



THE FUTURE: eXtreme data warehousing

Technology may help you win the data race, but what do you do about the social issues?

The motivation behind our idea of eXtreme Data Warehousing (X-DW) is to consider where the industry trends will take us a few generations ahead of where we are now. What will happen when organizations begin to deploy data warehouses with scalability and service-level requirements 10 times beyond any system currently in existence?

Imagine millions of users accessing petabytes of data, with response times for decision-making measured in milliseconds. Immediate visibility to business events will be required. And the data will be integrated to the extreme—knowledge workers will be able to exploit all business relationships, internal and external, when driving decisions from the data warehouse. The granularity of data collected in today's data warehouses will increase by multiple orders of magnitude in

future generations. Atomic data capture will evolve to subatomic levels of granularity.

A future involving X-DW raises many interesting technical and social issues. We believe that technical concerns related to X-DW deployment are well-within the bounds of our future engineering capabilities. The social implications, on the other hand, will be more difficult to address.

Bigger, faster, fresher

There was a time when some predicted that Moore's Law would catch up with data warehouse business requirements and that in-memory databases would become the standard.

Quite frankly, we never believed it.

Demand for information is seemingly insatiable. As the cost of storing data continues to drop while the value of information continues to increase simultaneously, there is an economic imperative to leverage increasing volumes of information in data warehouses. This demand is outpacing Moore's Law, and it is clear that higher and higher density disk drives will be required to store this data. >

BY STEPHEN BROBST & RICHARD HACKATHORN

Over the next two years, it is predicted that the total amount of data available for decision support will quadruple vs. today's volumes. This is more data than has been created in 40,000 years.¹ However, we should not confuse data and information.

It is a much more simple task to collect data than it is to create information. To create information from data requires integration, cleansing and access. The computing power to facilitate the delivery of information to a large enterprise is enormous and will continue to grow.

Luckily, technological advances will help us. Massively parallel processing will become even more massive with blade technology. This high-density computing technology packages a "server on a card," with each blade having its own collection of processors, memory and I/O capability.

While the I/O bandwidth in a blade server using today's implementations is not sufficient for high-end data warehouse workloads, advances in technology will make this type of computing power very attractive within the next two to three years.

Within the foreseeable future, it will be possible to deliver huge numbers of CPUs with terabytes of memory for data warehousing using commodity blade components. Virtualization using grid computing as a framework—but with much more sophistication than in today's implementations—will allow cost-effective management of computing, memory, I/O and storage resources.

The ability to leverage enterprise data warehousing capabilities has expanded beyond the traditional knowledge workers in the corporate ivory tower. Active data warehousing has enabled human and software agents on the frontlines of an organization to access analytic services for tactical decision support. (For more on this topic, read "The five stages of an Active Data Warehouse evolution" in the Spring 2001 issue of *Teradata Magazine*.)

Business intelligence plays a critical role in the real-time enterprise to ensure that decisions are not just fast, but "intelligent" as well.

These tactical decision-making capabilities complement traditional uses of data warehousing by enabling more effective execution of the business strategy with better access to information. Moreover, new organizational structures involving "virtual" organizations, with increasingly closer integration of an enterprise with its suppliers and customers, require much more efficient sharing of information outside of traditional organizational boundaries.

The emergence of tactical decision-support workloads on the data warehouse demands more aggressive service levels in the area of data freshness. The phenomenon known as "real-time" data warehousing is part of a much bigger trend.² Business intelligence plays a critical role in the real-time enterprise to ensure that decisions are not just fast, but "intelligent" as well. Gartner predicts that any participant in a competitive marketplace that has not embraced real-time enterprise capability by 2006 will have difficulty sustaining a leadership position.

To facilitate the real-time enterprise, data warehouses must be able to capture and deliver information in near real-time. Message-based data acquisition using enterprise application integration (EAI) infrastructure is quickly supplanting batch-oriented file processing for capturing time-sensitive data into the data warehouse.³

The reach of enterprise data warehousing is expanding to a global scale. As businesses continue the trend toward globalization, information requirements are spanning international borders.

When one considers the need for information to manage the entire value chain for product or service delivery—from raw materials, through manufacturing, distribution and consumption, and then beyond to retirement or disposal—it is clear that data warehousing is becoming a borderless entity. Organizations such as Dell, Vodafone, 3M, Wal-Mart, DHL and others are pursuing worldwide data acquisition and delivery from their data warehouses.

How will the use of information change with the evolution of X-DW? Take a look at some case studies that share what the future might hold.

Future case study: extreme retailing

In today's retail data warehouse deployments, atomic data consists of market-basket detail. Stores acquire information about every market basket, along with details about individual items in the baskets. Price, cost, coupons, tender types, time of checkout, lane, etc. are all captured at a detailed level. That's not new.

What is new is that tomorrow's subatomic detailed data will encompass the total shopping experience. Of course, the market-basket detail will be captured, but so will the order and timing of item placement in the market basket.

The shopping route through the store can also be recorded. This additional information will help store owners, floor managers and even manufacturers understand customer buying behaviors and more accurately segment their customer.

RFID technology can easily be used to track individual items as they pass from a store shelf into a shopping cart and through the checkout line. A shopping cart also can be tracked in its movement throughout a store using the same type of technology. When the customer goes through the checkout line, the RFID information can be directly attached to the traditional market-basket data for



later analysis inside the data warehouse.

In an even further extreme in retail data warehousing, the shopping experience can be captured in a video recording along with structured data. Analysis of facial patterns can be used to glean information about customer reactions to new products or price changes.⁴

Video surveillance is already in place in most retail outlets for security reasons, but it would need to be expanded in most cases if the full shopping experience is to be captured. The technology already exists for tracking individuals with digital-image matching and can even be used to identify repeat visits through facial recognition using a digital-image database of previous shoppers. In fact, this technology is already in place in many airports for security reasons.

The big issue is privacy. The industry has generally agreed that RFID tracking would be disabled as soon as items leave a retail outlet. Most people are fully aware that video surveillance is done within the retail environment for security purposes.

But is it acceptable to engage in this kind of tracking within a retail store for purposes of customer analytics? While this information may be very useful from a customer-analytics perspective, the implications of “Big Brother” analyzing every shopping nuance for the purpose of understanding how to better influence shopping behaviors may be a bit unnerving to some.

Another interesting set of possibilities exists in the area of pricing and promotional execution. In the near term, electronic shelf labels allow retail enterprises

to use data warehousing analytics to automatically drive markdown pricing with centralized synchronization between shelf pricing and point-of-sale (POS) devices. Future scenarios in X-DW will allow efficient pricing and promotions at an individual customer level.

Customers who participate in a retail loyalty program would be able to receive in-store delivery of personalized offers via hand-held devices such as PDAs. Differentiated pricing by individual customer would be possible through data warehouse analytics.

Such pricing strategies could be used to influence behaviors such as cross-sell, up-sell and loyalty. This approach to promotional pricing obviates the need for traditional, often ineffective, coupons.

Of course, it also means that customers will be trading privacy for preferential pricing. There are also issues of fairness that come into play when differentiated pricing is put into place. These issues already exist with traditional implementation of loyalty programs, but they will gain visibility as automation facilitates more widespread deployment.

Future case study: extreme healthcare

The integration of medical history for every patient to include all treatments and drug prescriptions, outcomes, reactions, etc., has great promise for more effective delivery of healthcare. In addition to the capture of traditional structured data in codified form, the integration of unstructured data, such as X-rays and CAT scans, could be included in X-DW.

While it is desirable to have the medical history for each individual patient, even better would be to have access to the medical history for the complete family tree of each individual as well as DNA encoding for each individual.

This depth of information would allow better understanding of patient predisposition toward certain ailments in order to facilitate proactive health checks and early intervention. Health care becomes more effective with complete information.

Patient-administered testing devices are already on the market for a number of different ailments, such as hypertension and diabetes. Technology to enable data collection directly from such devices via wireless connections or analog phone lines also is well-within technological reach. This opens the possibility of obtaining the results from at-home testing directly into a data warehouse in addition to in-hospital data collection. >

The challenge is that there are significant privacy issues involved in X-DW. How far should the use of information be allowed to go? —Richard Hackathorn



Continuous and near-real-time data acquisition with event monitoring efficiently allows for proactive intervention. By combining up-to-date data with historical data, it will be possible to detect opportunities for early diagnosis and intervention in treating patients. Data mining and other deep analytics can be used to identify important events and actions to take when they are encountered.

Global access to healthcare information also is becoming more important. As society becomes more mobile, the likelihood

of requiring healthcare outside the reach of a primary care physician is dramatically higher. A doctor in another city or country would benefit significantly by having access to a complete medical history, rather than having to try to reconstruct it or do without.

The use of information in a data warehouse to assist in healthcare delivery will necessarily require extremes in data quality management. Strict controls and processes for managing data quality and creating accountability for accurate content in the data warehouse will be critical.

There is also a question of ownership. Does the patient get to decide who has access to his or her medical history and how much the recipient is allowed to see? Is the patient qualified to make such decisions? Who will pay for the construction of such a data warehouse: a government organization, a health insurance carrier, a health care provider? How should privacy be implemented? How are such controls enforced?

There are many social questions that need to be addressed in the sensitive area of health care data management.

Future case study: extreme automotive insurance

Insurance is all about using information to better understand and assess risk. In a world with a lack of information about individual driving habits, risk is assessed with actuarial calculations using data that is available for independent verification.

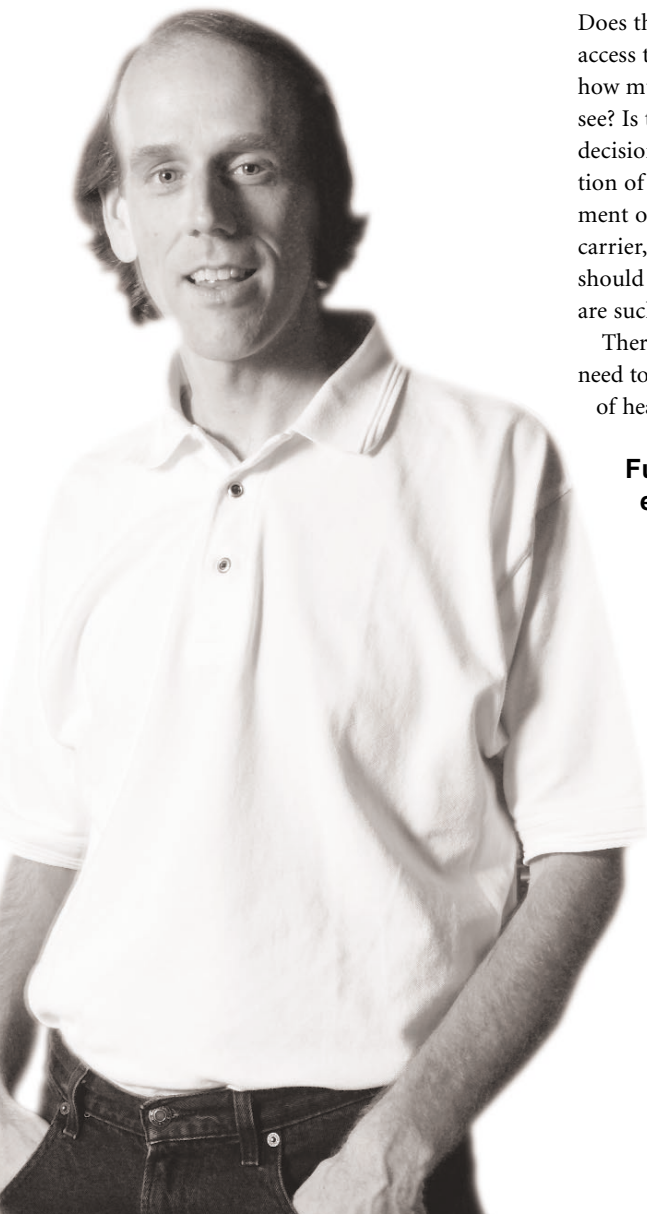
Insurance premiums typically are based on factors such as automobile type (engine size, color, model, etc.), geography in which the vehicle is driven and driver demographics (age, gender, academic performance, smoker status, etc.). As an individual accumulates a driving history, including accidents and speeding tickets, the risk models adjust for these factors.

The inputs just described provide a basis for understanding the risk of individual drivers based on the statistical performance of peer groups with similar characteristics. What this means, however, is that “good” drivers within a peer group end up subsidizing the “poor” drivers. After all, not all 18-year-old males who drive red sports cars are reckless maniacs on the road (just most of them). Better information about the driving habits of each individual within the peer groups facilitates more accurate risk assessment.

The X-DW scenario to address this involves monitoring driving habits to facilitate premium adjustments based on actual driving behaviors. Such driving behaviors can be observed via satellite collection of data during automobile operation.

To do this, an individual must be identifiable when getting into the driver’s seat. That way, subsequent driving habits can be collected during vehicle use. Those who drive beyond the speed limit and weave in between lanes will pay higher premiums than those who do not. The data warehouse is used to analyze the vast amount of data from these monitoring activities and provide individual premium pricing recommendations.

We should not confuse data and information. It is a much more simple task to collect data than it is to create information. —Stephen Brobst





Monitoring will be optional by customer choice, but those who do not submit will be assumed to be “worse case” drivers and therefore subject to higher insurance premiums. The goal behind this use of X-DW is to make insurance pricing more “fair” by limiting the subsidies from good drivers to poor drivers.

The challenge is that there are significant privacy issues involved in the approach. How far should the use of this information be allowed to go? Would law enforcement agencies have access to it? How about the parents of a teenage driver? Of course, there are already many examples where driving habits are being monitored by external devices (as victims of the congestion tax for driving in the London city center can readily attest), but this future scenario takes it to another level.

Tough issues

These X-DW examples demand service levels and increase scale far beyond current practice. However, the challenges to the success of these applications are not issues of technology. The tough issues involve the organizational, political, legal and ethical aspects.

A concise way to describe these issues was outlined by Richard Mason.⁵ He was the first to propose a framework for IT ethics, and his work has been the standard for nearly 20 years. Mason identified the following four IT ethics categories:

- > Privacy: personal identification and confidentiality
- > Accuracy: data quality and liability
- > Property: ownership and control
- > Accessibility: haves and have-nots

Privacy deals with disseminating information that is personal to an individual, such as name, location, finances, health and the like. The Web is enabling the global distribution of personal information with a simple click, and X-DW will enable its analysis with similar ease.

Lawmakers and privacy organizations are struggling to strike the right balance between the right to privacy vs. the right to know.

The issue of privacy has reached a high level of visibility with the increasing use of data warehousing and data mining for security applications. Organizations such as the Technology and Privacy Advisory Committee (TAPAC)⁶ are proposing approaches for analysis and data mining that involve requirements for anonymization of customer identifying information. It is clear that lawmakers and privacy organizations are struggling to strike the right balance between the right to privacy vs. the right to know in every society.⁷

Maintaining the accuracy of information in a large data warehouse is a costly undertaking. Responsibility for the accuracy of individual data fields is a huge issue. Equally important are the accuracy of join-paths to integrate disparate sources of data and the accuracy of analytics performed upon the data.

These areas, unfortunately, do not receive nearly enough attention in most data warehouse deployments. In addition, the legal liability and the ethical responsibility for data accuracy are confusing areas. When incorrect data causes loss or injury, the determination of responsibility is often difficult.

Ownership implies the treatment of information as property. That notion begs the question about who pays for and controls the data.⁸ As we hurtle into a world of wireless digital information in every human activity, the traditional pillars of intellectual property rights including trademarks, copyrights and patents begin to crumble. In X-DW with global deploy-

ment and data sourcing, information may be subject to conflicting laws and regulations from various countries and regulatory agencies. It may be difficult to determine whether an enterprise has rights and obligations to collect, maintain and use specific information for specific purposes across global boundaries.

Accessibility means the ability to use information for one’s own purposes. Information is power. Access to that information is an allocation of power. The ability to access the information is part of the equation, but the skill to understand and manage information, along with the tools to enable this understanding, are equally important. The digital divide that separates the world into the “info haves” and “info have-nots” will take on deeper meaning as the emergence of X-DW may separate the world into the “BI haves” and “BI have-nots.” As X-DW scales to global levels, data without analytics will become worthless.

While the applications of X-DW hold great promise for society, the tough issues surrounding X-DW hold great controversy for society. As an industry, we need to seek solutions that are creative and compassionate and avoid those emerging out of ignorance or malice. **T**

1 Lyman, P. and H. Varian. How Much Information. University of California at Berkeley, www.sims.berkeley.edu/how-much-info-2003, 2003.
 2 Flint, D. and M. Raskino. The Real-Time Enterprise. Gartner Report K-18-9848. January, 2003.
 3 Brobst, S. Active Data Warehousing and Enterprise Application Integration. Proceedings of Data Warehousing 2002: From Data Warehousing to the Corporate Knowledge Center. Physica-Verlag Heidelberg. November 12-13, 2002. pp. 15-23.
 4 Coutts, M. and D. Schrader. E-Motions: Kiosks that See and React. NCR Self-Service Strategic Solutions Annual Research Report. 2001. pp. 71-78.
 5 Mason, R. Four Ethical Issues of the Information Age. Management Information Systems Quarterly. Volume 10, Number 1. March, 1986.
 6 Go to <http://www.sainc.com/tapac> to read final report.
 7 Lane, C. Naked in Cyberspace. Pemberton Press. 1997.
 8 Branscomb, A. Who Owns Information? Harper Collins. 1995.

Stephen Brobst is Teradata’s chief technology officer. Richard Hackathorn is president and founder of Bolder Technology, Inc.