InformationWeek
THE BUSINESS VALUE OF TECHNOLOGY

3 Big Data Challenges: Expert Advice

The "big" part of big data doesn't tell the whole story. Let's talk volume, variety, and velocity of data--and how you can help your business make sense of all three.

By Doug Henschen, InformationWeek
October 18, 2011
URL: http://www.informationweek.com/news/software/bi/231900914

12 Top Big Data Analytics Players
(click image for larger view and for slideshow)
You don't need petabytes of information to play in the big data league. The low end of the threshold is more like 10 TB, and, in fact, "big" doesn't really tell the whole story. The many types of data and the speed at which data changes are, along with sheer volume, daunting challenges for businesses struggling to make sense of it all. Volume, variety, velocity--they're the hallmarks of the big data era we're now in.

Variety comes in the form of Web logs, wirelessly connected RFID sensors, unstructured textual information from social networks, and myriad other data types. Velocity breeds velocity. Fast-changing data drives demand for deep analytic insights delivered in hours, minutes, or, in extreme cases, seconds, instead of the weekly or monthly reports that once sufficed.

How are IT organizations coming to grips with data volume, variety, and velocity? Specialized databases and data warehouse appliances are part of the answer. Less heralded but also essential are information management tools and techniques for extracting, transforming, integrating, sorting, and manipulating data.

IT shops often break new ground with big data projects as new data sources emerge and they try unique ways of combining and putting them to use. Database and data management tools are evolving quickly to meet these needs, and some are blurring the line between row and column databases.

Even so, available products don't fill all the gaps companies encounter in managing big data. IT can't always turn to commercial products or established best-practices to solve big data problems. But pioneers are proving resourceful. They're figuring out how and when to apply different tools--from database appliances to NoSQL frameworks and other emerging information management techniques. The goal is to cope with data volume, velocity, and variety to not only prevent storage costs from getting out of control but, more importantly, get better insights faster.

Big data used to be the exclusive domain of corporate giants--Wall Street banks searching for trading trends or retailers like Wal-Mart tracking shipments and sales through their supply chains. Now the challenge of quickly analyzing massive amounts of information is going mainstream, and many of the technologies used by early adopters remain relevant. In the early 1980s, for instance, Teradata pioneered massively parallel processing (MPP), an approach now offered by IBM Netezza, EMC Greenplum, and others. MPP architectures spread the job of querying lots of data across tens, hundreds, or thousands of compute nodes. Thanks to Moore's Law, processing capacity has increased exponentially over the years, as cost per node has plummeted.

A second longstanding technique for analyzing big data is to query only selected attributes using a column-store database. Sybase IQ became the first commercially successful column-oriented database following its launch in 1996. Newcomers like the HP Vertica, Infobright, and ParAccel databases now exploit the same capability of letting you query only the columnar data attributes that are relevant--like all the ZIP codes, product SKUs, and transactions dates in the database. That could tell you what sold where during the last month without wading through all the other data that's stored row by row, such as customer name, address, and account number. Less data means faster results.

As an added bonus, because the data in columns is consistent, the compression engines built into column-store databases do a great job--one ZIP code, date, or SKU number looks like any other. That helps column stores achieve 30-to-1 or 40-to-1 compression, depending on the data, while row-store databases (EMC Greenplum, IBM Netezza, Teradata) average 4-to-1 compression. Higher compression means lower storage costs.

One big change, under way for several years, is that the boundaries between MPP, row-store, and column-store databases are blurring. In 2005, Vertica (acquired this year by HP) and ParAccel introduced products that blend column-store databases with support for MPP, bringing two scalability and query-speeding technologies to bear. And in 2008, Oracle launched its Exadata appliance, which introduced Hybrid Columnar Compression to its row-store database. The feature doesn't support selective columnar querying, but, as the name suggests, it does offer some of the compression benefits of a column-store database, squeezing data at a 10-to-1 ratio, on average.

In the most recent category-blurring development, Teradata said in late September that its upcoming Teradata 14 database will support both row-store and column-oriented approaches. IT teams will have to decide which data they'll organize in which way, with row-store likely prevailing for all-purpose data warehouse use, and column-store for targeted, data-mart-like analyses. EMC Greenplum and Aster Data, now owned by Teradata, have also recently blended row-store and column-store capabilities. The combination promises both the fast, selective-querying capabilities and compression advantages of columnar databases and the versatility of row-store databases, which can query any number of attributes and are usually the choice for enterprise data warehouses.

Any time data volumes start getting really big, compression is key because it saves on storage, which remains a significant component of data management expense, despite the continuing decline of hardware costs when measured by the terabyte.

Consider Polk, a household name in the auto industry, which sells online subscriptions to data about vehicle sales and ownership to automakers, part suppliers, dealers, advertising firms, and insurance companies. Polk surpassed 46 TB of storage last year before it launched a long-term upgrade to eventually move from conventional Oracle RAC (clustered) database deployments to an Oracle Exadata appliance. As of this summer, the project was halfway done, and databases that formerly held about 22 TB had been compressed down to about 13 TB, using Oracle's Hybrid Columnar Compression.

Polk's Exadata migration is still in progress, but to date it has consolidated nine databases down to four and eliminated eight of 22 production database servers. The cost of a new Exadata deployment is about $22,000 to $26,000 per terabyte, before discounts, according to independent analyst Curt Monash. If Polk's storage efficiencies hold up through the rest of the project, 46 TB will be trimmed to about 28 TB. Using Monash's estimate, the 18-TB difference could trim the deployment's cost by as much as $400,000.

As companies begin managing big data, crafty IT pros are finding that some old tricks have renewed value when applied to such large volumes. Steps to improve compression and query performance that may not have seemed worth the effort can become more valuable. This is where technical capabilities end, and experience and ingenuity begin.

Wise use of best practices like data sorting lets companies improve the performance, and prolong use, of their current database platforms. And, for those that do move up to one of the latest and greatest big-data platforms, data management discipline instilled from the start will optimize performance and prolong the life of that investment. Sorting, for example, is a relatively easy way to optimize compression. Just as consistency of columnar data aids compression, sorting brings order to data before it's loaded into a database; that makes it easier for compression engines to do their work.

ComScore, the digital-media measurement company, has been using tricks like sorting since its first deployment back in 2000. Sybase IQ has been the company's primary database platform from the beginning, and it makes the most of the product's selective querying and compression capabilities. But with more than 56 TB in its store, ComScore also applies techniques such as sorting to help the database platform do a better job.

ComScore uses Syncsort's DMExpress data-integration software to sort data alphanumerically before loading it into Sybase IQ. While 10 bytes of the raw clickstream data that ComScore typically examines can be compressed to 3 or 4 bytes by Sybase IQ, 10 bytes of sorted clickstream data can often be crunched down to 1 byte, according to ComScore CTO Michael Brown.

Sorting also can streamline processing, and that improves speed as well as lowering storage costs. For example, ComScore sorts URL data to minimize how often the system has to look up the taxonomy that describes, say, ESPN.com as a sports site, Ford.com as an auto site, Google News as a news site, and Facebook as a social network. Think of someone who spends a Sunday afternoon bouncing across those sites, checking scores, reading news, browsing for a car, and posting on Facebook.

Instead of loading the URLs visited during that Web session in the order they were visited, possibly triggering a dozen or more site lookups, sorted data would lump all visits to the same sites together, triggering just four lookups. "That saves a lot of CPU time and a lot of effort," Brown says.

Polk also relies on sorting to cut processing time in a slightly different way. The Oracle database has built-in indexing capabilities that can help improve query performance, but the feature may not work if it can't spot obvious groupings of data. Sorting helps Polk force indexing to happen the way it's most useful. "If you can lump the data that goes together, the index knows exactly where to find the data you're after," says Doug Miller, Polk's director of database development and operations.

Polk subscribers often do queries by region, so the company sorts auto sales data by ZIP code. If a car manufacturer wants to know which models were the best sellers in Seattle last month, the database knows just where to find that data and won't waste time querying data tied to nonrelevant ZIP codes.

Polk is also making extensive use of "materialized views," which effectively store often-requested query results for rapid recall. Exadata's compression has helped reduce the size of materialized views, which lets Polk do more sophisticated analyses, since it can hold more views in cache and thus speed up performance when exploring multiple dimensions of data.

"If a customer wanted to look across multiple dealer zones and then start bringing in customer demographics and comparing lease transactions versus purchases, that would have taken as long as two to five minutes in the old environment," Miller says. "In Exadata, these sorts of queries are running in 10 seconds."

The critical point is that these tricks for managing data volume are about more than cutting storage costs. Getting faster, more relevant insight is really the name of the game with big data.

Data variety is such a key part of big data it has spawned an entire computing movement--NoSQL. While the name suggests an all-or-nothing showdown, think of it as "not only SQL." The movement is about alternatives when you don't have conventional structured data that fits neatly into the columns and rows of relational databases such as Greenplum, IBM DB2 or Netezza, Microsoft SQL Server, MySQL, Oracle, or Teradata. NoSQL databases can handle semistructured data or inconsistent, sparse data. That accounts for a lot of data growth coming from sources such as Web log files used by Internet marketers, remote sensor data like that used in emerging smart-meter utility applications, or security log files used to detect and thwart hacking and identity theft.

Some companies are also processing unstructured information such as text comments from Facebook and Twitter, mining the data for customer-sentiment analysis.

More than a dozen, mostly open-source, products are associated with the NoSQL movement, including Cassandra, CouchDB, Membase, and MongoDB. But the one getting the most attention is Hadoop.

Hadoop is a collection of open-source, distributed data-processing components for storing and managing large volumes of structured, unstructured, or semistructured data. Clickstream and social media applications are driving much of the demand, and of particular interest is MapReduce, a data-processing approach supported on Hadoop (as well as in a few other environments) that's ideal for processing big volumes of these relatively new data types. MapReduce breaks a big data problem into subproblems, distributes those onto hundreds or thousands of processing nodes on commodity hardware, then combines the results for an answer or a smaller data set that's easier to analyze.

Internet marketers and e-commerce players were the first to recognize the importance of clickstream data and social media sentiment, but it's now rare to find a company with a prominent brand that isn't paying close attention to online sales and marketing trends and trying to gauge the social media buzz around its products, categories, and company. Whether consumer products companies like Procter & Gamble, automakers like Ford and Toyota, or clothing manufacturers like Levi Strauss, they're analyzing where Internet users spend time, spotting what marketing messages draw their attention, and gauging their mood. They're mining this data to predict product demand, sense new-product acceptance, gauge competitive threats, and detect threats to brand reputations.

Hadoop runs on low-cost commodity hardware, and it scales up into the petabyte range at a fraction of the cost of commercial storage and data-processing alternatives. That has made it a staple at Internet leaders including AOL, eHarmony, eBay, Facebook, Twitter, and Netflix. But more conventional companies coping with big data, like JPMorgan Chase, are embracing the platform.

Data provider Infochimps relies on Hadoop to parse data from Facebook, Twitter, and other social sources and create new data sources. Infochimps' popular "Twitter Census: Trst Rank," for example, provides metrics on the influence of Twitter users. This helps companies with a Twitter presence gauge their followers' clout based on how many other Twitter users they interact with and how many people pay attention to them. This, in turn, helps these organizations know what their most influential customers are saying about their brands and products.

Why is Hadoop processing necessary? First, the data being studied is often semistructured or text-centric information, and second, it's really big. Infochimps has been collecting Twitter data since 2008, and the entire set includes nearly 7 billion tweets and more than 1 billion connections among users.

Hadoop is attracting attention from commercial vendors. In May, EMC Greenplum announced its own distributions of Hadoop software (one open-source and a commercially supported enterprise edition). And in September EMC added a modular version of its Greenplum Data Computing Appliance that will let IT organizations run the Greenplum relational database and Hadoop on the same appliance, albeit with separate processing and storage capacity dedicated to each environment. This is a stab at making it easier to address the variety of information that's part of big data by supporting it all on a single computing platform.

Another vendor bridging the SQL and NoSQL worlds is Aster Data. Acquired last year by Teradata, Aster Data is best known for supporting MapReduce processing within its relational database. This makes MapReduce accessible to SQL-literate analysts so they can do pattern detection, graph analysis, and time-series analysis on clickstreams and social media data.

IBM has embraced Hadoop by way of InfoSphere BigInsights, an analytics platform based on Hadoop. The company has also exploited Hadoop in internal projects and development work, most notably in the development of its Jeopardy-playing Watson supercomputer.

Oracle and Microsoft are the latest vendors to join the Hadoop bandwagon. Oracle confirmed at Oracle OpenWorld in October that it will introduce a Hadoop software release and related Big Data appliance, though it didn't say when. As recently as last year, company executives had told financial analysts that mainstream commercial customers weren't asking for unstructured-data-analysis capabilities.

Microsoft announced in October that it will introduce a beta Hadoop processing service through its Azure cloud by year end. And in 2012 it's promising an open-source-compatible software distribution that will run Hadoop on Windows servers.

For now, many practitioners and vendors are content to let SQL and NoSQL systems coexist. Most data warehousing platforms and many business intelligence suites now offer integration with Hadoop. So practitioners can do their large-scale MapReduce or data-transformation work in Hadoop, then move result sets into more familiar and accessible data warehousing and BI tools. An Internet marketing firm might use MapReduce to spot Web sessions relevant to an ad campaign from huge volumes of clickstream data, then bring that result set into an SQL environment for segmentation or predictive analysis.

Online retailer Ideeli is applying this blended approach, using Hadoop to store and process large volumes of Web log clickstream and email campaign data and using Pentaho for BI.

The company sets up members-only "flash sale" sites where it sells small quantities of high-fashion items, fueled by email and social media promotions. The sales typically last a day or two before the inventory is gone, and the boutique is taken offline. Ideeli studies Web traffic to understand which of its 5 million members are responding to a campaign, the traits of lookers versus buyers, and so on.

The trouble with an all-Hadoop approach, Ideeli found, was that Apache Hive--the data summarization, query, and analysis tool that runs on top of Hadoop--was too slow, taking several minutes to handle demanding queries, says Paul Zanis, director of data services at Ideeli. The choice of Pentaho for BI is perhaps no surprise, given that Pentaho has support for Hadoop, including the ability to design MapReduce jobs, extract data from Hadoop, and support scheduled reporting and ad hoc analysis from Hadoop tools.

Ideeli is still building the data warehouse it needs to support the new approach, but the idea is to use Pentaho's data-integration software to extract and transform end-of-day batch loads of clickstream and campaign data. From there, Pentaho's OLAP capabilities will automatically generate new cubes for rapid analysis.

"Once that's in place, we'll be able to explore high-level, summarized data within seconds versus trying to run a Hive query, which would take several minutes," Zanis says.

But there's a limitation today on Hadoop and other NoSQL environments: scarce expertise. Schools, vendors, and companies have spent decades teaching SQL, but Hadoop software distributions have only been available since 2009. Efforts like EMC's Hadoop initiatives are aimed at making it easier to deploy and manage big-data-oriented relational and Hadoop environments side by side, but you'll still need Hadoop expertise to deploy and manage that separate environment. Until these platforms gain larger pools of expertise, data management pros will have to find ways to deliver results through fast and familiar tools.

The velocity aspect of big data is tied to growing demand for fast insights. That's relative, of course, but according to The Data Warehousing Institute's Big Data Analytics survey, released in September, 13% of analyses are now rerun or rescored "every few hours" or in real time, compared with 35% monthly, 14% weekly, and 24% daily.

There are many examples of data that might demand analysis in real time or near real time, or at least in less than a day. RFID sensor data and GPS spatial data show up in time-sensitive transportation logistics. Fast-moving financial trading data feeds fraud-detection and risk assessments. Marketing analyses, too, are increasingly time sensitive, with companies trying to cross-sell and up-sell while they have a customer's attention.

Combine marketing with mobile delivery, and you've entered the fast-moving domain of Bango Analytics. Bango started out in the mobile payment business, but it discovered some companies were setting the price to zero on its tools, simply to track access to mobile content. So two-and-a-half years ago it started a separate Bango Analytics service promising near-real-time insight.

Bango measures traffic to mobile Websites and ads and the use of mobile apps. The content might be articles in the case of media sites or storefront pages in the case of online retailers. Bango's custom tracking app runs on Microsoft SQL Server, so the company uses SQL Server Integration Services (SSIS) as the workflow engine for overall integration, starting with extraction. It applies transformations, rules, and precalculations using queries and scripts written in SQL, a step that minimizes processing demands in the data warehouse environment. SSIS puts the final results into CSV text files and loads the data into an Infobright-powered data mart.

To keep its transactions running smoothly, Bango copies transactional information into an operational data store on a different tier of servers, and SSIS extracts the data from there. To get up-to-date information into the company's Infobright analytic data store as quickly as possible, it keeps batches small, so it only takes five to six minutes from a click on a mobile device until the interaction is made available in Infobright. Reports and dashboards are delivered to customers through a Web interface. Near-real-time data is "a key selling point of our product, so the faster we can load, the better," says Tim Moss, Bango's chief data officer.

Bango processes billions of records each month, says Moss, yet the company's total 13-month data store has yet to break into the double-digit terabyte range. Here again, the compression supported by Infobright's column-oriented database helps keep the storage footprint small. Nonetheless, Moss says Bango is preparing for high-scale and high-velocity demands, knowing that the amount of mobile content and the number of mobile campaigns and apps will grow, and with it, pressure to show what's catching on.

Marketers have spoken on this point. The weekly reports that were good enough two years ago must now be generated daily, and that's led to the development of near-real-time dashboards. This trend has made once-exotic information management techniques such as micro-batch loading and change-data-capture much more common.

Taken together, volume, variety, and velocity are emerging as the three-headed beast that must be tamed as IT teams look to turn big data from a challenge into an opportunity. But old demons haven't gone away. Complex data such as supply chain records or geospatial information can prove to be more of a bottleneck than varied data. And a large number of users (1,000 plus), many queries, or complex queries that call on multiple attributes or need complex calculations--all can lead to performance problems. Fail to anticipate demands along any one of these dimensions, and you may outgrow your data warehousing platform much sooner than expected.

The technology for handling big data will get better. In terms of the information management that must be done before the data gets into the warehouse, we're still in the early days. Expect to see new tools, services, and best-practices developed to address the thorniest problems. Of course, your business partners want results now, so draw on your experience and exploit the tools you know. But it's also time to begin experimenting with new approaches. Big data isn't getting any smaller.