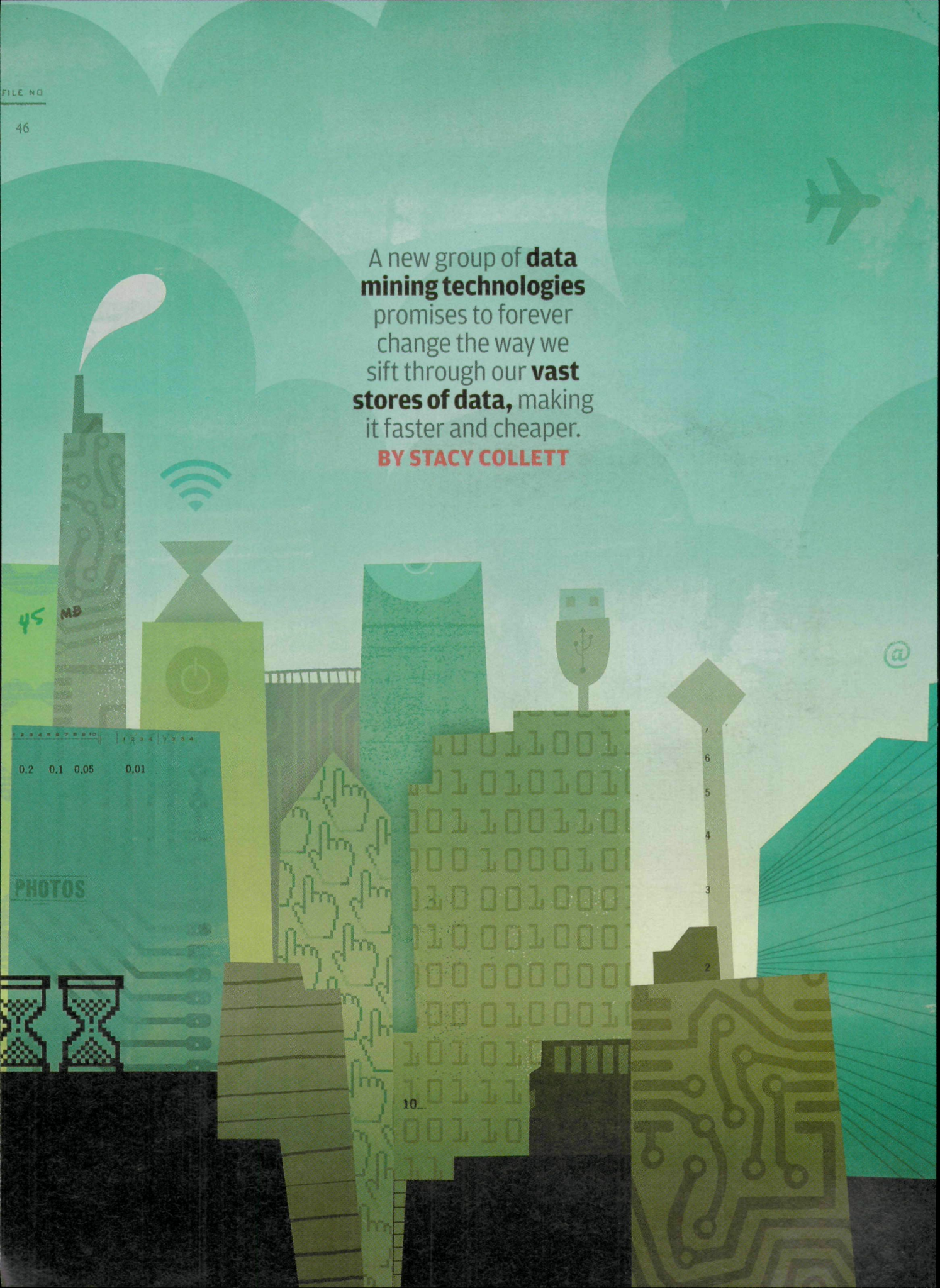


A new group of **data mining technologies** promises to forever change the way we sift through our **vast stores of data,** making it faster and cheaper.

BY STACY COLLETT

DANTE TERZIGNI

WHY BIG DATA IS A BIG DEAL

WE'VE ALL HEARD the predictions: By 2020, the quantity of electronically stored data will reach 35 trillion gigabytes, a forty-four-fold increase from 2009. We had already reached 1.2 million petabytes, or 1.2 zettabytes, by the end of 2010, according to IDC. That's enough data to fill a stack of DVDs reaching from the Earth to the moon and back — about 240,000 miles each way.

3 BIG DATA MYTHS

There's a good deal of confusion over what big data is and what it can do. Here are three myths about big data:

DANTE TERZIGNI

1 Relational databases can't scale to very large volumes and therefore can't be considered big data technologies. (Not true.)

2 Hadoop or, by extension, any MapReduce environment is the best choice for big data regardless of the workload or use case. (Also not true.)

3 The era of schematic database management systems is over. Schema development only gets in the way of big data deployment. (Laughably untrue.)

SOURCE: IDC, "THE BIG DEAL ABOUT BIG DATA," FEBRUARY 2011 (CARL W. OLOFSON)

For alarmists, this is an ominous data storage doomsday forecast. For opportunists, it's an information gold mine whose riches will be increasingly easy to excavate as technology advances.

Enter "big data," a nascent group of data mining technologies that are making the storage, manipulation and analysis of reams of data cheaper and faster than ever. Once relegated to the supercomputing environment, big data technology is becoming available to the enterprise masses — and it is changing the way many industries do business.

Computerworld defines big data as the mining of huge sets of structured and unstructured data for useful insights using non-traditional data-sifting tools, including but not limited to Hadoop.

Like the cloud, big data has been the subject of much hype and a lot of uncertainty. We asked analysts and big data enthusiasts to explain what it is and isn't, as well as what big data means to the future of data mining.

Setting the Stage for Big Data

Big data for the enterprise has emerged thanks in part to the lower cost of computing power and the fact that the systems are able to perform multiprocessing. Main memory costs have also dropped, and companies can process more data "in memory" than ever before. What's more, it's easier to link computers into server clusters. Those three factors combined have created big data, says Carl Olofson, a database management analyst at IDC.

"We can not only do those things well, but do them affordably," he says. "Some of the big supercomputers of the past involved heavy multiprocessing of systems that were linked together into tightly knit clusters, but at the cost of hundreds of thousands of dollars or more because they were specialized hardware. Now we can achieve those kinds of configurations with commodity hardware. That's what has helped us be able to process more data faster and more cheaply."

Not every company with vast data warehouses can say it's using big data technology. To qualify as big data, IDC says, the technology must first be affordable, and then meet two out of the three criteria that IBM describes as the three V's: variety, volume and velocity.

Variety means data comes in structured and unstructured forms. Volume means the amount of data being gathered and analyzed is very large. And velocity refers to the speed at which the data is processed. It "isn't always hundreds of terabytes," Olofson says. "Depending on the use case, a few hundred gigabytes could be quite large because of the third dimension, which is speed or time. If I can perform an analytic process against 300GB in a second, and it used to take an hour, that greatly changes what I can do with the results, so it adds value. Big data is the affordable application of at least two out of three of those."

The Open-Source Connection

"A lot of people consider Hadoop and big data to be synonyms. That's a mistake," Olofson says. Some implementations of Teradata, MySQL and "clever clustering technologies" that don't use Hadoop can also be considered big data, he explains.

Hadoop, an application environment for big data, has drawn the most attention because it's based on MapReduce, an approach common in supercomputing circles but simplified and made elegant by a project largely funded by Google. Hadoop is the predominant implementation of a mix of closely related Apache projects, including the HBase database found in the MapReduce environment.

Software developers have responded by coming up with all

Continued on page 22

Continued from page 20

kinds of techniques to exploit Hadoop and similar advanced technologies — many of them developed in open-source communities. "They've created a dizzying variety of so-called noSQL databases, which are mostly key-value paired databases that optimize either on throughput or variety or size with various techniques," says Olofson.

The open-source technologies aren't commercially supported, "so those things are going to have to evolve for a while and will have to shake out, which could take several years. That's the nascent aspect of big data that won't come to fruition for a while" in the general marketplace, he adds. In the meantime, IDC expects at least three commercial vendors to offer some type of support services for Hadoop by year's end. Also, several vendors, such as Datameer, will release analytics tools with Hadoop components that let enterprises develop their own applications. Cloudera and Tableau already use Hadoop in their offerings.

Upgraded RDBMS

Industry-watchers disagree on whether upgraded relational database management systems should also be considered big data technology. "I think it satisfies the criteria of faster, bigger, cheaper," Olofson says. Teradata, for instance, has made its system more affordable, and it's a scalable, clustered environment, he adds.

But others disagree. "The processing that you ordinarily do using an RDBMS in general, using standard BI tools — that's not really big data," says Marcus Collins, a data management analyst at Gartner. "That processing has been around for a long time."

So, who is really doing big data analytics?

A year ago, the primary users of big data technology were large Web companies, such as Facebook and Yahoo, that wanted to analyze clickstream data. But today, "it's moving outside the main Web properties into just about any company that you can think of that's got large volumes of data," Collins says. Banks, utilities, the intelligence community — all of them are jumping on the big data bandwagon.

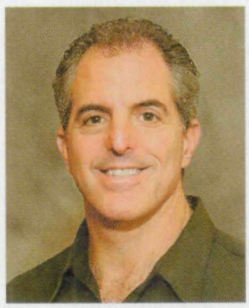
Some of the technologies are actively being used by people who are on the bleeding edge and need the technology now, like those involved in creating Web-based services that are driven by social media. They're also heavily contributing to these projects.

In other industries, businesses are realizing that much more of their value proposition is information-based than they had previously thought, so they'll likely become big users of big data technologies before long, Olofson says. Couple that with affordable hardware and software, and enterprises find themselves in a perfect storm of business transformation opportunities.

New York-based TRA helps organizations measure the value of TV advertising by matching the advertisements received in a given home via TVs and DVRs with buying behavior at the retail checkout counter. The company gathers data from cable provider DVRs and grocery store loyalty card programs to make these correlations. TRA's big data system processes reams of data that represents the second-by-second viewing habits of 1.7 million households — a feat that would have been impossible without big data technology. It deployed Kognitio's WX2 database, which allows the company to load, profile and analyze data quickly, collect granular ad-viewing information from DVRs, integrate it with detailed point-of-sale data, and produce customized reports.

"Kognitio has an in-memory solution, so a full half of our current entire database can be in memory, which means the response time when a customer of ours runs a query literally can be seconds as opposed to hours and days," says TRA's CEO, Mark Lieberman.

> **[We're] giving advertisers more confidence that TV is a good place to advertise. That's the big step — and it's all about big data analytics.**
>
> **MARK LIEBERMAN,** CEO, TRA



The database runs on commodity hardware, and TRA uses its own front-end application built in .Net Visual Studio. "We still use a little bit of MySQL, and the user interface was developed with DevExpress," Lieberman adds.

He says that big data has the potential to revolutionize the $70 billion TV ad buying business. Traditional methods of measuring viewership required installing special set-top boxes in a sampling of as few as 20,000 households nationwide. Today, data can be analyzed in detail from 2.5 million DVR and cable boxes.

"We're injecting accountability into that $70 billion — giving advertisers more confidence that TV is a good place to advertise," Lieberman says. "That's the big step, and it's all about big data analytics."

Greg Belkin, an analyst at Aberdeen Group, says the tools used by TRA and others have the requisite velocity, volume and variety to be labeled big data. "This is very poignant in retail because you have a lot of exploding sources of data that haven't traditionally been mined," such as social media sites, DVR boxes and grocery store loyalty card data, Belkin says. "It's data that is so enormous and complex that it can't be analyzed using traditional database methods, so retailers are turning toward these big data platforms."

Similarly, big data technology has revolutionized business at Catalina Marketing. The St. Petersburg, Fla.-based company runs a huge customer loyalty database of 2.5 petabytes that has years of purchasing history data for more than 190 million U.S. grocery shoppers. Its largest single database houses a staggering 425 billion rows of data, and the company manages 625 million rows each day in that one database.

By analyzing the data, Catalina helps major consumer goods manufacturers and large supermarket chains predict what customers are likely to buy and who will be interested in new products.

"We wanted to bring the technology to the data and not the

Continued on page 24

Continued from page 22

data to the technology," says Eric Williams, executive vice president and CIO at Catalina. "The technology exists now that allows companies like SAS to move their [analytics] technology into the database. That has exponentially changed the entire corporation. We were doing these things before but had serious limitations that would not allow us to get where we wanted to go. We had to use homegrown tools, and they were very rudimentary in what they could accomplish. Bringing big data technology to the forefront has changed our entire organization."

In addition to some open-source software in its proprietary systems, Catalina uses SAS Analytics on a Netezza data warehousing appliance platform.

Companies are "developing technology to operate on generic, Intel-based hardware, which makes it possible to operate secondary and tertiary products — like SAS Analytics' scoring solution — directly on the Netezza [software] that's running the database," Williams says. "Being able to take that technology and operate it directly on the database meant that Catalina could speed up our data mining solutions from weeks to a matter of hours."

Big data is fundamentally changing the way Bank of America does business, according to Abhishek Mehta, formerly Bank of America's managing director for big data and analytics, who spoke at last year's Hadoop World. "I look at Hadoop today as what Linux was 20 years ago. We all have seen what Linux has done in the enterprise software space. It has been massively disruptive. Hadoop will do the same. It's not a question of if, but a question of when."

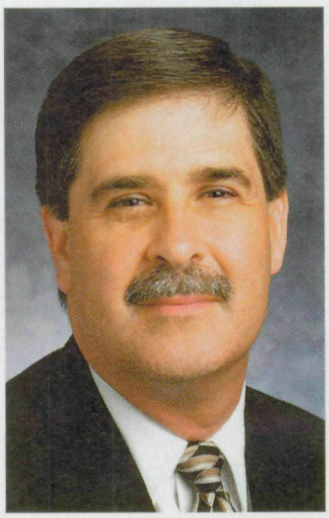
Beyond clickstream and transaction analysis, Hadoop allows Bank of America to quickly solve business problems. "Now, as a bank, I can think of eliminating fraud," Mehta says. "Now I can build a model looking at every incidence of fraud going back five years for every single person, rather than sampling it now, building a model, realizing there is an outlier that breaks the model, and then rebuilding the model. Those days are over."

The utilities industry is just beginning to understand the vast amounts of data at its fingertips and the value it holds. One Midwest utility uses Hadoop to analyze input from its "smart meters," which are primarily used to automate the billing process, but which also collect information on any fluctuations in amperage on the line. "If you collect this information and look for patterns, you can identify where a transformer is going to fail before it fails," IDC's Olofson says. "Or if a power outage happens, it causes fluctuations [in amperage], and they can detect [the outage]."

Down the road, he foresees utilities using big data to improve service to customers and to reduce operational costs through electrical grid monitoring, problem detection and the ability to do micro-adjustments against the grid — but it may require significant upgrades to the aging infrastructure.

Brand marketers are experimenting with Hadoop for "sentiment analysis" in social media. There are emerging service providers that use Hadoop to sift through Twitter on behalf of clients to discover what tweeters are saying and thinking about specific products.

> **We wanted to bring the technology to the data and not the data to the technology.**
>
> **ERIC WILLIAMS,** EXECUTIVE VP AND CIO, CATALINA MARKETING



Proceed With Caution

Big data technology is evolving rapidly. The companies that use it have IT staffs that are exceptionally tech-savvy and can adapt to changes in the technology and their companies' requirements.

"If you're not in a position to do that, then work with a service provider — maybe a cloud service — or wait until these things have reached a point where there's a number of established software products and services that are supported," Olofson suggests. "You'll have something that your business people will understand."

No doubt, data mining has changed forever. But analysts say that big data technology won't completely replace today's data warehouse and data mining tools.

"Today, data mining is really about building relatively sophisticated models with not very much data," Gartner's Collins says. "Now, big data gives you huge volumes of data — so it could well be that you don't need as sophisticated a model anymore. That may [mean] a shift in the way data mining is done."

"My view is that it will actually augment [the data warehouse market]," Olofson says. "They'll use a technology like MapReduce, whether Hadoop or some other commercial augmentation, to generate interesting business intelligence data they never could have gotten at before. Then, in order to reuse it and track historical patterns, they will put it in the data warehouse and actually expand its use."

Scale represents another challenge, Collins says, along with "the fact that there are not established architectural patterns on deploying and using this technology. We're learning as we go."

Some of the challenges are dissipating with the arrival of prepackaged tools, but the technology is still very much a programming interface — which is a step backward for BI, Collins says. For instance, he says, "Hadoop is a pretty techie system, and the drive in business intelligence has been to push it down in the enterprise and onto the desktop with a very user-friendly interface. We've taken a step back with Hadoop, but new vendors will help put it back into the user community where it needs to be.

"[Big data technology] needs to leap out of being in IT, and we have to put tools in the hands of users" in the business units, Collins adds. "That hasn't happened yet." ◆

Collett *is a* Computerworld *contributing writer. You can contact her at stcollett@aol.com.*