



Published in DM Review in June 2004.
Printed from DMReview.com

The BI Watch: eXtreme Data Warehousing

Richard Hackathorn

Technology is No Longer the Limiting Factor for Data Warehousing

Hackathorn wishes to thank Stephen Brobst for his contributions to this month's column.

At the recent Teradata Universe conference in Paris, we explored a fresh perspective on the future of data warehousing - eXtreme Data Warehousing or X-DW.¹ The session was intended to be provocative, pushing the audience to think out of the box. We realized that the session was more than provocative; it was prophetic, surfacing the really tough issues that will impact our profession.

Energized by more than a decade of tangible business success, industries are hurtling ahead with fundamental changes to their data warehousing (DW) systems. A new generation is upon us with real-time data collection, operational applications and second-by-second decision making. Technology barriers are rapidly and relentlessly being eroded. We can build DW systems with 10x or more of the capability of last year, but should we?

X-DW is thinking about data warehousing beyond current implementations and imagining the implications if each of the following dimensions was pushed to the extreme:

- Performance: Response times measured in milliseconds.
- Scalability: Millions of users accessing petabytes of data.
- Availability: Up all the time, without planned or unplanned downtime.
- Data freshness: Business events visible immediately.
- Comprehensiveness: If the data is not there, you do not need it.
- Integration: Exploiting all meaningful business relationships in the data.
- Granularity: Atomic data becomes subatomic.

X-DW drives us to move beyond enterprise DW into truly global DW. It strives toward a single version of the truth that encompasses multiple companies across multiple countries. It considers the entire value chain from raw materials through manufacturing and distribution to the end consumer, and then beyond to retirement and disposal.

The purpose of X-DW is to challenge our thinking. Are we heading in the right direction? And, what is the business imperative for doing so? Should we be the first to push the technology to its limits? Or, should we be the first to warn others not to do so? With DW success comes DW responsibility. The next generation of data warehousing has great promise for health and safety, customer care, efficiency, optimization and diversity; but it also raises many difficult social and ethical issues having potential for volatility and corruption.

Let's consider a series of examples to highlight various aspects of X-DW.

Consider eXtreme Retailing. Current retail analysis focuses on the market basket (i.e., items purchased). Future retail analysis will attempt to capture the entire shopping experience, from the time the need is recognized to the order and timing of placing items into the basket, routing through the store and even video of the complete shopping experience. RFID (radio frequency identification) technology will track the movement of all items including the shopping cart. With electronic loyalty cards and consumer self-identification, pricing will be dynamically calculated to maximize profit to the store and value to each individual customer. There are many privacy concerns that will need to be resolved.

A manufacturing example is eXtreme Quality Control. Every machine and every step in a manufacturing process

is monitored in real time, implying more than 1 billion test results flowing into the X-DW per day. Sophisticated regression analysis can predict trends in component failures and product defects, while correlating across assembly lines, factory locations, part suppliers and the like.

In healthcare, an example is eXtreme Health Monitoring where vital signs for the majority of a population are recorded every minute and fed into a globally accessible data warehouse. Proactive intervention can be quickly initiated for a person with a critical health problem, and proactive maintenance of population segments can be established if long-term health signs are declining.

Another example is eXtreme Transport, in which all shipments of packaged goods are managed in a quick, reliable and inexpensive manner. Consider a system that spans 200 countries with a 24x365 operation using air, truck, rail and sea. One approach is to start with a very large shipping firm and scale by 10x for global reach. This results in 40 billion shipments per year, or approximately 100 million per day. Assuming an average of five days for delivery, the system must track approximately a half of a billion packages at any one time. Customer tracking of shipments will result in tens of millions of queries against the global data warehouse on a daily basis. Assuming that 2KB is captured at the origin, 1KB at each of 10 waypoints and 1KB at the destination, the system must collect 520TB per year with a data rate of 16MB per second. This implies that the system must handle 3PB for five years of raw data. Although these requirements are large numbers for current technology, such a system will soon be possible and even economical.

A final example is eXtreme Energy, where a global system manages electricity generation and distribution worldwide. The goals are to provide adequate energy for residences and industries into the future inexpensively, reliably and sustainably. Electricity currently constitutes 40 percent of global energy and will increase to 70 percent by 2050. Consumption is doubling roughly every 20 years. Using statistics from the European Union and EPRI, a rough estimation of requirements was made based on the year 2000. In that year, the world consumed 3 terawatts, requiring 13,300 generation plants, 39,900 substations and 26,600 high-voltage lines, totaling 80,000 objects to be managed. Assuming that 1GB is needed to characterize each plant and 1MB is needed for each substation and line, the requirements for a DW are 13TB for the primary grid.

The benefit of X-DW comes not from actually building such systems, but from probing the four big challenges with X-DW: business justification (Why would a business want such a system?), technical feasibility (Can we create/operate such a system economically?), organizational-political matrix (Who owns/controls the system?) and legal-ethical imperative (Should we do so?).

The really tough challenges are the last two, which can be referred to as "PAPA" challenges. This term comes from Richard Mason, who suggests examining the following four ethical issues: Privacy (personal identification and confidentiality), Accuracy (data quality and liability), Property (ownership and control) and Accessibility (haves and have-nots).²

Privacy deals with disseminating information relating to a specific individual. The Web is enabling the global distribution of personal information with a simple click. *Naked in Cyberspace* by Carole Lane outlines various techniques for harvesting that information but first explores the balance between the right to privacy versus the right to know. This balance is a continuing struggle. The criteria for personally identifying data are not as obvious as one would expect. In a DW, retrieval of certain fields may be obviously identifiable, but aggregations may also be with high probabilities. Progress is being made in the elements of a proper privacy policy, having a notice about what and how personal data is being used, a choice to opt-in and to opt-out, and a review of one's data for correction. Security is being recognized as the flip side of privacy, for without security, any privacy policy is inadequate.

Managing data quality is costly and involves a decision about how much accuracy is adequate in a certain business situation. Responsibility for the accuracy of individual data fields (e.g., person's last name is misspelled) receives considerable attention; however, the equally important accuracy of join paths (to integrate disparate sources) and reliability of analytics receive little attention. The legal liability and the ethical responsibility for data accuracy is a vague area. Who will be held accountable if incorrect data causes loss or injury?

Property implies treating information as property. In other words, who owns and controls the data? *Who Owns Information* by Anne Branscomb paints a complex and insightful look at this question. In X-DW where the data spans many jurisdictions, information may be subject to conflicting laws and regulations by various countries and regulatory agencies.

Finally, accessibility means the ability to use information for one's own purposes. Access to information is an allocation of power. Permission 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 future is today in each of our companies throughout the world. As DW professionals, we must think beyond the traditional boundaries, understand the implications and help others decide on the proper policies and directions. Technology is no longer the limiting factor for data warehousing. Increasingly, our actions impacting the PAPA issues will be ones that limit our use of DW technology. Hopefully, we will decide on the proper policies, avoiding ignorance and malice in the use of this technology.

References:

1. With Rosemary O'Mahony, global managing partner in Resources at Accenture.
2. Mason, Richard. "Four Ethical Issues of the Information Age." *Management Information Systems Quarterly*, Volume 10, Number 1, March, 1986: 5-12.

*Dr. Richard Hackathorn is president and founder of Bolder Technology, Inc., a 12-year-old consultancy in Boulder, Colorado. He has more than 30 years of experience in the IT industry and is a well-known technology innovator and international educator, conducting professional seminars in 18 countries. He has written three textbooks entitled **Enterprise Database Connectivity, Using the Data Warehouse** (with W.H. Inmon), and *Web Farming for the Data Warehouse*. Hackathorn earned his B.S. degree from the California Institute of Technology and his M.S. and Ph.D. degrees from the University of California, Irvine. To contact him, send e-mail to richardh@bolder.com.*

Copyright 2004, Thomson Media and DM Review.