FALL 2012 VOL.54 NO.1

MITSloan
Management Review

Thomas H. Davenport, Paul Barth and Randy Bean

How ‘Big Data’ is Different

Please note that gray areas reflect artwork that has been intentionally removed. The substantive content of the article appears as originally published.

REPRINT NUMBER 54104

[DATA AND ANALYTICS]

How ‘Big Data’ Is Different

These days, lots of people in business are talking about “big data.” But how do the potential insights from big data differ from what managers generate from traditional analytics?

BY THOMAS H. DAVENPORT, PAUL BARTH AND RANDY BEAN

Big data encompasses everything from clickstream data from the Web to genomic and proteomic data from biological research and medicine.

These days, many people in the information technology world and in corporate boardrooms are talking about “big data.” Many believe that, for companies that get it right, big data will be able to unleash new organizational capabilities and value. But what does the term “big data” actually entail, and how will the insights it yields differ from what managers might generate from traditional analytics?

There is no question that organizations are swimming in an expanding sea of data that is either too voluminous or too unstructured to be managed and analyzed through traditional means. Among its burgeoning sources are the clickstream data from the Web, social media content (tweets, blogs, Facebook wall postings, etc.) and video data from retail and other settings and from video entertainment. But big data also encompasses everything from call center voice data to genomic and proteomic data from biological research and medicine. Every day, Google alone processes about 24 petabytes (or 24,000 terabytes) of data. Yet very little of the information is formatted in the traditional rows and columns of conventional databases.

Many IT vendors and solutions providers use the term “big data” as a buzzword for smarter, more insightful data analysis. But big data is really much more than that. Indeed, companies that learn to take advantage of big data will use real-time information from sensors, radio frequency identification and other identifying devices to understand their business environments at a more granular level, to create new products and services, and to respond to changes in usage patterns as they occur. In the life sciences, such capabilities may pave the way to treatments and cures for threatening diseases.

Organizations that capitalize on big data stand apart from traditional data analysis environments in three key ways:

- They pay attention to data flows as opposed to stocks.
- They rely on data scientists and product and process developers rather than data analysts.
- They are moving analytics away from the IT function and into core business, operational and production functions.

1. Paying attention to flows as opposed to stocks There are several types of big data applications. The first type supports customer-facing processes to do things like identify fraud in real time or score medical patients for health risk. A second type involves continuous process monitoring to detect such things as changes in consumer sentiment or the need for service on a jet engine. Yet another type uses big data to explore network relationships like suggested friends on LinkedIn and Facebook. In all these applications, the data is not the “stock” in a data warehouse but a continuous flow. This represents a substantial change from the past, when data analysts performed multiple analyses to find meaning in a fixed supply of data.

Today, rather than looking at data to assess what occurred in the past, organizations need to think in terms of continuous flows and processes. "Streaming analytics allows you to process data during an event to improve the outcome," notes Tom Deutsch, program director for big data technologies and applied analytics at IBM. This capability is becoming increasingly important in fields such as health care. At Toronto's Hospital for Sick Children, for example, machine learning algorithms are able to discover patterns that anticipate infections in premature babies before they occur.

The increased volume and velocity of data in production settings means that organizations will need to develop continuous processes for gathering, analyzing and interpreting data. The insights from these efforts can be linked with production applications and processes to enable continuous processing. Although small "stocks" of data located in warehouses or data marts may continue to be useful for developing and refining the analytical models used on big data, once the models have been developed, they need to process continuing data streams quickly and accurately.

The behavior of credit card companies offers a good illustration of this dynamic. In the past, direct marketing groups at credit card companies created models to select the most likely customer prospects from a large data warehouse. The process of data extraction, preparation and analysis took weeks to prepare — and weeks more to execute. However, credit card companies, frustrated by their inability to act quickly, determined that there was a much faster way to meet most of their requirements. In fact, they were able to create a "ready-to-market" database and system that allows a marketer to analyze, select and issue offers in a single day. Through frequent iterations and monitoring of website and call-center activities, companies can make personalized offers in milliseconds, then optimize the offers over time by tracking responses.

Some big data environments, such as consumer sentiment analysis, are not designed for automating decisions but are better suited for real-time monitoring of the environment. Given the volume and velocity of big data, conventional, high-certitude approaches to decision-making are often not appropriate in such settings; by the time the organization has the information it needs to make a decision, new data is often available that renders the decision obsolete. In real-time monitoring contexts, organizations need to adopt a more continuous approach to analysis and decision-making based on a series of hunches and hypotheses. Social media analytics, for example, capture fast-breaking trends on customer sentiments about products, brands and companies. Although companies might be interested in knowing whether an hour's or a day's changes in online sentiment correlate with sales changes, by the time a traditional analysis is completed there would be a raft of new data to analyze. Therefore, in big data environments it's important to analyze, decide, and act quickly and often.

However, it isn't enough to be able to monitor a continuing stream of information. You also have to be prepared to make decisions and take action. Organizations need to establish processes for determining when specific decisions and actions are necessary — when, for example, data values fall outside certain limits. This helps to determine decision stakeholders, decision processes and the criteria and timeframes for which decisions need to be made.

2. Relying on data scientists and product and process developers as opposed to data analysts Although there has always been a need for analytical professionals to support the organization's analytical capabilities, the requirements for support personnel are different with big data. Because interacting with the data itself — obtaining, extracting, manipulating and structuring it — is critical to any analysis, the people who work with big data need substantial and creative IT skills. They also need to be close to products and processes within organizations, which means they need to be organized differently than analytical staff were in the past.

"Data scientists," as these professionals are known, understand analytics, but they also are well versed in IT, often having advanced degrees in computer science, computational physics or biology- or network-oriented social sciences. Their upgraded data management skill set — including programming, mathematical and statistical skills, as well as business acumen and the ability to communicate effectively with decision-makers — goes well beyond what was necessary for data analysts in the past. This combination of skills, valuable as it is, is in very short supply.

As a result, some early adopters of big data are working to develop their own talent. EMC Corporation, for example, traditionally a provider of data storage technologies, acquired Greenplum, a big data technology company, in 2010 to expand its capabilities in data science and promptly started an educational offering for data scientists. Other companies are working with universities to train data scientists.

Early users of big data are also rethinking their organizational structures for data scientists. Traditionally, analytical professionals were often part of internal consulting organizations advising managers or executives on internal decisions. However, in some industries, such as online social networks, gaming and pharmaceuticals, data scientists are part of the product development organization, developing new products and product features. At Merck, for example, data scientists (whom the company calls statistical genetics scientists) are members of the drug discovery and development organization.

3. Moving analytics from IT into core business and operational functions Surging volumes of data require major improvements in database and analytics technologies. Capturing, filtering, storing and analyzing big data flows can swamp traditional networks, storage arrays and relational database platforms. Attempts to replicate and scale the existing technologies will not keep up with big data demands, and big data is changing the technology, skills and processes of the IT function.

The market has responded with a broad array of new products designed to deal with big data. They include open source platforms such as Hadoop, invented by Internet pioneers to support the massive scale of data they generate and manage. Hadoop allows organizations to load, store and query massive data sets on a large grid of inexpensive servers, as well as execute advanced analytics in parallel. Relational databases have also been transformed: New products have increased query performance by a factor of 1,000 and are capable of managing the wide variety of big data sources. Statistical analysis packages are similarly evolving to work with these new data platforms, data types and algorithms.

Coming to terms with big data is prompting organizations to rethink their basic assumptions about the relationship between business and IT — and their respective roles.

Another disruptive force is the delivery of big data capabilities through "the cloud." Although not yet broadly adopted in large corporations, cloud-based computing is well suited to big data. Many big-data applications use external information that is not proprietary, such as social network modeling and sentiment analysis. Moreover, big data analytics are dependent on extensive storage capacity and processing power, requiring a flexible grid that can be reconfigured for different needs. Cloud-based service providers offer on-demand pricing with fast reconfiguration.

Another approach to managing big data is leaving the data where it is. So-called "virtual data marts" allow data scientists to share existing data without replicating it. eBay, for example, used to have an enormous data replication problem, with between 20- and 50-fold versions of the same data scattered throughout its various data marts. Now, thanks to its virtual data marts, the company's replication problem has been dramatically reduced. eBay has also established a "data hub" — an internal website to make it easier for managers and analysts to serve themselves and share data and analyses across the organization. In effect, eBay has built a social network around analytics and data.

Coming to terms with big data is prompting organizations to rethink their basic assumptions about the relationship between business and IT — and their respective roles. The traditional role of IT— automating business processes — imposes precise requirements, adherence to standards and controls on changes. Analytics has been more of an afterthought for monitoring processes and notifying management about the anomalies. Big data flips this approach on its head. A key tenet of big data is that the world and the data that describe it are constantly changing, and organizations that can recognize the changes and react quickly and intelligently will have the upper hand. Whereas the most vaunted business and IT capabilities used to be stability and scale, the new advantages are based on discovery and agility — the ability to mine existing and new data sources continuously for patterns, events and opportunities.

This requires a sea change in IT activity within organizations. As the volume of data explodes, organizations will need analytic tools that are reliable, robust and capable of being automated. At the same time, the analytics, algorithms and user interfaces they employ will need to facilitate interactions with the people who work with the tools. Successful IT organizations will train and recruit people with a new set of skills who can integrate these new analytic capabilities into their production environments.

A further way that big data disrupts the traditional roles of business and IT is that it presents discovery and analysis as the first order of business. Next-generation IT processes and systems need to be designed for insight, not just automation. Traditional IT architecture is accustomed to having applications (or services) as "black boxes" that perform tasks without exposing internal data and procedures. But big data environments must make sense of new data, and summary reporting is not enough. This means that IT applications need to measure and report transparently on a wide variety of dimensions, including customer interactions, product usage, service actions and other dynamic measures. As big data evolves, the architecture will develop into an information ecosystem: a network of internal and external services continuously sharing information, optimizing decisions, communicating results and generating new insights for businesses.

***Thomas H. Davenport** is a visiting professor at Harvard Business School and President's Distinguished Professor of Information Technology and Management at Babson College in Wellesley, Massachusetts. **Paul Barth** and **Randy Bean** are the cofounders and managing partners of NewVantage Partners, a Boston-based management consulting firm. Comment on this article at http://sloanreview.mit.edu/x/54104, or contact the authors at smrfeedback@mit.edu.*

Reprint 54104.