Big Data and Analytics

by **Farhad Hussain**

Technical Specialist (e-government), Leveraging ICT for Growth, Employment and Governance Project, BCC

Big Data and Analytics are hot topics in both the popular and business press. Today, many organizations are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as “Big Data” because of its volume, the velocity with which it arrives, and the variety of forms it takes. Big Data is creating a new generation of decision support data management. Businesses are recognizing the potential value of this data and are putting the technologies, people, and processes in place to capitalize on the opportunities. A key to deriving value from Big Data is the use of Analytics. Collecting and storing Big Data creates little value; it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value.

Big Data and Analytics are intertwined, but Analytics is not new. Many analytic techniques, such as regression analysis, simulation, and machine learning, have been available for many years. Even the value in analyzing unstructured data such as e-mail and documents has been well understood. What is new is the coming together of advances in computer technology and software, new sources of data (e.g., social media), and business opportunity. This confluence has created the current interest and opportunities in Big Data Analytics. It is even spawning a new area of practice and study called “data science” that encompasses the techniques, tools, technologies, and processes for making sense out of Big Data.

Big Data is creating new jobs and changing existing ones. A 2011 study by the McKinsey Global Institute predicts that by 2018 the U.S. alone will face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze Big Data and make decisions. Because companies are seeking people with Big Data skills, many universities are offering new courses, certificates, and degree programs to provide students with the needed skills. Vendors such as IBM are helping educate faculty and students through their university support programs.

At a high level, the requirements for organizational success with Big Data Analytics are the same as those for business intelligence (BI) in general. At a deeper level, however, there are many nuances that are important and need to be considered by organizations that are getting into Big Data Analytics. For example, organizational culture, data architecture, analytical tools, and personnel issues must be considered. Of particular interest to information technology (IT) professionals are the new technologies, platforms, and approaches that are being used to store and analyze Big Data.

Governments and companies are able to integrate personal data from numerous sources and learn much of what you do, where you go, who your friends are, and what your preferences are. Although this leads to better service (and profits for companies), it also raises privacy concerns. There are few legal restrictions on what Big Data companies such as Facebook and Google can do with the data they collect.

Examples of Big Data Analytics

Let us consider several examples of companies that are using Big Data Analytics. The examples illustrate the use of different sources of Big Data and the different kinds of Analytics that can be performed.

Drilling for Oil at Chevron

Each drilling miss in the Gulf of Mexico costs Chevron upwards of $100 million. To improve its chances of finding oil, Chevron analyzes 50 terabytes of seismic data. Even with this, the odds of finding oil have been around 1 in 5. In the summer of 2010, because of BP’s Gulf oil spill, the federal government suspended all deep water drilling permits. The geologists at Chevron took this time to seize the opportunity offered by advances in computing power and storage capacity to refine their already advanced computer models. With these enhancements, Chevron has improved the odds of drilling a successful well to nearly 1 in 3, resulting in tremendous cost savings.

Targeting Customers at Target

Target received considerable negative attention in publications such as the New York Times and Forbes for mining data to identify women who are pregnant. The negative press began when a father complained to a Target store manager in Minneapolis that his daughter had received pregnancy-related coupons. He felt that the coupons were inappropriate and promoted teen pregnancy. Little did he know that his daughter was pregnant. He later apologized to the store manager and said that there had obviously been some activities going on in his household of which he was unaware.

How did Target identify pregnant women? To build its predictive models, Target focused on women who had signed up for the baby registry—an excellent indicator that they were pregnant. They then compared the women’s purchasing behavior with the purchasing behavior of all Target customers. Twenty-five variables were found useful for identifying this market segment—pregnant women—and when their babies were due. The variables included buying large quantities of unscented lotions; supplements such as calcium, magnesium, and zinc; scent-free soaps; extra large bags of cotton balls; hand sanitizers; and washcloths. Using these variables, pregnancy predictive

▶

models were built and used to score the likelihood that a woman was pregnant and when she was likely to deliver. For example, pregnant women tend to buy hand sanitizers and washcloths as they get close to their delivery date. Target used these predictions to identify which women should receive specific coupons.

The story continues, however, with another public-relations nightmare. . Soon afterward, Target received unfavorable press for predicting engagements. Target was sending out invitations to join its bridal registry before sons and daughters told their parents they were engaged. In response to the negative press, Target no longer sends out only pregnancy-related coupons, but mixes in others, such as for lawnmowers. Target is also much more guarded in what information it shares about its data mining activities. While Target's data mining is legal, it strikes many people as creepy, if not inappropriate.

Characteristics of Big Data

Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, Analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize Big Data: volume, variety, and velocity, or the three V's. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data. Data volume is the primary attribute of Big Data. Big Data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Because of its potential benefits, some people add a fourth *V* to the characteristics of Big Data: *high value.* This value is realized, however, only when an organization has a carefully thought out and executed Big Data strategy.

Additionally, one of the things that make Big Data really big is that it's coming from a greater variety of sources than ever before, including logs, click streams, and social media. Using these sources for Analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as eXtensible Markup Language (XML) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to Big Data. Thus, with Big Data, variety is just as big as volume.

The Benefits of Big Data Analytics

As has been discussed, collecting and storing Big Data does not create business value. Value is created only when the data is analyzed and acted on. As the Starbucks, Chevron, and U.S. Xpress examples show, the benefits from Big Data Analytics can be varied, substantial, and the basis for competitive advantage.

Research shows the benefits of using data and Analytics in decision making. One study of 179 large publicly traded firms found that companies that have adopted data-driven decision making have output and productivity that is 5% to 6% higher than that of other firms. The relationship extends to other performance measures such as asset utilization, return on equity, and market value. In 2010, the MIT Sloan Management Review, in collaboration with the IBM Institute for Business Value, surveyed a global sample of nearly 3,000 executives. Among the findings were that top-performing organizations use Analytics five times more than do lower performers and that 37% of the respondents believe that Analytics creates a competitive advantage. A follow-up study in 2011 found that the percentage of respondents who reported that the use of Analytics was creating a competitive advantage rose to 58%, which is a 57% increase. Although these studies do not focus exclusively on Big Data, they do show the positive relationships between data-driven decision making, organizational performance, and competitive position.

There are also potential benefits from governments' use of Big Data. A TechAmerica report of 2012 described the following scenario of a world that is benefiting from Big Data Analytics:

Imagine a world with an expanding population but a reduced strain on services and infrastructure; dramatically improved health care outcomes with greater efficiency and less investment; intensified threats to public safety and national borders, but greater levels of security; more frequent and intense weather events, but greater accuracy in prediction and management. Imagine a world with more cars, but less congestion; more insurance claims but less fraud; fewer natural resources, but more abundant and less expensive energy. The impact of Big Data has the potential to be as profound as the development of the Internet itself. This scenario may be optimistic, but it suggests uses of Big Data Analytics that are being aggressively pursued.

Big Data Analytics Tools and Methods

With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analyzing such data. Having piles of data on hand is no longer enough to make efficient decisions at the right time. Such data sets can no longer be easily analyzed with traditional data management and analysis techniques and infrastructures. Therefore, there arises a need for new tools and methods specialized for Big Data Analytics, as well as the required architectures for storing and managing such data. Accordingly, the emergence of Big Data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions.

Big Data Storage and Management

One of the first things organizations have to manage when dealing with Big Data is where and how this data will be stored once it is acquired. The traditional methods of structured data storage and retrieval include relational databases, data marts, and data warehouses. The data is uploaded to the storage from operational data stores using Extract, Transform, Load (ETL), or Extract, Load, Transform (ELT), tools which extract the data from outside sources, transform the data to fit operational needs, and finally load the data into the database or data warehouse. Thus, the data is cleaned, transformed, and catalogued before being made available for data mining and online analytical functions. However, the Big Data environment calls for Magnetic, Agile, Deep (MAD) analysis skills, which differ from the aspects of a traditional Enterprise Data ▶

Warehouse (EDW) environment. First of all, traditional EDW approaches discourage the incorporation of new data sources until they are cleansed and integrated. Due to the ubiquity of data now a days, Big Data environments need to be magnetic, thus attracting all the data sources, regardless of the data quality could the Internet make to productivity growth?

Hadoop is a framework for performing Big Data Analytics which provides reliability, scalability, and manageability by providing an implementation for the MapReduce paradigm, which is discussed in the following section, as well as gluing the storage and Analytics together. Hadoop consists of two main components: the HDFS for the Big Data storage, and MapReduce for Big Data Analytics. The HDFS storage function provides a redundant and reliable distributed file system, which is optimized for large files, where a single file is split into blocks and distributed across cluster nodes. Additionally, the data is protected among the nodes by a replication mechanism, which ensures availability and reliability despite any node failures. There are two types of HDFS nodes: the Data Nodes and the Name Nodes. Data is stored in replicated file blocks across the multiple Data Nodes, and the Name Node acts as a regulator between the client and the Data Node, directing the client to the particular Data Node which contains the requested data.

Big Data Analytic Processing

After the Big Data storage, comes the analytic processing. There are four critical requirements for Big Data processing. The first requirement is fast data loading. Since the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time. The second requirement is fast query processing. In order to satisfy the requirements of heavy workloads and real-time requests, many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increase. Additionally, the third requirement for Big Data processing is the highly efficient utilization of storage space. Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing, and issues on how-to store the data so that space utilization is maximized be addressed. Finally, the fourth requirement is the strong adaptivity to highly dynamic workload patterns. As Big Data sets are analyzed by different applications and users, for different purposes, and in various ways, the underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns.

How Big Data Analytics Is Used Today

As the technology that helps an organization to break down data silos and analyze data, business can be transformed in all sorts of ways. According to Datamation, today’s advances in analyzing Big Data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook. Another example comes from one of the biggest mobile carriers in the world. France’s Orange launched its Data for Development project by releasing subscriber data for customers in the Ivory Coast. The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users. Researchers accessed the data and sent Orange proposals for how the data could serve as the foundation for development projects to improve public health and safety. Proposed projects included one that showed how to improve public safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment.

Enterprises are increasingly looking to find actionable insights into their data. Many Big Data projects originate from the need to answer specific business questions. With the right Big Data Analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management. Webopedia parent company, QuinStreet, surveyed 540 enterprise decision-makers involved in Big Data purchases to learn which business areas companies plan to use Big Data Analytics to improve operations. About half of all respondents said they were applying Big Data Analytics to improve customer retention, help with product development and gain a competitive advantage. Notably, the business area getting the most attention relates to increasing efficiency and optimizing operations.

Conclusion

Big Data has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. In the information era we are currently living in, voluminous varieties of high velocity data are being produced daily, and within them lay intrinsic details and patterns of hidden knowledge, which should be extracted and utilized. Organizations are gaining unprecedented insights into customers and operations because of the ability to analyze new data sources and large volumes of highly detailed data. This data is bringing more context and insight to organizational decision making. Success with Big Data is not guaranteed, however, as there are specific requirements that must be met. Organizations should start with specific, narrowly defined objectives, often related to better understanding and connecting with customers and improving operations. There must be strong, committed sponsorship. For some companies (e.g., Google), alignment between the business and IT strategies is second nature because Big Data is what the business is all about. For others, careful consideration needs to be given to organization structure issues; governance; the skills, experiences, and perspectives of organizational personnel; how business needs are turned into successful projects; and more. There should be a fact-based decision-making culture where the business is run by the numbers and there is constant experimentation to see what works best. The creation and maintenance of this culture depends on senior management. Big Data has spawned a variety of new data management technologies, platforms, and approaches. These must be blended with traditional platforms such as data warehouses in a way that meets organizational needs cost effectively. The analysis of Big Data requires traditional tools like SQL, analytical workbenches like SAS Enterprise Miner, and data analysis and visualization languages like R. All of this is for naught, however, unless there are business users, analysts, and data scientists who can work with and use Big Data ◘