



# DATABASE

## TRENDS AND APPLICATIONS

The Journal of Information Integration and Management • Published by Unisphere Media, L.L.C.

Volume 21, Number 9 • September 2007

September 2007 Issue

# Data as Demanded

**By Joe McKendrick**

A couple of decades back, when a company was preparing for a merger or acquisition, lawyers typically submitted a large file folder's worth of background documents for government agency review. Today, if most documents had not moved from paper to imaging and electronic documents, it would take truckloads to transport.

"These days, the requests are massive - a recent merger had about 4.6 million documents to be reviewed and analyzed," said

John Tredennick, CEO of Catalyst Repository Systems. Tredennick's company maintains an online repository in which relevant case documents are stored on an on-demand basis for law firms and corporate customers. The growth of such a service reflects the growing reliance on online data, as well as the growing inability for companies to manage these large data stores on their own. "Until recently, most corporations and law firms had their own systems for litigation support," he pointed out. "That

worked okay for lower volumes. But as the volumes started growing out of control, and

### THE JAZZ ABOUT DATA ON DEMAND

- ➔ Data on demand cuts to the bottom line.
- ➔ Queries are more complex and diverse.
- ➔ Multiple data sources are regularly accessed.

the switch went from just images to native files, suddenly, you need very big, very sophisticated systems to handle them. And they need to be delivered over the Web, to connect people all over the world to access these files."

Such is the fast-evolving nature of data on demand. Today's enterprise user is not only seeking relational transaction data reports, but also documents in a wide variety of formats and contexts. Queries are more diverse and more complicated. "There are many challenges associated with the

quest for on-demand data," Jonathan Wu, senior principal with HP's Information Management Practice, told DBTA. "They begin early on, when tough questions arise while assessing business requirements: What data is needed? When is it needed? What level of accuracy, or data quality, is expected? How will the data be used or presented?"

### Multiple Data Sources

Add the need for access to real-time or near real-time data, and things get even more complicated. A recent survey among 342 members of the Oracle Applications Users Group (OAUG), conducted by Unisphere Research in partnership with GoldenGate Software, revealed that companies are challenged to manage the growing volumes of data coming out of their applications. Close to half reported that typical reports require access to three or more applications - usually ERP and financial systems. At the bottom line "effective business operations are impossible without

*Data On Demand page 24*

## Data On Demand

*Continued from cover*

ready access to data across the enterprise,” Michael Lazar, technology director at GemStone Systems, told DBTA. “Business processes and users must have rapid access to a variety of data sources that aggregate supporting technologies such as databases, application servers, and data transport layers.”

The OAUG study found that most of the data demanded needs to be less than 24 hours old, which creates technical challenges in terms of integrating new data with archived sources. “Many current systems are unable to route relevant data to users in near-real-time based on their requirements,” Lazar said. They also “face additional obstacles as they attempt to correlate the various real-time information sources with historical data.”

HP’s Wu agreed that “the greatest technical challenge will be in the integration of the data from disparate sources, and then cleansing and transforming it in near real time. This challenge is reduced if the transformation and cleansing is minimal. However, integration of data from disparate systems is always challenging if the data cannot be easily linked together.”

## System Performance

Other technical challenges also surface, particularly in terms of systems performance - many existing infrastructures simply were not designed to handle the tremendous volumes of data flowing through corporate networks. For example, close to six out of 10 organizations in the OAUG survey reported suffering from server performance degradation as a result of end-users running reports against these live production environments. More than half of the organizations surveyed have data warehouses, but few have been able to deploy these environments to offload reporting requirements. “While this is a great idea in concept, many data warehouses and data marts have fallen short in reality,” Andy Palmer, CEO and founder for Vertica Systems, told DBTA. “Why? Because most are built on 30-year-old database technology originally built for business transaction processing - the primary goal of which was to get data into these systems quickly and reliably. The result is very poor performance for analytics, often at very high cost.”

Palmer advises looking into newer technologies such as massive parallelism using grid-based commodity hardware, which replaces more expensive, proprietary hardware with massive collections

of commodity PCs connected by standard networking technology. "This approach has been widely used at Google," he pointed out, adding, "database queries are easily parallelized on such hardware, resulting in a linear speedup as nodes are added." Other approaches include column-oriented database architectures, which enable SQL queries to access only the columns of data they need, versus every row of data - as in traditional row-store databases. Data compression is another option, Palmer added.

The greatest technical challenge is integrating data from disparate sources, and cleansing and transforming it in near real time.

Managing intense growth of data is difficult. For example, iBasis, a global VoIP company and one of the world's 10 largest carriers of international voice traffic, experienced a dramatic increase in data volumes in the back-office systems that run its business, with an annual run rate of more than nine billion minutes of international traffic and an annual growth of more than 45 percent. The company was challenged to accelerate data load cycles and optimize system performance to enhance both customer experience and profitability. "We have several groups with very different needs," Mark Saponar, vice president of information systems for iBasis, told DBTA. This includes external customers - telecommunications providers - that contract for the iBasis capabilities. Another is the market analysis group, and a third is the company's sales group.

### **Innovative Solutions**

The company maintained a 12-terabyte data warehouse to fulfill the various reporting needs of these groups, Saponar said. Recently, the company migrated from the warehouse environment to

offload queries to data warehouse appliance from Netezza. "The main reason for migration was we had a need for analyzing massive amounts of data. Every call generates five to six different transactions, which we're analyzing in order to determine quality, margins, and revenues. It is a massive compilation of analytics on data to determine what's the best route for traffic; what's the best quality; what's the best margins; and what's the best profits overall."

Catalyst, the case document management company, turned away from traditional database structures and implemented a combination of search technology from FAST and database technology to capture and deliver data to end-users. Database-based information "gets very expensive, particularly as the sizes get bigger, and you start hitting RAM limits. We hit the problem with 250,000 docs, which is nothing," said Tredennick. With a variety of file types being accessed, traditional databases could not handle the load in a speedy fashion. "When you start having fields, and text, or text and fields mixed up, performance degrades substantially - to a two-to-three-minute response time. That was unacceptable for us. We have users that put together 1,000, 1,500, or 2,000 queries with a wild mix of fields, proximities, date ranges, and text, in no particular order. We're talking about 20 million documents running these kinds of searches and bringing them back in under a second, that need to be delivered back in subsecond search response. The database just cannot do that yet."

For many companies, the ability to deliver data as rapidly and efficiently as possible to end-users that need it goes right to the bottom line. In the case of iBasis, for example, the deployment of its data warehouse appliances enables decision-makers to execute complex analyses against increasingly detailed data, which further enhances the company's management of call routing and rating to optimize gross profit. In addition, providing cost-efficient communications services represents a tremendous revenue growth opportunity, Saponar said.