Data Management: Lasting Impact of the Wild, Wild, Web

Reed M. Meseck
Consultant
Irvine, CA
(949) 854-2172
rmeseck@attglobal.net

ABSTRACT
This paper describes some of the ways the Internet and World Wide Web have affected databases and data warehousing and the lasting impact in these areas.

General Terms
Management, Performance, Design.

Keywords
Data Management, Data Warehousing, Database, Internet, WWW.

1. INTRODUCTION
The Internet has changed the time dimension and scale dimension of databases and data warehouses in a manner from which there will be little retreat. Two significant expectations have been set: for customers the expectation of immediacy, and for business and marketing executives the expectation of detail. The impact on databases is twofold: the frequency of transactions has increased and the number of updates or inserts has increased, as well.

2. THE TIME DIMENSION
A lasting influence of the Internet is the concept of "Net Time", a world in which timeliness is everything and where change is both rapid and constant. Where opportunity abounds, competition is fierce, and because of the fundamentally low barriers to entry the Internet presented in many areas, velocity and surefooted execution often became the key differentiators between winners and losers in the marketplace.

Another legacy of the Internet is that the life cycle of designs and software in general has become far more dynamic. Business models evolved rapidly and sometimes completely, leading to an environment where "legacy" data and applications were anything more than a week old. Often software was not modified but discarded and rewritten completely.

ACM SIGMOD 2001 May 21-24, Santa Barbara, California, USA

The challenge to data modelers and data warehouse designers is to produce designs that both meet the business requirements and which are malleable and can quickly adapt to changes in the business model. Database and data modeling tools vendors need to assist database architects and developers by providing tools that better support round-trip design, development and deployment.

2.1 Customer Expectations and Immediacy
Customers of web sites have become accustomed to the immediacy of the Internet. With a few clicks of the mouse you can order a book from Amazon.com [1] or check your account balance at WellsFargo.com [2]. If you have a question about your bank account at Wells Fargo at 2 AM, you can view the history of your account and see when a transaction cleared. These transactions occur with near instantaneous access. The customer has come to expect near instant access with up-to-date data across a variety of industries.

2.1.1 Banking ATMs and Customer Expectations
When banking ATMs first appeared on the scene many years ago, transactions performed at ATMs were not connected in real-time to the account databases, but were batch processed at the end of the day. Running update transactions against a "large" accounts database was unheard of. This worked and was acceptable to customers for a couple of reasons.

First, ATMs represented the same transaction and processing model as the teller inside the bank: walk inside and the balance from the teller was the same. Second, there was an "end of day" and thereby a batch processing window. The batch model could be understood by the customer and was tolerated.

2.1.2 Real-time Updates
The *24x7x365* global model of the Internet and the customer expectation of immediacy pushes us closer to where updates to a database occur in real-time or near-time, and where a batch processing window is narrow or nonexistent. While databases and systems have been evolving in this direction, the customer expectations fostered by the Internet make real-time or near-time updates a requirement. The architectural, operational and computational impact of this shift is non-trivial and the costs potentially substantial.

3. THE SCALE DIMENSION
Another lasting impact of the Internet is the volume, detail and frequency of data. A major website can record in excess of 500 million page events per day, providing significant information

about the site and its visitors. As Kimball [3] points out, "The Web presents us with a new and unprecedented data source."

At the macro level, one can obtain statistics about what types of pages were visited, how many of each page type were visited, etc. At the micro level, the objects served to assemble a page may be available as well. In addition, all of the pages a visitor accessed and the order in which they were visited may be available, as well.

3.1 Scale – But at What Cost?

Scale is also potentially unpredictable and more than one web site has been caught without adequate processing power in place. The Internet presents a unique elasticity with respect to scale since the number of users accessing a site at any moment is largely unpredictable, with the only guidance being prior site statistics.

We have witnessed that a major event, such as a press release, can cause spikes large enough to effectively deny access to a web site. The fear of such outages has led companies to stockpile capacity, perhaps to the point of excess.

Companies need to consider the ongoing cost of carrying excess capacity. Vendors, ISPs and researchers need to come up with innovative ways to provide incremental capacity on demand, to pool excess capacity in a shared reserve or to utilize the idle capacity for productive purposes.

3.2 The Problem with Detailed Data

Terabyte data warehouses has been in existence for some time; however, they have generally consisted of customer activity over some relatively long span of time, in most cases years and in others decades. The rate of growth is relatively slow.

By contrast, if data is collected and warehoused for a high volume web site at the detail level, the data warehouse expands to multi-terabyte proportions very quickly. For example, if the grain of the fact table is the page event, and the fact table is a modest 40 bytes wide, the fact table alone contributes 600 gigabytes of data to the database in one month.

For many corporations simple summarization of page events, page types and the like is adequate. However, for activities such as data mining or user behavior modeling, summary data is of limited use and so detailed data must be used. Once detailed data is required, the scale of the data increases tremendously.

3.2.1 Scalability Through Parallelism

Processing power, I/O bandwidth and storage all have to scale accordingly to meet the task, and as a result so does the cost. Parallel processing and clustered database environments quickly become necessary to handle the requirements of processing detailed data at the rate of 500M records per day.

A significant challenge is that while today's databases may leverage parallel processing, the processes that feed these very large databases often do not adequately leverage parallelism since this activity takes place outside of the database. Database vendors need to consider extending the benefits of parallel processing to tasks outside of the database or at the edge of the database.

3.3 Examining the Business Case

While the influences of the Internet may be lasting in terms of immediacy and detailed data, it is prudent to examine whether a business case exists to support such immediacy and detail. One could argue that the demise of the dotcoms [4] has proved otherwise.

While business owners may desire detailed data they must be prepared to justify the investment required to support such efforts and show an adequate return on that investment. Customers may demand instantaneous response and up-to-the-second data but may be unwilling to bear the cost.

Real-time updates against a primary production database may be technically possible, however, it may not be economically feasible. Likewise, processing and warehousing 500M records per day may not be cost effective. However, cost effective alternatives may exist which provide the desired results in an approximated manner.

3.3.1 Alternatives

If, for instance, it is desired to track the behaviors and nuances of a community of users over time, it may be possible to identify a representative set of users, collect, retain and process highly detailed data for that subset of users, and produce statistically accurate results.

With respect to immediacy, introducing a messaging layer may allow update requests to be processed in near-time as opposed to real-time and provide a more cost effective solution while still providing the illusion of immediacy to the customer. By using creative alternatives, we can provide solutions that meet the requirement and do so in a cost effective manner.

4. CONCLUSION

The Internet changed the time and scale dimensions of data management in significant ways, and the challenge is managing these dimensions. The dotcom era often produced solutions that worked where an unlimited budget was available. Perhaps the most valuable lesson to learn in the aftermath of the dotcom era is that being cost effective is as important in the solution as any technical criteria. Cost matters.

5. ACKNOWLEDGMENTS

I would like to thank the many individuals at IBM Silicon Valley Lab and Toronto Lab. My thanks also to Ralph Kimball for his pioneering work in the area of the Data Webhouse.

6. REFERENCES

[1] Amazon.com Home Page. http://www.amazon.com

[2] Wells Fargo Bank Home Page. http://www.wellsfargo.com

[3] Kimball, Ralph, and Merz, Richard. The Data Webhouse Toolkit. John Wiley& Sons, Inc. 2000.

[4] Dotcom Demise and Layoffs. http://www.idg.net/english/channel_content/channel_dotdemise_news.html.