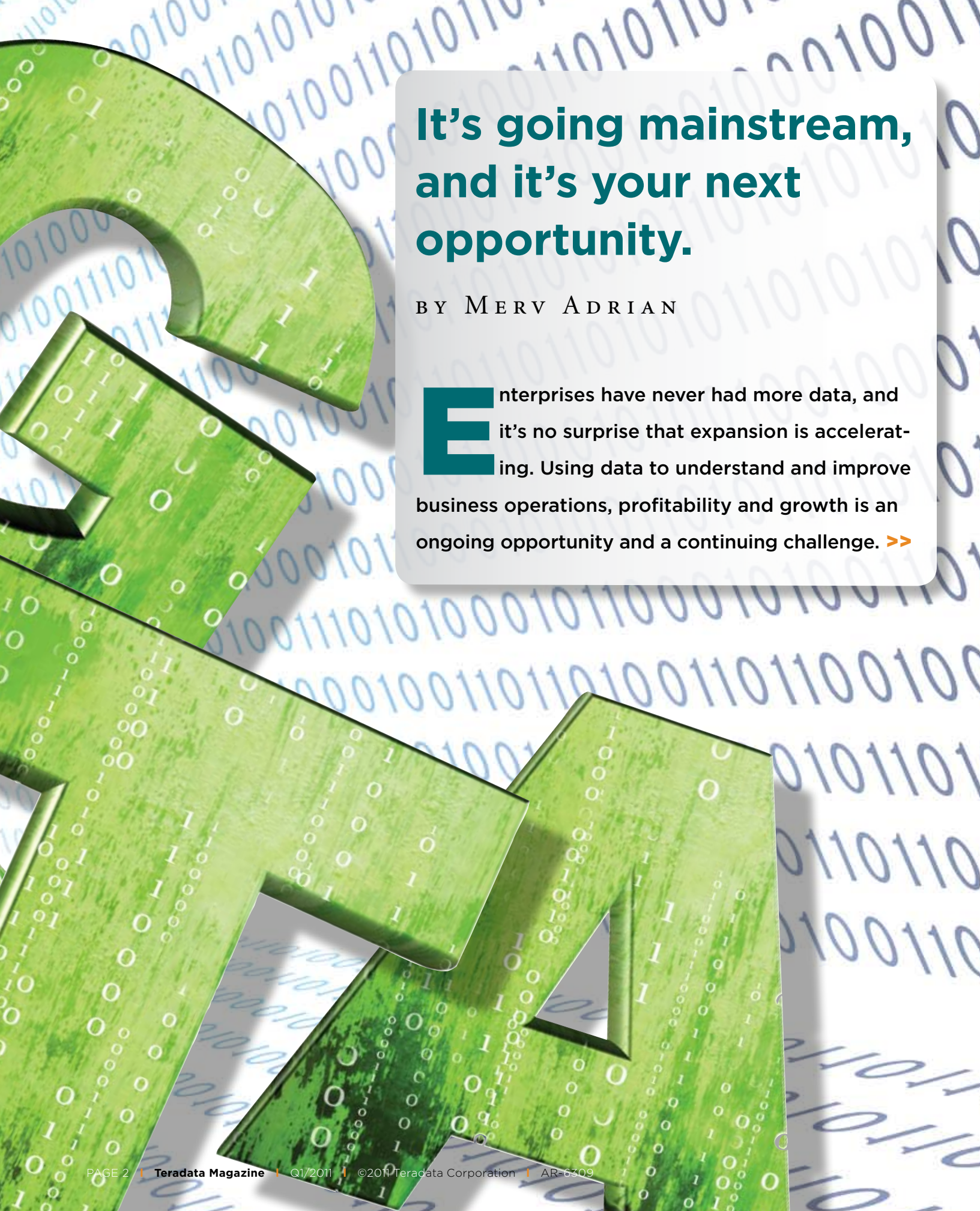

It's going mainstream, and it's your next opportunity.

BY MERV ADRIAN

Enterprises have never had more data, and it's no surprise that expansion is accelerating. Using data to understand and improve business operations, profitability and growth is an ongoing opportunity and a continuing challenge. >>

Decades of collection, design and construction of data warehouses and data marts have created enormously valuable assets. But in an October 2010 TechTarget study, "Analytic Platforms: Beyond the Traditional Data Warehouse, BI Research and IT Market Strategy," 53% of respondents indicated that they routinely perform business analysis on data not maintained in a relational database management system (RDBMS). Data flooding in from Web behavior, customer loyalty programs, remote sensor technology, call center records and social media has left organizations unsure about how to maximize the business value of all that data.

Welcome to the world of big data: new questions, new models and new opportunities for actionable insights.

What Is Big Data?

Traditionally, the term "big data" has been used to describe the massive volumes of data analyzed by huge organizations like Google or research science projects at NASA. But for most businesses, it's a relative term: "Big" depends on an organization's size.

The point is more about finding new value within and outside conventional data sources. Two-thirds of the firms Tech Target surveyed are keeping more than a year's worth of data online—43% have more than three years' worth.

Pushing the boundaries uncovers new value and opportunities, and "big" depends on where you start. Consider this description:

Big data *exceeds the reach of commonly used hardware environments and software tools to capture, manage and process it within a tolerable elapsed time for its user population.*

The Challenges

For big data problems, size really is a relative measure, a constantly moving target depending on where you start. But 23% of survey respondents are already managing more than 10TB for analytics; one-third expect 100 or more concurrent users. Big data, and big usage, are already here.

Several other elements contribute to the size challenge:

- The **user population** is growing. The story often focuses on a few Ph.Ds. writing highly complex programs to run over petabytes of data while business users wait for answers. But this is far from the truth. Big data applications drive operations in finance, retail, social networking and telecommunications to serve hundreds or thousands of users—both people and processes—inside and outside an organization. High rates of concurrent use are no longer out of reach.
- **Processors** and the memory associated with them can contain just so much data at a time. Until recently, enormous financial commitments were required to acquire hardware larger than immediate needs to have "headroom." Today, however, blades with processors, memory and storage fit in racks and snap in on demand. They provide a more gradual scaling model that maps costs more closely to usage and enable "bigger" data volumes more easily and affordably. At the same time, software is catching up, leveraging memory better and using storage more intelligently.

Big data is driving innovative approaches that change as frequently as the problems do. Remaining open to new ideas, and frequent experiments, is the order of the day.

Retail Recommendations Rule

Customers of large e-commerce sites often utilize search to start their purchase selection. A category-based approach to search is too difficult to navigate and risks losing customer interest when a request for blue shirts turns up thousands of variations. But understanding consumer behavior, using analytics based on profiles and demographics, can address that.

Leveraging MapReduce and exported warehouse data, an online retailer can process behaviors and create search indexes personalized dynamically and matched to available inventory. If an individual always buys designer gear, search indexes can be constantly revised in the recommendation engine. A Hadoop-based system scrubs Web clicks and most popular search indexes, while the data warehouse provides several years of integrated historical data. Customers then see attractive choices in the fewest possible clicks.

—M.A.

What Are the Tasks?

Although input data volume and user numbers define many projects thought of as big data, they aren't the only criteria. Some kinds of analysis are themselves big: computations and analyses that create enormous amounts of temporary data, join the results back to themselves several times and execute multiple computational passes.

A telco analysis might ask: "What percentage of the people called by last quarter's most frequent users of our conference calling option are themselves users of the service? Has that proportion changed notably in the past several quarters in the regions where our price promotion was tested? And how does it relate to the pattern in the regions where it has not?"

Responding to such questions involves a complex process. Even if the incoming data volume is not particularly large, it often requires enormous resources, such as multiple writes to temporary stores and reads—in memory or on disk.

Many such challenges are posed by analysts tasked to find new value, test ongoing marketing programs, optimize new ones and so on. But a similar analysis might well be automated and inform a call center screen designed to ensure the best offer is presented to a customer. For such big data issues, a data warehouse has been the right home.

Will it remain so? The answer depends in large part on how often the data will be used. For infrequently used, rarely changed data, distributed file systems on inexpensive hardware may be an effective option now that tools such as MapReduce are emerging to process it efficiently.

Hadoop and MapReduce Offer an Alternative

Documents, unstructured text, Web log data and other content not typically stored inside an RDBMS contain valuable information. Such data changes very rarely, if at all. A new class of products, often collectively referred to as NoSQL, is emerging for working with this information.

The Hadoop Distributed File System is gaining popularity to store data on a massively parallel collection of inexpensive commodity hardware. At Internet firms like Yahoo and Facebook, data farms grow steadily on servers, and developers write applications in Java, Python and Perl.

Typically these approaches involve MapReduce, a programming model for processing and generating large data sets that is automatically parallelized and executed on a large cluster of commodity machines. The resulting programs are relatively specialized and may have very limited use. They may be tested and discarded when newer approaches come along, but they can be very cost-effective.

Use cases for MapReduce include extract, transform and load (ETL) processing, "sessionization" of Web logs and various types of data mining. In such cases, the result set may be imported into a data warehouse for further processing with SQL. Alternatively, MapReduce may be run inside databases. And finally, an emerging approach is the use of Hive, a SQL-like layer, atop Hadoop. Which approach to use, and when, has as much to do with an organization's skills and resources available as it does with the technology.

Big data is driving innovative approaches like these that change as frequently as the problems do. Remaining open to new ideas, and frequent experiments, is the order of the day.

>>

Inside and Out

Although much of the big data conversation centers on entirely new workloads with data outside the warehouse, the fact is that data warehouses are handling problems of greater scale and complexity than many of the new use cases. What drives your best architecture?

Start with where the needed data resides. If it's already inside a data warehouse, emerging techniques for processing inside, "close to the data," allow you to leverage the features of an RDBMS and the platform running it. New techniques using MapReduce—as well as user-defined functions and data types—are emerging and may let you innovate. Some products offer "sandboxing" to build temporary places to experiment inside your data warehouse and tear them down when finished.

That's not the only possible scenario, though. Much data resides in file systems—outside the warehouse—and that data may well be growing faster than what is inside. If it's mostly appended data and rarely updated, it may be less costly to leave it in file systems on inexpensive servers. If it's used only rarely for analysis and reporting, much of the value—and cost—of a DBMS may be unnecessary. For that data, new approaches like Hadoop and MapReduce are an obvious fit.

Some problems will require the use of both kinds of data—and joining them will be accomplished most effectively inside the data warehouse. For those applications, process the data where you can before importing. It may even make sense to export data from the warehouse to combine with external data, run some complex scoring, time series or graph analysis outside, and bring the results back into the data warehouse.

Don't be afraid to push the envelope: Big data will change everything. Experiment—and be willing to fail before you succeed.

Push the Envelope

When getting started, consider where your computational capacity lies. Infrequent jobs on large sets of data can be a drain on even the most powerful systems if other workloads must run at the same time. If you have available capacity on inexpensive servers outside the data warehouse, leverage it.

Finally, assess the skills needed to solve the business problems you are targeting. Analytical specialists often have tools of choice—and SQL is not always at the top of their list. Their preferred tools can be enabled for big data in many ways—from analytic tools inside the data warehouse to programming jobs outside with open-source libraries to run on a Hadoop cluster.

As with any project, where you begin, what's available and what you're comfortable with will converge to form the most obvious solution. But don't be afraid to push the envelope: Big data will change everything. Experiment—and be willing to fail before you succeed. T

Merv Adrian, a vice president with Gartner, was the principal of IT Market Strategy, an independent consultancy, when he contributed this article.

Find Fraud Fixes

Analytical sandboxes can enable financial quantitative analysts, or "quants," to rapidly iterate data mining scenarios for fraud detection. Financial institutions can use Hadoop in applications to sift through massive amounts of data, increasing accuracy and ensuring that all relevant data is looked at. Outliers can often be fraudsters who can easily be washed away and missed if small samples of data are used instead.

Building effective models in data mining is also hastened by making changes to the data on the fly—changing segmentation, realigning columns, adding scores—while production data needs to be left alone. Moving the data to Hadoop gives the quants the flexibility to learn, fail fast, try again and finally get a machine learning algorithm (such as a neural net) to create a model for pattern detection. Scoring models can then be moved back into the data warehouse where they are applied to production data.

—*M.A.*