



UBM
TechWeb

The road to cloud standards 16 | Browsers are still vulnerable 31
SLAs done right 34 | UC's elusive ROI 38 | Dr. Dobb's: More secure development 41

InformationWeek

THE BUSINESS VALUE OF TECHNOLOGY

ELECTRONICALLY REPRINTED FROM AUG. 9, 2010



You have terabytes aplenty. Now you need speed: faster queries, real-time insights. Here's how to get there. **p.22**

By Doug Henschen

ALSO

Bob Evans on why Larry Ellison will rock the Big Data world **p.6**

Art Wittmann on why Oracle's customers have their doubts **p.56**



Big Fast *and*

If you want to understand the challenges of the Big Data era, hang around Catalina Marketing, a global marketing firm that works with a who's who of consumer packaged goods companies and retailers.

Catalina's data warehousing environment shot past the petabyte mark seven years ago and today stands at 2.5 PB. Its single largest database contains three years' worth of purchase history for 195 million U.S. customer loyalty program members at supermarkets, pharmacies, and other retailers. With 600 billion rows of data in a single table, it's the largest loyalty database in the world, Catalina maintains.

At the cash registers of Catalina's retail customers, real-time analysis of that data triggers printouts of coupons that shoppers are handed with their receipt at checkout. Each coupon is unique—two shoppers checking out one after the other, with identical items in their carts, will get different coupons based on their buying histories, combined with third-party demographic data.

Few companies operate at Catalina's scale, but most every company is living in its own version of the Big Data era. Two forces define this era: size and speed. And those forces are driving companies to consider new choices for how they deal with data.

Size is relative—by some estimates, 90% of data warehouses hold less than 5 TB. But it's the pace of growth that has companies rethinking their options. Nearly half (46%) of organizations surveyed last year by the Data Warehousing Institute said they'll replace their primary data warehousing platform by 2012.

Speed is sometimes about pure performance, as in how quickly a system answers a query, but more important is the broad notion of "speed to insight." That's about how much time people—often statistician-analyst-type people—must spend loading data and tuning for performance. The pressure is on IT to get insights out of ever-larger data sets—faster.

This Big Data era got rolling way back in the dot-com days. Since then, a number of alternatives have emerged to challenge the conventional relational databases from Oracle, IBM, and Microsoft. Those options fall into two camps: systems supporting massively parallel processing (MPP), and those harnessing column-store databases.

Success in the Big Data era is about more than size. It's about getting insight from these huge data sets more quickly

By Doug Henschen

It's now almost a given that new deployments headed north of 10 TB will feature MPP, column-store architectures, or both. Oracle now delivers MPP through its Exadata V2 and IBM through its Smart Analytic System, both introduced last year. And Microsoft will join the MPP camp this fall with its Microsoft SQL Server Parallel Data Warehouse.

But the story doesn't end there. Practitioners handling complex analytics, really big data sets, real-time analysis, or all three will need much more. If you're using or expanding your use of advanced analytics such as predictive models, for example, you need to explore emerging options for in-database processing. The right choice could save your team countless hours of data extraction and prep work.

If swamped with truly massive data, consider powerful, Google-inspired query techniques such as MapReduce. Need to query huge stores of text and other nonrelational data? Open source options such as Hadoop might be the ticket to flexible-yet-affordable analysis. And if competition is driving your business toward real-time responses, in-memory analysis is without doubt in your future. As IT leaders from Catalina to Barnes & Noble to Cabela's show that standing still isn't an option in the Big Data era.

Catalina's data warehouse is accessed from more than 44,000 retail locations, and it collects data on more than 250 million transactions per week. So Big Data is its business. Catalina started building grid computers on its own, before it became the No. 3 customer of Netezza back in 2003. Since then, Netezza's MPP architecture and its proprietary data-filtering capabilities have helped Catalina keep pace with soaring data stores.

Why So Big?

Other companies have multiple warehouses that may not sound like Big Data

individually, but they sure do when added up. Consolidating them can yield analytic insights and system management efficiencies as well as cut hardware and software costs.

Bookseller Barnes & Noble has "dozens and dozens" of terabytes, says Marc Parrish, VP of retention and loyalty marketing, and until recently that data was spread across nine Oracle data warehouses. One warehouse handled point-of-sale data from 730 retail stores. Another handled 630 college bookstores. Another handled the Web site. And so on.

Yet one of the single most important insights Barnes & Noble needs, as e-books and e-readers take off, is how readers interact across those channels. It's no accident that its new CEO, William Lynch, ran the retailer's Web site before taking the helm.

Shortly after Barnes & Noble entered the e-reader market last year, with the Nook device and iPhone and Android e-readers, its executives backed an investment in a consolidated enterprise data warehouse. The company finished migrating to an Aster Data nCluster database, running on commodity MPP hardware, this spring and started using it for analysis within the last month. Parrish says Barnes & Noble already is doing a better job of cross-channel analysis, which was next to impossible with silos.

"Before, when somebody visited us online, we only knew about their online purchases," he says. "Now that all our data is in one place, we can understand their interactions across our entire ecosystem." Barnes & Noble gets better understanding of customer reading interests, as well as insight into the dynamics among e-reading, online activity, use of in-store cafes, and store purchases.

Some companies end up with Big Data because of fast growth or a need to dig into historical information. Hutchison 3G, a British mobile network operator with 6 million sub-

scribers, is dealing with both. Its customer ranks are growing, its call volumes are swelling as subscribers give up landlines, and mobile Web usage is soaring via smartphones.

Hutchison chose IBM's Smart Analytic System to take on both network performance optimization and customer behavior analysis. It cut dropped-call rates by spotting bad network interconnects and poor antenna performance. Hutchison is also looking at customer use of text, voice, and Web, plus calls to customer service. The usual business objectives apply: spotting cross-sell and up-sell opportunities, best customers, and customers likely to leave.

Hutchison's new warehouse holds 33 TB, including a full year of call detail records. The company expects to reach 60 TB within a year as it aggregates older data for historical trend analysis. With that, Hutchison expects to get "better detail on our subscriber segments and a clearer understanding of the complete life cycle of contracts," says Darren Silvester, Hutchison's information management architect.

Need For Speed

Companies increasingly expect to be able to do advanced analytics on Big Data, and they expect results fast.

Advanced analytics go hand-in-hand with data warehousing because you need lots of data to analyze customer behavior (for sales and marketing), understand risk (think finance), or optimize performance (think IT or IP networks). In many cases, massively parallel processing and column-store databases alone can't tackle those big analytic challenges.

The problem, until fairly recently, was that advanced analytic functions had to be handled outside the database, so MPP and columnar querying and compression couldn't speed things along. Procedures and models in SAS Institute's analytics software, for instance, are

typically developed and scored on the SAS platform. In that approach, data must be copied and moved from the data warehouse. Between disk-reading and network-bandwidth constraints, it takes a long time to get Big Data out of a warehouse. You also have to wait for the comparatively slow (flat-file-based) SAS platform to process. In a final step, the derived result from SAS usually must be loaded back into the data warehouse.

That's all changing thanks to in-database analytics. Lots of data warehousing vendors have added it to their feature list. The crucial question is whether they support the types of analytics you need. Most vendors now support a range of analytic queries that can be written in or converted to SQL. But a few vendors have also added support for running SAS procedures and models, and for analytics written in C/C++, Java, Python, Perl, R, and other languages inside their database.

Cabela's, a sporting goods direct marketer and superstore retailer, was an early advocate of in-database processing. When it moved from a conventional IBM DB2 data warehouse to a Teradata one in 2005, Cabela's used every trick available to handle SAS procedures inside the Teradata database. It used various optimization approaches whereby SAS procedures were converted to SQL.

"We had to retrain our statisticians to think outside of standard SAS and be more SQL-based, but they were flexible and made that transition," says Dean Wynkoop, Cabela's senior manager of data management and a member of the SAS/Teradata advisory council.

Now, thanks to its in-database approach, a reporting process that took four days of extraction, data prep, and processing in SAS takes about an hour inside Teradata. Translation: Cabela's needs only 1.5 full-time equivalents to handle a direct-mail campaign that previously required seven full-time statisti-

cians. "We wanted to make the most of these people as statisticians, not data management people handling data extracts and running jobs," Wynkoop says.

Rather than lay those people off, Cabela's has them doing analytics—not data handling—on sales forecasting, product allocation optimization, and even potential store locations. Next up: digging into Web clickstreams.

SAS and Teradata took the hint from Cabela's and other customers and co-developed tools, released in 2008 and since expanded, that offer a slightly different approach to in-database processing. Teradata adapted its database

Catalina Marketing

2.5 PB of storage

Challenge: Quickly sample its 195 million consumer data profiles for research

Solution: In-database modeling lets it do 10 times the number of models with the same staff

to run key SAS functions so statisticians can stick with the code they're accustomed to. A tool for building data sets lets statisticians build SAS models using sample data directly from Teradata. Statisticians use the SAS Scoring Accelerator for Teradata to score models within the database. Depending on the scale of the data, some procedures are said to run 40 to 50 times faster in Teradata than on the SAS platform.

Modeling On Steroids

There's a theme in that example: customers driving vendors into partnerships. Three years ago, Catalina was among the customers that got Netezza

and SAS working together on in-database processing. Statisticians typically sample 10% to 15% of a customer base to build a model. With a loyalty database of 195 million consumers, Catalina in some cases needs hundreds of terabytes just to build a model—more data than most companies have in their entire data warehouse.

The two vendors collaborated on a Scoring Accelerator for Netezza, which they introduced early this year and which Catalina adopted. "Before, we were lucky if we could develop 50 to 60 models per year," says Eric Williams, CIO and executive VP at Catalina. "Because of the in-database technology, we believe we'll be able to do 600 models per year with the same staff."

Teradata, Netezza, and Aster Data appear to have the most advanced efforts to support SAS within their databases, but SAS is also working with Hewlett-Packard, IBM, and Greenplum (recently acquired by EMC). IBM says its SPSS analytics platform supports in-database analytics by way of a server that can push SQL versions of SPSS code into DB2, Oracle, and Microsoft SQL Server.

SAS and SPSS are the analytics leaders, but practitioners need to run other forms of analytics more efficiently, such as open source R and applications written in other languages. Aster Data, Greenplum, Netezza, and others are making more non-SQL code understood and operable in their databases.

If you know what you need now or expect to use in the future, by all means call out those specifics in your RFP. Expect the list of in-database processing options to expand. If the data you're analyzing doesn't fit the SQL mold—often the case for text—consider options such as Hadoop.

In-Memory Is Your Future

One trend that will impact all data warehousing practitioners, sooner or later, is in-memory analysis. Disk drives

will become extinct, replaced by solid-state memory. With solid-state drives (SSD) as much as 150 times faster than spinning disks on sheer data input and output, the speed-of-analysis advantages that in-memory will deliver will be impossible to resist.

BNP Paribas, the French financial services giant, recently deployed Oracle's Exadata V2 appliance as the foundation for an analytic trading floor application that formerly ran on an Oracle 10g RAC deployment. Exadata V2 is one of a handful of platforms now available with solid-state memory. IBM, Kognitio, and Teradata, among others, have added SSD options.

Exadata, with flash memory from the Sun F5100 array, supports the high-speed end of BNP Paribas' data access scheme, which stores data based on high-, medium-, and low-speed needs. The most recent week's trading floor data is on the flash cache, as is any data that can be directly accessed at the presentation layer, including a query-oriented internal Web site and data accessed by its BusinessObjects tools.

"The Web site is now, conservatively, five times faster than it used to be, and we also have staging tables in flash that give us real-time application performance statistics," says Jim Duffy, BNP Paribas' data warehouse architect.

SAP was among the first to demonstrate the potential of in-memory analysis, offering its SAP Business Warehouse Accelerator more than three years ago. The derivative SAP BusinessObjects Explorer appliance, introduced last year, can tap multiple data sources beyond data warehouses.

SAP rightfully caused a stir this spring

when chairman Hasso Plattner and CTO Vishal Sikka outlined the company's long-term vision for handling transaction processing and data analysis with in-memory speeds on a single platform. That's different from Oracle Exadata V2; Exadata also supports both transaction processing and data warehousing but handles the two as distinct environments. The High-Performance Analytic Appliance that SAP is promising—perhaps by year's end—isn't expected to cross into transactions. If it delivers, it will be interesting to see just how much memory it brings to analytic processing.

SAS also is promising a product by year's end that could shake up this market. SAS plans to introduce in-memory applications for financial risk assessment, retail pricing, and product-mix applications, packaged to run on HP server blades. Such an app running on an in-memory platform would presumably outperform the same app supported by an MPP warehouse with in-database processing using old-fashioned spinning disks.

Catalina CIO Williams has tested a few of the latest in-memory options and found them four to eight times faster than platforms using fast-spinning disks. So why isn't everyone on in-memory? The cost, he says, is as much as 10 times higher than for conventional platforms.

Know Your Needs

We're years away from seeing large data warehouses routinely stored completely in memory. But for practitioners under pressure in this Big Data era—to handle huge scale, do advanced analytics, and deliver real-time insights—

the good news is those demands are sparking new technologies, new strategies, and lower costs.

With lowball pricing from the likes of Netezza now in the \$20,000-per-terabyte range, a deployment that would have busted budgets even five years ago might not even reach seven figures today. Among the few vendors that publicly reports such figures, Netezza's average deal size in the fourth quarter was \$1.1 million, and smaller deployments land in the \$300,000 range. Other vendors quote prices in the \$70,000- to \$100,000-per-terabyte range, and large data volumes are known to trigger steep discounts.

The most expensive deployment, though, is the one that doesn't meet your needs and has to be replaced. So know your requirements and long-range expectations. By all means, ask for a pilot test with your sample data and queries. Not all vendors will oblige. Others may charge fees applicable to a purchase. Another option might be testing the platform in the cloud or in a hosted environment.

The theme we keep hearing from leaders in this Big Data era is to explore new options. "My team goes out and takes a look at all new technologies and vendors," says Catalina's Williams. "We're going to stay where we're at for the next six to nine months, but we'll continue to make sure we have the best product for the price point."

Six to nine months. If you're not exploring your options that often, you're probably falling behind.

*Write to Doug Henschen at
dhenschen@techweb.com*