BIG DATA: CONCEPTS, TECHNOLOGIES, AND APPLICATIONS

ABSTRACT

We have entered the big data era. Organizations are capturing, storing, and analyzing data that has high volume, velocity, and variety and comes from a variety of new sources, including social media, machines, log files, video, text, image, RFID, and GPS. These sources have strained the capabilities of traditional relational database management systems and spawned a host of new technologies, approaches, and platforms. The potential value of big data analytics is great and is clearly established by a growing number of studies. There are keys to success with big data analytics, including a clear business need, strong committed sponsorship, alignment between the business and IT strategies, a fact-based decision-making culture, a strong data infrastructure, the right analytical tools, and people skilled in the use of analytics. Because of the paradigm shift in the kinds of data being analyzed and how this data is used, big data can be considered to be a new, 4th generation of decision support data management. Though the business value from big data is great, especially for online companies like Google and Facebook, how it is being used is raising significant privacy concerns.

I. INTRODUCTION[1]

Big data and analytics are "hot" topics in both the popular and business press. Articles in publications like the *New York Times*, *Wall Street Journal* and *Financial Times*, as well as books like *Super Crunchers* [Ayers, 2007], *Competing on Analytics* [Davenport and Harris, 2007], and *Analytics at Work* [Davenport, et al., 2010] have spread the word about the potential value of big data and analytics.

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

[1] This tutorial is based on a presentation with the same title given at the America's Conference on Information Systems in Seattle, WA, August 2012. The slides from the presentation are available on the Teradata University Network (www.teradatauniversitynetwork.com). A version of this tutorial is published in the *Communications of AIS* in 2014.

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 email 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 changing existing jobs and creating new ones. For example, market researchers must now be skilled in social media analytics. Data management professionals must be able to store massive amounts of data of any structure. The job of data scientist, the “high priest” of big data analytics, has emerged. Because many 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 (e.g., 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 for business intelligence in general [Williams, 2004]. 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, there are organizational culture, data architecture, analytical tools, and personnel issues that need to 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. They aren’t “your mother’s BI architecture” [Watson, 2012].

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 [Clemons, 2014]. There are few legal restrictions on what big data companies such as Facebook and Google can do with the data they collect.

In this tutorial, we will first consider the nature and sources of big data. Next, we look at the history of analytics, the various kinds of analytics, and now they are used with big data. Starbucks, Chevron, U.S. Xpress, and Target are used to illustrate various uses of big data analytics. Current research is documenting the benefits of big data and provides a compelling argument for its use. The requirements

for being successful with big data are discussed and illustrated: including establishing a clear business need; having strong, committed sponsorship; alignment between the business and IT strategies; a fact-based decision-making culture; a strong data infrastructure; the right analytical tools; and users, analysts, and data scientists skilled in the use of big data analytics. Special attention is given to the technologies, platforms, and approaches for storing and analyzing big data. Privacy concerns about the use of big data are also explored.

II. WHAT IS BIG DATA

From an evolutionary perspective, big data is not new. A major reason for creating data warehouses in the 1990s was to store large amounts of data. Back then, a terabyte was considered big data.[2] Teradata, a leading data warehousing vendor, used to recognize customers when their data warehouses reached a terabyte. Today, Teradata has more than 35 customers (e.g., Walmart, Verizon) with data warehouses over a petabyte in size. eBay captures a terabyte of data per minute and maintains over 40 petabytes, the most of any company in the world.

So what is big data? One perspective is that big data is more and different kinds of data than is easily handled by traditional relational database management systems (RDBMSs). Some people consider 10 terabytes to be big data, but any numerical definition is likely to change over time as organizations collect, store, and analyze more data.

Another useful perspective is to characterize big data as having high volume, high velocity, and high variety – the three Vs [Russom, 2011]:

- *High volume* – the amount or quantity of data
- *High velocity* – the rate at which data is created
- *High variety* – the different types of data

In short, "big data" means there is more of it, it comes more quickly, and comes in more forms.

Both of these perspectives are reflected in the following definition [Mills, et al., 2012; Sicular, 2013]:

> *Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.*

[2] As a frame of reference, a terabyte can hold 1,000 copies of the Encyclopedia Britannica. Ten terabytes can hold the printed collection of the Library of Congress. A petabyte can hold approximately 20 million four-door filing cabinets full of text. It would take about 500 million floppy disks to store the same amount of data. See www.whatsabyte.com.

It is important to understand that what is thought to be big data today won't seem so big in the future [Franks, 2012]. There are many data sources that are currently untapped, or at least underutilized. For example, every customer e-mail, customer service chat, and social media comment may be captured, stored, and analyzed to better understand customers' sentiments. Web browsing data may capture every mouse movement in order to better understand customers' shopping behaviors. Radio frequency identification (RFID) tags may be placed on every single piece of merchadise in order to assess the condition and location of every item.

III. BIG DATA SOURCES

There are many sources of big data. For example, every mouse click on a *web site* can be captured in Web log files and analyzed in order to better understand shoppers' buying behaviors and to influence their shopping by dynamically recommending products. *Social media* sources such as Facebook and Twitter generate tremendous amounts of comments and tweets. This data can be captured and analyzed to understand, for example, what people think about new product introductions. *Machines*, such as smart meters, generate data. These meters continuously stream data about electricity, water, or gas consumption that can be shared with customers and combined with pricing plans to motivate customers to move some of their energy consumption (e.g., washing clothes) to non-peak hours. There is a tremendous amount of *geospatial* (e.g., GPS) data, such as that created by cell phones, that can be used by applications like Four Square to help you know the locations of friends and to receive offers from nearby stores and restaurants. *Image, voice, and audio* data can be analyzed for applications such as facial recognition systems in security systems.

IV. BIG DATA ANALYTICS

By itself, stored data does not generate business value, and this is true of traditional databases, data warehouses, and the new technologies for storing big data (e.g., Hadoop). Once the data is appropriately stored, however, it can be analyzed and this can create tremendous value. A variety of analysis technologies, approaches, and products have emerged that are especially applicable to big data, such as in-memory analytics, in-database analytics, and appliances (all discussed later).

What Is Analytics?

It is helpful to recognize that the analytics term is not used consistently; it is used in at least three different, yet related ways [Watson, 2013a]. A starting point for understanding analytics is to explore its roots. Decision support systems (DSS) in the 1970s were the first systems to support decision making [Power, 2007]. DSS came to be used as a description for an application and an academic discipline. Over time, additional decision support applications, such as executive information systems, online analytical processing (OLAP), and dashboards/scorecards, became popular. Then in the 1990s, Howard Dresner, an analyst at Gartner, popularized the business intelligence term. A typical definition is that "BI is a broad

category of applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help business users make better decisions" [Watson, 2009a]. With this definition, BI can be viewed as an umbrella term for all applications that support decision making, and this is how it is interpreted in industry and increasingly in academia. BI evolved from DSS, and one could argue that analytics evolved from BI (at least in terms of terminology). Thus, analytics is an umbrella term for data analysis applications. BI can also be viewed as "getting data in" (to a data mart or warehouse) and "getting data out" (analyzing the data that is stored). A second interpretation of analytics is that it is the "getting data out" part of BI. The third interpretation is that analytics is the use of "rocket science" algorithms (e.g., machine learning, neural networks) to analyze data. These different "takes" on analytics do not normally cause much confusion, because the context usually makes the meaning clear. The progression from DSS to BI to analytics is shown in Figure 1.

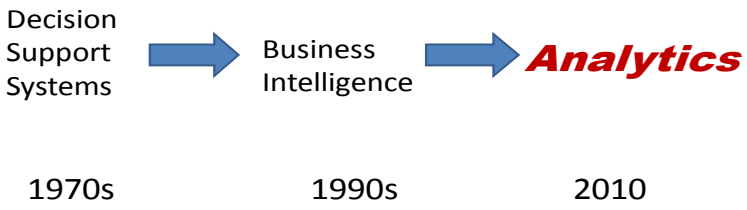


Figure 1: From DSS to BI to analytics.

Different Kinds of Analytics

It is useful to distinguish betwee three kinds of analytics because the differences have implications for the technologies and architectures used for big data analytics. Some types of analytics are better performed on some platforms than others.

Descriptive analytics, such as reporting/OLAP, dashboards/scorecards, and data visualization, have been widely used for some time, and are the core applications of traditional BI. Descriptive analytics are backward looking (like a car's rear view mirror) and reveal *what has occurred*. One trend, however, is to include the findings from predictive analytics, such as forecasts of future sales, on dashboards/scorecards.

Predictive analytics suggest *what will occur* in the future (like looking through a car's windshield). The methods and algorithms for predictive analytics, such as regression analysis, machine learning, and neural networks, have existed for some time. Recently, however, software products (e.g., SAS Enterprise Miner) have made them much easier to understand and use. They have also been integrated into specific applications, such as for campaign management. Marketing is the target for many predictive analytics applications; here the goal is to better understand customers and their needs and preferences.

Some people also refer to *exploratory* or *discovery* analytics, although these are just other names for predictive analytics. When these terms are used, they normally refer to finding relationships in big data that were not previously known. The ability to analyze new data sources (i.e., big data) creates additional opportunities for insights and is especially important for firms with massive amounts of customer data.

Golden path analysis is a new and interesting predictive or discovery analytics technique. It involves the analysis of large quantities of behavioral data (i.e., data associated with the activities or actions of people) to identify patterns of events or activities that foretell customer actions, such as not renewing a cell phone contract, closing a checking account, or abandoning an electronic shopping cart. When a company can predict a behavior, it can intercede (e.g., with an offer) and possibly change the anticipated behavior.

Whereas predictive analytics tells you what will happen, p*rescriptive analytics* suggests what to do (like a car's GPS instructions). Prescriptive analytics can identify optimal solutions, often for the allocation of scarce resources. It, too, has been researched in academia for a long time but is now finding wider use in practice. For example, the use of mathematical programming for revenue management is increasingly common for organizations that have "perishable" goods (e.g., rental cars, hotel rooms, airline seats). For example, Harrah's Entertainment, a leader in the use of analytics, has been using revenue management for hotel room pricing for many years.

Organizations typically move from descriptive to predictive to prescriptive analytics. Another way of describing this progression is: what happened – why did it happen – what will happen – how can we make it happen? This progression is normally seen in various BI and analytics maturity models [Eckerson, 2004].

V. 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.

Introducing a New Coffee Product at Starbucks

Starbucks was introducing a new coffee product but was concerned that customers would find its taste too strong. The morning that the coffee was rolled out, Starbucks monitored blogs, Twitter, and niche coffee forum discussion groups to assess customers' reactions. By mid-morning, Starbucks discovered that although people liked the taste of the coffee, they thought that it was too expensive. Starbucks lowered the price, and by the end of the day all of the negative comments had disappeared.

Compare this fast response with a more traditional approach. An alternative would be to wait for the sales reports to come in and notice that sales are dissapointing. A next step might be to run a focus group to discover why. Perhaps in several weeks Starbucks would have discovered the reason and responded by lowering the price.

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 succesful well to nearly 1 in 3, resulting in tremendous cost savings.

Monitoring Trucks at U.S. Xpress

U.S. Xpress is a transportation company. Its cabs coninuoulsy stream more than 900 pieces of data related to the condition of the trucks and their locations [Watson, 2011]. This data is stored in the cloud and analyzed in various ways, with information delivered to various users (e.g., drivers, senior executives) on tablet computers (e.g., iPads). For example, when a sensor shows that a truck is low on fuel, the driver is directed to a filling station where the price is low. If a truck appears to need maintenance, drivers are sent to a specific service depot. Routes and destinations are changed to ensure that orders are delivered on time.

Trucks experience necessary and unavoidable idle time. An example of the former is when a truck is stuck in traffic and nothing can be done about it (unless it is routed around a traffic delay). An example of the latter is when a driver stops for lunch in the winter and keeps the truck running in order to keep the cab warm. By monitoring its trucks, U.S. Xpress can tell which is which, and has saved millions in fuel costs and reduced emissions into the environment by incenting its drivers to reduce avoidable idle time.

Targeting Customer at Target

Target received considerable negative attention in publications such as the *New York Times* [Duhigg, 2012] and *Forbes* [Hill, 2012] 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 pregant. 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 (i.e., 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 predicitons to identify which women should receive specific coupons.

The story continues, however, with another public relations nightmare. Soon afterward, there were headlines such as "Target hears you say 'yes' before you do." Based on predictive analytics models, Target was sending out invitations to join its bridal registry before sons and daughters told their parents they were engaged [Hill, 2012].

In response to the negative press, Target no longer sends out only pregnancy-related coupons, but mixes in others, such as for lawnmovers. Target is also much more guarded in what information it shares about its data mining activties. While Target's data mining is legal, it stikes many people as creepy, if not inappropriate.

VI. THE BENEFITS OF BIG DATA ANALYTICS

As has been said, collecting and storing big data does not create business value. Value is only created 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. Because of its potential benefits, some people add a fourth V to the characteristics of big data – *High value.*

Research is showing the benefits of using data and analytics in decision making. One study of 179 large publically traded firms found that companies that have adopted "data-driven decision-making" have output and productivity that is 5-6% higher than other firms. The relationship extends to other performance measures, such as asset utilization, return on equity, and market value [Brynjolfsson, et al., 2011]. 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 [LaValle, et al., 2010]. Among the findings were that top-performing organizations use analytics five times more than 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% (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 [Miller, 2012] describes 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.

VII. THE REQUIREMENTS FOR BEING SUCCESSFUL WITH BIG DATA ANALYTICS

The requirements for success with big data analytics, such as executive support and sponsorship, are largely the same as with most projects, including analytics and BI in general [Williams, 2004; Watson, 2013]. The differences are in the details, and some of the details, such as the storage and analysis platforms, are very important.

A Clear Business Need

It is common knowledge that projects should be business rather than technology driven. They should address a business need, such as solving a problem or seizing an opportunity. While the media attention

given to big data has created an awareness of its potential, it has also led some executives to push big data projects without clearly defined goals. Below are opportunities for big data analytics in different industries [Franks, 2012; Smart, et al., 2012].

- *Automobile insurance* – pricing, client risk analyis, fraud detection, faster claims processing
- *Telecommunications* – analysis of patterns of services across social networks, profitability of customers' social networks, churn minimization
- *Manufacturing, distribution, and retail* – tracking shelf availability, assessing the impact of promotional displays, assess the effectiveness of promotional campaigns, inventory mamagement, pricing, advanced clickstream analysis
- *Transportation and logisitics* – real-time fleet management, RFID for asset tracking
- *Utlities* – analysis of smart grid data to determine variable pricing models, smart meters to forecast energy demand, customized rate plans for customers
- *Gaming* – game play analysis to provide feedback to game producers, opportunities for in-game offers
- *Law enforcement* – identifying people linked to known trouble groups, determining the location of individuals and groups

In many organizations, the initial business case for big data analytics focuses on customer-centric objectives and uses existing and newly accessible internal sources of data [Smart, et al., 2012]. Big data analytics can be especially helpful for companies that seek to understand customers better, develop meaningful relationships with customers, and improve operations that enhance the customer experience [Schroeck, et al., 2012]. Whatever the focus, successful big data initiatives should start with a specific or narrowly defined set of objectives rather than a "build it and they will come" approach [Miller, et al., 2012].

Strong, Committed Sponsorship

Without solid sponsorship, it is difficult to succeed with any IT project, and this includes big data analytics projects. If the project is departmental, sponsorship can reside at the departmental level. However, projects that are more strategic and enterprise wide should have senior management suppport.

An IBM study [Schroeck, et al., 2012] found that in the early stages of big data adoption, the CIO is often the sponsor, but as the technology infrastructure is put in place and business opportunities are identified, sponsorship tends to shift to a function-specific executive, such as a CMO or CFO, or even the CEO.

Alignment between the Business and IT Strategy

It is important to make sure that big data analtyics projects support the business strategy. This is why most projects should be driven by business people rather than IT. In analytics-based organizations, the

alignment is especially close; in fact, it may be impossible to separate the business and IT strategies. Without IT as an enabler, the business strategy cannot succeed.

Large online retailers such as Amazon.com and Overstock.com are great examples of analytics-based organizations [Watson, et al., 2009b]. The most visible example of analytics at work are the product recommendations that appear when customers use their web sites. Recommendations are the result of recommendation engines that consider the search terms entered, previous mouse-clicks, market basket analysis of other shoppers' purchases, the availability and profitability of various products, and what the shopper has considered or purchased in the past. Less visible, but equally important BI applications include reporting, dashboards/scorecards, demand forecasting, pricing, product return analysis, market segmentation analysis, campaign management, and search engine optimization.

In a conversation with Patrick Byrne, the CEO of Overstock.com, the question was asked: "How would you describe your company?" Byrne responded, "A BI (analytics) company." His answer reflects how important analytics is to the success of Overstock.com and the need for close alignment between business and IT strategies.

A Fact-Based Decision-Making Culture

To benefit from big data analytics, decisions must be based on "the facts" (generated by analytics) and there should be constant experimentation to see what works best. Changing the organizational culture associated for how decisions are made can be more challenging than solving technical issues. This was seen in *Moneyball,* which tells the story of the Oakland Athletics and general manager Billy Beane's use of analytics to make personnel and other baseball decisions [Lewis, 2003; Bennett, 2011]]. Beane had to overcome the authority and influence of dissenters with years of baseball experience in order to implement his new analytical approach. Now, every major sporting team relies on analytics for all kinds of decisions, such as when to try a two-point conversion in football.

Harrah's became a leader in the gaming industry through its use of analytics and its Total Rewards loyalty program [Watson and Volonino, 2000]. Prior to the introduction of analytics, decisions were typically based on "Harrahisms" – practices that managers believed worked well. After analytics became the norm, decisions had to be based on the facts. Today it is said that three things will get you fired at Harrah's: stealing, sexual harassment, and failing to make decisions based on the facts.

To create a fact-based decision-making culture, there are several things that senior management can do. First, recognize that some people can't or won't adjust and will have to be replaced. First American Corporation, a regional bank headquartered in Nashville, Tennessee, was in financial trouble and brought in a new management team [Cooper, et al., 2000]. After considering several strategies, management

decided on a customer intimacy strategy where the bank would use analytics to understand its customers especially well and design products and services that would meet customers' needs and preferences as well as increase bank profits. The strategy was highly successful but some people were not able to adjust to the change. Prior to the analytics based approach to running the bank, there were 12 people in the marketing department. Afterwards, there were still 12, but none of the original people were still in marketing. They had either gone to other positions or left the bank. As the CEO explained: "Their idea of marketing was giving balloons and suckers along the teller line and running focus groups, but marking has become very analytical."

There are other things that senior management can do to change the culture. A survey by the Economist Intelligence Unit [2012] found the top strategies in promoting a data driven culture are top down guidance and/or mandates from executives, promotion of data-sharing practices, increased availability of training in data analytics, and communication of the benefits of data-driven decision-making. Other management strategies not included in the survey are stressing that outdated methods must be discontinued, asking to see what analytics went into decisions, and linking incentives and compensation to desired behaviors.

A Strong Data Infrastructure

Data is critically important to BI and analytics. When a strong data infrastucture is in place, applications can often be developed in days. Without a strong data infrastructure, applications may never be completed. IT understands the importance of the data infrastructure, but the business units sometimes assume it is a "given" and don't fully appreciate what is required to create and maintain it.

Technology Advances

At the turn of the century, companies were struggling to store big data. Fortunately, improvements in storage and CPU capabilities, all at a lower cost, saved the day.

Of particular significance is the emergence of the scale-out architecture. With it, hundreds or thousands of low-cost, commodity servers are placed in parallel. Each server has multiple CPUs and large, shared memory caches. Data is spread across the servers and processing takes place in parallel. This approach allows vast quantities of data to be stored and analyzed quickly. If more storage and processing power are needed, additional servers are added to this massively parallel processing (MPP) architecture. Big data platforms rely on a scale-out architecture.

In-memory and solid-state disks are other important technological advances. Each improves response times by storing data in memory rather than on hard disk drives. The major bottle neck to performance (i.e., response time) is the time it takes to access and return data from disk drives; these new technologies greatly reduce this impediment.

The need to store and analyze big data has spawned a variety of technologies, approaches, and platforms. Many of them are complementary. Special attention will be given to Hadoop/MapReduce because of the considerable attention that it is receiving and its potential importance.

Data Warehouses

For many organizations, data warehouses provide the "single version (or source) of the truth" for decision support data. The data is extracted from source systems (e.g., operational systems, ERPs), transformed (e.g., consistent formats), integrated (e.g., around a common key, such as a customer ID), and loaded into the data warehouse. The data can be thought of as "squeaky clean" because of the care taken to insure its accuracy. Users and applications access the data from the warehouse to support decision making.

Data warehouses are primarily designed for the storage and analysis of structured data (i.e., data easily stored in the rows and columns of relational databases). They employ a MPP architecture to provide massive, scalable storage capacity and powerful analytical capabilities. The data is used for queries, reporting, online analytical processing (OLAP), dashboards/scorecards, data visualization, and regulatory and compliance requirements. It is the workhorse for descriptive analytics but also supports predictive and prescritive analytics.

The four major BI companies – IBM, Oracle, SAP, and Microsoft – offer data warehousing products, as does Teradata, an especially important player in the upper end of the market. IBM and Teradata have customers with many of the largest data warehouses in the world.

Data Mart Appliances

Appliances provide an integrated "stack" of hardware, software, and storage in a "single box." They are built from the ground up for speed to process queries very quickly. Appliances can be used in a variety of ways. For example, they can be a standalone system that meets a smaller organization's or a department's needs. Or, they can be tied to a data warehouse and used to offload data and specific applications. Some companies use appliances as "sandboxes" to develop and test new applications before moving them to the warehouse. Some appliances are designed to handle unstructured data by incorporating Hadoop/MapReduce. Others are fine tuned for specific applications, such as retail data and call data records in telecommunication firms.

Appliances deliver high performance in a variety of ways. They use an MPP architecture and are highly scalable. Some have databases that are columnar rather than row based (e.g., Vertica). Some use solid state disks to store data (e.g., Teradata). Others integrate Hadoop/MapReduce into their architecture (e.g., Teradata Aster).

Many people say that Teradata's data warehouse products were the first appliances because they were the first purpose-built systems designed specifically for storing and processing large amounts of data, but the company never described its products as appliances. The first company to use the appliance term was Netezza (now an IBM company, and recently renamed IBM PureData for Analytics) in the early 2000s. Today, all of the major BI vendors offer, as well as other companies such as HP (e.g., Vertica) and EMC (e.g., Greenplum) offer appliances.

Analytical Sandboxes

Advanced analytics can be very computational intensive and create performance problems for descriptive analytics such as reports and dashboards when competing for computing resources. Query managers (part of the RDBMS) can help by prioritizing the order in which queries are processed, but do not provide a complete solution. Another approach is to create an analytic sandbox where modelers can "play" with advanced analytics without impacting other users.

Sandboxes can be real or virtual. With a real sandbox, a separate platform, such as an appliance, is used. Data for the sandbox is sourced from the data warehouse, and possibly augmented by other data that the modelers add. In a virtual sandbox, a partition of the data warehouse is loaded with the data the modelers need. The data warehouse and sandbox reside in the same database software but operate as separate systems. A virtual sandbox requires data warehousing software that supports its creation and use.

In-Memory Analytics

An impediment to fast response times for queries is the time required to find data on disk and to read it to memory. This time can be reduced anywhere from 10 to 1,000 times by in-memory analytics where the data is stored in random access memory (RAM) rather than on a physical disk [Read, 2013]. There is no need to page data in and out of disk storage with this technology.

In-memory technology comes in two forms: either on the platform or with the BI tool. When implemented on the platform, the server stores the data in-memory, and data is accessed by the BI tools. SAP's Hana is an example of in-memory analytics on the platform. Some desktop BI tools, such as QlikView, provide in-memory analytics by maintaining data on the desktop computer's memory. Up to a teradabyte of data can be stored on computers with 64-bit operating systems. In-memory BI tools also move data between disk (e.g., in the data warehouse) and the local, desktop memory so that the most frequently used data (so called "hot" data) is available in memory.

Some applications are especially well suited for in-memory analytics. For example, with OLAP (often incorporated in reports and dashboards/scorecards), users want to "slice and dice" data to look at the business from different perspectives, such as comparing this and last year's sales in different locations.

When all of the data is in memory, this analysis can be done very quickly. Vendors talk about this as "analysis at the speed of thought."

Not all applications require or are well suited for in-memory analytics. Some applications, such as a market basket analysis that is run weekly or applications that require more data than can be provided by current in-memory technologies, are not good candidates. While the cost and reliability of in-memory technology continues to improve, it is still relatively expensive and prone to failure.

In-database Analytics

A change is taking place in where analytics are performed. In the past, data was moved to a server (think of a sandbox) and the analysis was performed there. The trend is to make analytics part of the database software so that data does not have to be moved. With this approach, the analytic capabilities are part of the database software. SAS, a leading provider of advanced analytics software, has partnered with Oracle and Teradata to integrate SAS analytics into their products.

There are several advantages to in-database analytics. First, it eliminates the need for a separate server. It also makes all of the warehouse data available for analysis rather than the typical subset that is used. This improves model accuracy. When the final model is created, it can be used easily with warehouse data. For example, a propensity-to-buy predictive model might be created and then customers can be scored to assess whom to target in a marketing campaign.

Columnar Databases

Historically, RDBMS have stored records in rows; see Figure 2. This is very efficient for entering, updating and deleting records. It is less efficient, however, for analytics where only a few columns are needed (think of the typical WHERE clause in a SQL query) and the table has perhaps hundreds of columns.

In response, Sybase IQ (now a SAP company) offered the first columnar database (in the mid 1990s) that reversed the rows and columns; see Figure 2. This approach provides greater processing speed for queries and opportunities for data compression. A columnar database is a RDBMS, but with the rows and columns reversed. Columnar databases are used by Vertica and ParAccel in their appliances and Teradata in its data warehouse and appliances.

EmpID	LName	FName	Salary
1	Billie	Stevens	80000
2	Mark	Jones	75000
3	Susan	Wilson	83500

Row oriented

1,Billie,Stevens,80000;
2,Mark,Jones,75000;
3,Susan,Wilson,83500;

Column oriented

1,2,3;
Billie,Mark,Susan;
80000,75000,83500;

Figure 2: Row and column-oriented databases.

Streaming and Complex Event Processing (CEP) Engines

We are entering the "Internet of Things" where devices (e.g., cars, utility meters) automatically send out data across the Internet. For businesses, there is value in receiving this data, processing it in real time, and taking immediate action. High-profile applications include automatic stock trading, credit card fraud detection, supply chain management, and equipment monitoring.

Streaming and complex event processing (CEP) engines (e.g., Tibco StreamBase and BusinessEvents) provide continuous intelligence by ingesting large amounts of real time data; accessing historical data from high performance databases, appliances, and data warehouses; making calculations and correlations; detecting patterns and anomalies; applying business rules to incoming data streams; supplying information to users; and automating decision making. With streaming event processing, there is only a single data source, while with CEP there are multiple sources [Eckerson, 2011].

Credit card monitoring provides a good example of a streaming application. A common fraudulent sequence of events is a $5 charge for gas at a convenience store (to see if the credit card is good), followed by the purchase of thousands of dollars of electronic equipment at a big box store. When this stream is detected, store personnel are alerted to possible fraud.

Cloud-based Services

The cloud is now in the mainstream of computing. With the cloud, computing resources are virtualized and offered as a service over the Internet. The potential benefits of the cloud include access to

specialized resources, quick deployment, easily expanded capacity, the ability to discontinue a cloud service when it is no longer needed, cost savings, and good back up and recovery. These same benefits make the cloud attractive for big data and analytics.

Cloud-based services come in a variety of forms. Public clouds are offered by third-party providers while private clouds are implemented within a company's firewall. Concerns about data security is a primary reason that private clouds are sometimes preferred over public clouds. We will discuss public clouds although the same approaches and technologies are used with private clouds.

Cloud services are available as software-as-a-service (SaaS), platform-as-a-service (PaaS), or infrastructure-as-a-service (IaaS), depending on what software is provided. Cloud services are all similar in that a company's data is loaded to the cloud, stored, analyzed, and the results are downloaded to users and applications.

With SaaS, the vendor provides the hardware, application software, operating system, and storage. The user uploads data and uses the application software to either develop an application (e.g., reports) or simply process the data using the software (e.g., credit scoring). Many BI and analytics vendors offer cloud services versions of their software, including Cognos, Business Objects, MicroStrategy, and SAS. SaaS is a particularly attractive option for firms that lack the financial or human resources to implement and maintain the software and applications in-house.

PaaS differs from SaaS in that the vendor does not provide the software for building or running specific applications; this is up to the company. Only the basic platform is provided. The benefits of this approach include not having to maintain the computing infrastructure for applications that are developed; access to a dependable, highly scalable infrastructure; greater agility in developing new applications; and possible cost savings. Examples of PaaS include Oracle Cloud Computing, Microsoft Windows Azure, and Google App Engine.

With IaaS, the vendor provides raw computing power and storage; neither operating system nor application software are included. Customers upload an image that includes the application and operating system. Because the customer provides the operating system, different ones can be used with different applications. IaaS vendors' offerings include Amazon EC2 (part of the Amazon Web Services offerings), Rackspace, and Google Compute Engine.

Consider several interesting examples of cloud-based services that involve big data. Amazon introduced Amazon RedShift in spring 2013 as an offering within Amazon Web Services (AWS). RedShift manages the work of setting up and operating a data warehouse in the cloud. Once the data warehouse is in place,

data can be accessed through SQL queries and analytic applications. What is particularly interesting is the cost -- storing a terabyte of data is only $1,000 per year [Imhoff 2013]. Also in 2013, Jaspersoft, a pioneering open source BI vendor, offered BI Professional for AWS [Kavanagh, 2013]. This cloud-based service is easily set up, costs less than $1 per hour, includes connectivity to RedShift, and can be started, stopped, and restarted.

The second example involves Zynga, a pioneer in online social gaming with games like FarmVille and Mafia Wars, and more recently, Candy Crush Saga. Zynga provides an unusal, but effective, example of a cloud-based business strategy [Babcock, 2011]. Most companies use the cloud as an extension of their data center, shifting work to the cloud when in-house data storage capacities are exceeded. Zynga, because of the nature of the online gaming industry, turns this strategy around.

When a new game is introduced, there is considerable uncertainty as to how many players it will attract. It is important to have sufficient capacity in case the game quickly becomes popular. If Zynga isn't prepared for the demand, players will drift away. Because of this, Zynga launches games on Amazon's EC2 infrastructure-as-a-service, and only after the demand is well understood, is the game brought in-house to Zynga's own Z Cloud. At one time, Zynga's mix of Amazon EC2 and Z Cloud was 80/20, but this has flipped to 20/80 as Zynga has learned how to reap efficiencies from a cloud that is custom-designed to meet its specific needs [Babcock, 2013].

Online gaming is highly dependent on big data and analytics [Rudin, 2010]. Zynga collects a tremendous quantity of data. Upwards of 60 million people play every day and every mouse click is recorded. Like most big data, however, only a small portion of the data merits long-term storage and analysis. Alerts are sent within minutes when there are problems, such as a game going down for a group of players. Reports are generated that show metrics such as the number of players, how many signed up today, how often they play, how many haven't played in 30 days, and the amount of revenue generated (nearly all from the purchase of virtual goods) for every game. Game designers make changes to the games weekly and analysts work with them to investigate how to make the games more engaging and profitable.

Other gaming companies use big data analytics too. For example, EA found that 80% of the players who bought grenades (a virtual good) blew themselves up in the popular Battlefield game 3 [Cifio and Meley, 2011]. This was not fun for the players or good for the sale of virtual goods. In response, EA sent training emails to users about how to use the grenades and feedback to game developers about this problem with the game.

Non Relational (NoSQL) Databases

Relational databases have been a computing mainstay since the 1970s. Data is stored in rows and columns and can be accessed through SQL queries. By way of contrast, non-relational (i.e., NoSQL)

databases are relatively new (1998), can store data of any structure, and do not rely on SQL to retrieve data (though some do support SQL and are perhaps better called “not only SQL databases”). Data such as XML, text, audio, video, image, and application-specific document files are often stored and retrieved “as is” through key-value pairs that use keys to provide links to where files are stored on disk. There are specialized non-SQL databases that are designed for specific kinds of data such as documents and graphs and use their own storage and retrieval methods. Non-relational databases tend to be open-source (e.g., Apache Cassandra, MongoDB, and Apache Couchbase); store data in a scale-out, distributed architecture; and run on low-cost, commodity servers. Hadoop/MapReduce, which is discussed next, is one example of a non-relational database. Because non-relational databases are often relatively new and open source, they are not as well supported as established RDBMS. They also are weaker on security, which can limit their usefulness for some applications (e.g., financial).

Hadoop/MapReduce

Of all the platforms and approaches to storing and analyzing big data, none is receiving more attention than Hadoop/MapReduce. Its origins trace back to the early 2000s, when companies such as Google, Yahoo!, and Facebook needed the ability to store and analyze massive amounts of data from the Internet. Because no commercial solutions were available, these and other companies had to develop their own.

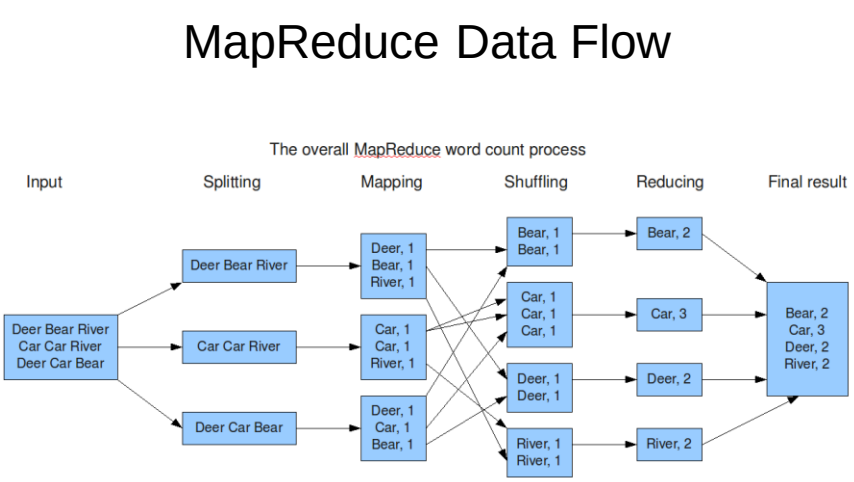
Important to the development of Hadoop/MapReduce were Doug Cutting and Mike Cafarella who were working on an open-source Web search engine project called Nutch when Google published papers on the Google File System (2003) and MapReduce (2004). Impressed with Google’s work, Cutting and Cafarella incorporated the concepts into Nutch. Wanting greater opportunities to further his work, Cutting went to work for Yahoo!, which had its own big data projects under way. With Yahoo!’s support, Cutting created Hadoop (named after Cutting’s son’s stuffed elephant) as an open-source Apache Software Foundation project [Harris, 2013].

Apache Hadoop is a software framework for processing large amounts of data across potentially massively parallel clusters of servers. To illustrate, Yahoo has over 42,000 servers in its Hadoop installation. Hadoop is open source and can be downloaded at www.apache.org. The key component of Hadoop is the Hadoop Distributed File System (HDFS), which manages the data spread across the various servers. It is because of HDFS that so many servers can be managed in parallel. HDFS is file based and does not need a data model to store and process data. It can store data of any structure, but is not a RDBMS. HDFS can manage the storage and access of any type of data (e.g., Web logs, XML files) as long as the data can be put in a file and copied into HDFS.

The Hadoop infrastructure typically runs MapReduce programs (using a programming or scripting language such as Java, Python, C, R, or Perl) in parallel. MapReduce takes large datasets, extracts and transforms useful data, distributes the data to the various servers where processing occurs, and

assembles the results into a smaller, easier to analyze file. It does not perform analytics per se; rather, it provides the framework that controls the programs (often written in Java) that perform the analytics. Currently, jobs can only be run in batch, which limits the use of Hadoop/MapReduce for near real-time applications. Although Hadoop and MapReduce are discussed and typically used together, they can be used separately. That is, Hadoop can be used without MapReduce and vice versa.

Figure 3 illustrates how processing occurs with Hadoop/MapReduce [von Groningen, 2009]. This is a simple processing task that could also be done with SQL and a RDBMS, but provides a good example of Hadoop/MapReduce processing. At the left is a data file with records containing Deer, Bear, River, and Car. The objective is to count the number of times each word occurs. The first step is to *split* the records and distribute them across the clusters of servers (there are only three in this simple example). These splits are then processed by multiple *map* programs (e.g., Java, R) running on the servers. The objective in this example is to group the data by a *split* based on the words. The MapReduce system then merges the *shuffle*/*sort* results for input to the *reduce* program, which then summarizes the number of times each word occurs. This output can then be input to a data warehouse where it may be combined with other data for analysis or accessed directly by various BI tools (e.g., Tableau, MicroStrategy).



Source: van Groningen, 2009

Figure 3: Hadoop/MapReduce processing flow.

There are many related Apache projects that are part of the Hadoop ecosystem. For example, Pig is a high-level parallel processing programming language that is used to write MapReduce programs to run within the Hadoop framework. HBase is a distributed columnar database that gives Hadoop a data

storage option for large tables. Hive is used for SQL-like queries and data summarization. Mahout is a library of data mining algorithms for clustering, classification, and filtering. Collectively, these and other Apache projects provide an ever-growing set of capabilities for processing and analyzing big data. The particular projects (i.e., parts of the Hadoop ecosystem) that are implemented depend on the intended applications and form a BI or analytics stack.

While you can download Apache Hadoop and other parts of the ecosystem for free, one consultant said: "It's like downloading a bag full of razor blades." The various projects are independent and often have competing functionality, separate release schedules, and aren't well integrated. Because of this, companies like Cloudera, Hortonworks (a spinoff from Yahoo!), and MapR have written and offer software that integrates the various parts; provide additional capabilities and administrative tools; and offer consulting services, training, and support.

There are three major ways that companies use Hadoop/MapReduce [Eckerson, 2011]. With the first, companies employ Hadoop as an online archive because of its expandable storage capacity and low cost. The second is as a source system for a data warehouse. In this situation, Hadoop/MapReduce processes data of any structure and then passes the output file to a data warehouse where it can be analyzed along with other data. With this usage, Hadoop/MapReduce complements a data warehouse. In the third use, either the MapReduce programs or data analysis tools that work with the output file are used to analyze the data. This latter use is growing as vendors see the opportunity to capitalize on companies' desire to analyze semi- and unstructured data.

Hadoop is considered to be fault tolerant. Data is always replicated on three separate servers, and if a node fails, is unavailable, or simply slow, another node takes over the processing of the data. When a server recovers or is added, the system automatically recognizes and adds it. The one weakness is the NameNode that keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept. If it fails, it brings the Hadoop cluster down [Russom, 2013].

Over time, Hadoop/MapReduce will become easier to use as various open-source and commercial vendors introduce enhancements and complementary products. Currently, however, a good knowledge of Linux and Java is needed to work effectively with Hadoop/MapReduce. Prominent BI tool vendors like MicroStrategy and Tableau have extended their products to work directly with Hadoop/MapReduce, and new ones such as Karmasphere and Datameer can be used for reporting and dashboards/scorecards.

Which Platform Is Best

The phrase "courses for horses" means that some horses perform better on certain kinds of tracks (e.g., short/long, dry/wet) than others. The same is true for big data analytics. Some kinds of work are better

done on some platforms than others. For example, reporting and dashboards/scorecards normally rely on warehouse data because of the “squeaky clean” data stored there.

There is no “formula” for choosing the right platforms; however, the most important considerations include the volume, velocity and variety of data; the applications that will use the platform; who the users are; and whether the required processing is batch or real time. Some work may require the integrated use of multiple platforms. The final choices ultimately come down to where the required work can be done at the lowest cost.

Integrating the Various Platforms

An increasing number of organizations are using multiple platforms to realize the value from big data. Which platforms are added depends on the platforms, the applications that use the platforms, and the organization’s maturity in working with the various platforms. For example, a company might add an appliance to off-load some computationally intensive applications (e.g., predictive and prescriptive analytics) from a data warehouse. Or it might turn to SaaS for particular applications (e.g., data visualization and analytics).

In many organizations, there is an evolutionary path with greater integration of the platforms over time. For example, Hadoop/MapReduce may initially be implemented in isolation to RDBMS-based systems. Because of its newness, firms often want to test the technolgy and its potential value. It this goes well, Hadoop/MapReduce is put to work on the tasks that it does best. It usually becomes apparent, however, that there is great value in analyzing all of the data together. For example, it is useful to be able to look at both sales figures and what customers are saying about products, and this is likely done best by using Hadoop/MapReduce as a source system for the data warehouse.

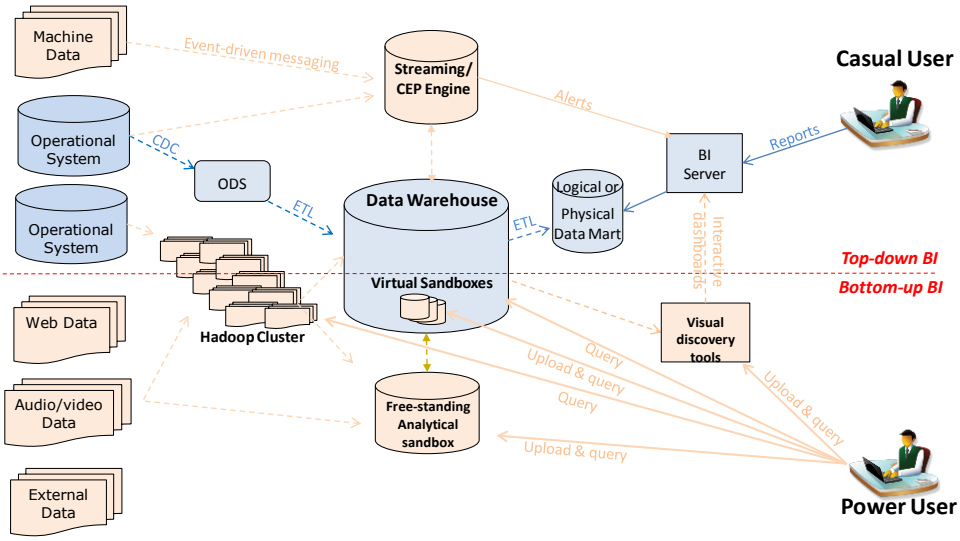
There should be fast, seamless interaction and collaboration among the different platforms. For example, a report run on one platform should mirror a report run on another (the data needs to be synchronized). An analysis run on a specialized platform should be able to access information from the data warehouse (which is a source system for a specialized platform like Hadoop/MapReduce). If needed, the results of an analysis run on a specialized platform should be stored in the data warehouse (the specialized platform is a source system for the warehouse). Vendors recognize the importance of integrating the component parts and are including software solutions for doing this. For example, Teradata is currently emphasizing its Unified Data Architecture that ties its family of products together. However, this integration does not include non-Teradata platforms.

The use of SQL and SQL-like query languages is an important trend for integrating the various platforms and may ultimately be the foundation for creating single logical systems. SQL is a powerful query language that is well known by many IT and BI professionals. Many companies who implement

Hadoop/MapReduce quickly turn to Hive and its HiveQL language because it is learned easily by people who know SQL. Both platform and BI tool vendors are quickly evolving their products to be able to combine data of any structure and to access and analyze it using some variant of SQL.

Figure 4 shows how the various platforms might be integrated and used. The line through the middle of the figure divides what Eckerson [2011] calls the top-down and bottom-up architectures. The top is the traditional BI architecture while the bottom is the new big data architecture. Casual users (i.e., business users) use BI tools to access reports, dashboards/scorecards, and data visualizations based primarily on structured data in the data marts and warehouse (and the new, streaming/CEP engines). Power users (i.e., analysts, data scientists) access a wide variety of data sources, including big data, on a variety of platforms, in a variety of ways, including SQL and analytical workbenches.

Notice that Hadoop and the data warehouse co-exist; one is not a replacement for the other [Russom, 2013]. Each is best for certain kinds of data and processing tasks, and they complement one another.



Source: Eckerson 2011

Figure 4: An integrated analytics architecture.

The Right Analytical Tools

While traditional BI vendors claim their tools support data mining/predictive analytics, this is not always true. Slicing and dicing data and data visualization are not data mining. Data mining requires tools that incorporate algorithms and processes that are designed specifically to find hidden relationships in data.

SAS and SPSS (now an IBM company) have long been leaders in this space, with products like SAS Enterprise Miner and IBM SPSS Modeler. Each product provides a "workbench" where the analysis process is designed using a drag and drop visual interface and execution of the process is automated by the workbench. The most popular tool for data mining is R, a programming language and software environment for statistical computing and graphics. It is also at the core of many open-source data mining products.

Excel has long been denigrated by BI vendors, but continues to be extremely popular with users and business analysts alike, including for use with big data [Healy, 2012]. Excel can handle a million rows of data, can source data from virtually any database or BI product, and has powerful native or third party plug-in analytical features. The plug-ins include PowerPivot (for advanced pivot table capabilities) and Data Mining (which requires SQL server) from Microsoft, and Analyse-IT (for statistics), Excellent Analytics (for importing Web analytics data from Google Analytics), and Unistat (for statistics).

People Skilled in the Use of Analytics

Big data is creating jobs. Gartner [2012] predicts that by 2015 the need to support Big Data will create 4.4 million IT jobs globally, with 1.9 million of them in the U.S. For every IT job created, an additonal three jobs will be generated outside of IT.

Big data is also creating a high demand for people who can analyze and use big data. 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 [Manyika, et al., 2011]. This shortage is significant because inadequate staffing and skills are the leading barriers to Big Data analytics [Russom, 2011].

When thinking about big data analytics, it is useful to consider a continuum of users, anchored at one end by end users, with analysts in the middle, and data scientists at the other end; see Figure 5 [Watson, 2013a]. Each group requires different skills when it comes to working with big data, including a mixture of business, data, and analytics expertise.

Business users access big data-related information through reports, OLAP, dashboards/scorecards, and data visualization tools. They are information consumers rather than information creators. Two things

stand out about big data analytics in comparison to analytics in general. The first is the ability to analyze and display more kinds of information than ever before. For example, think of the fleet managers at U.S. Xpress who monitor the conditions of their trucks on iPads and route them in for servicing when sensor data indicates that it is necessary. The second is the ability to provide more context for decision making by combining information from multiple sources. To illustrate, rather than just providing sales data on a new product, it is possible to analyze what consumers are saying about it (think of the Starbucks' example).

Business users should have extensive business domain knowledge and understand the potential of big data analytics. They need to understand what data is available and be able to access it, manipulate it in simple ways, and use it to create business value. They don't need to be experts on the details of the algorithms and models used to analyze big data.

There are two major categories of analysts. BI analysts are part of a BI or analytics department and work throughout the organization. Business analysts, on the other hand, work in a business unit (e.g., marketing) and perform their analytics work there. Both categories represent information producers rather than information consumers. BI analysts typically understand the organization's data and available tools better than the business analysts. For example, they might implement an enterprise-wide scorecarding system. MIS graduates are good fits for this position. The business analysts, however, understand their business unit's use of analytics better. For example, a supply chain analyst might work to optimize supply chain processes, from sourcing raw materials to distributing products to the point-of-sale. Many organizations have both kinds of analysts who work both separately and collaboratively on projects.

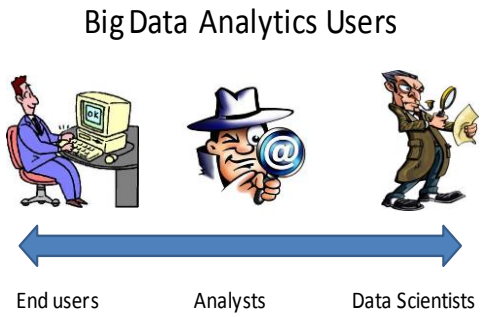


Figure 5: A continuum of big data analytics users.

Data scientists are at the far right of the continuum. These are the highly trained, skilled, and experienced professionals who discover new insights in big data. The term "data scientist" is relatively new and is often attributed to D.J. Patil and Jeff Hammererbacher, who are in charge of data and analytics at

LinkedIn and Facebook, respectively. [Davenport and Patil, 2012]. Data scientist has been called "the sexiest job of the 21st century" [Davenport and Patil, 2012].

The job of data scientists is to discover patterns and relationships that no one else has seen or wondered about, and turn these discoveries into actionable information that creates value for the organization. To do this requires a rich mixture of skills. Data scientists need to understand the different types of big data and how they can be stored (e.g., RDBMS, Hadoop), write code (e.g., Java, Python, R), access data (e.g., SQL, Hive), analyze it (e.g., regression analysis, social networks), and communicate findings to management in business terms (e.g., briefings, reports). Data scientists are typically very curious, like to solve difficult problems, and have advanced degrees (often PhD's) in analytical fields, such as statistics, management science/operations research, computer science, and mathematics. Because of these requirements, data scientists are in short supply and command high salaries. Fortunately, organizations don't need many of them.

Many data scientists don't have degrees in business. To be effective, however, they need to understand the specific industry and organization in which they work. To make up for this potential deficiency, it is common to have data scientists work closely with people in the organization who have business domain knowledge.

Big data analytics users make up a continuum. The reason is that there are users all along the continuum. For example, some end users become power users in their departments and perform analyst roles, such as developing reports for other users. There are analysts who further develop their skills and are able to perform some of the tasks associated with data scientists, especially when provided analytical workbenches like SAS Enterprise Miner and IBM SPSS Modeler.

Meeting the Demand

As the McKinsey Global Institute and Gartner studies show, there is a growing demand for people who can work with analytics and big data. Universities, companies, and vendors are responding in a variety of ways.

Many universities are adding a required course in analytics (including big data analytics) to their undergraduate and graduate programs. From a historical perspective, this is interesting because most business schools used to require a course in management science or quantitive methods but phased it out when the market place did not require it. Obviously, businesses now want graduates with analytical skills.

Another significant change is the rapid emergence of undergraduate degree programs, certificates, MBA concentrations, and graduate degree programs in analytics. To illustrate the growth in analytics offerings,

a survey in fall 2012 found 59 universities offering a business intelligence/business analytics degree, with 22 at the undergraduate level. Only two years earlier, a similar survey found only 12 schools that offered a BI/BA degree [Watson, et al., 2013b].

Graduate analytics offerings are located across campus, including in business, engineering, and statistics. The instructional delivery varies from on-campus to online. One of the first and best known programs is the Master of Science in Analytics at North Carolina State University. SAS has been an important contributor to the program, which is offered through the Institute for Advanced Analytics and has its own facility on campus. Deloitte Consulting partnered with the Kelly School of Business at Indiana University to offer a certificate in business analytics for Deloitte's professionals. Last year, Northwestern University initiated an online Master of Science in Predictive Analytics offered through its School of Continuing Studies. Still unknown, however, is how many students will choose to study big data analytics, because it is an intellectually challenging field of study.

Companies are responding in several ways. One approach is to create educational opportunities for people who are already on board (e.g., business and BI analysts) and have the interest and aptitude to further develop their skills through in-house programs, conferences, and college courses. In the past, there have been dire forecasts of shortages of people with specific skills (even COBOL programmers), and people acquired the needed skills and took advantage of the opportunities [Healy, 2012].

For companies that want to hire data scientists, Davenport and Patil (2012) have useful ideas for how to proceed, including scanning the membership rolls of user groups dedicated to data science tools like R; looking for them on LinkedIn; engaging them at conferences like Strata, Structure:Data, and Hadoop World; and hosting a competition on Kaggle or TopCoder, the analytics and coding sites.

Leading BI and analytics vendors like IBM, Oracle, SAP, Microsoft, SAS, and Teradata sponsor university alliance programs that make software, case studies, research reports, and more available either free or at minimal cost to universities. In analytics classes, it is important for students to have hands-on experiences with the software they will encounter in the workplace, and the alliance programs are especially good for meeting this need. For example, Oracle makes its database as well as reporting and analysis software (e.g., Hyperion) available. SAS offers its data mining/predictive analytics software (e.g., SAS Enterprise Miner). IBM provides Cognos (reporting and analysis) and SPSS Model (data mining/predictive analytics).

For the past 13 years, the author has been involved with the Teradata University Network (www.teradatauniversitynetwork.com), a free portal for faculty and students with interests in analytics, business intelligence, data warehousing, and database. Through the portal, faculty and students can access software (e.g., Teradata, MicroStrategy, Tableau, and SAS), large data sets (offered through the

University of Arkansas), articles, web seminars, cases, assignments, course syllabi, and more. Plans are in place to add Teradata Aster, a big data platform, to the portal.

VIII. BIG DATA AND PRIVACY

The collection, storage, and mining of big data will only increase. A big data issue that is gaining attention and will become more important is individual privacy: What data should the government and organizations be allowed to collect and what safeguards should be in place about how it is used? The Target story provides a glimpse into the uses that make some libertarians uneasy. Other people see no problem with this use of big data because it results in better customer service and appealing offers.

In Summer 2013, there were numerous news stories about Edward Snowden, a former CIA employee and NSA consultant, who revealed secrets about how the U.S. government was tracking private individuals' phone calls and using this information. Many people, especially in the government, viewed Snowden as a traitor who jeopardized the country's national security in the fight against terrorism, while others saw him as a whistle-blower who informed the public about a practice that threatens civil liberties.

Clemons, et al. [2014] identify three different ways to characterize privacy and online invasions of privacy. The first way is an uninvited intrusion into a user's personal space. This includes online marketing, spam advertising, pop-ups, and sponsored sites around the edges of a Web page. In their study, most people see this is as the most salient invasion of privacy invasion, even though its potential consequences are the least harmful. The most serious threats are fraudulent e-commerce transactions and identity theft. Although these are concerns, people are not worried that these kinds of activities would perpetuated by big data companies like Google and Facebook. The third kind of invasion is personal profiling for commercial advantage. This occurs when companies like Google, Facebook, and Yahoo! combine hundreds or thousands of pieces of data from different sources (i.e., data blending) to understand who you are, where you live, where you go, who your friends are, what you buy, and the like. This blended information may be used to simply make offers that are likely to appeal to you, or for less benign purposes such as knowing if you engage in risky hobbies and should be charged higher insurance rates [Clemmons, et al., 2014].

Research shows that most people have very little understanding and concern about how organizations are using big data [Clemmons, et al., 2014]. However, as individuals understand the potential uses better, their concerns increase quickly. This suggests that as companies increasingly use big data analytics on customer data, the public is likely to become more concerned.

Although there are laws that limit the activities of telecommunications companies (e.g., they can't listen to phone calls), there are few regulations and laws applicable for the new digital age and they are largely

non-existent for Internet firms. These companies own privacy policies largely serve their commercial interests rather than protecting individuals' privacy. We need laws about individual privacy that people think that are consistent, reasonable, transparent, and easy to understand [Clemons, et al., 2014].

VIIII. CONCLUSION

From a historical perspective, big data can be viewed as the latest generation in the evolution of decision support data management [Watson and Marjanovic 2013c]. The need for data to support computer-based decision making has existed at least since the early 1970s with *DSS*. This period can be thought of as the first generation of decision support data management. It was very application-centric with data organized to support a single decision or a set of related decisions. By the 1990s, there was a need to support a wide variety of BI and analytic applications (e.g., reporting, executive information systems) with data. Having separate databases (i.e., independent data marts) for each application was costly, resulted in data inconsistencies across applications, and failed to support enterprise-wide applications. The outcome was the emergence of *enterprise data warehouses* (the second generation), which represented a data-centric approach to data management. The next generation (the third) was *real-time data warehousing*. Technology had improved by 2000 so that it was possible to capture data in real time and trickle feed it into the data warehouse. The significance of this evolution is that it changed the paradigm for what kinds of decisions could be supported. With real time data, operational decisions and processes could be supported. *Big data* is the fourth generation decision support data management. The ability to capture, store, and analyze high-volume, high-velocity, and high-variety data is allowing decisions to be supported in new ways. It is also creating new data management challenges.

For many years, companies developed data warehouses as the focal point for data to support decision making. This is changing, as new data sources, platforms, and cloud-based services have emerged. As a result, data is becoming more *federated*; that is, data is stored and accessed from multiple places. Adding to this trend are business units such as finance and marketing that have the business need, resources, and political clout to acquire their own platforms, services, and tools. In many organizations, IT is losing some control over data management. This is not bad if it leads to more agility and better organizational performance. The downside, however, includes data silos that don't share data, data inconsistencies, inefficiencies in storing data, and duplication of resources. Organizations are accepting that data federation is going to exist, at least in the short to medium term, and are instituting greater controls over their data management practices. Some are putting more emphasis on data governance (e.g., data stewards, metadata management, and master data management). They are also creating BI or analytics centers of excellence to provide strategic direction for the use of data and analytics, prioritize projects, provide shareable resources, establish guidelines and standards, participate in tool selection, troubleshoot problems, and more.

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 [Russom, 2012]. 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. Depending on the project(s), the sponsorship can be departmental or at the senior executive level. The CIO is typically responsible for developing and maintaining the big data infrastructure. 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 (e.g., data warehouses) in a way that meets organizational needs cost effectively. The analysis of big data requires traditional tools like SQL, analytical workbenches (e.g., 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. As organizations make greater use of big data, it is likely that there will be increased concerns and legislation about individual privacy issues.

REFERENCES

Ayers, I. (2007) *Super Crunchers*, New York: Bantam Books.

Babcock, C. (2011) "Zynga's Unusual Cloud Strategy Is Key to Success," *Information Week*, July 1. Available at http://www.informationweek.com/cloud-computing/infrastructure/zyngas-unusual-cloud-strategy-is-key-to/231000908

Babcock, C. (2013) "Zynga, Cloud Pioneer, Must Fix Revenue Woes," *Information Week*, June 4. Available at http://www.informationweek.com/cloud-computing/software/zynga-cloud-pioneer-must-fix-revenue-woe/240156007

Brynjolfsson, E., L. Hitt, and H. Kim"(2011) "Strength in Numbers: How does data-driven decision-making affect firm performance?," *Social Science Research Network,* April. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486

Cifio, J., and C. Meley (2011) Presentation at the Teradata Universe Conference, Barcelona, April 11.

Clemons, E., J. Wilson, S. Barnett, F. JIN, and C. Matt (2014) “Investigations into Consumers Preferences Concerning Privacy: An Initial Step Towards the Development of Modern and Consistent Privacy Protections Around the Globe,” *Proceedings of the Hawaii International Conference on Systems* Sciences, Big Island, Hawaii, January.

Cooper, B.L., H.J. Watson, B.H. Wixom, and D.L. Goodhue (2000) "Data Warehousing Supports Corporate Strategy at First American Corporation," *MIS Quarterly*, (24)4, pp. 547-567.

Davenport, T.H. and J.G. Harris (2007) *Competing on Analytics: The New Science of Winning,* Boston: Harvard Business School Press.

Davenport, T. H., Harris, J.G. and R. Morison (2010) *Analytics at Work: Smarter Decisions, Better Results,* Boston: Harvard Business School Press.

Davenport, T.H. and D.J. Patil (2012) “Data Scientist: The Sexiest Job of the 21st Century,” *Harvard Business Review*, October, pp. 2-8.

Duhigg, (2012) “How Companies Learn Your Secrets,” *New York Times,* February 16. Available at http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0

Eckerson, W. (2004) “Gauge Your Data Warehousing Maturity,” *DM Review* (14)11, pp. 34.

Eckerson, W. (2011) “Big Data Analytics: Profiling the Use of Analytical Platforms in User Organizations,” TDWI, September. Available at http://tdwi.org/login/default-login.aspx?src=%7bC26074AC-998F-431B-BC99-4C39EA400F4F%7d&qstring=tc%3dassetpg

Economist Intelligence Unit (2012) “Fostering a Data-driven Culture,” *The Economist*, October. Available at http://www.managementthinking.eiu.com/sites/default/files/downloads/Tableau_DataCulture_130219.pdf

Franks, B. (2012) *Taming the Big Data Tidal Wave*, New York: Wiley.

Gartner (2012) “Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015,” Gartner Press Release, October 22. Available at http://www.gartner.com/newsroom/id/2207915

Harris, D. (2013) “The history of Hadoop: From 4 nodes to the future of data,” Gigaom, March 4. Available at http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/

Healy, M. (2012) “6 Big Data Lies,” Information Week Research Report, November. Available at http://twimgs.com/audiencedevelopment/JT/OE/LPs/SLP1_IBM/Webcast_SLP1_research-big-data-smart-data_59742.pdf

Hill, K. (2012) “How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did,” *Forbes*, February 16. Available at

http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

Imhoff, C. (2013) “Seeing RedShift: How Amazon Changed Data Warehousing Forever,” *Inside Analytics*, April 9. Available at http://archive.constantcontact.com/fs108/1104983460042/archive/1112865946329.html

Kavanagh, E. (2013) “Throwing Down the Gauntlet: Cloud BI for $1/hr,” *Inside* Analytics, July 22. Available at http://insideanalysis.com/2013/07/throwing-down-the-gauntlet-cloud-bi-for-1hr/?utm_source=Throwing+Down+the+Gauntlet%3A+Cloud+BI+for+%241%2Fhr&utm_campaign=Advance&utm_medium=email

Lewis, M. (2003) *Moneyball: The Art of Winning an Unfair Game,* New York: W.W. Norton & Company.

Manyika, et al., (2011) “Big Data: The Next Frontier of Innovation, Competition, and Productivity,” McKinsey Global Institute, May. Available at http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Miller, B. (2011) *Moneyball*, Columbia Pictures.

Miller, S., S. Lucas, L. Irakliotis, M. Ruppa, T. Carlson, and B. Perlowitz (2012) “Demystifying Big Data: A Practical Guide to Transforming the Business of Government,” Washington: TechAmerica Foundation. Available at http://breakinggov.com/documents/demystifying-big-data-a-practical-guide-to-transforming-the-bus/

O’Brien, J. (2012) Presentation at The Data Warehousing Institute conference, Las Vegas. February 2012.

Power, D.J. (2007). “A Brief History of Decision Support Systems,” *DSSResources.com*, Available at http://DSSResources.COM/history/dsshistory.html, version 4.0.

Read, K. (2013) “Is ‘In-Memory’ Always the Right Choice?” *Business Intelligence Journal* (18)1, pp. 46-50.

Rudin, K. (2010) “Actionable Analytics at Zynga: : Leveraging Big Data to Make Online Games More Fun and Social,” TDWI BI Executive Summit, San Diego, August. Available at http://tdwi.org/videos/2010/08/actionable-analytics-at-zynga-leveraging-big-data-to-make-online-games-more-fun-and-social.aspx

Russom, P. (2011) “Big Data Analytics”. TDWI Best Practices Report. Seattle: The Data Warehousing Institute, Fourth Quarter. Available at http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx

Russon, P. (2013) “Integrating Hadoop into Business Intelligence and Data Warehousing,” TDWI Best Practices Report. Seattle, The Data Warehousing Institute. Second Quarter. Available at http://tdwi.org/research/2013/04/tdwi-best-practices-report-integrating-hadoop-into-business-intelligence-and-data-warehousing.aspx

Schroeck, M., R. Schockley, J. Smart, D. Romero-Morales and P. Tufano, P. (2012) *Analytics: The real-world use of big data,* IBM Institute for Business Value. Available at http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html

Sicular, S. (2013) "Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s," *Forbes*, March 27. Available at http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/

Sondergaard, P. (2012) Gartner Symposium/ITxpo, Orlando. Available at http://www.gartner.com/newsroom/id/2207915

van Groningen, M. (2009) "Introduction to Hadoop," *TRIFORK Blog*, August 4. Available at http://blog.trifork.com/2009/08/04/introduction-to-hadoop/

Watson, H.J. and L. Volonino (2000) "Harrah's High Payoff from Customer Information," W. Eckerson and H.J. Watson (eds.) *Harnessing Customer Information for Strategic Advantage: Technical Challenges and Business Solutions*," TDWI. Available at www.teradatauniversitynetwork.com

Watson, H. J. (2009a) "Tutorial: Business Intelligence – Past, Present, and Future," *Communications of the Association for Information Systems* (25)39. Available at: http://aisel.aisnet.org/cais/vol25/iss1/39

Watson, H.J., J.A. Hofer, and B.H. Wixom (2009b) "RetailStore.com," Teradata University Network. Available at http://teradatauniversitynetwork.com

Watson, H.J. and T. Leonard (2011) "U.S. Xpress: Where Trucks and BI Hit the Road," *Business Intelligence Journal*, (16)1, pp. 4-7.

Watson, H.J. (2012) "This Isn't Your Mother's BI Architecture," *Business Intelligence Journal,* (17)1, pp. 4-6.

Watson, H.J. (2013a) "All about Analytics," *International Journal of Business Intelligence Research* (4)2, pp.13-28.

Watson, H.J., Wixom, B.H, and T. Ariyachandra (2013b) "Insights on Hiring for BI and Analytics," *Business Intelligence Journal* (18)2, pp. 4-7.

Watson, H.J and O. Marjanovic (2013c) "Big Data: The Fourth Data Management Generation," *Business Intelligence Journal* (18)3, pp.4-7.

Williams, S. (2004) "Assessing BI Readiness: A Key to BI ROI," *Business Intelligence Journal*, Summer 2004.

ABOUT THE AUTHOR

Dr. Hugh J. Watson is a Professor of MIS and a holds a C. Herman and Mary Virginia Terry Chair of Business Administration in the Terry College of Business at the University of Georgia. Hugh is a leading scholar and authority on business intelligence and analytics, having authored 24 books and over 150

scholarly journal articles. He helped develop the conceptual foundation for decision support systems in the 1970's, researched the development and implementation of executive information systems in the 1980's, and for the past twenty years has specialized in data warehousing, BI, and analytics. He is a Fellow of the Association for Information Systems and TDWI and is the Senior Editor of the *Business Intelligence Journal.* Hugh is the founding Director of the Teradata University Network, a free portal for faculty who teach and research BI/DSS, analytics, data warehousing, and database management.