into one record that can be used in generating the model. With the budgetary policies in place at INL, each month's budget variance is calculated and managers are required to justify their spending if their variance exceeds a threshold. Having attributes such as previous month variance, number of months out of variance, or average variance for the current fiscal year would be important in predicting future spending patterns as managers work to bring their budgets back in line.

The Last Three Steps

The first three steps of the process can take as much as 80 percent of the overall predictive analytics project life cycle and often have to be revisited once the project moves into later stages. This is because the results of these steps will effectively determine the model's success in achieving business objectives.

With the advances in analytics software over the last 20 years, much of the work of building out the model is a black box using a plethora of advanced mathematical and statistical algorithms. The process of actually building the model has three major parts that comprise the final three steps of the CRISP-DM method: training the model, evaluating the model, and deploying the model.

Activity #4: Training the Model (Modeling)

With many model development algorithms available, developing a behavior-based model does not require significant coding but does require oversight for training the model. Training involves taking a historical set of data with both independent and dependent variables and running it through a commercial or open source solution, which in turn builds and optimizes the resulting model, which will be the final product to be launched into production to predict future values.

In the data understanding and data preparation phases, the goal was to identify a pool of potential independent variables that could be used to represent behaviors that have an influence on how the organization uses and manages a budget in the performance of work. During model development, it is important to determine which of these independent variables are most highly correlated with the dependent variable and which act merely as noise. Noise from unnecessary data elements dampens the ability for the model to predict the dependent variable.

This is where the art aspect of predictive analytics is the most critical. The model needs enough information to be effective at predicting outcomes, but no more. There might be hundreds or thousands of potential independent attributes representing different behaviors associated with budgetary decision making, but ultimately the best performance is gained by a small subset of those attributes.

With analytic modeling algorithms, there is no silver bullet that works in all situations. Based on the nature of the data elements and the nature of the problem at hand, logistical regression, linear regression, decision trees, neural networks, support vector machine (SVM), and many other algorithms can be used. Exercise them

against the training data to identify which will perform optimally within the context of your business problem.

When training the model, your first step is to striate the existing set of data into two or more subsets. The first subset is the training set of data, which must represent the entire population of data because it will train the model to perform in all instances. (Note that some representative data must also be reserved for evaluating the model. More on this below.)

For our INL project, we looked at the budget forecasting, examined those attributes that were most highly correlated to the dependent variable, and made sure that our training set fully represented the data in that domain. For example, we found that the department to which the budget was allocated was highly correlated to the manner in which the budget was managed. High-level managers had different beliefs with respect to budget management, and these feelings permeated those managers' departments. To account for this, we made sure that a representative sample of records from each department was present in the training data and did not merely collect a random sample.

Sizing the training data is also a balancing act and depends significantly on the amount of historical data available. Too little data will not allow the models to reach maturity and be able to respond adequately to a full set of given production inputs. Too much data will overfit the model. Overfitting occurs when the model optimizes itself so closely to the training data that it performs very well with the training set but fails when used on another set of data.

At INL, after much testing and many trials, we determined that the optimal solution was a decision tree algorithm with the classification of multiple subsets of data using attributes such as work package size, organization, expenditure type, and type of work (internal versus external) to define budgetary buckets of behavioral similarities. Within each of these subsets of data, we were able to use linear regression to determine a predictive forecast for future time periods.

Activity #5: Evaluation

Once a model has been developed with its formula or set of formulas, it must be evaluated to validate its performance.

During the identification of the training set of data, it is important to set aside a portion of data to be used in the evaluation phase. The evaluation data set(s) and the training data must be separate and distinct, because the model was developed and optimized using the training data. The true test of its effectiveness comes when other data, which the model has never encountered, is used to produce a forecast. Ensure your test data set embodies a good representation of the independent attributes so that it can fully test a majority of scenarios that will be found in production.

Unlike production data, test data has known independent and dependent variables. It is similar in composition to training data, but during evaluation the known dependent variables are withheld from the model. The model is presented with the independent variables and generates a predicted dependent variable, which is compared to the actual dependent variable to determine how well the model performed on the test records. Many solutions on the market perform the evaluation phase and produce reports showing how well each model performed against the known set of dependent variables. This is often measured in terms of lift or gain against a base level.

If the initial steps of business understanding, data understanding, and data preparation are done well, the process of training and evaluation can run through multiple models using multiple sets of independent attributes to quickly identify an optimal prediction model. There will rarely be a perfect formula that will match 100 percent of the time, but multiple very good formulas will emerge. This is the point where other artistic choices have to be incorporated.

For example, two formulas might be close in performance: a decision tree and a neural network. Neural networks often perform well but are truly black boxes that are hard to explain to management; decision trees are intuitive and can be graphically represented and

explained. The decision tree might be the optimal choice in this case because it is easier for senior management to adopt, especially in the nascent stages of developing an organization's analytic capability when the organization is still uncertain of its full potential.

Even within a single algorithm type, two models with different sets of independent variables may exist that perform at similar levels, but one might use an attribute that is computationally expensive to achieve during the data preparation phase. The overall cost of running the model would indicate a preference for the model with more highly accessible data points.

Activity #6: Deployment (Production)

Once the model is derived and vetted, it can be moved into production. At this point, the dependent variable will be completely unknown and the model will rely solely on independent variables to predict outcomes. The model will take those independent variables and calculate a forecasted dependent variable.

Since the model was optimized using a set of behaviors that existed in the organization in the past, changes to the organization in the future will reduce the model's effectiveness. The competitive landscape, government policies, management structure or composition, economic factors, and public sentiment will evolve over time. The behaviors of work managers will evolve as well.

To monitor this, capture the predicted values of the dependent variable and match them against the actual dependent variable once the prediction time period is complete. When the formula deteriorates to the point where it is no longer predicting with the necessary accuracy, start the process again and develop a new, more accurate model reflecting the changes in organizational behavior. This is represented by the outlining circle of arrows in the CRISP-DM diagram (Figure 1).

Our organization's ability to accurately predict how future actuals will fall out each fiscal year based on budgets is paramount. Idaho National Laboratory's contractual requirements about how budgets are managed meant that we had to move beyond simply using time and money as the only factors in our prediction model. We needed to decompose and measure past behaviors and utilize them as components in our model to more accurately estimate future outcomes. Through the use of CRISP-DM and predictive analytics, we were able to derive a model that will meet our needs today and have flexibility into the future. ■

References

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth [1999]. *CRISP-DM 1.0*, SPSS. Retrieved July 23, 2012, from ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

Han, Jiawei, Micheline Kamber, and Jian Pei [2012]. *Data Mining: Concepts and Techniques*, Morgan Kaufmann (Elsevier).

Miner, Gary, John Elder IV, Thomas Hill, Robert Nisbet, Dursun Delen, and Andrew Fast [2012]. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press (Elsevier).

Nisbet, Robert, John Elder IV, and Gary Miner [2009]. *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press (Elsevier).

Rathburn, Tony [2012]. "Predictive Analytics and Data Mining: Strategic Implementation," three-day workshop, The Modeling Agency. http://the-modeling-agency.com/strategic-implementation/

BI Case Study

Analytic Platform Provides Fast Performance on Big Data

Linda L. Briggs

Customer relationship marketing agencies work with massive amounts of data as they help their customers create personalized marketing campaigns from many customer lists. After expanding its use of an analytic platform, one industry-leading agency can now process more data faster, allowing it to better compete.

As one of the country's largest and fastest-growing customer relationship marketing agencies, Merkle helps *Fortune* 1000 companies segment their customer data to create personalized marketing campaigns.

Predictably, the customer databases Merkle handles often involve huge numbers of records and a wide variety of source data. Furthermore, the company keeps each of the many data sets submitted or requested by its customers completely separate as it analyzes the data, resulting in multiple versions of each customer record. The end result: billions upon billions of customer records. For its largest customer, for example, Merkle regularly processes 2.5 billion records in a reference database that is constantly being added to, matched, and increased.

The company had an older, Linux-based, flat-file architecture, explains VP of technology Peter Rogers, but wanted to move to a faster, more efficient, and more dynamic data processing environment. Earlier this year, the company—already a ParAccel user—expanded its use of the ParAccel Analytic Platform, which was at the core of its customer data integration application.

Rogers works with a team of 160 in the technology services group. The team's focus is working with large marketing databases containing reams of demographic and other information about client customers to allow the companies they work with to better segment and target their customers.

Merkle, an early customer of ParAccel, uses the product to interface with Merkle's customer data integration (CDI) platform. CDI platforms, which allow entity recognition—recognizing that different e-mail addresses in different databases belong to the same person, for example—are a core technology in Merkle's industry and an essential tool for remaining competitive.

With the company growing at close to 20 percent a year, the ability to scale is important. ParAccel's massively parallel architecture allows Merkle to predict growth, scale in a linear fashion, and add new clients with node-by-

node expansion. Merkle can thus incrementally add nodes to its multiple clusters as needed—unlike the approach of an appliance vendor, where increased performance calls for an additional piece of hardware. "ParAccel gives us the ability to add nodes on top of our existing clusters, so that's a nice scaling feature," Rogers says.

Along with improving performance and ensuring scalability as the company grows, Merkle also wanted to find a cost-effective tool that returned great performance for the price. "At the end of the day," Rogers says, "ParAccel's ability to scale, and the price-to-performance [ratio] that we got, made it the best decision for us." Today, Merkle is spending about $5,000 per terabyte in its ParAccel cluster, Rogers says, estimating that a comparable database appliance would cost between $20,000 and $40,000 per terabyte.

The expanded application now stores 200 TB of raw data, compressed down to 50 TB on ParAccel.

A Billion Records a Day

The new system has made a tremendous difference in processing speed. "We can now process a billion records a day on this platform, something we weren't able to do before," in part because the expanded platform allows processes to be run that much faster, Rogers says.

Another challenge for Merkle is dealing with the many data sources submitted by the companies it works with, which buy data from Merkle and other third-party sources as well as submit their own customer data. "I couldn't even tell you the number of different sources that we process," Rogers says. Merkle aggregates and performs calculations on customer data from all the various sources while keeping each database essentially separate.

A ParAccel plus that Rogers specifically mentions is the solid relationship his company has with the vendor; he cites a circumstance from several years ago as an example. Merkle is largely a SQL Server shop. As the company started running into size limits on SQL Server, it decided to incorporate other, more scalable technologies. ParAccel worked with Merkle, adding a SQL Server interface to the ParAccel engine. "That was something we requested, and they went and built it right into their product for us. I'm sure other customers are taking advantage of it," Rogers says, "but I know that was something Merkle requested, and they put it into their road map and built it for us."

That accommodation was important to Merkle because staff members in data services had specific SQL Server expertise: "We have all these people that know SQL Server here," Rogers says. "We have code that we could move from those SQL Server boxes right onto this platform; [ParAccel was] able to build that for us."

The columnar technology inherent in ParAccel was also a plus. In dealing with so much data, a column-oriented database management system is helpful because it stores data tables as sections of columns of data rather than as rows of data. This has advantages for ad hoc query systems where aggregates are computed over large numbers of similar data items. In Merkle's case, it allows the company to retrieve just those data elements needed for a particular analytics case. By not reading every row, the system can retrieve a subset of the data, avoiding parsing through data unnecessarily. That in itself can provide a huge performance boost.

More Data Sets on Less Hardware

With the expanded platform, Merkle can also run more customer data sets on less hardware, allowing the company to process much more data than before. "We can process more data, and we can process more customers on less hardware," Rogers says. With the previous environment, a server was often required for each new customer to keep data separate. Although the processing load could sometimes be spread over multiple servers, Rogers says, the process wasn't as straightforward as it is with ParAccel. Now, especially for smaller customers, the company can use a multi-tenant environment, thus saving on servers.

In fact, Rogers says Merkle overall has seen a 25 percent decrease in servers—and that in the face of handling 300 percent greater data volumes.

With its increased firepower, Merkle can now offer its clients better

matching options, fine-tuning its targeted marketing and improving conversion rates. The agency is able to generate a targeted list of consumers based on more than 100 specific attributes for each client's marketing objectives and audience. Merkle has added new regions, increased the number of analysts accessing its system, and moved to daily updates rather than the previous weekly updates. "We can now offer more functionality in the product," Rogers says. "Our matching logic is more robust because we have more data." The additional computing power in the new platform also allows Merkle to use data in more granular form, including more complex algorithms.

A final benefit Rogers cites: ParAccel gives Merkle access to every node on the cluster, yielding tremendous backup speed. When Merkle started dealing with big data, Rogers says, it was a struggle to back up some of the appliances—all data had to go through the leader node, which could bring the system to its knees. "ParAccel gives us the ability on the back end to be able to access all the nodes and to do the backups in a much more efficient fashion," he explains.

As analytics becomes increasingly critical to the ability of companies to target their customers appropriately and pull value from vast amounts of data, Merkle's use of its analytic platform provides a clear example of how the right software can offer significant competitive benefits. ■

Linda Briggs writes about technology in corporate, education, and government markets. She is based in San Diego. lbriggs@lindabriggs.com

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.

Visit tdwi.org/journalsubmissions for the *Business Intelligence Journal's* complete submissions guidelines, including writing requirements and editorial topics.

Submissions

tdwi.org/journalsubmissions

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

Upcoming Deadlines

Volume 18, Number 3
Submissions Deadline: May 17, 2013
Distribution Date: September 2013

Volume 18, Number 4
Submissions Deadline: August 9, 2013
Distribution Date: December 2013

The Database Emperor Has No Clothes

Hadoop's Inherent Advantages over RDBMS in the "Big Data" Era

David Teplow has been a consultant for Integra Technology Consulting since 2000.
dteplow@integratc.com

David Teplow

Background

Relational database management systems (RDBMS) were specified by IBM's E.F. Codd in 1970, and first commercialized by Oracle Corporation (then Relational Software, Inc.) in 1979. Since that time, practically every database has been built using an RDBMS—either proprietary (Oracle, SQL Server, DB2, and so on) or open source (MySQL, PostgreSQL). This was entirely appropriate for transactional systems that dealt with structured data and benefitted when that data was normalized.

In the late 1980s, we began building decision support systems (DSS)—also referred to as business intelligence (BI), data warehousing (DW), and analytics systems. We used RDBMS for these, too, because it was the de facto standard and essentially the only choice. To meet the performance requirements of DSS, we denormalized the data to eliminate the need for most table joins, which are costly from a resource and time perspective. We accepted this adaptation (some would say "misuse") of the relational model because there were no other options—until recently.

Relational databases are even less suitable for handling so-called "big data." Transactional systems were designed for just that—transactions; data about a point in time when a purchase occurred or an event happened. Big data is largely a result of the electronic records we now have about the activity that precedes and follows a purchase or event. This data includes the path taken to a purchase—either physical (surveillance video, location service, or GPS device) or virtual (server log files or clickstream data). It also includes data on where customers may have veered away from a purchase (product review article or

comment, shopping cart removal or abandonment, jumping to a competitor's site), and it certainly includes data about what customers say or do as a result of purchases or events via tweets, likes, blogs, reviews, customer service calls, and product returns. All this data dwarfs transactional data in terms of volume, and it usually does not lend itself to the structure of tables and fields.

The Problems with RDBMS

To meet the response-time demands of DSS, we pre-joined and pre-aggregated data into star schemas or snowflake schemas (dimensional models) instead of storing data in third normal form (relational models). This implied that we already knew what questions we would need to answer, so we could create the appropriate dimensions by which to measure facts. In the real world, however, the most useful data warehouses and data marts are built iteratively. Over time, we realize that additional data elements or whole new dimensions are needed or that the wrong definition or formula was used to derive a calculated field value. These iterations entail changes to the target schema along with careful and often significant changes to the extract-transform-load (ETL) process.

The benefit of denormalizing data in a data warehouse is that it largely avoids the need for joining tables, which are usually quite large and require an inordinate amount of machine resources and time to join. The risk associated with denormalization is that it makes the data susceptible to update anomalies if field values change.

For example, suppose the price of a certain item changes on a certain date. In our transactional system, we would simply update the Price field in the Item table or "age out" the prior price by updating the effective date and adding a new row to the table with the new price and effective dates. In our data warehouse, however, the price would most likely be contained within our fact table and replicated for each occurrence of the item.

Anomalies can be introduced by an update statement that misses some occurrences of the old price or catches some it shouldn't have. Anomalies might also result from an incremental data load that runs over the weekend and selects the new price for every item purchased in the preceding week when, in fact, the price change was effective on Wednesday (which may have been the first of the month) and should not have been applied to earlier purchases.

The benefit of denormalizing data in a data warehouse is that it largely avoids the need for joining tables.

With any RDBMS, the schema must be defined and created in advance, which means that before we can load our data into the data warehouse or data mart, it must be transformed—the dreaded "T" in ETL. Transformation processes tend to be complex, as they involve some combination of deduplicating, denormalizing, translating, homogenizing, and aggregating data, as well as maintaining metadata (that is, "data about the data" such as definitions, sources, lineage, derivations, and so on). Typically, they also entail the creation of an additional, intermediary database—commonly referred to as a staging area or an operational data store (ODS). This additional database comes with the extra costs of another license and database administrator (DBA). This is also true for any data marts that are built, which is often done for each functional area or department of a company.

Each step in the ETL process involves not only effort, expense, and risk, but also requires time to execute (not to mention the time required to design, code, test, maintain, and document the process). Decision support systems are increasingly being called on to support real-time operations such as call centers, military intelligence, recommendation engines, and personalization of advertisements or offers. When update cycles must execute more frequently and complete more rapidly, a complex, multi-step ETL process simply will not keep up when high volumes of data arriving at high velocity must be captured and consumed.

Big data is commonly characterized as having high levels of volume, velocity, and variety. Volume has always been

a factor in BI/DW, as discussed earlier. The velocity of big data is high because it flows from the so-called "Internet of Things," which is always on and includes not just social media and mobile devices but also RFID tags, Web logs, sensor networks, on-board computers, and more. To make sense of the steady stream of data that these devices emit requires a DSS that, likewise, is always on. Unfortunately, high availability is not standard with RDBMS, although each brand offers options that provide fault resilience or even fault tolerance. These options are neither inexpensive to license nor easy to understand and implement. To ensure that Oracle is always available requires RAC (Real Application Clusters for server failover) and/or Data Guard (for data replication). RAC will add over 48 percent to the cost of your Oracle license; Data Guard, over 21 percent.[1]

Furthermore, to install and configure RAC or Data Guard properly is not simple or intuitive, but instead requires specialized expertise about Oracle as well as your operating system. We were willing to pay this price for transactional systems because our businesses depended on them to operate. When the DSS was considered a "downstream" system, we didn't necessarily need it to be available all the time. For many businesses today, however, decision support is a mainstream system that is needed 24/7.

Variety is perhaps the biggest "big data" challenge and the primary reason it's poorly suited for an RDBMS. The many formats of big data can be broadly categorized as structured, semi-structured, or unstructured. Most data about a product return and some data about a customer service call could be considered structured and is readily stored in a relational table. For the most part, however, big data is semi-structured (such as server log files or likes on a Facebook page) or completely unstructured (such as surveillance video or product-related articles, reviews, comments, or tweets). These data types do not fit neatly—if at all—into tables made up of fields that are rigidly typed (for example, six-digit integer, floating point number, fixed- or variable-length character string of exactly X or no more than Y characters, and so on) and often come with constraints (for example, range checks or foreign key lookups).

Like high availability, high performance is an option for an RDBMS, and vendors have attempted to address this with features that enable partitioning, caching, and parallelization. To take advantage of these features, we have to license these software options and also purchase high-end (that is, expensive) hardware to run it on—full of disks, controllers, memory, and CPUs. We then have to configure the database and the application to take advantage of components such as data partitions, memory caches and/or parallel loads, parallel joins/selects, and parallel updates.

A New Approach

In December of 2004, Google published a paper on MapReduce, which was a method it devised to store data across hundreds or even thousands of servers, then use the power of each of those servers as worker nodes to "map" its own local data and pass along the results to a master node that would "reduce" the result sets to formulate an answer to the question or problem posed. This allowed a "Google-like" problem (such as which servers across the entire Internet have content related to a particular subject and which of those are visited most often) to be answered in near real time using a divide-and-conquer approach that is both massively parallel and infinitely scalable.

Yahoo! used this MapReduce framework with its distributed file system (which grew to nearly 50,000 servers) to handle Internet searches and the required indexing of millions of websites and billions of associated documents. Doug Cutting, who led these efforts at Yahoo!, contributed this work to the open source community by creating the Apache Hadoop project, which he named for his son's toy elephant. Hadoop has been used by Google and Yahoo! as well as Facebook to process over 300 petabytes of data. In recent years, Hadoop has been embraced by more and more companies for the analysis of more massive and more diverse data sets.

Data is stored in the Hadoop Distributed File System (HDFS) in its raw form. There is no need to normalize

[1] Based on the "Oracle Technology Global Price List" dated July 19, 2012.

(or denormalize) the data, nor to transform it to fit a fixed schema, as there is with RDBMS. Hadoop requires no data schema—and no index schema. There is no need to create indexes, which often have to be dropped and then recreated after data loads in order to accelerate performance. The common but cumbersome practice of breaking large fact tables into data partitions is also unnecessary in Hadoop because HDFS does that by default. All of your data can be readily stored in Hadoop regardless of its volume (inexpensive, commodity disk drives are the norm), velocity (there is no transformation process to slow things down), or variety (there is no schema to conform to).

As for availability and performance, Hadoop was designed from the beginning to be fault tolerant and massively parallel. Data is always replicated on three separate servers, and if a node is unavailable or merely slow, one of the other nodes takes over processing that data set. Servers that recover or new servers that are added are automatically registered with the system and immediately leveraged for storage and processing. High availability and high performance is "baked in" without the need for any additional work, optional software, or high-end hardware.

Although getting data into Hadoop is remarkably straightforward, getting it out is not as simple as with RDBMS. Data in Hadoop is accessed by MapReduce routines that can be written in Java, Python, or Ruby, for example. This requires significantly more work than writing a SQL query. A scripting language called Pig, which is part of the Apache Hadoop project, can be used to eliminate some of the complexity of a programming language such as Java. However, even Pig is not as easy to learn and use as SQL.

Hive is another tool within the Apache Hadoop project that allows developers to build a metadata layer on top of Hadoop (called "HCatalog") and then access data using a SQL-like interface (called "HiveQL"). In addition to these open source tools, several commercial products can simplify data access in Hadoop. I expect many more products to come from both the open source and commercial worlds to ease or eliminate the complexity inherent in MapReduce, which is currently the biggest inhibitor to Hadoop adoption. One that bears watching is a tool called "Impala," which is being developed by Cloudera and allows you to run SQL queries against Hadoop in real time. Unlike Pig and Hive, which must be compiled into MapReduce routines and then run in "batch mode," Impala runs interactively and directly with the data in Hadoop so that query results begin to return immediately.

Summary

Relational databases have been around for more than 30 years and have proven to be a far better way to process data than their predecessors. They are especially well suited for transactional systems, which quickly and rightfully made them a standard for the type of data processing that was typical in the 1980s and 1990s. We soon found ways to adapt RDBMS for decision support systems, which we've been building for about the past 20 years. However, these adaptations were unnatural in terms of the relational model, and inefficient in terms of the data staging and transformation processes they created. We tolerated this because it achieved acceptable results—for the most part. Besides, what other option did we have?

When companies such as Google, Yahoo!, and Facebook found that relational databases were simply unable to handle the massive volumes of data they have to deal with—and necessity being the mother of invention—a new and better way to process data for decision support was developed. In this age of big data, more companies must now deal with data that not only comes in much higher volumes, but also at much faster velocity and in much greater variety.

Relational databases are no longer the only game in town, and for decision support systems, they are no longer the best available option. ■

BI Experts' Perspective

A Golden Opportunity or a Risky Move?

Rob Armstrong, Jim Gallo, and Steve Williams

Rob Armstrong is director of data warehousing for Teradata. rob.armstrong@teradata.com

Jim Gallo is national director of business analytics for Information Control Corporation. jgallo@iccohio.com

Steve Williams is president of DecisionPath Consulting. steve.williams@decisionpath.com

Eric McCarthy graduated with a computer science degree 10 years ago and worked as a database specialist for three years. He then earned an MBA and has worked as a BI specialist ever since. Eric heads the BI application development team for his current company.

A couple of months ago, a headhunter contacted Eric about a BI director position. Eric had an initial interview with the company and thinks that he will be offered the job. Although he is interested in becoming a BI director, he is uncertain whether this is the right opportunity.

BI is well developed in Eric's current company, which is well along the BI maturity curve. In contrast, the BI program at the other firm is still in its infancy. Decision support data is scattered across several independent data marts, and most analysis is performed using Excel. Eric doesn't mind working hard, but he doesn't want to take a job where the chances for success aren't good.

Please help Eric think through this opportunity by answering these questions:

1. **What are the factors or conditions that are important for BI success?**
2. **Which of these factors are "deal breakers" and which ones can be overcome with hard work?**
3. **What are some of the questions that Eric should ask (of himself or of the new company) before making a decision?**
4. **If Eric decides he wants the position, what are the key points (about the BI environment, the company's plans, expectations, and so on) that should be agreed upon with management before he accepts?**

ROB ARMSTRONG

This is an interesting question; I have been in just this position twice in the past! The fact that I am still with Teradata indicates that there was a "deal breaker" in each situation. With that said, let's explore some of Eric's questions.

1. What are the factors or conditions that are important for BI success?

Several major factors should be considered. Number one is commitment from the new company and executives to the BI program. There should be a defined vision that is supported across the entire executive team.

Part of the vision includes identifying what the new company sees as the purpose or scope of "business intelligence." Is it simple reporting and analysis? Does it extend into real-time data that is interrogated with predictive tools? Are the enterprise's operational processes driven by actionable data? Does the new company envision a goal of self service from the business users, or simply an IT project to reduce costs and consolidate systems? The answers to these questions will provide important signals about the job that lies ahead.

Another factor to consider is the tools and skill sets of people at the new company. Although the current environment is in its infancy, there may be qualified individuals who can help Eric succeed. However, even the best people cannot make up for inadequate tools.

2. Which of these factors are "deal breakers" and which ones can be overcome with hard work?

The deal breaker is the level of executive commitment to giving Eric the support he needs. Without a strong executive sponsor who can champion the funding of BI efforts, Eric is only setting himself up for failure, regardless of how hard he works. One indication of this support is the reporting structure (whether on the business or technology side) that Eric would be a part of.

Without a strong executive sponsor to champion funding of BI efforts, Eric is only setting himself up for failure.

Not quite a deal breaker, but close, is the existence (or lack) of a cohesive, agreed-upon goal state. If the new company is frustrated with its current situation and is looking for a new direction, then executives should be able to document the gaps between what they have now and where they want to be.

The issue of skill sets and tools can be overcome with time and effort. In some situations, it is not even a matter of hard work—it's just executing on a plan. These are not deal breakers as long as Eric has strong commitment from management to resolve the shortcomings.

3. What are some of the questions that Eric should ask (of himself or of the new company) before making a decision?

Eric should start by asking why the current BI director is leaving. Is she being promoted, indicating she has done a good job? Is she being let go, indicating the current solution is viewed as unsuccessful? Is she retiring or simply moving to a new opportunity? The answer may indicate how the company views the current environment from a plan and execution perspective.

Eric also needs to understand how the new company prioritizes its funding, because he will need time and money to succeed. How is the budget already allocated and how was the total budget determined? Is the current solution seen as providing return on investment (ROI), and how is that money being reinvested? What ROI do the business and executives expect from their BI environment, and how much is that worth to them?

Eric should also understand his ability to hire new skill sets or to reorganize his team to remove doubters and those unable to execute his plan. If he is going to have such authority, then Eric should consider his network of other BI professionals and how well he can attract new people to this company.

On the personal side, Eric must consider whether this new opportunity is a challenge for him and something that will make a difference in his career, or whether it is simply "the next job." He should also investigate whether he could get the same opportunity with his current employer, thereby keeping his seniority and any company benefits he may have due to his tenure.

4. If Eric decides he wants the position, what are the key points (about the BI environment, the company's plans, expectations, and so on) that should be agreed upon with management before he accepts?

As mentioned, one key agreement needs to be the vision that Eric is expected to deliver. This will either need to be created or refined depending on the visioning work that has already been done at the new company.

I would go as far as recommending that Eric present a goal state prior to taking the job. Eric can lay out a proposed plan, and if it is accepted, he can hit the ground running. If the vision already exists, then Eric should be able to read through that plan to ensure it agrees with what he would want to champion.

The other aspect that needs to be agreed to up front would be the time frame that Eric has to show success and how that success will be measured.

Finally, although there are no indications in the scenario about Eric's personal life, he does need to consider the impact this new job will have on his family. That can be a whole new topic in itself!

JIM GALLO

Several key factors and conditions are important for BI success:

- Business executives are willing sponsors of the BI program and have a fundamental belief that the solution will drive real value
- BI content is driven from the top down and is based on a set of core metrics that are shared both vertically and horizontally across the organization in such a way that day-to-day tactical decisions are aligned with strategic objectives and drive individual accountability
- Executives are purposeful in articulating the importance of the BI solution and are willing to drive the organizational change needed to treat data as an enterprise asset that is not "owned" by any individual or department
- The organization recognizes that BI solutions are not a series of independent projects but rather a holistic program that requires, to some degree, a continuous and centralized funding model for things such as technology acquisition and salaries for a core team of employees
- Business users are willing to be active participants in solution delivery and data governance activities
- The IT organization accepts that there are fundamental differences in the way BI solutions are built and does not try to force-fit BI delivery methods into a systems development life cycle (SDLC) that was designed for application development

"Deal breakers" include a lack of business sponsorship at the executive level, an unwillingness on the part of the business leaders to drive organizational change, and a poorly defined funding model. These are clearly out of Eric's control and need to be agreed upon before he accepts the position.

It's clear that there's a lot of hard work ahead of Eric if he accepts the job. The following items will require his attention.

First, he will need to define the BI vision and road map, and, if necessary, he may help institute a framework for both program and data governance. He will need to sell the importance of a metrics-driven organization where data is viewed as a shared asset and where responsibility and accountability play a critical role in the organization's success.

Finally, he will need to formulate a delivery model that allows the business to be active participants in the solution without becoming burdensome. I suggest that Eric

work to establish an agile delivery framework that addresses the lack of clarity that is typical in most BI solutions. He should use a business-driven approach that articulates the pronounced differences between designing and building software applications versus BI/DW solutions.

Assuming he gets the answers he's looking for from the hiring organization so far, Eric should have an honest dialogue with himself, specifically asking:

Is he ready to assume the mantle of leadership?

Moving from head of BI applications to a director's position requires a higher degree of technical and interpersonal skills. On the technical front, Eric needs to be confident that he has the ability to create a technical vision that is both evolutionary and practical and that he has the technical wherewithal to articulate a solution that is both scalable and sustainable.

Does he have the interpersonal skills and confidence in his ability to set expectations and sell the solution?

Eric will need to sell the ideas, concepts, and value proposition of a holistic solution—particularly to those individuals who are used to doing things their own way. Change management and expectation setting are imperatives that Eric will have to deal with.

Any road map takes time and money to bring to fruition. Eric needs to assure himself that he can deal with the duality of a long-term program that also needs to show short-term value.

He will need to build a team of like-minded individuals whose goal is to deliver value every step of the way. Eric needs to ensure he'll be able to recruit and hire at least some BI/DW veterans and will not have a team consisting only of BI neophytes.

> Eric will need to formulate a delivery model that allows the business to be active participants in the solution without becoming burdensome.

Finally, Eric will need to play politics. Unfortunately, this may be the skill he'll need to draw upon the most. He must be able to identify and deal with the naysayers who believe they're losing control and clout when an enterprise solution moves to the forefront, as well as with the myriad of product vendors who will do their best to outmaneuver Eric every step of the way.

If Eric decides he wants the position, what are the key points (about the BI environment, the company's plans, expectations, and so on) that should be agreed upon with management before he accepts?

First, Eric should find out the name of the individual who will serve as his executive sponsor and who will be an active participant in helping him drive the necessary organizational changes. If possible, there should be an agreement that the sponsor's compensation will depend, at least in part, on the success of the BI program.

Eric and management should agree on the organization structure. To whom will Eric report? The job title of "director" has different meanings depending on the organization. Eric should make sure that he will not be buried within another IT organization such as application development or architecture. At best he should seek to report directly to the CFO or COO and at worst the CIO. If this is not the case, then he should question his decision-making authority and ability to directly foster and maintain relationships within the organization.

Next, Eric should seek agreement on the company's willingness to allocate sufficient budget for initial capitalization of the infrastructure and to support a core team of BI professionals. Because the organization will be moving away from spreadmarts and Excel, Eric needs to make sure there is sufficient funding to purchase the core elements of the solution, including hardware, database, data integration, and analytics tools. Ideally, he should

seek a commitment regarding the actual amount that the business has budgeted for the BI program for both infrastructure and staff.

Another important point is the breadth and depth of his authority for making key decisions, including architecture, selecting products, and instituting the appropriate delivery methodology.

Finally, he should identify the roles that will be contained within his group versus working within a matrix-based model. Eric should insist that, at a minimum, the following roles be full-time members of the BI group: BI architect, data modeler, business analysts, ETL developers, and BI tool developers. He needs to make it clear that this isn't about "empire building," but rather about the need to direct resources that do not have conflicting priorities and organizational objectives.

Once Eric has come to an understanding of all these items, he needs to find out how both he and the BI program will be measured. That is, he needs to make sure that his authority and responsibilities will be in alignment.

If Eric is confident in his ability to drive change and to wear a salesman's hat, and he gets the business commitment he needs for success, he should jump at the chance with the new employer. If he gets half-hearted answers from the hiring organization, or if they are unwilling or unable to provide the commitments he's looking for, he should walk away from the offer.

If, after all is said and done, he's still a bit uncertain about which way to go, then he should follow what his gut is telling him. After more than 30 years in the industry, I've found this to be the best indicator of what lies ahead.

STEVE WILLIAMS

My BI strategy work with leading companies has uncovered many reasons why BI and analytics initiatives come up short. Some are technical, but many are business or organizational reasons. For Eric, I'll focus on five strategic barriers to BI success that I wrote about in *Strategic Finance* magazine in July 2011.

> Eric should look for evidence that key people in the new company understand that BI is an important profit improvement tool.

The road to BI success is much like that of many other enterprise performance improvement initiatives. Whether the goal is improved financial results, enhanced customer service, reduced operating costs, or any of the many other improvement initiatives companies undertake, success demands very skillful general management and change management. Accordingly, it is important for BI initiatives to identify key barriers to success and make plans to overcome them. These barriers include:

1. **Confusing BI terminology and/or unclear value propositions.** If the top people in a company do not understand what BI is and how it can create business value, they won't fund it at a level that is conducive to success. Eric should look for evidence that key people in the new company understand that BI is an important profit improvement tool. If there are BI or analytics goals in the company's strategic plan, that is a good sign. If not, Eric may be able to overcome this barrier through executive education and prototyping.

2. **Unclear BI mission.** Companies compete in a variety of ways, and targeted BI uses should be consistent with business strategies and competitive dynamics. Some industries are complex and information-intensive, which enables innovators/leaders to create a competitive advantage based on superior use of BI and analytics.

   Eric would be well served to investigate the industry and determine whether the new company is trying to: (a) achieve a BI-based competitive advantage; (b) achieve competitive parity; (c) simply enhance its own strategic execution; or (d) simply make

sure that its BI capabilities do not impede strategic execution. Without a clear BI mission, the company is likely to under-invest and/or meander around the world of BI without a clear purpose. Eric may be able to overcome this barrier through industry analysis, competitor analysis, company analysis, and executive education aimed at obtaining a clear BI mission and charter.

3. **No clear link between business strategy and business processes.** BI and analytics create value by improving management, revenue generation, and operating processes that drive incremental revenues and/or reduce operating expenses. These linkages can be identified and defined through a systematic BI opportunity analysis.

   Eric should determine whether the new company has done such an analysis or would be willing to do so if he accepts the expected offer. If the answer to those questions is "no," Eric may be able to overcome this by forming informal alliances with key business executives who are forward thinking and desire to leverage BI to improve their results.

4. **No burning platform.** Many companies have succeeded for years without much in the way of BI—except for the usual hodgepodge of departmental databases, spreadsheets, and manually intensive, one-off analyses that take weeks instead of seconds. Eric would be well served to investigate the new company's industry position, financial performance, and business leader attitudes toward BI. Absent a burning platform, BI ambitions may be limited, funding may be tight, and BI projects may be limited in scope and potential business impact. This is probably the hardest barrier to overcome, and unless Eric is a great salesman, he may want to consider this factor to be a deal breaker.

5. **Gaps in business-IT alignment.** Rare is the company where business leaders and managers report that IT is easy to work with and gets work done quickly and inexpensively. Equally rare is the company where those same leaders and managers engage to the degree that is required for the BI team to design and build what the business units desire. The challenge, then, is to design the appropriate processes and mutual expectations to overcome the usual business-IT alignment gaps.

   Eric should explore this subject with both business and IT people so he can determine the extent of the gap in the new company (assuming there is one) and make a judgment as to whether that gap can be overcome. Eric may be able to overcome this gap by building alliances with business counterparts who are hungry for BI and analytics.

More broadly, a fundamental question for Eric to ask himself is whether he is looking for career stability (stay with current employer) or a career challenge (the new company). If he is looking for a challenge, he is right to make sure he has a reasonable chance of success. Agreements he should seek include budget, staffing latitude, direct access to business stakeholders, BI development methodology (and waivers from standard SDLC if needed), and control over IT resources. ■

Is "In-Memory" Always the Right Choice?

Selective use of "in-memory" technology is key to finding the right balance between managing exponential data growth and supporting in-time business decisions.

Katrina Read is a business analytics solution architect for IBM.
katrina.read@au1.ibm.com

Katrina Read

Abstract

Data volumes are growing exponentially. Whether you're trying to make sense of the 12 terabytes of tweets created each day to better understand campaign effectiveness or collecting 350 billion annual meter readings to better forecast power consumption, there is no question that the amount of data available to us is increasing at a rate faster than our ability to consume and derive insight.

At the same time, business users are demanding instantaneous answers to their analytical questions. Organizations are encouraging decision makers to substantiate their intuition with fact- and data-driven insight—a futile effort if information is not available as needed to make key decisions. Although executives want to support fact-based decision making, they know they can't do it with complex, hard-to-use tools that require extensive IT handholding. For time-sensitive processes such as responding to a customer who has contacted the call center, data must be accessed and analyzed within minutes to be able to take the right action. The traditional approach of storing information in a data warehouse for retrospective analysis simply won't cut it.

Faced with these opposing forces, it's easy to fall into the trap of thinking that "in memory" is the answer to everything. Promising faster query response times and insight at the speed of thought, many vendors tout in-memory analytics as the silver bullet to performance issues often associated with traditional data warehouses. Coupled with the recent decline in the cost of memory, there has been a renewed interest from enterprises of all sizes as they strategize how in-memory will feature in their business intelligence road maps.

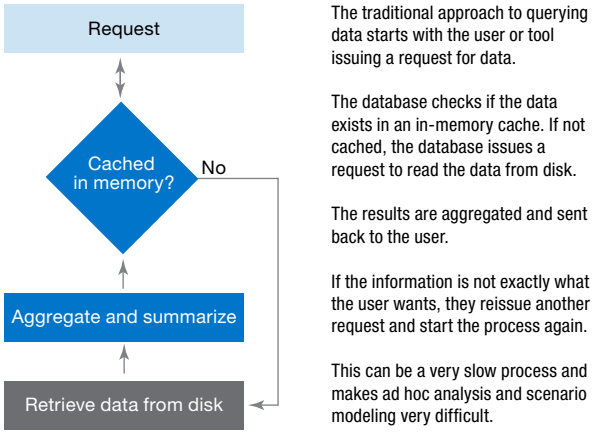

Figure 1: Traditional query approach.

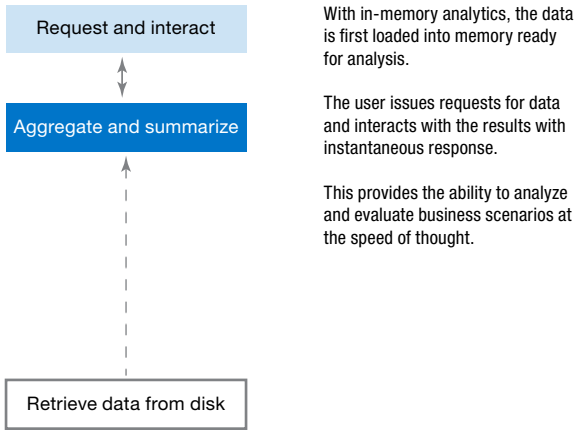

Figure 2: In-memory query approach.

In-Memory Analytics Defined

In-memory analytics is an approach to querying data that resides in a computer's random access memory (RAM), as opposed to querying data that is stored on physical disks. Essentially, in-memory analytics works by bringing data physically closer to the central processing unit (CPU) and eliminating the need to run expensive disk-seek operations each time a data look-up is requested. There is no need to page data in and out of memory—it waits patiently at your fingertips, ready to be interrogated, serving significantly faster query performance.

When a user runs a query against a traditional data warehouse, the query goes to a database that reads information from multiple tables stored on a server's hard disk and processes and aggregates the data in memory (see Figure 1).

Most existing database management systems provide some form of in-memory caching to help reduce the number of disk reads needed to service user queries. In this basic form, data blocks are kept in memory based on caching rules such as when they were last accessed or how often they are accessed. As users interact with and interrogate data across multiple dimensions and time horizons, only a subset of the data required might be found in the data cache at any one time, reducing the performance gains that can be delivered using in-memory technology.

By contrast, in-memory analytics first loads all information into memory. As a result, all user queries are serviced in memory without requiring calls to retrieve data from physical disks. This provides significant performance gains—anywhere from 10 to 1,000 times faster than the traditional query approach. See Figure 2.

There are multiple flavors of in-memory analytics available—from relational-based systems to in-memory online analytical processing (OLAP) platforms. Each focuses on answering different types of questions, but all can benefit from performance improvements when processed in memory instead of on disk. However, in-memory multidimensional platforms store information in a form that provides superior analytical performance—above and beyond what would be experienced by simply loading existing structures into memory. When coupled with write-back capabilities, multidimensional platforms can also provide the foundation for "what-if" scenario modeling and forecasting at the speed of thought.

In-memory analytics can be supported at the platform level or within a BI tool on the end user's desktop. When

supported on the desktop, a subset of data is deployed to the end user's PC to provide lightning-fast response for simple reports and ad hoc analysis. Performance gains are achieved by minimizing and/or eliminating unnecessary disk reads, as well as by avoiding network and server requests for additional data. This type of deployment is generally suited only to small data sets, as it is restricted by the memory resources available on the end user's PC.

Today, in-memory technology provides support for terabytes of data, with the potential for loading an entire data mart.

More commonly, in-memory analytics is deployed to the server or platform responsible for supporting the BI tools. This approach supports larger volumes of memory (and therefore larger volumes of data that can be stored in memory), provides easier maintenance and control over which data sets are supported using in-memory technology, and delivers performance gains to a broad range of users as opposed to a single end user. The focus of this article is on providing enterprise-level, in-memory analytics, delivered via a platform approach.

In-memory analytics also benefits from loading data into memory using advanced compression algorithms. Such compression provides much-needed scalability that was missing from initial in-memory platforms. Decompressing the data for requests requires processor cycles, but the impact is still far less than the processing capacity that would be required to retrieve and process the same data from disk.

An additional benefit of managing analytical data in memory is the elimination of some traditional maintenance activities. Unlike with a disk-based data warehouse, faster performance is not dependent on data indexing or precomputing aggregate data. This simplifies design and allows faster implementation of business intelligence and analytic applications. It also reduces the time and effort required to maintain the ongoing performance of the system.

"In Memory" Is Not Always the Right Answer

With superior performance and reduced maintenance effort, it's easy to understand why many would turn to in-memory technology to support all of their analytical needs. However, "in memory" is not always the answer.

First and foremost, organizations with low data volumes and basic reporting requirements may not have the business case to justify investment in in-memory technology. If traditional methods of storing and accessing data are meeting current business needs, there is no reason to introduce unnecessary complexity into the business intelligence footprint.

Size does matter. The introduction of 64-bit technology delivered the ability to address more than 4 GB of memory and provided much needed scalability. (In fact, if your organization is not yet using 64-bit technology, don't even consider in-memory analytics.) Today, in-memory technology provides support for terabytes of data, with the potential for loading an entire data mart.

However, the implications of growing your database one bit more than the memory on which it's run can be catastrophic—requiring system recovery, redesign, and reprioritization of the data that will be made available to the business. There are other technologies on the market more appropriate for dealing with big-data solutions that analyze petabytes of data.

Providing high availability and disaster recovery are still, to some extent, untested in the realm of in-memory analytics. Some vendors have managed to provide persistent data stores in case of failure, as well as incorporate hardware failover solutions. However, with few tried and tested implementations in the real world, all would agree that the ability to provide a fault-tolerant solution requires meticulous planning and architecture design.

Memory is cheap; deploying in-memory analytics across your entire data infrastructure is not. Although the cost of memory has fallen significantly over the past few

years, and will continue to do so, the cost of in-memory technology has not necessarily followed the same trend. Some vendors have only just released in-memory technology to the market, meaning their products are priced at a point to recover research and development costs. Pricing can range anywhere from $25,000 for an in-memory analytic and planning solution to over $25 million for an in-memory transactional database. Given the high cost of some solutions, the return on investment may be outweighed by the total cost of ownership.

Furthermore, in-memory analytics doesn't make sense for some traditional business intelligence functions. For example, scheduled reports that summarize vast amounts of historical information don't need to be delivered in such aggressive response times and can continue to be serviced from a data warehouse. The same argument can be applied to compliance reporting activities that are conducted annually and compiled over a period of weeks or months.

There are clear and significant benefits to using in-memory technology. Organizations looking to deliver insight about different business groups in time to take action need to consider the use of in-memory analytics to complement—not necessarily replace—traditional data warehouses and marts. This begs the question: How do you know which data is best suited to in-memory? Feedback from business groups is often a good indicator of areas that could benefit from performance gains. However, we also need to consider which subject areas could deliver the most business value if they were made available with instantaneous response.

Recency, Frequency, Monetary (RFM)

To understand which data candidates are suitable for in-memory analytics, we employ a method that is traditionally used to rank customers in the world of marketing—RFM analysis (recency, frequency, monetary). The goal of RFM data analysis is to identify the best candidates for deploying and managing in memory to support in-time business decisions within the constraints of infrastructure budgets. In this case, RFM stands for:

- **Recency:** How recently was the data updated at the source?
- **Frequency:** How often is the data needed to support business decisions?
- **Monetary:** What is the financial impact of this data not being available at the time a decision is made?

Recency deals with the currency of data. Information that needs to be kept up to date with source systems is not always a good candidate for in-memory analytics. For example, when an automated teller machine requests the balance on a customer's account, it is imperative the number received is up to date with all transactions. If this information is being managed in memory, there is a risk that transactions in the past hours/minutes/seconds have not yet been reflected in the balance, making it easier for the customer to withdraw money to which he or she is not entitled.

Frequency accounts for how often the data is needed to support business decisions. The argument for in-memory is that we can improve overall performance by loading data that is accessed most often in memory and increasing the number of times a user request can be serviced without accessing physical disks.

However, recency and frequency on their own can create a bias for regular reporting activities and draw attention away from high-value insight. Typically, they are repeatable reports that can be scheduled and processed using a traditional data warehouse, then published for access by business groups.

The monetary component accounts for the financial impact of the data not being available in time to take action and is arguably the most important of the three. For example, an analyst might evaluate the opportunity cost if information is not available in time to make an informed decision and the wrong decision is made, or if a better decision would have been possible. These subject areas tend to be associated with analytical decision support systems and are the focus of innovation and gaining a competitive edge.

Data subjects can be scored against each of these three categories and the intersection ranked from best candidate for in-memory analytics (most recent, greatest frequency, and highest value to the business) to least valuable (least recent, least frequent, and lowest value). This provides a priority list that identifies where the most value can be gained from leveraging in-memory analytics—starting from the top and working down until the IT budget is spent.

In this process, we find that subject areas related to customer insight, performance scorecard, and planning and forecasting activities provide the highest value to the business when supported using in-memory technology. Subject areas supporting compliance reporting provide a less compelling business case for in-memory investment.

We also distinguish between the predictive models created for analytical decision support and the customer insight derived from such models. This is because they have different data requirements—the former requiring an extensive range of subject areas across long time horizons to create the predictive model, and the latter combining the results of the model with the customer's profile.

It's also important to note that RFM data analysis results can vary over time. Subject areas that are frequently accessed cyclically may provide good candidates for in-memory analytics during times of peak access, reverting to traditional disk-based systems during off-peak periods so that other subject areas can make use of in-memory performance improvements.

Summary

As data volumes grow, decision makers are demanding instantaneous insight to help them make more informed decisions—even as they are supported by IT departments with diminishing budgets. "In memory" is not the answer to every analytical business problem. However, with in-memory scalability increasing, and the cost of memory in decline, in-memory analytics should play a role in every organization's analytical road map.

The real value of in-memory analytics is delivered to the people who have to make decisions quickly and don't have time to wait for disk reads. Vastly shortened query response times put much-needed insight in the hands of decision makers at the time they need to make a decision—whether in the boardroom or at the point of contact. To understand which subject areas are the best candidates for in-memory analytics, we need only look to the RFM data analysis technique to understand the value of having that information readily available at our fingertips.

When considering an investment in in-memory technology, look for solutions that provide complementary use of in-memory and disk-based analytics to provide flexibility for IT and a consistent experience for business groups. The future of analytics lies in technologies that leverage the benefits of both disk-based and in-memory processing to deliver the full spectrum of analytical insight to business groups in time to take action. ■

The Philosophy of Postmodern Business Intelligence

Frank Buytendijk is a TDWI Europe Fellow. This article builds on his TDWI conference keynotes in Europe and in the U.S., and is based on his latest book, *Socrates Reloaded: The Case for Ethics in Business and Technology*. Follow him on Twitter: @FrankBuytendijk. f.a.buytendijk@planet.nl

Frank Buytendijk

An article about philosophy in TDWI's *Business Intelligence Journal*, a magazine for practitioners, might seem out of place. I would argue it is not. The faster the rate of innovation, the greater the need for a philosophical approach. Today, big data, mobile, social media, and cloud computing trends are having an enormous effect on how we view our profession, run our businesses, and even live our lives. These trends move so fast that we have trouble organizing for them. Sometimes it seems these technologies develop autonomously and all we can do is react and pick up the pieces. How do we deal with that?

Best practices describe solutions for yesterday's problems; they don't help. Given a lack of external help for new problems, we need to rely on what we feel is best and organize the debate to come to new common points of view. This is the essence of philosophy.

Becoming Better Professionals

Philosophy also helps us become better BI/DW professionals for several reasons. First, philosophy teaches us how to think: how to judge a line of argument, evaluate objections, and reason in a logical way. In the words of Socrates, "If I can follow good arguments wherever they lead, then my thinking perhaps improves, and I may reduce the degree to which I fool myself."

Concepts such as deduction and induction were derived from philosophy. Deduction is the process of taking a general idea or rule and applying it to certain circumstances; induction is the opposite approach—collecting as many observations as possible on a certain phenomenon and trying to create an overall theory that explains all observations.

It is practical to know when to apply each style of thinking. For instance, applying a methodology for building a business case is deductive; requirements gathering requires inductive thinking. Although IT people are often strong conceptual thinkers, studying philosophy would help them stand on the shoulders of giants.

Second, philosophy teaches us that there are different ways of thinking and that several ways might be equally logical, convincing, and even valid. Consider the competing factions in organizational goal setting. One school of thought is that the ultimate goal of every business is to create and optimize shareholder value because the business is owned by its shareholders. Another school of thought reasons that the goal of an organization is to sustain itself, like any other living organism as part of an ecosystem. Therefore, it should focus on stakeholder value and view profits as merely the oxygen required. Neither of these logical positions is easy to disqualify.

Finally, once we understand that there are multiple ways of thinking, we can examine and question our own thinking as well. We may discover that long-held preconceptions don't hold up, and we may have to change our fundamental beliefs as a consequence. This can be unsettling, but it leads to a higher level of thinking: meta-thinking. Philosophy helps us think about *how* we think. Once we master that, or at least apply it, our ideas will be more creative and multidimensional, and we'll be able to test them against multiple views and scenarios.

Different Ways of Thinking

It is surprisingly difficult to accept that you can separate a person from his or her ideas. The story of Carneades confirms this. Carneades, one of the heads of the Academy originally founded by Plato (in 385 BCE), was sent on a mission to Rome in 156 BCE. He decided to combine the mission with a series of lectures. Because Greek philosophy was popular in Rome in those days, there was considerable interest. During the first lecture, Carneades explained the views of Aristotle and Plato, which people were thrilled to hear. In the second lecture, Carneades reasoned the complete opposite perspective and was equally convincing.

His point was not to prove Aristotle and Plato wrong but to show the skeptics that there are different ways of thinking. This caused consternation among Roman politicians, especially senator Marcus Cato, who felt such independent thinking was a bad influence on Rome's youth. He complained to the Senate, claiming that it was better for the people to simply obey the law. As a result, Carneades was sent home.

Fast forward to the 21st century. Surely, we must have progressed? A few years ago, I participated in what we called a "dialectical debate" as a conference keynote for TDWI Europe. My co-presenter, Wayne Eckerson, and I came up with the idea of taking opposing views in reaction to propositions from the moderator of the conference. Then we would switch sides and equally vigorously argue our opponent's view, bringing new arguments to the table. We even visualized this by holding up red or green cards to indicate if we were arguing for or against a proposition. Although the attendees were amused (at least that much has improved in two millennia), they were also confused. They insisted that we share what we *really* believed. What did we believe? Well, all of it!

Let's apply some philosophical schools of thought to the world of business intelligence and data warehousing. I will concentrate mostly on one school of thought: postmodernism. Postmodern philosophers believe there is no truth, there is only perception. They don't believe in definite categories, just in many shades of gray.

I have chosen the postmodern point of view for two reasons. First, our society is (still) thoroughly postmodern, so everyone should feel comfortable with this style of thinking. Second, its contrast with information management best practices couldn't be bigger. Just like the philosophers of the Enlightenment, IT professionals are still searching for truth, objectivity, and optimization.

DW Architecture Leads to Ethical Issues

In April 2011, European newspapers reported that police in the Netherlands were using data collected from TomTom navigation devices to plan speed traps. TomTom collects driver data in real time and uses the information to notify subscribers about traffic jams.

TomTom also states in the terms and conditions for its service that it is allowed to sell the collected data in an aggregated and anonymous form.

The authorities in charge of highways and roads have found good uses for the TomTom data. By looking at average speed, officials can see where road improvements are needed to eliminate recurring traffic jams or minimize those caused by ongoing roadwork. That didn't pose any problems, but then the data landed in the hands of the Dutch police, who used it to calculate how *fast* people were driving and to position cameras to catch speeders.

Was that use of the data appropriate? The data was purchased legally and contained no identifiable information, so citizens did not have to give consent for the police to use it. Furthermore, the data wasn't used to find and punish speeders after the fact—it was used to catch people at the actual moment of speeding. In fact, this type of data-based decision making is an efficient use of taxpayers' money. Yet TomTom's immediate reaction was to stop the practice. Facing negative feedback from customers, the company decided that the police use of its data was bad for business. Customers pay extra for premium services such as dynamic traffic-jam monitoring, and they enable those services by supplying the required data. They are supposed to benefit from that, not be punished as a result.

Could this outcome have been predicted? Perhaps, but best-practice data warehouse architecture stands in the way. It dictates that data warehouses should be application neutral; that data should be usable for many purposes, most of which are not even foreseen at the time the data warehouse is built.

Postmodernists would frown on this objective approach. They would point out that data is never neutral. People observe through their senses, which are designed to capture a specific style of data. We see, we hear, we smell, we feel, and we taste. Based on the combination of those partial observations, each colored by the limitations of the specific sense, the brain constructs a reality. A data warehouse is not different. The data is not neutral; it is colored by the system that collected it. By not capturing what the data was originally intended to show (i.e., by making the data warehouse application neutral), we blind ourselves to questionable uses. In fact, there is a rule here. *The more a certain use of data is removed from the original goal and the original measurement instrument, the bigger the chance that issues will arise.*

Postmodernists would argue for including more metadata in the data warehouse to describe the purpose of use and ban or flag queries that are too far removed from that purpose. The data warehouse should not be a slave to the query and simply respond. It should have a "mind of its own" that can carefully interpret the query and decide what the best response is—which is not always the direct answer. This is a very different principle indeed.

Big Data Leads to Less "Truth"—Not More

Big data is the best thing that ever happened to information management. Now we know it all. Once we have all the data, we can measure everything, and fact-based decision making will become a reality. This is the vision BI/DW professionals are trying to sell: more truth, more objectivity, and better decisions.

Again, the postmodernists would have a different point of view. They would point to infrastructure. Big data databases may be able to harness incredibly large data streams, but most predictions are not optimistic about the infrastructure's capacity to copy such data volumes. Big data warehousing increasingly becomes federated or virtualized.

In short, if data sets become too big to be copied within reasonable time frames, you effectively cannot copy them anymore. They become unique. Data collections become *individuals* in the literal sense of the word: they exist just once. Two collections of data may be similar or related, like siblings, but can never be identical. With a little bit of imagination, you can argue that data sets become like persons.[1] They grow and mature over time. Data sets develop unique behaviors that they display when you interact with them. They can even develop dysfunctions

[1] I would like to recognize Roland Rambau, a colleague when I worked at Oracle, for coming up with this idea.

and disorders as they are trained by the data and the analyses that systems perform. (My career recommendation for the years to come is to become a "data therapist.")

Furthermore, their complexity in terms of volume, variety, and velocity is such that it cannot be fully understood anymore by ordinary human beings. This complexity drives us to trust the answers the systems give because the moment we try to audit the answers, the data has already changed. Effectively, like people, systems offer a subjective point of view that is sometimes hard to verify. They express, for all intents and purposes, opinion. Managers need to think for themselves and interpret the outcome of querying different sources, forming their particular picture of reality—not based on "the numbers that speak for themselves" or on fact-based analysis, but by synthesizing multiple perspectives to construct a story.

According to this perspective, information managers are further away from the "one version of the truth" they strive for than ever before. Perhaps information managers should leave the Era of Enlightenment behind. Perhaps the idea that there is a single truth, and that we simply need to discover and roll it out, is unrealistic.

The Fallacy of the One Version of the Truth

Professionals concerned with defining key performance indicators, putting together organizational taxonomies, and building data warehouses have been looking for a single version of the truth since the advent of the information management discipline. Most organizations have fundamental alignment issues in defining the terminology they use. In fact, I have formulated a "law" that describes the gravity of the problem: *The more a term is connected to the core of the business, the more numerous are its definitions*. There typically are many definitions of what constitutes revenue in a sales organization, what a flight means to an airline, or how to define a customer for a mobile telephone provider.

Few organizations have been successful in reaching one version of the truth. Business managers have fiercely resisted. Machiavelli might have pointed to the political motives of business managers, because a single version of the truth would limit their flexibility to choose the version of the truth that fits their story best. However, IT professionals say business managers should see that satisfying their own goals is less important than the satisfaction of contributing to the success of the overall organization. In fact, ignoring less important needs for the benefit of higher pleasures, or for the benefit of others, is a hallmark of human civilization. So much for civilization if we can't even achieve this in the workplace.

To explain our failure to reach a single version of the truth, postmodernists would point back to IT professionals themselves, saying they are simply misguided. By taking a postmodern approach to the single-version-of-the truth problem that has been troubling information management professionals for such a long time, it simply disappears. The reason why all these versions of the truth, often under the same name, exist in isolation is the vertically aligned setup of the management structure. Each business domain reports up to strategic objectives, and most of the reporting is "self reporting"—the domain reporting based on its own data.

However, in a process-oriented approach, multiple versions of the truth—horizontally aligned and next to each other—actually make sense. The various departments or business units each have a different relative position in the value chain and, therefore, a different view of the current revenue or the number of customers. This doesn't mean that every single definition is valid and should be preserved—in fact, many definitions may be redundant. The real question is: How does an organization decide which definitions are valid and which are not? Multiple valid definitions, placed in the right order, constitute "one context of the truth."

Consider the following real-life example. "What is a train?" is a more complicated question than you might think. Different stakeholders have different views. For a passenger, it is the means for the journey to their destination. From a regulator's perspective, a timetabled train is a line that runs multiple times per day based on budget and policies on public transportation. The planning department would also add maintenance movements and empty trains scheduled to travel to new departure points. Traffic control would look at actual—including unplanned—train movements.

The infrastructure department might actually count slots (a time window in which a train is supposed to travel), and this might include other operators' trains as well.

With a horizontal alignment approach that organizes all definitions in a single report, the definitions become more transparent and comparable. There is value in analyzing the differences; it is important to minimize the differences in planning efficiencies and number of incidents and accidents between the demand plan and the operations. The closer the number, the more optimized the plan. Next, the difference between operations and the staffing plan can be minimized, allocating scarce human resources as efficiently as possible.

This is a perfect example of the power of philosophy. A different school of thought may make a problem that seemed unsolvable suddenly look very different.

We Are All Philosophers

IT professionals and philosophers have a lot in common. Both professions focus on thinking things through and using analysis to understand the essence of things and the way they work. IT professionals speak of functional decomposition to describe all elements of an envisioned system. Both professions have a conceptual and logical view of the world, which is another reason it is strange that IT people do not have more appreciation for philosophy. Wasn't it Socrates who laid much of the groundwork for today's logic, rivaled only much later by Spinoza, Heidegger, and Russell? Weren't Euclid, Pythagoras, and Descartes some of the main contributors to mathematics?

IT professionals, like most philosophers, enjoy discussing definitions. There are endless debates about what cloud computing or enterprise 2.0 really means, and these debates are not likely to be resolved soon. In fact, IT professionals have been arguing for the last 20 years about the exact definitions of BI and knowledge management. Maybe these endless debates are caused by the conceptual nature of IT. After all, have you ever heard two grocers debate how to define an orange? I haven't!

As in philosophy, in IT there are always multiple—often conflicting—schools of thought. Remember the endless Inmon or Kimball debates? Having multiple schools of thought is not unique; it happens in every conceptual discipline, such as macroeconomics (e.g., how to deal with a crisis) and strategy (Henry Mintzberg even defines 10 distinct schools of thought).

IT professionals and philosophers are known for their ability to create frameworks and models to describe reality and to share certain views. Think of different frameworks to describe an enterprise architecture, a project management methodology, or function-point analysis to estimate the complexity of certain projects. These frameworks come with jargon, another commonality between IT and philosophy. Why are these difficult terms needed? Actually, jargon is part of every profession. Doctors have jargon and so do carpenters. There is a need for jargon—not just to create an aura of wisdom, but to be able to communicate ideas concisely and precisely.

Another similarity is that both IT professionals and philosophers tend to position themselves slightly outside of reality. IT professionals frequently refer to the rest of the organization they work for as "the business," and philosophers talk about "life" and "society" as if they weren't part of them. Both professions feel that taking an abstract view positions them above the matters they discuss. It gives them oversight and insight and provides them with a deeper understanding. Some IT professionals go as far as to call themselves *business architects*.

At the same time, there are many differences between IT professionals and philosophers; most noticeably, IT people are paid better.

One more story: Aristotle wrote about Thales (c. 624–547 BCE), one of the first known Greek philosophers, physicists, and mathematicians who, during winter, read from the weather and the stars that the next year would have a great olive harvest. He made a fortune by buying up all the olive presses he could get and renting them out when it was harvest time.

Philosophy can be a highly practical discipline. It's time that we all started practicing more philosophy. ■

BI StatShots

TDWI Technology Survey: Emerging Technologies and Methods in BI

The Technology Survey that TDWI circulated at its recent World Conference in Orlando presented a list of 30 emerging technologies and methods (ETMs) and asked attendees to identify those they have no plans for using, those they are already using, and those they'll adopt within three years.

A third of ETMs will see very aggressive adoption. The ETMs in Group 1 (Figure 1) were each selected by approximately 50 percent of respondents as techniques they are not using today, but will be using within three years. The ETMs in Group 1 vary from very new techniques (big data analytics, text analytics, mobile BI, social BI) to techniques that have been with us for years but are just now emerging in terms of brisk user adoption (real-time operation, master data management, and advanced data visualization).

The newest ETMs are set for the most growth. These show the greatest difference between "using today" and "within three years": big data analytics, social media analytics, text analytics, and clouds for BI/DW.

A few ETMs will be adopted by most organizations. Very small percentages of survey respondents selected "no plans" for MDM, self-service BI, predictive analytics, agile BI, and data services, which means that these are high priorities for most organizations.

Some of the least-used ETMs today will see appreciable adoption. For example, cloud BI is in use today by a mere 10 percent of respondents, yet a whopping 40 percent anticipate adopting it within three years. Other relatively new and obscure ETMs will similarly ascend into popularity, namely NoSQL DBMSs, MapReduce, social media analytics, and Hadoop.

—Philip Russom, TDWI Research Director for Data Management

Which of the following ETMs is your organization using for business intelligence (BI), data warehousing (DW), or data management (DM)?

GROUP 1
Approximately 50% of respondents will adopt the following ETMs within 3 years.

	No plans for using	Already using today	Not using today; will within 3 years
Big data analytics	28%	18%	54%
Real-time BI/DW	20%	27%	53%
Mobile BI	18%	29%	53%
Social media analytics	32%	16%	52%
Text analytics	27%	21%	52%
Unstructured data	20%	28%	52%
Predictive analytics	10%	40%	50%
Master data mgt	9%	43%	48%
Unified data mgt	32%	21%	47%
Advanced data visualization	23%	29%	47%

GROUP 2
Approximately 40% of respondents will adopt the following ETMs within 3 years.

	No plans for using	Already using today	Not using today; will within 3 years
Hadoop	42%	17%	41%
Clouds for BI/DW	50%	10%	40%
Complex event processing	40%	20%	40%
Data federation	31%	29%	40%
In-memory analytics	25%	35%	40%
Data virtualization	27%	34%	39%
MapReduce	47%	15%	38%
Streaming data	43%	19%	38%
Self-service BI	9%	54%	37%

Figure 1: Based on 139 respondents. Values in the table represent percentages of respondents. The table is sorted by the "Not using today; will within 3 years" column.

WHAT SETS **YOU** APART FROM THE CROWD?

Distinguishing yourself in your career can be a difficult task. Through TDWI's CBIP (Certified Business Intelligence Professional) program, we help you define, establish, and set yourself apart professionally with a meaningful BI certification credential.

Become a Certified Business Intelligence Professional today!

To find out what CBIP exams you should take, how to prepare, and where you can take the exams, **visit tdwi.org/cbip**.

tdwi
THE DATA WAREHOUSING INSTITUTE

TDWI Partners

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

tdwi PARTNER