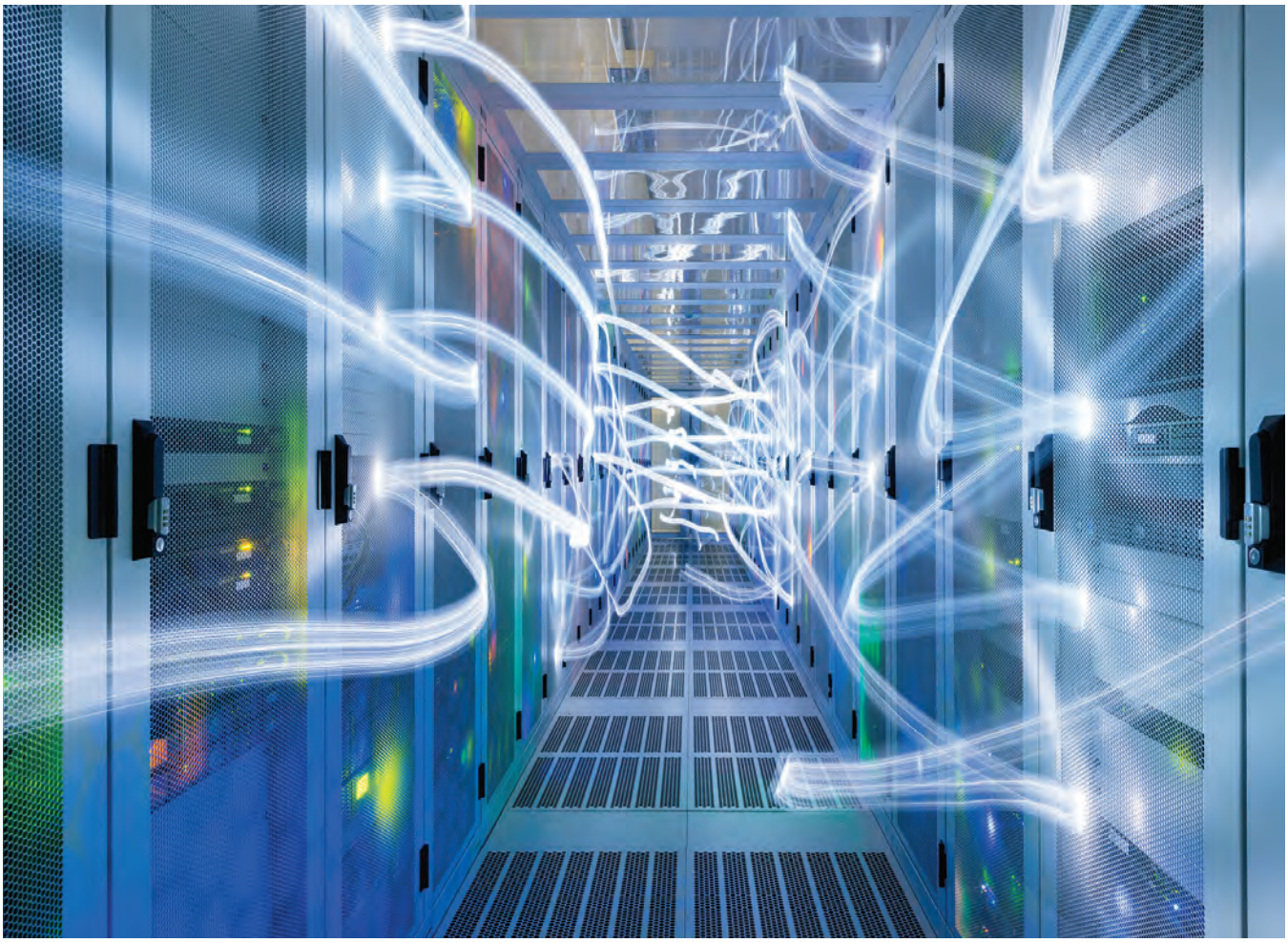

What CPAs Need to Know about Big Data

Opportunities for Businesses and Concerns about Privacy

By P. Paul Lin

Big data is an emerging IT tool that allows businesses to analyze vast collections of financial and nonfinancial information in order to create competitive advantages. CPAs can use big data analytics to improve performance in numerous areas, including sales, operation efficiency, risk management, and fraud detection (Chris Eaton, Dirk deRoos, Tom Deutsch, George Lapis, and Paul Zikopoulos, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data,* IBM, 2012, http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF).

Big Data in the Spotlight

Big data has been utilized by large retailers to understand customer behavior and drive sales. Sears Holdings used big data techniques to combine huge amounts of customer, product, and promotional data from its Sears, Craftsman, and Lands' End brands in order to generate more personalized promotions. This would have taken about eight weeks using traditional data-processing methods—far too long to be useful, given that some of the products would be out of season by the time the analysis was complete; however, big data reduced the processing time to one week (Andrew McAfee and

Erik Brynjolfsson, "Big Data: The Management Revolution," *Harvard Business Review*, October 2012). Similarly, using big data to analyze the purchasing patterns of about 25 products, Target assigned each shopper a "pregnancy prediction score" and found out about a teenage girl's pregnancy before her father did (Charles Duhigg, "How Companies Learn Your Secrets," *New York Times*, Feb. 16, 2012). Although such a use reveals the power of big data, it also raises potential privacy concerns, which will be discussed later in greater detail.

In addition to affecting retailers, big data has also changed how farmers work. For example, the *Wall Street Journal* recently reported that agricultural companies take advantage of big data to roll out "prescriptive planting" for farmers in the United States. These companies contend that farmers can use big data techniques to increase the harvest of crops, such as corn and soybeans. One of the leading agricultural companies has even estimated that data-driven planting techniques could increase worldwide crop production by about $20 billion per year (Jacob Bunge, "Big Data Comes to the Farm, Sowing Mistrust: Seed Makers Barrel into Technology Business," *Wall Street Journal*, Feb. 25, 2014).

Big data also has applications in the healthcare sector. For example, using clusters of search terms by region in the United States, Google could predict flu outbreaks faster than—and as accurately as—the Centers for Disease Control and Prevention (CDC), which relies on hospital admission records (Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patei, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature*, vol. 457, Feb. 19, 2009, pp. 1012–1014). Another researcher used cell phone data to study outbreaks of communicable diseases, such as cholera in Africa. Specifically, the researcher used the time and locations of mobile phone records to map people's daily and weekly commuting patterns in Rwanda. The mapped patterns predicted outbreaks in areas that recently experienced floods, because floods wash away roads and thus lead to increased susceptibility to a short-term communicable disease outbreak (Jonathan Shaw, "Why 'Big Data' Is a Big Deal: Information Science Promises to Change the World," *Harvard Magazine*, March–April 2014).

As shown by these examples, big data can have a major impact on businesses. CPAs need to adapt to this data-driven environment. In a Chartered Global Management Accountant (CGMA) study, 85% of accountants believed that increasing their ability to work with big data will enhance their career and employability ("From Insight to Impact: Unlocking Opportunities in Big Data," Chartered Institute of Management Accountants, 2013, http://www.cgma.org/Resources/Reports/Pages/insight-to-impact-big-data.aspx). The ensuing discussion may help CPAs to understand exactly what big data is and what it means for the accounting profession.

What Is Big Data?

Much of the confusion about "big data" starts with its definition. There is no consensus definition of big data today.

EXHIBIT 1
Possible Myths about Big Data

Myths	Facts
Given the "accounting entity concept," accountants should focus on their own transactional data and financial reporting.	Connecting existing transactions with external data can shed more light on the "connected" data in order to generate critical insights for businesses' enhancement.
Given the "monetary unit concept," accountants should focus on structured financial data.	Big data goes beyond "debits and credits." The unstructured data (e.g., location and time stamp of text messages) and streaming data on social media can be a gold mine for business analytics.
Big data is for customer profiling only.	Big data can be used for a variety of applications, including personalized marketing, logistics, operational efficiency (e.g., optimizing wind turbine operations at a power plant), public health (e.g., reacting to an epidemic), hospital patient care (e.g., faster diagnoses and treatments), farming, risk management, and fraud detection.
Big data is for computer scientists, not accountants.	Understanding the business and financial data, accountants can play an important role in interpreting critical insights to generate actionable plans for businesses. In addition, accountants can provide ongoing evaluations to make sure that big data meets the preset project objectives.
Big data will replace accounting data someday.	Big data platforms consist of a cluster of servers that host vast amount of data, including the accounting data that is indispensible for daily business operations.

EXHIBIT 2

Cases of Using Proxy Measures for Business Enhancement

Entity	Objective	Original Approach	Alternate Approach	Breakthrough	Results
Amazon	Increase sales	Providing book reviews and critiques to help customers	Providing personalized recommendations based on customers' shopping preferences	Using the associations between products to develop "item-to-item" collaborative technique	Recommendations based on these analytics contribute to one-third of sales at Amazon.
Fair Isaac Corporation (FICO)	Make sure patients take required medications	Offering uniform reminders (e.g., instructions)	Identifying patients who are more likely to need a reminder	Using a wealth of variables to come up with "Medication Adherence Scores" for various patients	Targeted reminders for a smaller group of patients helps healthcare providers save money.
Experian	Know or verify a customer's income level	Obtaining a copy of tax return from the IRS for about $10	Using less expensive information	Discovering correlations between credit histories and income levels	Information is available for less than $1.
Aviva (insurance)	Identify applicants at higher risk of illness	Performing lab tests ($125 per person)	Using lifestyle data as proxy indicators	Using daily life variables, along with estimated incomes, to identify applicants' health risks	Potential customers can apply for health insurance without lab tests.
UPS	Maintain a fleet of 60,000 vehicles in the United States	Scheduling maintenance (even when some parts are fine)	Replacing parts when necessary to save money	Monitoring individual parts on trucks to predict the proper maintenance time	UPS saves millions annually.
University of Ontario Institute of Technology and IBM	Care for premature babies	Monitoring babies and prescribing treatments if necessary	Developing early warning systems for caring premature babies	Tracking 16 different real-time data streams, which amount to about 1,260 observations per second, including heart beat, respiration rate, temperature, blood pressure, and blood oxygen level	The systems can signal the onset of infections 24 hours before symptoms appear, which calls for more proactive and lighter medical interventions. It also alerts doctors sooner if a treatment seems ineffective.
MasterCard	Operate credit card business	Processing transactions for fees	Analyzing transactions and selling the "new" information to generate additional revenues	Discovering, for example, that people who fill up their gas tank around 4:00 pm are likely to spend $35–$50 at a grocery store in the next hour	MasterCard sells the information to a marketer who can print coupons for a nearby supermarket on the back of the gas receipt.
Accenture and Metro St. Louis, Mo.	Find the optimal time for engine overhauls on public buses	Using "check engine light" (CEL) warning systems	Developing early warning systems to replace the CEL warning	Using wireless sensors to monitor bus engines and predict breakdowns	Project evolved into sensors-embedded "smart buses" at additional cost of $10,000 per bus, which can save as much as $100,000 over each bus's life.*

Source: Information compiled from *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, by Viktor Mayer-Schönberger and Kenneth Cukier, Houghton Mifflin, 2013

* Ken Leiser, "'Smart' Metro Buses Roll Extra Miles in St. Louis between Breakdowns," *St. Louis Post-Dispatch,* http://www.stltoday.com/news/traffic/along-for-the-ride/smart-metro-buses-roll-extra-miles-in-st-louis-between/article_c5dcc5c1-1d79-5f81-a6cb-0a1caeef1fd2.html

Although the volume of data must be large, its size or scale is not the only criterion. Generally, three parameters are relevant to big data: velocity, volume, and variety. Velocity refers to the speed of data coming in—big data always has a high rate of collection. Facebook is a prime example of high velocity: there are 757 million Facebook active daily users, and they upload 350 million photos on an average day (Craig Smith, "By the Numbers: 170 Amazing Facebook User and Demographic Statistics," Digital Market Ramblings, Oct. 4, 2014). High data velocity often results in a large data volume, though the term "large" is relative, and the sizes of big data vary by application. Finally, unlike the uniform transaction data that accountants deal with, big data contains a variety of structured and unstructured information collected via various methods, including humans, sensors, machines, and social media feeds.

Big data takes advantage of the "distributed storage" and "distributed computing" models published by Google in the early 2000s, using a "divide and conquer" approach to handle enormous volumes of data (http://static.googleusercontent.com/media/research.google.com/en/us/archive/bigtable-osdi06.pdf). The big data platform consists of a cluster of stacked servers that perform data processing independently and parallel to one another. Data redundancy improves fault tolerance by maintaining several copies of the same data or files across the cluster. If a file, hard disk, or server fails, the data can be automatically replicated from another server. This data redundancy might consume more IT resources, but it enhances system stability, which is a critical requirement for businesses. Meanwhile, different servers can perform the same data-processing tasks, and companies use the results from the first server to respond in order to take action faster.

Another important feature of big data is its scalability. Companies can start small and increase their project's scale by adding more servers. The existing operating systems will work seamlessly with any newly added servers, regardless of whether the increase is 20% or tenfold. Ultimately, a lower barrier to entry will attract more organizations to join the big data bandwagon.

The vast amount of information collected on the Internet makes big data possible, and some businesses are already monitoring society's daily activities in order to generate it. For example, one's activities on social media sites like Facebook or Twitter, searches on Google, shopping on Amazon, and the time and location of mobile phone calls and text messages are all recorded on servers. Companies can use big data technologies to analyze the collected information for batch or real-time business analytics in order to figure out what people like, who their friends are, how they live, and much more.

Exhibit 1 presents some popular myths about big data and offers explanations to help understand it. It is worth mentioning that, in addition to operational efficiency, big data analytics also serve accountants well in more challenging areas, such as risk management and fraud detection.

Big Data Analytics

Traditionally, a popular approach to problem solving has been to use data to identify the causes of problems and fix them accordingly; for example, accountants try to identify cost drivers, with the aim of controlling costs. But a focus on causes and effects might prevent people from considering alternatives, especially if they are confident in their expertise and knowledge, and the failure to think outside the box might have negative consequences. To make matters worse, causal links between two accounts and activities are difficult to prove statistically.

In contrast with traditional methods, big data focuses on the statistically easy-to-see "what," but stays away from the difficult-to-prove "why." Specifically, it uses the correlations that exist in the variables in order to generate valuable insights for improvement. In addition, the robustness of correlations—if any—can be easily and statistically identified and proven by powerful analytics software. Consequently, big data uses the correlations between variables to come up with cheaper proxy measures that help businesses solve problems quicker than is possible with traditional methods.

The abundance of public and personal data on the Internet (e.g., public posts on Facebook) further facilitates big data analytics. For example, two researchers at Carnegie Mellon University revealed that it is possible to predict people's Social Security numbers (SSN) with minimal data that can be found on the Internet (Alessandro Acquisti and Ralph Gross, "Predicting Social Security Numbers from Public Data," Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 27, Jul. 7, 2009, http://www.pnas.org/content/106/27/10975.full). Using only publicly available data, they identified a correlation between individuals' SSNs and their date of birth and birthplace; this correlation facilitates the statistical inference of SSNs, especially for younger cohorts. These inferences are facilitated by the Social Security Administration's Death Master File database and the availability of personal information from various sources.

Exhibit 2 reports some examples of big data analytics and explains how proxy measures were created in order to achieve various goals, including increasing revenues, reducing costs, or improving operational effectiveness and efficiency. (To learn more about the development of analytics projects, see the aforementioned *New York Times* article by Charles Duhigg, which covered real cases at Target and Procter & Gamble.)

Today's Big Data Practices

In November 2013, Enterprise Management Associates, Inc. (EMA) and 9sight Consulting conducted a study on big data strategies and practices ("Operationalizing the Buzz: Big Data 2013," http://wwwcdn.actian.com/wp-content/uploads/2014/02/EMA-BigData-2013-Operationalizing-the-Buzz.pdf). To offer an unbiased view, the subjects of the study consisted of 51% business stakeholders and 49% IT professionals and consultants; study participants were involved with a total of 597 active big data projects.

Exhibit 3 reveals that the majority (69%) of big data projects involve operational analytics and operational processing in order to confer competitive advantages, which aligns well with the aforementioned 2013 CGMA survey results. For example, the study revealed that better data quality and analysis are the most beneficial in the following five business areas:

- Identifying opportunities to increase efficiency or save costs
- Developing and monitoring key performance indicators

- Engaging in driver-based forecasting
- Monitoring external risks
- Increasing revenues.

The EMA/9sight study also revealed that higher-level business professionals (line executives versus analysts) are more likely to use big data tools, which might result in more business departments, rather than IT departments, sponsoring big data projects (49.8% versus 21.8%). About one-third (34.7%) of big data projects were deployed using existing hardware and software. Another study by TDWI Research in 2013 found that one-quarter of surveyed organizations were able to scale up preex-

EXHIBIT 3

Survey of Big Data Practices

Project Goals and Characteristics	Frequency
Operational analytics (e.g., fraud analysis, risk assessment, customer relationship management, staff scheduling, logistical planning, campaign optimization, and market basket analysis)	47.9%
Operational processing (e.g., billing, rating, point of sale, and customer care)	21.1%
Relationship and behavioral analyses (e.g., clustering, grouping and relationship analysis, path analysis, and customer churn analysis)	16.3%
Social brand/sentiment analyses (e.g., opinion mining, sentiment analysis, and social brand management analysis)	14.7%
Users of Big Data Projects	
Line-of-business executives	28.3%
Business analysts (e.g., marketing or finance)	21.6%
Database administrators and data analysts	16.2%
Data scientists*	10.7%
External users (e.g., customers)	10.2%
Report writers and dashboard builders	7.4%
Application developers	5.5%
Project Funding Sponsors (Top 5 Only)	
IT	21.8%
Finance	15.1%
Marketing	14.1%
Sales	12.6%
Corporate CEO	8.0%
Required Computing Resources for Big Data Projects	
Employing existing hardware and software	34.7%
Purchasing additional on-premises hardware	19.1%
Purchasing additional on-premises software	16.1%
Using additional cloud-based infrastructure as a service (IaaS)	15.1%
Operating additional cloud-based platforms as a service (PaaS)	8.4%
Using additional cloud-based software as a service (SaaS)	6.7%
Average Budget for One Big Data Project	
From $275,000 to $2.5 million	

*Data scientists are professionals who have the knowledge, training, skills, and curiosity to induce critical insights out of big data.

Source: Enterprise Management Associates, Inc., and 9sight Consulting, "Operationalizing the Buzz: Big Data 2013," http://wwwcdn.actian.com/wp-content/uploads/2014/02/EMA-BigData-2013-Operationalizing-the-Buzz.pdf

isting IT infrastructure in order to handle their big data needs (Philip Russom, "Managing Big Data," Oct. 1, 2013, http://www.tdwi.org/research/2013/10/tdwi-best-practices-report-managing-big-data.aspx). By using existing IT resources for big data projects, some companies can enjoy their benefits without incurring additional costs up front.

Implications for CPAs

Identifying the correlations in data is simple for computers, but understanding the implications of those correlations is more critical and difficult. Correlations by themselves are not actionable plans; they must be investigated further in order to provide insights that could lead to actionable solutions. But new insights often require more than conventional accounting data, making big data a viable method for businesses to improve their competitive advantages.

In the previously mentioned study, 86% of participants indicated that their companies were struggling to get valuable insight from data, and more than 90% agreed that finance professionals can play an essential role in helping businesses benefit from data-related projects. Thus, "data-driven predictions can succeed—and they can fail. It is when we deny our role in the process that the odds of failure rise" (Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*, Penguin Press, 2012, p. 9).

CPAs can use their extensive knowledge of a company's operations and operational data, coupled with their understanding of finance, to transform analytical insights into competitive advantages. Accountants need to recognize that, in addition to conventional financial information, big data also uses unstructured, external data to generate relevant analytical insights and provide valuable additional information that is lacking when analyzing only the structured internal data collected for financial reporting. Consequently, CPAs must go beyond the business-entity principle and monetary-unit principle in order to take advantage of the new possibilities presented by big data.

Although big data offers worthwhile opportunities for generating valuable insights that can help businesses improve their bottom line, it does integrate more personal information into a company's business analytics. Profiling can generate personalized recommendations and promotions, but it also creates additional parameters for customer classification that might expose a business to potential discrimination lawsuits. To return to the example offered at the beginning of the article, about one year after Target implemented its pregnancy-prediction model, a man walked into a Target store outside Minneapolis and demanded to see the manager because his teenage daughter had received coupons for baby clothes and cribs. Because he was not involved in the predictive analytics project, the manager did not know what the man was talking about. The manager apologized, and then called the man a few days later to apologize again. On the phone, the man revealed that he had later discovered that his daughter was pregnant after all. Target eventually realized that it needed to get its advertisements into expectant mothers' hands without seeming like the company was spying on its customers (Duhigg 2012). In fact, "gathering and combining big data from every conceivable source may enable the type of discriminatory practices based on race, religion, medical condition or sexual preference that have been illegal for years" (EMA and 9sight Consulting 2013).

Thanks to the scalability of big data, accountants can help businesses start small and focus on measurable outcomes.

In addition, big data might lead to additional concerns about consumer privacy. It will likely become more common for companies to work with the vendors of data services, which allows companies to access data from other sources in addition to their own. This privacy issue is an important concern, especially when a lot of personal information is included for profiling (Niko Karvounis, "What Should You Tell Customers about How You're Using Data?" *Harvard Business Review*, Oct. 9, 2012). As big data becomes more widespread, the issue will become more significant. Whereas businesses are interested in using the power of big data, consumers are beginning to inquire about their right to prevent the collection and use of their personal data (Alex "Sandy" Pentland, "Big Data's Biggest Obstacles," *Harvard Business Review*, Oct. 2, 2012).

Good data governance cannot be overemphasized. According to Data Management International (DAMA), data governance is the exercise of authority, control, and shared decision making over the management of data assets (http://www.dama.org/i4a/pages/index.cfm?pageid=1). Accountants can play a key role in enabling data governance and ensuring that it is aligned with an organization's overall corporate governance processes (http://www.aicpa.org/InterestAreas/InformationTechnology/Resources/BusinessIntelligence/DownloadableDocuments/Overview_Data_Mgmt.pdf).

Additional Considerations

Leveraging big data benefits not only large businesses, but also small and midsized ones. Although smaller businesses have fewer computing resources and fewer in-house IT skills, research suggests that they can tap into a growing number of easy-to-use applications and data services in order to reap the benefits of big data. Furthermore, their experience dealing with financial data and their knowledge of a business allows CPAs to play an important role in transforming critical insights into competitive advantages. Thanks to the scalability of big data, accountants can help businesses start small and focus on measurable outcomes, such as increased revenues, cost savings, or return on investment (ROI), in order to evaluate whether this emerging IT tool can improve their operations. CPAs who adapt to big data can transition from being data collectors and information providers to business enhancers. ❑

P. Paul Lin, PhD, *is an associate professor of accountancy at the Raj Soin College of Business, Wright State University, Dayton, Ohio.*