“With world Internet usage quintupling per decade, there is no upper limit on the number and value of new business opportunities for those who can bend the swelling flood of data to their purposes.”

— Ralph Hughes, Guest Editor

# Turning Big Data into Big Benefits

Opening Statement
by Ralph Hughes ................................................................. 3

Big Data Is a Solution — So Where’s the Problem?
by Richard Walsh, Richard O’Callaghan, and Sabine Yoffou .................. 6

Ensuring the Accuracy of Your Social Media Analysis
by Matthew Ganis and Avinash Kohirkar ...................................... 13

Traditional Data Warehousing Meets Big Data:
What Does It Mean for the Enterprise?
by Rich Johnson and Ron Zahavi ........................................... 19

A Semantic Web Strategy for Big Data
by Frank P. Coyle ............................................................... 28

Big Data Tools for Population Health Management
by Holly Korda, Ann Magee, and Lori Damiano ............................... 35

Big Data Analytics: Outsourcing vs. Crowdsourcing
by Saeed Lajami, Anson Mok, Mario Wahyu Prabowo, and Sara Cullen ....... 43
About Cutter IT Journal

Part of Cutter Consortium’s mission is to foster debate and dialogue on the business technology issues challenging enterprises today, helping organizations leverage IT for competitive advantage and business success. Cutter’s philosophy is that most of the issues that managers face are complex enough to merit examination that goes beyond simple pronouncements. Founded in 1987 as American Programmer by Ed Yourdon, Cutter IT Journal is one of Cutter’s key venues for debate.

The monthly Cutter IT Journal and its companion Cutter IT Advisor offer a variety of perspectives on the issues you’re dealing with today. Armed with opinion, data, and advice, you’ll be able to make the best decisions, employ the best practices, and choose the right strategies for your organization.

Unlike academic journals, Cutter IT Journal doesn’t water down or delay its coverage of timely issues with lengthy peer reviews. Each month, our expert Guest Editor delivers articles by internationally known IT practitioners that include case studies, research findings, and experience-based opinion on the IT topics enterprises face today — not issues you were dealing with six months ago, or those that are so esoteric you might not ever need to learn from others’ experiences. No other journal brings together so many cutting-edge thinkers or lets them speak so bluntly.

Cutter IT Journal subscribers consider the Journal a “consultancy in print” and liken each month’s issue to the impassioned debates they participate in at the end of a day at a conference.

Every facet of IT — application integration, security, portfolio management, and testing, to name a few — plays a role in the success or failure of your organization’s IT efforts. Only Cutter IT Journal and Cutter IT Advisor deliver a comprehensive treatment of these critical issues and help you make informed decisions about the strategies that can improve IT’s performance.

Cutter IT Journal is unique in that it is written by IT professionals — people like you who face the same challenges and are under the same pressures to get the job done. Cutter IT Journal brings you frank, honest accounts of what works, what doesn’t, and why.

Put your IT concerns in a business context. Discover the best ways to pitch new ideas to executive management. Ensure the success of your IT organization in an economy that encourages outsourcing and intense international competition. Avoid the common pitfalls and work smarter while under tighter constraints. You’ll learn how to do all this and more when you subscribe to Cutter IT Journal.

Please fill out the form below if you would like to subscribe to Cutter IT Journal.

Start my print subscription to Cutter IT Journal ($485/year; US $585 outside North America)

Name __________________________ Title __________________________

Company __________________________ Address __________________________

City __________________________ State/Province __________________________ ZIP/Postal Code __________________________

Email (be sure to include for weekly Cutter IT Advisor) __________________________

Fax to +1 781 648 8707, call +1 781 648 8700, or send email to service@cutter.com. Mail to Cutter Consortium, 37 Broadway, Suite 1, Arlington, MA 02474-5552, USA.
Interest in Big Data analytics (BDA) has certainly skyrocketed in the past few years to reach a fevered pitch, with the market for this technology projected to reach a 58% compounded annual growth rate over the next five years. Indeed, when I walked the vendor exhibit halls at several TDWI World Conferences during the past year, it seemed that nearly all the application vendors had introduced a new package offering a “Big Data” solution. At every booth, plenty of curious attendees lined up to hear about these new features. The vendors were certainly happy for the attention, but they also confided to me that they had grown tired of answering the same question day after day, namely “What is Big Data?”

I believe this lament is actually more emblematic of the state of BDA today than any particular solution being offered. When vendors rush to cater to needs that many customers do not yet understand, are we at risk of solving the wrong problem or cementing in place a basic strategy we will later regret? Perhaps at this early juncture we should carefully dodge the hype about Big Data and offer a sober appraisal of this new technology before acting.

LOOKING PAST THE HYPE

Industry pundits, in the area of data warehousing at least, take a jaundiced view of the buzz surrounding Big Data. “When haven’t business intelligence applications had to deal with ‘Big Data’?” they ask. Any type of data requires deliberate engineering to acquire, store, summarize, and present it in way that generates business insights. The cynics among us discount the fever over Big Data as a vendor-stoked overreaction to a few white papers by computer science wonks at Google and Yahoo! who found a couple of processing shortcuts while taming their own flood of Web stream data. These cynics see Big Data as a craze that will quickly fade. Such skepticism might be too extreme, however. New technologies do frequently follow quick lifecycles, but several considerations suggest that Big Data represents a sea change for enterprise information. With the cost of processing and data storage falling so rapidly each year, our society no longer seems constrained as to the amount of information it can create and retain. Today’s burgeoning numbers of online users now leave a trail of “digital exhaust” as they cruise social networking sites; e-commerce continues to grow at 35% per year; and RFID tags are steadily appearing on wholesalers’ pallets and manufacturers’ products. We are entering the “Internet of Things,” in which phones, cars, trains, and planes — plus process controllers, appliances, and medical devices — all transmit a steady stream of data for interested parties to mine. Even dairy cows now sport portable monitors announcing when they come into heat. The data our society generates in a single year recently surpassed a zettabyte (a trillion gigabytes), which is a hundred million times more information than is contained in the print collection of the US Library of Congress — and this onslaught is doubling every two years.

Naturally, people worry about how much of this data they should capture, manage, and analyze. We frequently read about creative entrepreneurs discovering riches hidden in this information. For example, companies can now measure customer sentiment toward their products by mining the comments, ratings, and even images shared on the Web. They can correlate these sentiment statistics with purchase records provided by loyalty programs at grocers and retail stores, empowering marketers to customize advertising campaigns for individual consumers. As we move between websites today, we encounter a sequence of offers that are so subtle they go unnoticed but are so aligned with our individual preferences and behavioral triggers that we are almost certain to buy. With world Internet usage quintupling per decade, there is no upper limit on the number and value of new business opportunities for those who can bend the swelling flood of data to their purposes. In this context, the frenzied interest in Big Data makes sense because the power of such analytics has been proven, and rational companies should be actively seeking to profit from it.

MAPREDUCE IS NOT OUR SILVER BULLET

Unfortunately, the best method of channeling this informational deluge is far from clear, because the term “Big
Data” has not yet been well defined. Big data analytics is frequently described as the management of information volumes much larger than our ordinary data management tools can handle. Pundits usually refer to Doug Laney’s “3Vs” — volume, velocity, and variety — which will be explored in the articles in this issue. Yet the 3Vs are only a description of the problem, one that leaves most of us searching for an industry standard approach proven to overcome the challenge. Such a search does not uncover a single direction, however, but instead a myriad of competing strategies. Despite the fact that experts have been discussing Big Data for over 10 years now, the field is still very new, and for all the urgency we feel, no silver bullet yet exists.

The most commonly cited solution for BDA involves a technology pioneered by the large Internet search engines, called “MapReduce” (MR). So frequently do Big Data conversations gravitate to MR that Hadoop, the open source implementation of MapReduce, is now a standard component of most mainstream databases. Yet MapReduce is not a universal solution to all Big Data problems, for several reasons. First, it solves only problems that can be formulated in terms of key-value pairs. This approach is capable of some powerful insights, but it has a distinct sweet spot that generally requires the input data to be already assembled into a flat file. Second, as anyone who has tried to join multiple tables using MR (or even wrestle it into printing “Hello, world!”) can tell you, MR is not a general solution to many common data management challenges. Third, the interface to MR data stores is fairly primitive in comparison to the standard DBMSs — a team must know Java well. Fourth, attempts to provide an SQL-like querying tool for MR still lack many ANSI SQL-92 commands and other common SQL extensions. Fifth, solid MR programmers are difficult to find, so the added cost and risk of building MR applications can far exceed the investment required by the many alternatives.

Because MapReduce is not the only solution available for high data volume, velocity, and variety, a solid Big Data strategy should look at the other technologies. There are many more columnar databases available today than there are MR implementations. Many of these columnar DBMSs are imbedded in data warehouse appliances that allow our existing business intelligence (BI) applications to handle very large volumes of data using a standard SQL interface. Furthermore, many columnar databases are more mature than MR, allowing Big Data applications to be designed and developed by developers with more typical skills. For organizations willing to consider newer offerings by smaller vendors, there are also the numerous types of Big Data solutions found in the NoSQL (“Not Only SQL”) universe, such as key-value pair databases that do not require MR programming; graphic databases that use “triples” rather than key-value pairs; and in-memory relational databases that settle for “eventual consistency” in the interest of very fast read-write operations. These products, too, often look more like our traditional tools, making them easier to work with, and several of them can tackle analytical questions that MR cannot begin to address.

REPLACING FEAR WITH DISCIPLINE

Given the limits of MapReduce and the presence of many alternative solutions, it is odd that so many conversations about Big Data turn instantly to Hadoop. This knee-jerk reaction is driven mostly by fear. Both business and IT executives feel threatened by the accelerating flood of data coming from a proliferating number of sources. They worry that they should be doing something creative and profitable with it today, before competitors blindside them with new capabilities. They naturally want to start storing everything now, even if they cannot articulate the value for this information, and they hope against all odds that grabbing hold of this information is going to be quick and easy. Indeed, Forbes notes that Big Data today is ill-defined, intimidating, and immediate (i.e., demanding action now) — all of which adds up to a set of “3 Is” that may be more important to consider than the 3 Vs.

A more sober view of the situation might suggest that data streams in the exabytes are only another chapter
in data management, just as terabytes and petabytes challenged us in previous decades. We must remind ourselves that new technologies frequently get over-hyped by the media and vendors, and that our search for a silver bullet often leads to profound disappointment. We will need time and discipline to see what Big Data can realistically offer. A disciplined approach should begin with compelling use cases that express clearly attainable business impacts. Only by articulating realistic objectives can we rationally choose a technical solution from the several competing Big Data technologies. Moreover, any Big Data solutions must integrate into our existing strategies for “not-so-Big Data,” so that the information flood from the coming “Internet of everything” calmly fills our carefully architected BI ecosystems with usable data rather than washing them away.

IN THIS ISSUE

The articles selected for this issue of Cutter IT Journal provide a handy opportunity to conduct that sober evaluation of Big Data technology. The discussion first provides a solid introduction to the world of BDA and then explores a set of important extensions of the technology. Richard Walsh, Richard O’Callaghan, and Sabine Yoffou start off our collection by systematically defining Big Data so that we can begin successfully planning a serious implementation effort. Next, IBM’s Matthew Ganis and Avinash Kohirkar examine one of the most common uses of BDA, namely mining social media discussions. Rich Johnson and Ron Zahavi of Microsoft then address the essential topic of incorporating this new style of analytics into our traditional data warehousing programs, so that we end up with well-integrated BI platforms.

The theme of extending Big Data technology begins with Frank Coyle, who discusses one of the primary competitors to MapReduce — the RDF triple, which will someday soon enable the Semantic Web. Holly Korda, Ann Magee, and Lori Damiano then explore Big Data’s potential in a specific industry, showing how it can be leveraged to bring transparency and accountability to the world of healthcare. Finally, Saeed Lajami, Anson Mok, Mario Wahyu Prabowo, and Cutter Senior Consultant Sara Cullen provide an interesting alternative for our solutions toolkit by advocating the use of crowdsourcing to solve Big Data challenges.

Together these articles introduce insights of breadth and depth into the new and quickly evolving world of BDA. We hope they will help you begin to explore and understand how this technology can solve what will be some of IT’s most pressing challenges for the foreseeable future.

ENDNOTES


Ralph Hughes is a Senior Consultant with Cutter Consortium’s Data Insight & Social BI practice. He also serves as Chief Systems Architect for Ceregenics, a DW/BI firm. As one of the world’s leading authorities on the methods, tools, and modeling techniques that enable incremental design and rapid delivery of enterprise BI systems, Mr. Hughes has led multiple BI programs and projects for Fortune 500 companies in aerospace, telecom, and pharmaceuticals. In addition, he is a certified ScrumMaster, a PMI project management professional, and a faculty member at The Data Warehouse Institute (TDWI). He has coached more than 1,000 BI professionals worldwide in the discipline of incremental and iterative delivery of large data management systems.

As the creator of the Agile Data Warehousing method and author of the book by that same name, Mr. Hughes specializes in the art of getting projects started quickly with streamlined requirements management techniques and then proving their success with automated and continuous integration testing. He is also a frequent keynote speaker and instructor at business analytics conferences, providing architectural reviews and program roadmaps for data warehousing departments worldwide. Mr. Hughes holds bachelor’s and master’s degrees from Stanford University in computer modeling and econometric forecasting. He can be reached at rhughes@cutter.com.
Big Data has the potential to transform the way organizations do business and how we run society. From the ability to harvest vast forests on a just-in-time, lean manufacturing basis to the efforts by the UN’s Global Pulse Unit to harness up to 2.5 quintillion\(^1\) bytes of Big Data to help enhance public policy making,\(^2\) it seems that Big Data and Big Data analytics (BDA) really can be the solution to many problems. Companies that fully realize this discipline can reap great rewards in terms of increased profitability, potentially by as much as 20\(\%\),\(^3\) thus helping them to distinguish themselves from their competitors. With a myriad of new Big Data–related technologies appearing, computing power and storage capacity costs decreasing daily, and new sources of Big Data from Internet traffic, smartphone applications, and sensors emerging almost daily, surely it has never been easier to begin exploiting Big Data. So where’s the problem with Big Data and, by extension, BDA?

To help us answer this question, we did some research. We began with a thorough review of current Big Data literature and followed that with a series of case study interviews of potential and actual users of Big Data. What we found was that while the era of BDA is clearly upon us, harnessing it will require companies to answer two key questions:

1. What activities must we undertake to exploit Big Data?
2. What capabilities (business, technological, HR, managerial) will we need to undertake these activities?

Many business managers do not know how to embrace Big Data technologies and techniques. As most companies are at the early stages of using analytics to support their business decisions, in this article we aim to meet the current needs of organizations in their quest to begin efficiently and intelligently harnessing Big Data and BDA. We drill down into the important activities and phases involved in adopting Big Data as a key business tool. We also provide useful checklists of key activities and examine associated barriers and the capabilities for overcoming them.

**DEFINING BIG DATA AND BIG DATA ANALYTICS**

The term “Big Data” has evolved over time and continues to evolve. At first, Big Data was defined only by its large volume, but with time, additional characteristics came to be associated with the term, such as velocity and variety of data. Today, the “Vs” of Big Data continue to expand to include veracity, viscosity, virality, and more. In fact, Big Data’s definition is becoming so broad that a unified definition is imperative, as the many existing variations of the term are causing confusion in both the trade and academic press.

While there are many acceptable and differing interpretations for the “Vs” of Big Data, we subscribe to the original three — volume, velocity, and variety — as defined by Doug Laney in his influential Meta Group research paper.\(^4\) Laney’s definition of the “3 Vs” gives a clear explanation of each V or “vector” of Big Data and continues to be widely accepted to this day.\(^5\)

In the course of our research, we interviewed Enda Keane, CEO of Treemetrics, an innovative company in the forestry industry that is using Big Data and Big Data analytics to manage and harvest forests more sustainably. It does so by combining and analyzing “big data” sets across all three Vs:

- **Large volumes** of data are captured. For example, a single, stationary, three-minute laser ground scan of a forest can capture 40 million data points, which can then be used to accurately model an area of forest conditions.
- **High-velocity** bursts of data are captured from hundreds of forestry harvesting machines. As the machines fell each log, they record diameter measurements every 10 cm along the length of the log. This streamed measurement data is fed back to centralized Big Data models and is used to more accurately record and model lumber quality.
- A **variety** of data sources — aerial and ground scans of forests, sensor readings from forestry harvesting machines, data collected on historical forest growth patterns — are all used as inputs.
That said, Big Data’s potential is severely limited without BDA. According to TDWI Research Director Philip Russom,6 Big Data analytics is, not surprisingly, a combination of two elements: Big Data (in the form of large data sets) and analytics (in the form of the powerful techniques shown in Figure 1). The term “analytics,” from a business point of view, is simply the analysis of data to gain knowledge that can be used to improve business processes and decisions. It includes the study of historical data to identify patterns and assess performance, the analysis of decisions and events to assess their impacts on business, and the prediction of future market behaviors.7

Thomas H. Davenport, director of research at the International Institute for Analytics, and Jeanne G. Harris, executive research fellow and a senior executive at the Accenture Institute for High Performance, describe the various types of analytics and divide them into three main groups (see Figure 1):

1. **Descriptive analytics** is the simplest form of analytics; it is the consolidation of historical data with the purpose of evaluating past performance and actions. Some examples of descriptive analytics are standard and ad hoc reporting, alerts, and dashboards.

2. **Predictive analytics** involves more complex methods of analysis, such as predictive modeling and simulation. It is used to anticipate future behaviors.

3. **Prescriptive analytics**, as used in our forestry example, determines how something can best be executed. It provides the highest level of understanding on a matter.8

Russom9 and Sara Philpott of IBM10 suggest that “advanced analytics” combines prescriptive, descriptive, and predictive analytics; it also includes data visualization, artificial intelligence, stream processing, and in-database analytics. Finally, according to Russom, BDA is “where advanced analytic techniques operate on big data sets.”11

**BDA IN ACTION**

Based on our analysis of companies that are successfully using Big Data, we contend that Big Data only generates true value for organizations when used in conjunction with advanced analytics. The aforementioned Treemetrics, which has built its business model around using BDA to change how forests are managed and harvested, is a case in point. Traditionally, foresters took a limited number of measurements from a small sample of trees to gauge the composition and quality of lumber in a forest. Using this traditional technique meant that the lumber quality of forests was not always properly understood and trees were often harvested in a non-optimal fashion. Treemetrics wanted to more accurately

![Figure 1 — Big Data analytical layers.](image-url)
profile the types of lumber being cultivated in forests, as this information would result in the right lumber in the right forest being harvested at the right time in order to best meet market needs.

Treemetrics began by capturing vastly more detailed forest measurements based on a new laser ground-scanning technique it developed. Over time it added additional Big Data sources, such as aerial scans and real-time data feeds taken directly from harvesting machines. These were successfully combined to form forest models and then analyzed. In effect, Treemetrics exploited all of Big Data’s 3 Vs.

More importantly, the company also began using Big Data analytical techniques, such as advanced data visualization, to more accurately model and predict each forest’s lumber content. Treemetrics has used these enhanced insights to carve out a new market for itself, advising forest owners and foresters on how best to manage their valuable and scarce resources. Using old forestry methods, up to 20% of a forest’s harvest would be wasted by inappropriately logging the wrong type of timber. With the use of BDA on large data sets, Treemetrics can prescribe for foresters which trees to harvest and when to harvest them. This brings, as Keane puts it, “lean manufacturing” to the world of forestry. Waste is greatly eliminated and thus value is created.

THE DOMAINS AND ACTIVITIES OF BDA

So what must companies do to begin harnessing Big Data? To answer this question, we identified and interviewed organizations that were either using or currently assessing the use of Big Data and BDA. In order to get a wide range of views from both the public and private sectors, our case research involved interviewing representatives from such industries as telecommunications, finance, forestry (i.e., Treemetrics), and Big Data IT vendors, as well as members of government and academic organizations. A summary of each of our seven case studies is found in Table 1.

While it is tempting to believe that Big Data and BDA revolve solely around technology, our research shows that to maximize the effectiveness of any Big Data–related investment, activities must be conducted across four key organizational domains (see Table 2):

1. The **business** domain refers to areas where the business as a whole must undertake activities if it is to successfully implement Big Data. The most common activity in this domain is preparing a business case for Big Data.

2. The **technology** domain refers to areas requiring direct input from technical staff or staff with strong technical and analytic skills. The most common activity here is choosing the right technology platform.

<table>
<thead>
<tr>
<th>Forestry Company</th>
<th>Academic Research Organization</th>
<th>Government Agency</th>
<th>Financial Organization</th>
<th>Telecom Equipment Vendor</th>
<th>Big Data Vendor 1</th>
<th>Big Data Vendor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>New business model based on Big Data and BDA</td>
<td>Researching Big Data techniques for use in modeling financial transactions</td>
<td>Evaluating Big Data technologies for use in gathering and publishing government and societal statistics</td>
<td>Evaluating Big Data technology to create a unified view of its customers</td>
<td>Combining Big Data sources to create new telecom services and identify new revenue streams</td>
<td>Supplier of Big Data infrastructure; migrating internal IT systems to a unified Big Data platform</td>
<td>Building new ecosystem of organizational users, consultants, and developers for its new Big Data platform</td>
</tr>
</tbody>
</table>

| Table 2 — BDAC Domains and Phases |
|-----------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Domain          | Exploration | Business Case | Proof-of-Concept | Implementation |
| Business        |              |                |                 |                |
| Technology      |              |                |                 |                |
| Human Resources |              |                |                 |                |
| Management      |              |                |                 |                |
However, given the relatively immature state of the technology and the rapid emergence of new innovations in this area, this may not be an easy task.

3. The human resources (HR) domain encompasses several areas, including staff training and upskilling and allowing staff to liaise with vendors or participate in vendor workshops. The most prevalent activity from this domain is upskilling staff in new technologies such as Hadoop.

4. The management domain involves activities that management must undertake in order to ensure the success of a Big Data project. The most commonly mentioned activities in this domain are establishing data governance surrounding the use of Big Data and changing the corporate culture to embrace decision making based on Big Data.

These four domains encompass 24 activities in total that need to be performed. Through careful analysis of the activities identified through our research case studies, and by gathering and categorizing these activities along the four domains described, we have derived a four-phase Big Data Adoption Checklist (BDAC) that organizations can follow to develop and adopt Big Data and BDA in their organization. The four sequential phases of the BDAC are:

1. Exploration
2. Business case
3. Proof-of-concept
4. Implementation

The comparative shading in Table 2 illustrates the relative number and importance of activities in each domain and phase. Each of the activities associated with these phases will now be explained in detail.

THE BDAC ACTIVITIES

Figures 2-5 illustrate each of the BDAC activities mapped across each of the four phases and four organizational domains.

The main focus of activities performed during the exploration phase, illustrated in Figure 2, is to carry out preliminary analysis into how Big Data and BDA could be used to assist the business. If we step back for a moment from these individual activities and focus on how the activities across each domain align, we can extract some valuable insights. The exploration phase is characterized primarily by the conjoint efforts of those
working across two main domains: business and technology. This alliance of functions and activities will help ensure the alignment of business and IT strategy. For the overall success of this early stage phase, however, it is highly advisable to include representatives of the other domains — namely, HR and management — to make sure that cross-functional teams can be employed.

Turning to the business case phase, Figure 3 shows that, in order to efficiently construct a business case, it is again imperative for the business and IT people involved in the project to work together, as the activity requires a deep level of expertise in both areas. At the same time, to ensure a fruitful deployment of BDA within the organization, executives should begin working to formulate corporate strategy around the use of Big Data. The earlier they do so, the higher the likelihood that lessons learned from the proof-of-concept phase can be properly assessed.

Moving to the proof-of-concept phase, illustrated in Figure 4, it is easy to see that the most critical aspects of this phase relate to the business and technology domains specifically. Generally speaking, any associated HR and managerial challenges should have been successfully addressed in earlier phases. The proof-of-concept phase should be used as a means to evaluate how BDA can be used to tackle and address specific business problems.

Returning to our forestry example, Treemetrics first had to determine whether the vastly increased amounts of data collected using its laser ground-scanning techniques would enable more accurate modeling of forests, and the lumber contained within, before implementing its business model. In this way, the proof-of-concept phase acts as a major stage-gate process where a go/no-go decision can be made on the intended Big Data implementation. Such is the importance of this phase that Big Data vendors such as Big Data Vendor 2 are establishing new programs to allow companies to test and develop their new ideas for BDA with direct vendor support and training before committing to a Big Data platform investment.

Finally, the implementation phase of BDA, illustrated in Figure 5, involves the active participation of all four domains: business, technology, HR, and management. During this final phase, purely business capabilities are less relevant. Instead, a broad spectrum of organization-wide activities must be performed. This does not mean that the activities of businesspeople are less relevant; rather, it means that other parts of the organization must actively support the business function to ensure that the Big Data project succeeds.

When implementing BDA, organizational responsibilities move clearly beyond simply using the latest technology and hiring data scientists to analyze the results. The organization as a whole must embrace the changes required to successfully extract value from Big Data. This means IT users and consumers must be educated as to why they should treat data as a valuable commodity and not waste precious data by guarding it in data silos. The entire organizational culture must embrace new Big Data technologies as a means of harvesting the information the organization creates or can utilize and move away from maintaining shadow IT systems.

**THE BARRIERS TO AND CAPABILITIES REQUIRED FOR BDA**

So far we have presented a checklist of 24 activities that need to be performed over four distinct phases. However, there can be barriers associated with each activity, and organizational capabilities are required to overcome these barriers. Our research identified a great number of activities organizations must conduct across each of the four domains. However, if we focus on the top five most commonly cited activities, along with their related barriers and capabilities, we may go a long way toward overcoming the major challenges to harnessing Big Data (see Table 3).

Realizing that all aspects of the organization, across all of the four described domains, must become involved
in the adoption of Big Data means that the organization must allocate significant time and resources to develop the capabilities necessary to exploit new Big Data–based opportunities. Being aware of the cross-domain organizational demands is an important insight to grasp, but this alone will not help organizations chart a path toward becoming a Big Data–centered organization. An organizational culture that embraces Big Data at every level must be fostered if Big Data is to realize its vision and overcome the associated barriers and resistance to change.

**CONCLUSION**

Big Data and BDA are a flourishing and emerging collection of technologies and techniques that enable organizations to enter a new era of advanced data analysis. As such, they have received much attention in the literature highlighting their many advantages — but also their current limitations. This, in turn, has prompted a call for researchers to propose innovative solutions to these challenges and thereby make the full power of analytics available to all companies.

Based on our case study analysis, we believe that our four-phase Big Data Adoption Checklist can greatly assist organizations in exploring, refining the business case for, evaluating, and implementing Big Data and BDA. Our research found that, from the initial exploratory phase through full implementation, a cross-functional “A team,” comprising people with skills from all four domains, must be formed to take control of Big Data and BDA deployment. The business function has a significant role to play here, as the activities necessary to implement Big Data and BDA are often business domain–specific.

However, barriers need to be overcome and capabilities developed in order to successfully exploit Big Data. Our research suggests that the barriers organizations confront when implementing BDA often lie mainly within the business and technology domains, whereas the capabilities needed for BDA can reside principally in the management domain. Management often needs to cultivate new organization-wide capabilities to truly exploit Big Data’s potential. Companies must carefully address issues related to all four domains — business,
technology, HR, and management — if they are to reap the benefits of Big Data and BDA.

ENDNOTES


7“Big Data – Is Your Data Warehouse a Dinosaur?” Wikibon, 21 June 2012 (http://wikibon.org/wiki/v/Big_Data_-_Is_your_Data_Warehouse_a_Dinosaur%3F).


9Russom. See 6.


11Russom. See 6.

Richard Walsh is CTO and cofounder of Acumenx, a telecom and technology consultancy firm specializing in the area of mobile network performance and solutions. With 19 years’ experience in telecom and IT as a software engineer, network engineer, solution architect, and project manager, Mr. Walsh recently completed a master’s degree in information systems for business performance from University College Cork (Ireland). He is currently focusing on the application of Big Data analysis techniques to improve the performance of mobile networks. He can be reached at richard.walsh@acumenx.com.

Richard O’Callaghan has an extensive background in sales and marketing within large organizations and SMEs, where he participated in many sales, promotional, and digital marketing/social media campaigns. Mr. O’Callaghan also carried out an extensive new product development project with University College Cork. He has recently completed an MBS in information systems for business performance from University College Cork and holds an undergraduate BSc degree, also from University College Cork. He can be reached at richard.ocalla@gmail.com.

Sabine Yoffou has studied economics and management at HEC Lausanne and at the University of Zurich in Switzerland. Ms. Yoffou previously worked for IBM Switzerland and has recently graduated from a master’s program in information systems for business performance at University College Cork. She currently works as a consultant in SAP support and service in Galway, Ireland. She can be reached at sabine.yoffou@gmail.com.
Ensuring the Accuracy of Your Social Media Analysis

by Matthew Ganis and Avinash Kohirkar

Companies are always searching for a competitive edge — a “leg up,” if you will, on their competition. As a result, many organizations today are looking to social media channels in an attempt to understand what consumers think of them or, better still, what those same consumers think about their competitors. In this article, we discuss the ramifications of drawing rapid conclusions from the analysis of social media discussions without a thorough understanding of the data. Drawing conclusions based on a faulty analysis can have an adverse effect on the marketing campaigns a company undertakes or the direction it chooses to follow. Given the investment that a firm may make based on the results of this consumer opinion, it is imperative that the analysis contain accurate, relevant data so decision makers can draw the right conclusions.

SOCIAL MEDIA AS BIG DATA

“Big Data” is a term that has caught the attention of many in the IT arena. It describes the phenomenal growth and availability of data (both structured and unstructured) on the Internet. A single day of traversing the Web by one user can produce large amounts of structured data (access patterns, site visits, etc.), and there are over 2 billion Web users worldwide.1 The amount of data generated from access logs alone can be incomprehensible, and the freely available data on the Internet continues to grow at a mind-bending rate.

Where is all of this Big Data coming from? It’s produced within the many social media applications by a wide variety of sources (people, companies, advertisers, etc.). Facebook recently announced that it reached the milestone of 1 billion users worldwide,2 and in the US it has over 167 million active users.3 As a result, if each of these users posts a daily update, an additional 167 million comments/suggestions/complaints would be created in a single day in the US alone. Couple that with the fact that on average every user has approximately 130 friends (each of whom could comment on a post), and the number of new data items — again, just in the US — quickly rises into the range of 21 billion per day!

It is estimated that, in a 20-minute period on Facebook, over 1 million links are shared, over 1.3 million photos are tagged, 1.8 million statuses are updated, and 10.2 million comments are made on various wall or fan pages.4

Although Twitter is not as large as Facebook, current statistics point to a user base of over 465 million registered users.5 While it appears that more than half of those users are “passive” (i.e., not posting, but following others), the number of tweets has grown from approximately 75 million in January 20106 to over 175 million as of February 2012.7

NEEDLES IN HAYSTACKS

These numbers are staggering and clearly representative of the term “Big Data.” For the purposes of this article, we assume Big Data is data created with a high velocity from a large number of sources. So how can you leverage this “crowd source” of information, opinions, and new insights? How do you look in this haystack of information for the needles (i.e., the conversations) that reveal people’s sentiments and opinions about your company or industry?

When looking at social media data, you need to think about the increasing volume, velocity, and variety of information (the so-called 3 Vs) that form this slice of the Big Data picture. In this torrent of data, your customers or clients are discussing a variety of issues ranging from their dealings with you as a provider, to their feelings about specific topics, to items that are going viral that may be important to your organization.

The question isn’t how you find these pieces of information (these needles in haystacks) but, more importantly, how you know that what you’ve found accurately articulates the opinions or feelings of your constituents. This is increasingly important if you consider mounting a marketing campaign based on the perceived views of consumers and later find out that the sentiment (i.e., the view or attitude of a group toward your company or product) was misinterpreted.
As with any large amount of information, uncertainties can arise due to an inability to fully grasp the depth and variety of the social media data on hand. Below we will explore the various pitfalls of social media analysis and the precautions organizations must take when attempting to look for insightful messages from this data source.

**EXPLORING SOCIAL MEDIA: OUR PROCESS**

Our process flow for evaluating social media data at IBM is shown in Figure 1. Mining for insights within a Big Data space resembles the iterative patterns of agile development techniques, where we start with a small recognizable deliverable and methodically build on it to completion. Within each step of the process, we go back and revisit the results or conclusions drawn, looking for clearer methods of analysis, richer data sources, or ways to clarify or narrow the result space.

Here are the five steps that we go through:

1. **Choose data sources.** Given the vast ocean of social media data, we must make a decision about where to look for the potential conversations. Do we want to focus on bloggers, individuals using Twitter, or review sites?

2. **Perform a structured search.** Once we have determined the data sources we will focus on, we then must determine how to search within these data sets to tease out relevant pieces of information.

3. **Create a data model.** Once we understand the data, we need to create a text analysis model that will allow for a deeper understanding of the data (such as understanding how the concepts within the data relate to each other). Here we begin the categorization and classification effort within our models.

4. **Perform analysis.** Once we have located the data for the analysis and constructed a model to represent the data, we perform an analysis that looks for common phrases and relationships between phrases/words.

5. **Interpret results.** Finally, once we have established the data sources and refined the search within the data sources, the results of the textual analysis need to be examined and “sanity checked” (preferably by a subject matter expert) to determine if we have to gather more data or make refinements to our model.

**BIG (NOT BAD) DATA**

When using all of this Big Data, you must make sure that the data you are going to analyze is indeed relevant to the question at hand. Too broad a search could result in irrelevant data being brought into your analysis. Thus, it is important to be looking at the right data.

Any attempt to mine information in a large user discussion space must start with a subset of the entire universe of data. Begin by asking:

- Where do we want to look for conversations?
- Are we interested simply in blogs, or are we interested in individual opinions and viral topics as well?

For example, at one point we were asked to understand the consumer sentiment around IBM’s BladeCenter offering. We chose to focus on Twitter and online discussion groups, since a secondary goal of the project
was to identify influencers (individuals whose opinions or comments influence the opinions and thoughts of others) and large “pockets” of discussion.

During our analysis, when we compiled the list of the top discussion sites, we were surprised to see a popular gaming site as a top location for discussion. After further investigation, we discovered that “Blade” and “Blade Center” are the names of a popular online gaming character. As a result, our sentiment analysis work needed to be corrected to exclude this site.

The other expected outcome of this project was the identification of influencers in the discussion surrounding IBM Bladecenter. It’s important to understand who has an influence over others when they make comments about your product or company. To achieve this, we relied on Klout, a service that measures an individual’s social clout, or ability to influence others. The resulting ranking is called a “Klout Score,” and those with high Klout scores are deemed influencers. These individuals can be your “brand advocates,” which, when properly engaged, can be important for driving sales.

In another of our projects, related to the US Open tennis tournament (of which IBM was a sponsor), we also encountered issues of “bad” data. In this case, we attempted to capture tweets related to the top men’s and women’s tennis players. Our business team provided a list of players’ names but also specified various nicknames for some of the players. In this case, a search rule with the nickname “Fever” was pulling in a large amount of irrelevant tweets (see Figure 2). Once we determined that the search term was too nebulous, we simply deleted the term, which eliminated all the extra noise. From then on our searches produced only relevant data, thus taking the “bad” out of Big Data.

Careful monitoring of your incoming data streams will allow you to spot irregularities in the data. When anomalies occur, it becomes prudent to validate the data long before any detailed (and potentially inaccurate) analysis is completed. The spike may very well be a valid topic that is going viral on the Internet, but for the sake of your long-term analysis, it’s better to be safe than sorry.

**DATA COVERAGE**

As we’ve seen, when looking to evaluate a specific claim (or answer any question using large data sources), we must perform a search over the data to limit it to pertinent material. That said, we must also make sure we have all the data we need. This is one of those concepts that are highly dependent on the use case. Let us review a couple of examples to illustrate this point.

A company we were working with was sponsoring an online discussion event with the aim of understanding the concerns and issues surrounding security. The company’s marketing department had contracted with some partners to promote the event in social media, and we were asked to determine if the promotion was adequate. To that end, we set up a listening model that captured conversations in Twitter and other social media sites prior to the event, during the event, and after the event. We found only a handful of mentions one week prior to the event and just a few more mentions a day before the event. Even though this was a small sample size in an absolute sense, for the use case it clearly implied that the social media promotion the business partners actually undertook was inadequate.

In another example, an IBM product group that conducts structured usability reviews of their products wanted to supplement this research with raw input from social media. They were trying to determine what usability issues the general user population was reporting. In this experiment, if only one person complained...
about the menu structure of a product, it was deemed an insufficient reason to change the user experience. Our team decided that for every issue captured, we would treat that feedback as significant if and only if there were at least 50 comments about that one issue. As a result, we continued “listening” until we either collected a sufficient amount of data or concluded that the number of comments would never be adequate. In the latter case, we would de-prioritize the issue.

**FALSE POSITIVE OR NEGATIVE RESULTS**

A false positive or negative is simply misinformation based on an incorrect set of assumptions or a misrepresentation. Many of the text analysis programs in use today do a fairly good job of assessing sentiment (positive versus negative opinions), but blindly trusting their results can be dangerous.

For example, depending upon the context in which it is used, the word “sick” can be viewed as a positive or a negative sentiment word. When a Gen Y person makes a comment such as “The user experience on this site is sick,” it could very well be interpreted as a positive sentiment. Of course, if the comment goes along the lines of, “This user interface is so convoluted that it makes me sick,” we can surmise that this is truly a negative view of the interface. Another example is the word “rad,” which may refer to “rapid application development” or suggest a positive sentiment.

Recently IBM showcased one of our technology breakthroughs by having a computer system named “Watson” compete on the popular television quiz show *Jeopardy*. Prior to this, the word “jeopardy” would have been treated as a largely negative sentiment word. However, after Watson’s stunning win, many references to “jeopardy” are seen as positive for IBM (depending of course, upon the topic being discussed).

Sarcasm can be viewed as an extreme case of a false positive. Take the example: “Gee, I got a great raise this year.” Because of the emphasis placed on the word “great,” the statement is no doubt referring to ill feelings about the raise the person received. While we can deduce the proper sentiment if we look at the emphasis of words or the presence of emoticons (or lack thereof), many of these “keys” (or signals) get lost when text is ingested into a text analysis engine. As a result, some items that could be flagged positive might instead be sarcastic negative statements. Care — and human interpretation — may be needed.

**KEYWORDS**

When dealing with Big Data, remember that not everyone thinks and speaks (or types) the same way. For example, what I call a “stoplight” others may call a “traffic light.” Understanding the keywords used for your analysis is crucial.

The increased internationalization of the social media landscape is also a factor here. In one instance, our team was attempting to understand the “social responsibility” issues facing employees of companies within certain growth geographies (e.g., India, China). After a careful inspection of our results, we found that we were missing a significant portion of our intended audience because of the term “employee.” In this particular use case, using phrases such as “worker” or “factory person” enabled us to pull in more content, thus enabling a richer analysis.

When evaluating social media, keep in mind that individuals making comments are not necessarily using polished “corporate-speak.” As you build a universe of data to explore, it is therefore imperative that you express terms and ideas in a common person’s language. At IBM, like many large corporations, we have our own company language. While we may proudly refer to our high-end “System/Z platforms,” consumers, architects, or C-level employees of our potential customers may refer to them as “mainframes” or “big iron” instead. The point is to understand how people are referring to your topics, or you may miss some vital insights, as in the “employee” versus “worker” example above.

Although it may not be particularly flattering, some of the derogatory terms that are used to describe your company can provide some useful insights. If we are looking to understand consumer sentiments and opinion, we have to be open to all options. Walmart may not appreciate being called “Wally World,” but if the nickname truly reflects how their patrons refer to them, they need to include it in their search terms.

Keep in mind that deriving the source of data for evaluation involves more than just accumulating an adequate set of keywords. Words used within a specific context may not only change their meaning, but also be critical to your analysis. For example, at one point we were

As you build a universe of data to explore, it is imperative that you express terms and ideas in a common person’s language.
keeping an eye on conversation around IBM as it pertained to a specific industry event called Sapphire Now. The keyword “sapphire” was problematic in that it has many meanings. In the obvious case, it refers to a gemstone, while in a completely unrelated area, it happens to be a popular brand of gin. The key was to look for the word “sapphire” in the context of IT and database terms. Expressing terms within a context is specific to a tool’s implementation. In our case, the use of simple regular expressions such as “IBM.{0,40}social” would look for the word “social” (as in social business) within a radius of 0 to 40 characters of the word “IBM.”

In our experience, keywords need to be tailored based on use case as well as on the geography where the conversations are taking place. For example, if we are trying to understand how different IBM employees are reacting to vacation policies, we need to remember to include the words “holiday” and “leave,” which are used in certain places to refer to vacation.

ANALYSTS

As mentioned previously, human analysts will have to iteratively revise and refine the first set of results from a typical analysis exercise to ensure accurate interpretations and follow-on actions. Several iterations are needed to make sure all snippets of information flagged as positive or negative are indeed positive or negative in the context of your business and the topic under discussion.

Of course, the end goal of these exercises is to derive some key insights from this data that could positively affect your business units. Simply looking at raw data provides some knowledge, but by aggregating the data and using various tools to analyze the text, hopefully you can move from raw data to something more organized — information. The next step in the transformation of data into insights involves social media analysts, with expertise in analytics, who can work iteratively with the analysis engine to extract knowledge, which can then be passed on to your business users. Analysts can further develop that information and knowledge into insights or conclusions (see Figure 3).

The value of analysts and/or a “listening office” is their ability to apply fuzzy logic to a situation. For example, in some cases the sample size of data you start off with may be too small to derive conclusions. A business analyst with domain and subject matter expertise could easily tell you if you don’t have a large enough sample size. On the other hand, if the sample size is too large, it is crucial to have a human show you that there are duplicate data points in the sample or summarize the information at a higher level.

The analyst also plays a key role in the iterative process of model evolution. For example, if you are conducting a sentiment analysis, the tool/dictionary may need to be customized to meet the needs of the particular topic or domain. You may have to incorporate additional results into the model to further refine or enhance its usefulness. This is also true when you need to include a fragment of information only when the context is right. For
example, in the case of IT security modeling, the keywords “hack” and “attack” — while very relevant to cyber security — can also occur in many non-IT conversations, thus requiring a change to your analysis models to ensure they are being analyzed in the proper context.

**CONCLUSION**

While the prospect of uncovering valuable insights in the deluge of social media is compelling, you must take care to ensure that you are getting an accurate portrayal of the data. In this article, we’ve discussed several pitfalls and areas of attention that we believe are crucial to bear in mind when utilizing this data.

**ENDNOTES**

1Internet World Stats (www.internetworldstats.com/stats.htm).

7Bennett. See 5.

Matthew Ganis is an IBM Senior Technical Staff Member (and Certified Scrum Professional) within the IBM CIO organization and part of the ibm.com site architecture team. He is the chief architect for IBM’s Internal Social Media Analysis Initiative. Dr. Ganis is a senior member of the ACM and cochair of the Agile New York City user group. He also serves on the editorial board of the International Journal of Agile and Extreme Software Development and has authored or coauthored a number of papers/books on his experiences with Agile methods, including the Practical Guide to Distributed Scrum. He can be reached at ganis@us.ibm.com.

Avinash Kohirkar is an Executive Project Manager within the IBM CIO organization. He has a strong technical and business background across various disciplines based on numerous projects executed within IBM as well as with IBM customers. He currently leads the business office associated with the CIO Internal Social Media Analysis Initiative being offered to IBM business units. Much of the material for this article has been derived from projects that he led, together with Dr. Ganis, over the past several years. He can be reached at kohirkar@us.ibm.com.
The need for businesses to maintain and use data in new ways is rapidly growing. Instead of removing data, businesses are adding and using more data than ever. The lower cost of storage is part of the reason, but regulation, competitiveness, and the digital age are also key drivers. Healthcare organizations have to track their patients to improve patient outcomes and to be eligible for reimbursement bonuses. Energy companies are converting decades-old data stored in paper or older digital formats so that they can maintain and integrate the data into new applications. Marketing organizations are trying to use the enormous amount of social media data to understand user behavior and sentiment, and financial services companies are trying to sift through huge amounts of transactions to find irregularities, ensure regulatory compliance, and/or learn about consumer patterns of use.

The above are just some of the examples of how organizations use their internal or external data. Traditionally, when organizations wanted to exploit data in these ways, they created databases or data warehouses where they stored the data, structured it, and then tried to analyze it by using business intelligence (BI) applications. But while data warehousing (DW) and BI approaches provide high value to businesses, it often comes at a large cost and time investment. Business units determine the questions they want to ask, and IT builds structured data stores to answer this limited set of questions. As the business needs grow, and with exponential growth in new data, many new questions arise and new types of analysis are required. As a result, these systems have to be updated and extended and projects launched to address them.

With today’s rapidly changing environment and rapid growth in data, businesses need the ability to react much quicker, utilize varying data sources (many with unstructured data), and answer new questions that may not have been previously considered. Big Data solutions provide the mechanism for storing any and all data in an unstructured store and enabling the business to discover new questions and answers.

At first glance, it may seem that the traditional, inflexible approach and the new flexible, unstructured approach do not fit together. Is there a place where both types of solutions can coexist and be integrated? Is there even the need to do so?

The answer to these questions is yes. Both approaches can be integrated so that the Big Data store feeds new information and insights into the traditional DW and BI platforms for analysis, and information from the BI and DW solutions feeds into the Big Data store so that it can be mixed with unstructured data and included in new analysis.

New opportunities are emerging for providing even higher value to organizations by utilizing cloud capabilities and external data sources. In this article, we explore the architecture and scenarios that drive traditional DW and BI systems, how Big Data systems and architecture compare, and how the two can be integrated with the emerging technologies to produce even greater value for businesses.

**CLASSIC DW AND BI PROCESSING**

Data warehousing and business intelligence have been around for the last few decades and are still very high on the list of CIO priorities. Tried-and-true methodologies and approaches for successfully designing and building DW/BI solutions have been around almost as long. With a well-trained team, an organization can typically design and develop the first iteration of a DW/BI solution in four to six months. There is a lot of merit to doing this in terms of managing an organization’s overall data assets and creating the corporate “one version of the truth.” The typical DW/BI solution provides a central integrated environment for reporting and analyzing a company’s data and information assets and offers several benefits needed by users of this data, including:

- Optimized schemas that make it easy for end users to write queries and attain good performance
- Data quality and confidence in the data, thanks to data-cleansing processes
- A robust and reliable method for acquiring and loading data from the source system into the data warehouse via the ETL (extract-transform-load) layer
- Data enrichment by combining data from one source system with data from another internal or external source
- Data integration that allows you to compare and contrast data from one source system to another
- Rich sets of reporting tools, from simple lists to advanced graphic visualizations and predictive analysis
- Collaboration, or the ability to publish, share, and act upon the results of data analysis
- Master data management and data stewardship for providing common definitions and taxonomies for an organization’s key master data (product lists, customer lists, facility lists, etc.)

A typical DW/BI solution (see Figure 1) has an ETL layer for extracting, transforming, and loading a company’s data sources into a centralized database. From there, other tools can be layered for reporting, data mining, collaboration, analysis, and decision making.

The majority of DW/BI solutions implemented today store structured data in a relational format in a series of dimension tables and fact tables — tables of well-formed columns and rows of data. Structured data, such as text and numeric values, can be stored in text files with column delimiters, relational database tables, Excel spreadsheets, and so on. Most data warehouses are built upon a relational database system. Even the source systems from which the data warehouse acquires the data are often structured files and databases. Structured data is fairly easy to understand, process, and load into a well-defined schema in the DW database.

Yet the structured, traditional approaches can be prohibitive, as we will see. Consider a typical scenario, in which a healthcare organization has many different data stores and applications. Historically, each silo department may have acquired different applications, or the patient information may be distributed across different types of systems — a patient registration application, physician order entry system, pharmacy application, insurance claim application, radiology imagery database, and/or billing application. These islands of data cannot be easily integrated, and often an additional data warehouse or patient electronic medical record application is added that pulls data from various stores so that it can be collocated and made more accessible.

It can take a multiyear effort before all these silos of data are loaded and integrated into the data warehouse. Most projects like this use a typical waterfall approach in which requirements are gathered and the solution is designed, built, and implemented. With each subsequent addition of a data source in the data warehouse, the process is repeated to include this new data. Any large organization committed to the data warehouse will have a high cost in terms of monetary investment, resources, time, and equipment to fully realize the benefits and overall value of the data warehouse.

Figure 1 — Classic DW/BI.
Imagine that the healthcare organization mentioned earlier needs information to demonstrate how patients are tracked and how their health is improving based on the use of a certain medical procedure. While the data is embedded somewhere in the systems, arriving at the answer is not straightforward. Patients may get different identifiers in each system or application, and existing business processes may not facilitate a continuous experience across different events (such as elective surgery and a visit to the emergency room for a broken arm).

When the business side of the organization has to answer a new question, such as how many patients have had a particular elective surgery, the development team jumps into action. A project is created with the businessperson, someone who manages the data, someone who knows the particular application (or applications), and a software developer and/or database developer. The team sits together and figures out where the data resides, how it can be accessed, and what changes may be needed to come up with the answer. The team may have to change a database table, add new fields, modify an application, and/or update a data warehouse and user interface.

The problem with the traditional approach is clearly evident on the faces of the project team when a physician walks into the room near the completion of the project and expresses how excited she is that she will be able to see which type of surgical procedure was used in the elective surgeries. While to the physician this is a natural progression from the initial question (“How many patients have had a particular elective surgery?”), to the team this is a completely new question, one they are not prepared to answer. The project comes to a halt and the cycle repeats in order to incorporate the new requirement.

BIG DATA PROCESSING

The term “Big Data” is applied to several types of data. Some use it to mean high velocity of data, such as that generated by social media or sensor data streams. Some use it to mean data that is very large, such as video or movie files. Some use it to mean large data storage, such as exabytes of data. What is common across all these definitions is that big or large data is very difficult to work with, does not fit well into structured databases, and requires nontraditional methods of analysis.

Often, analysts will refer to the “3 Vs” that characterize Big Data:

1. **Volume** — the ability to store up to many petabytes of information
2. **Velocity** — the fact we are generating and capturing large amounts of data very quickly
3. **Variety** — the ability to capture all sorts of data formats, including text, sensor data, documents, audio, image, video, log files, etc.

Hadoop to the Rescue

The Hadoop platform and its related components are often considered the Big Data solution. There are other database and data storage solutions out there that arguably could be called Big Data solutions as well, but for the purposes of this article, we’ll be discussing Hadoop as one of the most common implementations.

Hadoop, or Apache Hadoop, is an open source software platform for storing and analyzing the massive amounts of data, both structured and unstructured, that is distributed across a handful or even thousands of independent computers. Using independent, commodity computers allows storage of many petabytes of data, thus providing massive scale-out storage. After the data is loaded, thousands of independent computers — commonly referred to as a cluster — can be enlisted for massively parallel batch computations so as to analyze these large data sets.

At the center of the Hadoop platform is the Hadoop Distributed File System (HDFS) and MapReduce. HDFS is a file system for storing and managing large data files in a distributed manner. MapReduce is a programming engine used to process or query this data, primarily in parallel. It is distributed on multiple DataNode servers all under the control of the NameNode. There are a number of related products built on top of HDFS and MapReduce that provide additional features such as SQL-like query languages, data mining algorithms, schema definitions, and more.

MapReduce programs are written in Java. There are other higher-level programming components in the Hadoop ecosystem with odd names like Pig and Hive that provide simple scripting languages and SQL-like language commands that make it far easier for the
masses to query data in Hadoop versus writing Java programs. Pig and Hive scripts and commands are transparently compiled or translated into Java MapReduce programs at execution time.

Figure 2 provides a high-level overview of HDFS.¹ As data is loaded into HDFS, the HDFS system takes care of distributing large data sets across independent computers, called “DataNodes,” which make up the Hadoop cluster. By default, the data is replicated at least three times across at least three separate DataNodes. This provides a high degree of redundancy and fault tolerance for the cluster. Another server hosts the NameNode service, which contains the metadata about every file in the cluster and where its data is stored across the DataNodes. The BackupNode provides a backup of the NameNode metadata and can be used to manually failover and restart the cluster if needed. Automatic failover is planned for a future version of Hadoop.

MapReduce programs rely on the NameNode metadata to process or query data files in the cluster. Once data is loaded into HDFS, MapReduce programs process and query the data. Figure 3 provides a high-level overview of the MapReduce architecture. At the highest level, once a MapReduce program is submitted, it relies on master and worker nodes to manage the overall job. The Map step involves dividing the job or query up into smaller subproblems and distributing them to the TaskTracker nodes; this is the function of the JobTracker. Often the TaskTracker nodes are the DataNodes where the data is stored so that the computation is close to the data. Once the TaskTrackers process the smaller problem, they pass the answer back to the master node. In the Reduce step, the master node collects the answers from all the TaskTrackers and combines them into a single output or answer to the query that was originally asked.

One of the biggest benefits of Hadoop is its agility. It can ingest large amounts of data quickly in its raw format and begin analyzing that data right away. Hadoop
is unique in that it can quickly and efficiently manage the 3 Vs of Big Data, providing fast ingestion, scalability, and parallel computations across a variety of data types and formats.

In a typical relational database, a schema (a structured or partially structured definition of data such as a database table) is defined first before data is written to it. This is called “schema on write.” Defining a schema first often takes a lot of design effort and code in order to transform and load raw data from its original format into the structure of this predefined schema.

In the Hadoop world, the ability to write files and data in their original or raw format to the Hadoop file system introduces a concept called “schema on read.” In HDFS, raw data files are loaded as is. After that, when data is read, many MapReduce operations such as word or pattern counts can be performed on the data, often without a schema. A schema or multiple versions of a schema can be applied to the data after it is loaded. The important point to realize is that the original data is not lost and can be queried rather quickly using tools such as Pig and Hive. After the data is loaded, more structure can be applied to data files using tools such as HBase, which allows data to be queried more like a structured set of relational tables.

The primary workload for Hadoop is petabyte-scale data storage of unstructured content and massively parallel analysis of that data for queries and other advanced analysis. All types of queries and data mining capabilities exist to process this data, which will answer a lot of questions and also generate new ones. There is much to discover in this ocean of data, and the more organizations learn from it, the more competitive and successful they will become.

**INTEGRATING CLASSIC DW/BI WITH BIG DATA**

Big Data does not fit well into traditional relational databases, and because of this, many think that DW/BI and Big Data are incompatible and even competitors. Actually, the opposite is true. Not only is there a place for each, both technologies can play well together.

Table 1 summarizes some of the benefits of both approaches. The traditional DW/BI solution is used today to analyze structured, more-static data for decision support. A Big Data solution is more often used to analyze less structured and extremely large data and is very well suited to advanced data mining and predictive analytics. It’s important to realize that both the traditional DW/BI approach and the Big Data approach provide some of the same capabilities, but each has unique capabilities that allow it to more optimally handle certain types of analysis on varying types of data.

Organizations can derive huge benefits by extending the traditional forms of analysis to include Big Data. The unstructured data analysis enabled by Big Data lets organizations look for patterns, such as the purchasing behavior of consumers or incidents of fraud, while maintaining the analysis within a particular context. By combining the two approaches, analysts can gain new insights across a much more complete set of data. In particular, the impact will be profound on organizations that deal directly with people (such as consumers, patients, or citizens) and provide them with services or support. Using a combined approach to data analysis, such organizations will be better able to understand their customers and the behavior of those customers. This in turn will allow them to offer more customized services as well as better predict problems and address them ahead of time.

<table>
<thead>
<tr>
<th>Traditional DW/BI Supports</th>
<th>Big Data Supports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing business processes and mission-critical systems with defined data structures</td>
<td>Large volumes of unstructured data</td>
</tr>
<tr>
<td>Efficient and orderly collection and analysis of data</td>
<td>Quick search capability</td>
</tr>
<tr>
<td>Decision making and dashboards</td>
<td>Integration of social media and marketing data, and discovery of patterns</td>
</tr>
<tr>
<td>Solid analysis of existing data based on predefined algorithms</td>
<td>Ad hoc analysis based on newly discovered information</td>
</tr>
<tr>
<td>Self-service business intelligence with predictive analysis on well-structured data</td>
<td>Predictive analysis on a wider set of data</td>
</tr>
<tr>
<td>More interaction and responsiveness when working with refined data sets</td>
<td>More rapid turnaround from data capture to analysis</td>
</tr>
</tbody>
</table>
An Integrated Solution to a Law Enforcement Agency’s Data Woes

A good example of this is a law enforcement agency that maintains large amounts of data. It wanted to improve investigations and reduce manual procedures through improved BI and data mining capabilities. The agency had data sets with millions of records composed of both structured and unstructured data, including images, videos, and documents. Searching through the data took a long time, often requiring personnel to conduct field investigations that took as long as two years!

Instead of requiring officers to create specialized queries to correlate information, the agency wanted a system that would notify them of suspicious individuals or activities. The resulting solution, which integrates traditional database and BI capabilities with Hadoop Big Data analysis, enables the agency to relate data much quicker, identifying related information and completing investigations in just 15 days. The BI capability allows the organization to visualize data in a reporting tool and explore it, providing further detail about cases. In addition to images, videos, and documents, reports include analysis and statistics graphs.

The combined technology solution offers the following benefits:

- It is a complete solution, from database to reporting.
- Officers can produce their own reports.
- Optimized investigation increases speed and accuracy.
- It removes the limitations the agency previously encountered when working with large volumes of data.
- It extends the existing platform.
- It provides a base platform for the future.

The Cycle of Data Refinement

Figure 4 shows an integrated Big Data–DW/BI environment. In this new world, all data sources are captured and loaded into the Big Data cluster immediately, and some of this data can be transformed and loaded into the data warehouse as well, to provide more targeted and refined data. Using self-service and classic BI tools such as dashboards, standard reports, spreadsheets, and highly visual data reporting tools, end users can access the refined and more responsive data warehouse data to satisfy the majority of their queries and analysis. In addition, Hadoop client tools such as Hive and Hive connectors for spreadsheets provide access to the Big Data cluster. Over time these tools will further integrate and seamlessly link to the data warehouse and Big Data cluster.

In the center, the Big Data platform and classic data warehouse form a cycle of data refinement and continuous improvement. As knowledge and refined data are processed in the Big Data cluster, they can be fed into the data warehouse for more targeted and high-speed analysis by end users. Likewise, as knowledge and refined data are uncovered in the data warehouse, they...
can be fed back into the Big Data platform for historical storage as well as for the fine-tuning queries and analytics that occur there. Data mining and machine-learning algorithms in the Big Data cluster and the data warehouse benefit from this cycle of refinement and continuous improvement of the data.

Consider a retailer who tracks the monetary results of a new product line and discovers that specific products sell better in certain geographies. If the retailer exports and loads this product-geography relationship from the traditional data warehouse to HDFS, analysts running queries and predictive analytics on the Big Data platform can use this new information to fine-tune queries and discover new patterns or relationships in the unstructured data. Conversely, as new information and patterns are discovered in the Big Data solution, they can be exported to the traditional data warehouse solution. An example of this might be demographic information and how weather seasonality affects the sale of certain product types at different periods throughout the year. The retailer can export this information from the Big Data solution back to the traditional data warehouse and use it to create better sales forecasts on products and product types based on geography, demographics, and seasonality.

**ENRICHING HISTORICAL DATA AND USING EXTERNAL DATA SOURCES**

In Figure 4, one of the data sources is called “cloud services data.” This data source represents the ever-growing market of highly cleansed and enriched data sets that can be accessed in the cloud. Some examples of this data include accurate addresses, credit ratings, customer demographics, weather data, and more.

External companies that mine Internet data include social media sites and online retailers. They cleanse, aggregate, and score this data, which can then be purchased and fed into the Big Data and DW platforms for the further refinement of internal data. As this external data is further integrated with the existing data, it leads to better and more accurate decision making and competitive advantage.

**Combating a Pandemic with Historical and Real-Time Data**

An example where both approaches combine to provide significant value is bio surveillance and emergency response when quick action is needed to combat an emerging pandemic. Using traditional DW, BI, and analytics, we can study the spread of previous pandemics. We can view the number of cases, patterns of spread, maps, and results of actions taken. When a new pandemic suddenly occurs, the challenge is to determine how the old patterns relate. Where are the current cases emerging, and how are they validated? Where are stockpiles of pharmaceutical products needed to respond to the event?

By overlaying the data capabilities of the traditional historical analysis with new Big Data analytics, we can support scenarios that we couldn’t support previously.

A pandemic scenario requires a quick response, fast stand-up of infrastructure, and the ability to integrate analysis of historical data with real-time incoming data. The real-time data may originate from physicians’ offices about patients they are seeing and hospitals that may record pandemic incidents in their emergency rooms. The US Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) also track events and store them in databases for analysis and viewing. Lately, social media chatter — such as parents discussing how the kids on a soccer team are all out sick due to the flu — provide another source of potential data that needs to be validated.

By overlaying the data capabilities of the traditional historical analysis with new Big Data analytics, we can support scenarios that we couldn’t support previously. Using Big Data capabilities, we are able to integrate events from medical facilities with external sources such as CDC and WHO, as well as with data from the Internet. Based on historical data, we can score the likelihood of an event depending on the validity of the source (e.g., ER data scored higher than social media chatter). We can correlate the events to see whether, based on new and historical patterns, we are observing a real pandemic. For example, we can sift through unstructured social network data to see if schools and sporting events in certain areas are impacted (such as when H1N1 spread in 2009) and then compare the data to medical orders of thermometers and flu medicines in those areas. If we determine that an event is occurring, we can look at pharmaceutical supply chain contracting and electronic data interchange (EDI) data to determine where stockpiles of required medicine exist and how they can be best distributed to those in need.
LEVERAGING THE CLOUD FOR HIGHER VALUE

Hadoop and Big Data are all the talk these days, from journals to conferences for anyone involved in data analysis or enterprise information management. The technology is relatively new, and not many organizations have been able to take advantage of it yet given its newness and the costs and procedures involved in standing up and managing a Hadoop cluster. Many organizations don’t even have a basic understanding of the platform, nor have they given much thought to the applicability of the technology to their business. The majority of these organizations are wrestling with how to get started and realizing that implementing a Hadoop platform for their data assets is much more about a journey of discovery than it is about building to a known set of requirements.

As mentioned earlier, one of the key benefits of a Hadoop platform is the ability to quickly load all the corporate data in its raw form directly into HDFS and start querying data. That analysis will then lead to discovery of new meanings, patterns, and information. After learning from the data and extracting the most valuable information, organizations can think about moving some of this high-value data into the more traditional data warehouse or reporting database for broader and more rapid consumption by business analysts and other users.

So how do most organizations get started with Big Data? One answer is to take advantage of the public cloud Hadoop offerings on the market, which are evolving rapidly. Many vendors, such as Amazon, IBM, and Microsoft, are either in the game or have beta programs for their forthcoming Hadoop-in-the-cloud offerings. For many organizations, this is going to be the best way to get started, because they won’t have to acquire all the hardware, install all the Hadoop software, and manage this environment themselves on premises. They can simply request a Big Data solution of a certain size, and the cloud will automatically provision this environment and make it available. Some major benefits of a cloud-based Big Data platform include:

- Rapid deployment and automated provisioning
- High availability
- Optimal resource utilization
- Secure multi-tenancy
- Centralized data center monitoring and management
- Automated backup and recovery
- Elastic scale (the ability to bring up a 100-node cluster on the data one day and then bring up a 1,000-node cluster on the same data the next day)

By integrating Hadoop with the cloud, organizations can use large parallel computing resources only when they need them, thus reducing the cost and providing flexibility to handle special events. Such events might include a particular financial market change, a pandemic that occurs over a particular period of time, or an election.

To get started with Big Data or Hadoop, you need to:

1. **Think about your current BI solution and data warehouses and what data could benefit from the addition of unstructured Big Data.** Where is this data, and how would you collect it and allow your users to visualize it? Will you use the data to improve efficiency, cut cost, or help your clients (such as improving patient outcomes)?

2. **Determine what you want to extract from this data.** Are there patterns you want to explore and identify? Will you simply search for keywords and count them, or do you need to perform more sophisticated analytics? There are many data sources available; will you be searching across old emails or articles, available research or social media?

3. **Consider the team you will need to establish.** You will need a combination of staff who are familiar with the existing BI/DW solutions and architecture, as well as staff that can understand the new Big Data technologies and approach. You will need to provide some training on the new technologies and APIs and augment the team with a domain expert and someone who understands the statistical analytics that may be required.

4. **Think about the technologies.** You will need to consider the server(s) or cluster required to operate the Big Data solution. Will you operate the infrastructure; use software solutions, hardware solutions, and/or cloud offerings; or do a combination of these things? Big Data solutions, either those direct from open source or those offered by vendors, can typically operate on various OSs. You will need to select the OS and set of tools that you will use.

5. **Establish the environment and test it.** It is always smart to begin with smaller test data sets and expand from there. An iterative approach works best, since it allows you to implement the necessary infrastructure, processes, and governance and establish a foundation...
you can use as your operations and needs grow. You will need to consider issues such as the timing for data loads (initial, incremental, continuous flows, etc.), and how you will convert, store, and age the data.

In this article, we have tried to demonstrate how traditional data warehousing and business intelligence solutions do not compete with Big Data but rather are complementary. Together, the technologies provide organizations with opportunities to analyze data in new ways — giving them new insights and allowing them to create and deliver improved services. We hope that by following the approach we have outlined, you can take the first steps toward applying this integrated method of data analysis in your own organization.

ENDNOTES


Rich Johnson has spent over 25 years designing and building database, data warehouse, and BI solutions for a variety of customers across a range of vertical industries. He is currently a Business Intelligence and Data Warehouse Architect in Microsoft’s US Public Sector Services group. Mr. Johnson is a 15-year veteran of Microsoft and has worked in various roles there, including Development Manager, Consulting Manager, and Architect. He has worked as a Project Manager and Lead Architect for an end-to-end multiyear data warehouse solution for a national retail company. He has consulted with dozens of customers in the retail, healthcare, energy, public sector, and other verticals on the project methodologies, architectures, patterns, and practices for implementing enterprise-scale data warehouse and BI solutions built on Microsoft software products. Mr. Johnson has been a frequent speaker at internal Microsoft training events and public conferences, such as The Data Warehousing Institute (TDWI), The Professional Association for SQL Server (PASS), and Teradata Partners conferences. Mr. Johnson represents Microsoft’s US Public Sector Services in the Big Data/Analytics Group as part of the Executive Mosaic Group. He can be reached at richjohn@microsoft.com.

Ron Zahavi is currently a Senior Enterprise Strategy Architect with Microsoft. He has over 25 years of experience in all aspects of technology management and solution delivery. Prior to joining Microsoft, Mr. Zahavi held positions as CEO and founder of Auron Technologies LLC, his own consulting business, and Chief Business Architect at Unisys Corp. As a CTO/CIO, he managed technology across several companies and performed due diligence of potential acquisitions. His breadth of experience includes work with startups, large companies, government, and private equity firms. Mr. Zahavi has also worked in several business domains, including healthcare, pharmaceuticals, financial services, intelligence, and defense. He has served on the OMG Architecture Board, is certified as an OMG Expert in BPM, and is a member of several regional and national CIO and CTO councils. Mr. Zahavi is the author and coauthor of several books, including Business Modeling: A Practical Guide to Realizing Business Value. He holds a BSEE from the University of Maryland and an MS in computer science from Johns Hopkins University. Mr. Zahavi can be reached at ron.zahavi@microsoft.com.
The Semantic Web, a collection of technologies designed to add meaning and enable intelligent search across the Web, is now taking shape as a strategy to help navigate the challenges of Big Data, including how to find insights and leverage the ever-increasing volume of data available to us from sensors, social media posts, videos, pictures, and purchase transactions, just to name a few. But more than just size alone, Big Data is seen as the convergence of “3 Vs”: volume, velocity, and variety. Data comes at us not only in a larger-than-ever volume, but at unprecedented speeds and in a variety of formats. This heterogeneous nature of Big Data makes it difficult to devise a one-size-fits-all solution for companies seeking to use Big Data to gain a competitive edge.

Options for working with Big Data include data mining and warehousing, statistical analysis, machine learning, column-oriented analytics databases (such as those used by Zynga and Groupon), and the increasingly popular parallel processing options built around the open source Hadoop project and Google’s MapReduce frameworks for distributing processing across thousands of computationally independent computers. However, such options are most effective when the Big Data is homogeneous and can be efficiently partitioned across multiple computers. If there are multiple data sources, stored in a variety of formats, then a large-scale Hadoop MapReduce effort will not be able to uncover linkages, since each computer in a Hadoop cluster will only be searching a subset of the data. What is needed is a technology suited to making linked data connections. And for that, we turn to the Semantic Web.

SEARCHING FOR MEANING: THE SEMANTIC WEB

The word “semantics” comes to us from the Greek semantika, the study of meaning. Answering the question “What does all this data mean?” has taken on great importance as we are confronted with data coming at us from all sides and in a wide variety of forms. The reason behind this surge of data is the World Wide Web, perhaps one of the premier technological achievements of the late 20th century. The brainchild of Tim Berners-Lee, the Web as we know it today is the result of the confluence of simple technologies: the browser, HTML, and a text-based protocol for data transfer known as HTTP. Building upon an Internet infrastructure, Berners-Lee’s vision was that “by promoting interoperability and encouraging an open forum for discussion,” we will see “the technical evolution of the Web” by its three design principles of interoperability, evolution, and decentralization. However, Berners-Lee’s vision went beyond the Web itself to what some refer to as Web 3.0 — “an extension of the current Web in which information is given well-defined meaning, enabling computers and people to work in better cooperation.” For Berners-Lee, “most of the Web’s content today is designed for humans to read, not for computer programs to manipulate meaningfully,” and so as he sees it, the Semantic Web will provide an environment in which software will be able to help companies leverage Web data in new and interesting ways.

From this vision of the future, the Semantic Web was born. At the core of the Semantic Web is a data model called the “Resource Description Framework” (RDF), a simple language for stating simple facts about things. At the core of RDF are mini-sentences called “triples.” These mini-sentence triples consist of a subject (what we are talking about), a predicate (or property), and an object (or value). Thus, if we want to say “Marla is a female living in Sweden,” we can break this down into the following mini-sentences composed of RDF triples:


The advantage of RDF triples is twofold. First, they are simple, adhering to the W3C’s philosophy of embracing simplicity over complexity. This simplicity is reflected in the fact that three terms is the minimum needed to express information about things. With two terms, we can only say “:gender :female” or “:Marla :gender,” but that does not get us very far. In the Semantic Web, the RDF triple is the atomic data particle used to build more complex data structures. Second, their regularity and simple structure make triples amenable to processing by software, an important aspect of Berners-Lee’s vision of the next-generation Web.
RDF draws inspiration from an earlier classic conceptual modeling approach known as the entity-relationship (ER) model. In RDF, the subject denotes the resource (anything we want to say something about), and the predicate denotes traits or aspects of the subject, expressing a relationship between subject and object. For example, to represent the fact “The water has the color blue” the RDF triple would consist of a subject, “the water,” a predicate denoting “has the color,” and an object denoting “blue.” RDF differs from the ER model in that in RDF both predicates and objects can be turned around and used as the subjects of new triples, giving rise to a web of interconnections that can be used to model arbitrarily complex sets of links and interconnections. Thus, a collection of RDF statements intrinsically represents a labeled, directed multigraph. As such, an RDF-based data model is more naturally suited to certain kinds of knowledge representation than the relational model.

As illustrated in Figure 1, triples can be interconnected to form the equivalent of a network of facts and relationships.

**SPARQL: THE SEMANTIC QUERY LANGUAGE**

But RDF triples do not stand alone. Asking questions about RDF triple data is the province of SPARQL, an RDF query language capable of querying RDF databases. SPARQL (which stands for Simple Protocol and RDF Query Language) includes capabilities to retrieve and manipulate data stored in RDF format.

SPARQL was standardized by the W3C as an official Recommendation in 2008.

SPARQL provides capabilities similar to those of conventional relational database languages such as SQL. It includes constructs that can be mixed and matched, giving a wide range of query options. As illustrated in Figure 1, a search for “Individuals who know each other and like sushi” can be resolved by the following SPARQL query:

```sparql
SELECT ?x ?y
WHERE {
  ?x :likes :Sushi.
}
```

In the above query, the unknown variables ?x and ?y are prefixed by a “?” which triggers an internal match for all subjects that know one another. From this set of various ?x and ?y matches (our data set includes only one for simplicity) a search is made to determine if those subjects have matching triples that indicate the individuals who like sushi. The end result is the output of all the matches of ?x and ?y that satisfy the WHERE clause of the SPARQL query. The result is ?x = Joe and ?y = Sally.

In addition to the SELECT clause and the WHERE clause, a SPARQL query can include a FILTER clause to further refine a search. For example, to limit the above search to only those subjects below the age of 20, the following SPARQL query could be used:

```sparql
SELECT ?x ?y
WHERE {
  ?x :likes :Sushi.
  ?x :age 19.
}
```

Figure 1 — RDF triples give rise to a web of interconnections.
SELECT ?x ?y
WHERE {
  ?x :likes :Sushi.
  FILTER (?x :age < 20) && (?y :age < 20)
}

SPARQL also has strong support for querying data that may be partially incomplete. For example, in the above search for those under age 20 who know each other and like sushi, if the data about some of our subjects is missing age data, we will not retrieve those subjects in our result. However, we may want to retrieve all those we know are under age 20, exclude those we know are over 20, and include those for whom we don’t have age data, just so we don’t miss our target audience. SPARQL allows us to go ahead with the match, even if age data is not available. To do this, we make use of the OPTIONAL keyword, as in:

SELECT ?x ?y
WHERE {
  ?x :likes :Sushi.
  OPTIONAL {
    FILTER ((?agevalue1 < 20) && (?agevalue2 < 20))
  }
}

Together, RDF and SPARQL form a powerful combination for representing data and for doing complex queries across disparate data sources not sharing a single native representation. Because RDF represents all data as a collection of simple relations of the form subject-predicate-object, most data can be easily mapped to RDF and then queried and joined using SPARQL. Now let’s look at how this semantic technology can help us deal with Big Data.

THE CHALLENGE OF BIG DATA

Big Data — it’s a catchphrase, it’s a phenomenon, it’s a burgeoning industry. At the 2012 World Economic Forum Annual Meeting in Davos, Switzerland, Big Data was a featured topic, with a report entitled “Big Data, Big Impact.” In March 2012, the US federal government announced a $200 million research program for Big Data computing. Even Dilbert’s boss in the Scott Adams comic strip has gotten into the act, describing Big Data as “It comes from everywhere. It knows all.”

Big Data is the shorthand label that refers to the billions of bytes of data we are reaping from all the sensors and servers that are storing our every credit card swipe, registering climate and traffic data, capturing text from social media sites, tracking every click on every Web page we visit, and archiving cell phone pictures and videos uploaded to websites. According to IBM, “Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.” We are surfing an exponentially rising data curve.

The challenge for organizations is determining how they will integrate all this data, in all its variety, and leverage the knowledge and connections inherent in the data. We know there is insight and opportunity lurking in Big Data that can lead to a business advantage. Organizations can use Big Data to enhance the customer experience, track social networking sites to study the impact of marketing campaigns, and engage in real-time analytics to study user behavior, among other examples. The difficulty is connecting the dots.

Looking around we see companies like Google, Facebook, and Twitter capitalizing on the Big Data explosion. But Big Data storage and analysis is their business. Kalyan Viswanathan of Tata Consultancy sees a widening gap in the next three to five years between companies that understand how to exploit Big Data and those companies that are aware of Big Data but don’t know how to proceed. Companies that succeed in turning Big Data into actionable intelligence will have a clear competitive advantage. But how can an organization begin to evolve a Big Data strategy?

For starters, let’s begin with the word data. Data, the plural of datum, is derived from the Latin dare, “to give.” Thus, data is that which is “given to us.” The term “Big Data” gets our attention mainly because of the word “big,” which generally means “lots” but can also describe an exceptional situation. Consider: big problem, big mess, big trouble. Each of these uses of the word “big” implies something beyond the conventional meaning of the associated term. When a kid comes home from school and is told that he is in “big trouble,” or when the boss calls to let you know we have a “big problem,” it generally means we’ve moved to a dimension that requires some out-of-the-ordinary response or solution.
In the 1970s, 1980s, and well into the 1990s, data was, well, data. “Data processing” meant working hard just to get data into files and databases. Data entry clerks spent hours keypunching and green-screening data from forms into the bytes that found their way into files and databases. Transaction processing engines emerged to add reliability to the incoming flow. Volume, if a problem, could be solved by simply acquiring a larger disk or moving from a spreadsheet to a database. “Big” existed, but there was typically a ready solution to deal with the bigness.

Adam Jacobs, in an article entitled “The Pathologies of Big Data,” enumerates a litany of problems in working with Big Data. Jacobs recognizes Big Data as a useful catchphrase for our new data-centric world, but he also sees the term as a moving target, relative to existing capabilities. For example, if you attempt to load a million and a half records into a 2007 or 2010 Excel spreadsheet, you have Big Data. The Excel limit of 1,048,576 rows and 16,384 columns stops you cold. The solution, of course, is to move those records to a database that can handle the volume. And if you run up against database limits, move over to a data warehousing solution.

But as Jacobs points out, the problem is not so much getting data into a database. Companies have pretty much solved the problems of data entry and data storage. Storing large volumes of data is no longer a problem due to the availability of low-cost disk storage, and the problem of reliable data entry via transactions was solved by IBM with their CICS systems in the 1980s. What’s pushing the envelope is that the quantity problem now is exacerbated by our ability to continuously feed data into our systems from sensors and servers running 24/7. The Web logs millions of records a day, cell phone databases store time and location every 15 seconds for each of several million phones, and scientific measurements often have a time resolution of thousands of samples a second. No longer do we need to work hard to get data into our systems.

This combination of data quantity and the velocity at which it is funneled into databases creates problems not so much for data storage itself (we can always purchase more disk space), but for database queries that attempt to aggregate data from disparate sources with similar time stamps. Jacobs has found that processing times increase up to 10 times when doing queries on relational databases not optimized to execute time-based queries. For example, attempting to establish connections across multiple database tables (e.g., video data, audio data, and bank transactional data) based on separate time stamps from each table requires more complex database processing than queries based on traditional primary and secondary key fields such as user IDs, names, or ZIP codes.

BIG DATA BREAKDOWN
Beyond the performance issues associated with processing Big Data in relational databases, the reality is that much of the Big Data comes at us from a wide variety of sources not well suited for storage in relational tables. Structurally, the existing corpus of data in the world falls into three categories:

1. **Structured data** is the data we are most familiar with. It’s primarily the stuff of our relational databases. Traditional sources of structured data are transactional systems such as ERP, CRM, SCM, and other data management systems. Structured data has been neatly organized and formatted in ways that make it easy to manipulate and manage. In the relational

---

Figure 2 — Data falls into three categories: structured, semistructured, and unstructured.
database world, the Structured Query Language (SQL) is the workhorse for querying relationally structured data. Other data in the form of spreadsheets, fixed-format files, and formatted log files is widespread and has a structure that makes it easy to work with.

2. **Semistructured data** is a kind of structured data that does not fit neatly into the tables of the relational database model but does contain internal indicators (tags, labels, or keys) that provide semantic meaning. Semistructured data is also known as “schema-less” or “self-describing” due to its internal semantic tags. The volume of semistructured data has increased greatly due to the rise of Web services and data transfer on the Web. The two most widely used forms of semistructured data are XML and JSON (JavaScript Object Notation). XML contains tags that describe the data, as in “<date>2010-02-14</date>. JSON treats data as keys and values, as in (“date”：“2010-02-14”).

3. **Unstructured data** sits at the opposite end of the spectrum from structured data. Generally, unstructured data comprises the mass of information that does not fit easily into a set of database tables or XML or JSON. The most recognizable form of unstructured data is text — documents, articles, and slide presentations, as well as the text from email, blogs, SMS, social networking sites, and text fields. Unstructured data can also take the form of audio, video, and images. Estimates are that 80% of the data in organizations is unstructured, and according to analyst firms, “employees within an organization generate more than 3 GB of data each year, and this is set to increase by more than 600% over the next five years.”

How can companies leverage the knowledge contained in the disparate data sources to their advantage? One approach is to use semantic technology, specifically RDF triples and SPARQL.

**FROM DATA TO ACTIONABLE KNOWLEDGE**

Figure 3 illustrates how a central core of RDF triples can be used as the basis for analysis and querying using SPARQL. Since an RDF triple is a sentence about data in its most basic form, (i.e., subject-predicate-object), transforming that data, whether from a database table or from XML or JSON, can be done in a straightforward manner. As Figure 3 shows, structured data in the form of database tables or spreadsheets provides a direct path to RDF triples when entities are arranged in well-defined rows and columns. When each row defines an entity, the data can be broken down using column headers as RDF properties/predicates and column values as values of those properties. Looking at Figure 3, we see that each row in a table results in as many triples as there are columns in the table.
As noted above, semistructured data — data with a recognizable but flexible structure — is typically in the form of XML or JSON. Both XML and JSON are widely used to move data across the Web. XML includes well-defined XML element names that can be mapped to RDF properties/predicates, and element values can be mapped to RDF objects. For JSON, the name in a JSON name-value pair becomes the predicate and the value becomes the object of the RDF mini-sentence.

Unstructured data makes things a bit more complex, as its lack of associated semantic tags or indicators gives us little to work with. However, when the unstructured data has a Web URL, we can at least begin to talk about the data using our mini-sentence triples.

For example, consider a large collection of unstructured audio files stored on a server. Although we have no knowledge as to the actual content of each file, we do know:

- The URL of the file
- That it is an audio file (from the file type extension)
- The time and date of creation (from the timestamp)

Since the file is a Web resource, we automatically have an RDF subject, since any URL can serve as a subject. Then, using RDFS, the RDF schema technology for declaring resource type, we can add a triple that declares that our URL represents an audio file. And since we know the creation date from the file system, we can add another triple that puts a date on our URL, giving us:

```
:URL-of-audio :date "2012-02-14".
```

Since a triple store is basically a text file, this can be handled quite easily by a simple conversion program. Thus, even though we don’t know the content of the audio file, we can use SPARQL to search for “All audio files with a creation date of 14 February 2012.”

If the file contains text, we can drill even deeper, using a human analyst or one of a number of Semantic Web tools, such as OpenCalais, to parse the text and add additional meaning to our RDF triple database. With our data in a common RDF format, we can now utilize SPARQL to ask questions and uncover useful relationships.

**SEMANTIC DATA FOR FREE**

One side benefit of moving data to a common RDF triple format is the ability to make connections with other linked data sets stored as Semantic Web triples. One popular, widely accessed triple store is DBpedia, which is based on content extracted from Wikipedia. DBpedia allows users to query relationships and properties associated with Wikipedia resources, including links to other related data sets. As of August 2012, the English version of the DBpedia knowledge base describes 3.77 million things, including 764,000 persons, 573,000 places, 333,000 creative works (including 112,000 music albums, 72,000 films, and 18,000 video games), 192,000 organizations (including 45,000 companies and 42,000 educational institutions), 202,000 species and 5,500 diseases. The data set consists of 1.89 billion pieces of information (RDF triples), out of which 400 million were extracted from the English edition of Wikipedia; 1.46 billion were extracted from other language editions. From this data set, information spread across multiple pages can be extracted. For example, book authorship can be linked from pages about the work or the author.

In addition, the Data-gov Wiki project at Rensselaer Polytechnic Institute is working on translating open government data sets into RDF, linking them to a broader linked data cloud, and using the data to stimulate interesting applications that build on linked government data. While most of the data comes from the US government’s data.gov website, there is also data from other countries and nongovernment sources.

**DATA MODELS COME FULL CIRCLE**

At first glance, converting data from relational data tables and other structured sources into simple triples may seem like a radical idea. However, since triples combine to form a network of linked data, it’s instructive to look back at the history of database management systems (DBMSs) to see how the data graphs that arise from triples fit into the grand scheme of databases. Historically, databases fall into three major types: hierarchical, network, and relational. Hierarchical databases first arose in the 1960s as part of mainframe data processing. Their strength was that they mirrored the natural hierarchy occurring in many areas of data management. However, hierarchical databases were limited
in their ability to connect data elements across different hierarchies. This led to the emergence of network databases, which are similar to hierarchical databases except that they allow links between hierarchies, mirroring the graphs that arise from triples.

In the 1970s, the mathematical work of E.F. Codd and others created the now-ubiquitous relational database model. The basic idea was to break up natural hierarchies into tables in order to take advantage of Codd’s mathematical formulations. When first proposed, the notion of breaking up hierarchies and interconnected structures into relational tables was considered foolhardy from an efficiency standpoint. However, over time, database companies such as Oracle figured out how to optimize their database structures, computing power increased, and relational databases became mainstream, so much so that the word “database” has come to mean “relational database” for most business IT folks.

Today, the rise of Big Data is challenging the hegemony of the relational model. While effective for highly structured data, the relational data model has difficulty incorporating semistructured or unstructured data into its tables. Harking back to the early days of databases, it’s interesting to see the thinking about data models coming back around to the utility of the linked-data, network model. Made possible by the Web and the semantic vision of Tim Berners-Lee, the RDF triple — that basic, simple subject-predicate-object sentence — now holds the key to unlocking the secrets that lurk within the growing behemoth we call Big Data.

ENDNOTES


3Berners-Lee et al. See 2.


8Jacobs, See 7.

9Jacobs. See 7.


12OpenCalais (http://opencalais.com).


Frank P. Coyle is a Senior Lecturer in the Computer Science and Engineering Department in the Lyle School of Engineering at Southern Methodist University (SMU) in Dallas, Texas. He is the author of XML, Web Services, and the Data Revolution and is a frequent speaker at XML and Semantic Web conferences. He can be reached at coyle@smu.edu.
Managing population health entails a complex set of activities involving many interconnected parts: health and its multiple determinants, healthcare delivery systems, information technology, and analytics. Personal biology and genetics; diagnosis, testing, and treatment across multiple providers and settings; and a host of environmental factors all affect health outcomes and the quality and cost of care delivery.

Until recently, pulling together the big picture with sufficient detail to target and address issues for improvement was neither practical nor feasible, even in large national health systems. But things are changing fast. The widespread adoption of IT in healthcare and the massive stores of data — “Big Data” — it generates offer new analytic possibilities that can transform the focus and delivery of population health. Cloud databases, supercomputers, large centralized data warehouses, and distributed research networks, coupled with a growing cadre of powerful analytic applications, create the Big Data tools that make it possible to store and process Big Data for use in improving population health. The output of these tools can lay out the evidence, introduce checks and balances, and empower choice in healthcare markets. It can provide comparative information about the performance, outcomes, and prices of healthcare providers and plans; identify best practices and treatments; and bring to light gaps in quality care.

**SOURCES OF POPULATION HEALTH BIG DATA**

In order to achieve value-based healthcare’s quality improvement and cost-reduction goals, one of the most important functions of population health management (PHM) is managing chronic conditions. Big Data, digital infrastructure, and analytics can help pinpoint unwanted drug interactions, identify the most effective treatments, predict the onset of chronic disease, and identify inequities and gaps where standards of care are not met.

The emergence of Big Data for PHM is an outgrowth of the IT explosion, with the adoption of electronic medical records and patient-centric delivery models that integrate care across health and community services. It also depends on the availability and potential of enhanced, low-cost computing capabilities, including cloud-based and mobile computing. The growth of evidence-based medicine and data-driven healthcare decision making further fuel interest in Big Data and its application. Massive data sets generated by and integrated across the healthcare landscape provide the foundation for dynamic performance-based health systems. This places Big Data front and center of such concepts as the Learning Healthcare System, an approach to care delivery that uses metrics and performance dashboards to monitor and focus healthcare processes and activities for continuous improvement as they relate to patients, providers, and health system outcomes.

As the sampling of data sources in Table 1 illustrates, these strategies involve massive amounts of information from vastly different sources. These include electronic health records (EHRs), claims data from multiple providers and payers, and information about social determinants from large public data sets. Eventually, even genomic information could become commonplace as a guide to highly individualized care.

**DATA INTEGRATION TECHNOLOGIES**

Expanded IT capabilities, the availability of low-cost computing capabilities, and the growing demand for quality and cost data from new care-delivery models are giving rise to tools for managing Big Data as a way to improve population health. Nearly every interaction with the modern healthcare system is digitized, from diagnostic and screening tests to prescription drug ordering, dispensing, and management through pharmacy systems and EHRs. Healthcare is becoming increasingly integrated. Horizontal integration is happening through the consolidation of hospital and physician group networks, while vertical integration is occurring as health plans integrate providers across the healthcare continuum. This increased integration puts
additional pressure on healthcare systems to make
digitized information available and to use it intelli-
gently to manage care delivery.

Pulling together data from multiple sources enables
analysts to derive a more comprehensive picture of
patient health and the healthcare system in general.
Data aggregation approaches such as portals offer
ways to gather data on quality of life, functional status,
and self-management. An example is Epic Systems
Corporation’s MyChart, a consumer portal that lets

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Description</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electronic Health Record (EHR)</td>
<td>Digitized health record that includes information about an individual patient’s clinical history, medical visits, test results, and other pertinent facts.</td>
<td>Multiple facilities and agencies share the EHR, which helps streamline data entry, prevent duplicate tests, and make the most recent information available to providers.</td>
</tr>
<tr>
<td>Medical Image</td>
<td>Radiological photo of the human body (e.g., CT scan, MRI).</td>
<td>Providers can access the stored images through EHRs or imaging centers. Images enable a variety of geographically distant providers to review images for the same patient.</td>
</tr>
<tr>
<td>Personal Health Record (PHR)</td>
<td>Centralized electronic record of an individual’s health information, including complete medical history.</td>
<td>A PHR is accessible via the Internet and can be shared with designated others.</td>
</tr>
<tr>
<td>Pharmacy E-Data</td>
<td>Information captured electronically about prescribed medications.</td>
<td>Pharmacy e-data enables real-time access to information across providers, which reduces errors, drug interactions, and polypharmacy complications.</td>
</tr>
<tr>
<td>Registry</td>
<td>Centralized database that contains information about individuals diagnosed with a particular disease or condition.</td>
<td>Providers can use registries as a way to monitor patient groups.</td>
</tr>
<tr>
<td>Public Data Set</td>
<td>Contains secondary data such as census demographics, public health surveillance, income, environment, crime, housing, education, labor, and other factors.</td>
<td>Providers can combine this data with patient-level data to profile individuals in their social contexts.</td>
</tr>
<tr>
<td>Claims and Payer Data</td>
<td>Data about healthcare services that insurance companies request and process.</td>
<td>Analysts can use claims and payer data to compare patient and provider use and cost.</td>
</tr>
</tbody>
</table>
patients enter their own assessments, view lab results, make appointments, renew prescriptions, and share information with their providers. MyChart’s patient-centered approach also supports transparency and consumers’ engagement in their healthcare.

Most PHM programs are multipronged and involve the application of several approaches as part of a comprehensive, integrated quality improvement strategy. As a result, many large vendors offer an analytics platform with multiple capabilities. The GNS Healthcare REFS (Reverse Engineering and Forward Simulation)\(^1\) platform uses machine-learning algorithms run on cloud-based supercomputers to extract models from observational data as well as simulations to ask questions about interventions. Among other things, REFS enables healthcare professionals to:

- Compare the effectiveness of various treatments by simulating the range of possible outcomes from changing treatments or services.
- Predict adverse drug events and polypharmacy\(^2\) effects using data from pharmacy benefit managers.
- Optimize treatments based on patient demographics and clinical data, with biomarkers.

IBM also offers tools that can be used in a Big Data PHM effort:\(^3\)

- The InfoSphere Streams platform, a strong market player, enables real-time analytic processing of streaming data.
- InfoSphere BigInsights can be used to manage and analyze massive amounts of data.
- InfoSphere Warehouse allows its users to access information in real time.
- The IBM Smart Analytics System provides a modular portfolio of data management, hardware, software and services for business analytics, as well as related data warehouse and analytics products.

Intelligent data integration is central to the vision of PHM improvements. Data from divergent sources, now in silos, must be linked to enable the most effective treatment. Figure 1 shows some key PHM tools that use Big Data inputs and are instrumental in integrating data.

**BIG DATA ANALYTICS IN PHM**

Big Data analytics (BDA) — examining huge volumes of real-time or near-real-time data to identify patterns and correspondences — can provide strong support for PHM provisions that address wellness and prevention, care management, and integrated care delivery. Use of BDA can enhance the effectiveness of new delivery models like patient-centered medical homes, accountable care organizations, and risk-based payment schemes. These value-based approaches integrate and coordinate care delivery across providers and services and hold providers accountable for managing patient and population-based outcomes. IBM, Oracle, Microsoft, SAS, and McKesson are among the leaders in developing these new data mining applications.

With so many key healthcare tools relying on Big Data, there is a pressing need for new combinations and applications of digital information technology and analytics. Such developments will let analysts better understand integrated care across providers, detect gaps in care, identify testing to avoid duplication, and detect and address polypharmacy, which is a frequent concern in treating patients with several chronic conditions.

Multiprocessors are sophisticated enough to capture the huge data sets that healthcare systems generate, strip them of patient identifiers, and then analyze them using novel methods of discovering correlations and associations. Interoperability options are critical in supporting data exchange and information sharing — from specialist reports to discharge summaries to imaging studies. All these activities are integral to care coordination and PHM.

![Figure 1 — Big Data in population health.](image-url)
BDA draws on an extensive and evolving array of modeling, correlation, and risk stratification methods to recognize patterns in disease and treatment that enable faster innovation, better care, and lower cost. The healthcare industry has heavily invested in application development, and market competition is fierce.

Explorys, a spinoff of the Cleveland Clinic, uses an open cloud platform and an integrated framework for correlating clinical, operational, and financial events into actionable benchmarks, scorecards, and dashboards. Watson, the analytic engine developed by WellPoint for IBM’s supercomputer, enables providers to analyze patient care histories, test results, recent medical literature, clinical trials, and outcomes. GE Healthcare and Microsoft recently announced Caradigm, their joint venture to develop and market a real-time, open, interoperable technology platform and collaborative clinical application for PHM.

Large, centralized databases continue to support some PHM activities, but as Big Data gains a foothold, attention is shifting to a decentralized, networked technical architecture with original sources retaining control of the data they generate. One such architecture is the distributed research network, which, relative to a centralized database, provides more timely and nimble data access and offers a more practical solution for storing large volumes of data. Use of a common data model helps enforce consistent definitions and naming conventions. A common model also makes a distributed research network appealing because such a network simplifies analysis and allows protected health information and other proprietary data to remain in the original data holders’ possession.

Big Data methods and analytics expand possibilities for PHM beyond the disease- and service-specific silos common to fragmented fee-for-service delivery. Combining data across all providers and services, Big Data tools enable 3D visualization and mapping of patient health and healthcare patterns and trends, commonalities, and abnormalities; allow multiple variable analyses that incorporate biometric, behavioral, and attitudinal data on patient preferences and satisfaction; and support decision making and assessment of the comparative effectiveness of treatments based on patient characteristics. These approaches facilitate identification of patients with multiple risk factors for chronic conditions or those at risk for readmissions, pick up on gaps or treatment errors, and can be used to detect and manage individual, facility, and system-level provider performance, including clinical outcomes, care delivery processes, and patient satisfaction.

Adding data from outside the health sector — such as education, employment, and/or housing data — further expands this view to show individuals in their environment. Data aggregation approaches open new opportunities to include measures of quality of life, functional status, and self-management. There are also new opportunities to encourage consumer engagement and shared decision making, such as portals where patients can enter their own assessments and share information with their providers. Many predict that genomic data will soon be incorporated as part of the Big Data toolset, providing information essential to individualized medicine.

Analytic Approaches for PHM Efforts

Analytic approaches such as predictive modeling and comparative effectiveness research are the first steps in fulfilling the integration mission. The following analytic processes are vital to a PHM Big Data initiative:

- **Predictive modeling and risk stratification** detect trends and identify patients at risk, in treatment, and in need of care transition management. Once the province of sophisticated health systems and plans, these approaches are becoming mainstream thanks to the lower costs of technology and computing, expanded access to clinical information through EHRs, and development of and experience with predictive analytics. Public and private health systems, especially those serving older populations with costly chronic care conditions, are prime candidates for such applications. The German BVA (Social Health Insurance Office) is one example. When adjustments made to reflect the patient’s age, sex, and disability status were not considered sensitive enough to manage population-wide risk for Germany’s Sickness Funds, the BVA, working closely with Verisk Health’s German subsidiary, employed predictive modeling to support reimbursement adjustments to reflect the increased disease burden of its aging population. This more refined approach enabled assessment of possible costs within the context of the new system, both in the present and the future. Similar models are used for risk management in the US and other developed nations.
- **Comparative effectiveness research** identifies best treatment options and supports consumer choice and shared decision making. Harvard-affiliated Beth Israel Deaconess Medical Center’s recent launch of Clinical Query — a searchable patient data repository that includes more than 200 million data points on diagnoses, medications, lab results, and radiology images for over 2 million patients — is attracting widespread attention. Clinical Query offers physicians a way to compare treatment options, risk factors, and resource use for various diseases and conditions as well as the evidence needed for shared decision making.6

- **Profiling**, a quality monitoring technique, has been a standard tool for health plans, payers, and quality improvement organizations for several decades. Organizations can profile patients according to risk, utilization, and cost, and they can profile providers to monitor and report on performance, quality, and cost or value of care. While the technique is well established, a new generation of products using Big Data inputs and analytics brings profiling to new levels. For example, a new program to treat and prevent metabolic syndrome, developed by GNS Healthcare with Aetna, combines the GNS supercomputer-driven REFS platform with claims and other health information from Aetna to define an individual’s risk for developing the condition. For individuals with one or two of the disorder’s five indicators (large waist size, high blood pressure, high triglycerides, low HDL cholesterol, and high blood sugar — three or more of which define the condition), the model will predict which new indicator the patient will likely develop next and how quickly. The model then matches the individual with specific interventions that are most effective at reducing or eliminating risk factors for the condition.7

- **Registry review** for specific diseases or conditions supports panel management and quality improvement efforts. Disease or condition registries predate the widespread use of EHRs, providing a relatively straightforward and inexpensive quality improvement tool for targeting specific conditions, patients, and their treatment regimens. Big Data expands the scope and potential of this approach. De-identified information on more than 1 million people with diabetes, obtained from the EHRs of 11 integrated health systems, was used to create a large diabetes registry called “Surveillance, Prevention, and Management of Diabetes Mellitus DataLink,” or SUPREME-DM DataLink. Researchers analyzed inpatient and outpatient diagnosis codes, laboratory test results, and pharmaceutical distributions from 15.8 million EHRs to create a database designed to support epidemiologic surveillance, population-based care management studies, clinical trials, and other research. The database is maintained by 33 diabetes researchers and includes patients’ test results, prescription records, information on hospital and ambulatory visits, and vital statistics.8

### Clinical surveillance, including identification of gaps and errors in care, is an area of quality management readily enhanced by BDA.

- **Polypharmacy review** identifies medication errors. Polypharmacy, the use of multiple medications, may lead to harmful interactions or unnecessary pharmacy use, with consequent adverse quality and cost impacts. Such impacts are relatively common among elderly adults with multiple chronic conditions. Analytics to enable review of pharmaceutical use and prescribing practices are offered by many vendors, including IBM, SAS, GNS Healthcare, and others. Health iQ consultants in the UK have developed the Pharmasim modeling and simulation analytics tool to address a wide range of pharmacy scenarios. Pharmasim uses virtual models that input health system data and run simulations on it to examine interactions and processes across the system over time, targeting problems for follow-up, demonstrating outcomes, and communicating value.9

- **Data mining** uses ICD-9 code,10 pharmacy, cost, or utilization patterns to detect over- or underutilization and nonstandard care. EHRs are one of the richest sources of data here and have all but replaced administrative claims for mining large data sets for quality improvement. In 2010, use of Microsoft technology to scan patient records reportedly led to a significant drop in the rate of potentially fatal blood clots — by about a third — at New York–Presbyterian Hospital.11

- **Gaps identification** compares and contrasts services received and provided relative to evidence-based standards for care. Clinical surveillance, including identification of gaps and errors in care, is an area of quality management readily enhanced by BDA. It is also growing in importance as public and private health systems continue the move to performance-based payment and value-based care. New tools now enable
THE LEARNING HEALTHCARE SYSTEM

Big Data is a pillar of the Learning Healthcare System, which is:

- A healthcare environment that integrates data, technology, and care culture to generate new knowledge about effective prevention and treatment strategies that can be disseminated rapidly to clinicians and patients to improve the quality and efficiency of healthcare.¹

In this environment, the healthcare system learns quickly from its successes and failures.

The vision for the Learning Healthcare System — which stems from a 2001 critique of the US healthcare system, Crossing the Quality Chasm² — is to remedy healthcare’s shortcomings by consistently delivering the best outcomes. Ultra-large-scale systems or Big Data would support the Learning Healthcare System in this continuous improvement. Together they would provide the infrastructure for healthcare reform and the development of novel value-based delivery models to improve quality and reduce per capita healthcare costs.

In this vision, interprofessional teams access and share electronic information for clinical decisions that coordinate care. Patient-centered medical homes and accountable care organizations are examples of this coordinated care model. Sharing information and coordinating care reduce costs by eliminating service duplication and improving population health through the use of standardized guidelines that ensure appropriate treatment.

The Learning Healthcare System embraces the following key elements:³

- **Culture** — participatory, team-based, transparent, improving
- **Design and processes** — patient-anchored and tested
- **Patients and public** — fully and actively engaged
- **Decisions** — informed, facilitated, shared, and coordinated
- **Care** — starting with the best practice, every time
- **Outcomes and costs** — transparent and constantly assessed
- **Knowledge** — ongoing, seamless product of services and research
- **Digital technology** — the engine for continuous improvement
- **Health information** — a reliable, secure, and reusable resource
- **The data utility** — data stewarded and used for the common good
- **Trust fabric** — strong, protected, and actively nurtured
- **Leadership** — multifocal, networked, and dynamic

New quality measures build on the idea of integrated care across an expanded healthcare continuum — from disease prevention to end-of-life care. The continuum for care delivery might include support for care transitions and community-based prevention and wellness programs, as well as more conventional institutional and clinical care. The continuum is the context in which tools must evolve to analyze Big Data and to standardize measures that can translate that data into actionable information.

VALUE-BASED RESULTS FROM MARKET LEADERS

Some of the US’s leading health systems and providers, including Kaiser Permanente (KP), Geisinger Health System, and the Veterans Health Administration (VHA), have broad experience and demonstrated success in managing Big Data to advance value-based healthcare and improve population health. KP’s HealthConnect Systems uses EHRs across its facilities and clinicians to provide a comprehensive record of patient encounters. The system enables electronic prescribing and test ordering, disease registries, interaction with decision support tools, electronic referrals, personal health records, performance monitoring, and patient billing activities. KP has seen greater patient satisfaction, fewer medication administration errors in hospitals, and improved patient interactions with HealthConnect Systems.⁴

Geisinger reduced hospital admissions by 15%, finding savings for its Medical Home initiative, by leveraging MEDai’s Risk Navigator Clinical predictive modeling solution.⁵

The VHA developed its Veterans Health Information Systems and Technology Architecture (VistA) as an EHR-based enterprise-wide system. VistA, now an open system, has received numerous awards for quality improvement, with efficiency improvements of an estimated 6% per year.⁶

As these elite groups gain experience in using Big Data to improve population health, we expect that smaller organizations will begin duplicating these methods, possibly reaping similar benefits.

real-time surveillance and reporting for continuous monitoring of clinical care. IT experts at CSC, in partnership with the University of Kansas Hospital, developed the CareVeillance tool to integrate patient data from disparate systems, analyze the data against care standards, and report results to clinicians. The tool can also be used to identify best practices. CareVeillance is reported to have helped identify patients with sepsis early in their hospitalization — including some who were not detected through other means — thereby allowing immediate intervention.12

- Duplication of services review — designed to ferret out duplicative procedures such as diagnostic tests, imaging, and screening across providers — is used to reduce or eliminate waste. Care delivery inefficiencies resulting from poor care coordination and management are a major source of quality issues and excess spending in healthcare, especially when patient care involves multiple providers and services across diverse locations. Premier Healthcare Alliance, which includes more than 2,500 hospitals as members, targeted unnecessary care as part of its efficiency dashboard to pinpoint opportunities for savings. Using hospital data pulled from Premier’s comparative clinical, operations, and purchasing database, the alliance identified US $2.23 million per hospital per year in unnecessary lab testing such as blood, urine, or hemoglobin tests. Analyses revealed much of the duplication was associated with readmissions for circulatory, respiratory, and musculoskeletal system conditions. These problem areas could then be addressed by individual hospitals.13

MOVING FORWARD

Moving Big Data management forward in a PHM context requires overcoming some obstacles related to data quality, interoperability, data formats, common nomenclature, and the protection of personal health information. Protocols for incorporating patient-generated information can be barriers to aggregating and translating health data into useful decision support.14 Clinical data leaks and privacy concerns remain important security considerations. Data integrity compromises in current data use — incomplete participation, dirty data, and time lags — can hamper attempts to fully exploit Big Data’s possibilities as well.

Computing costs have dropped far enough to make Big Data management affordable as a way to improve population health, but many organizations may still find these costs prohibitive, particularly in light of the startup costs for the technology and software to generate analytics. Implementing these approaches also requires skilled informatics professionals.

Developing solutions to address these barriers will not be the work of a moment. Data governance across business enterprises will facilitate data sharing, but clearly it will take time to address the related legalities and politics that can make this governance a reality.

Big Data and the analytics that support it have been labeled a disruptive innovation. They are not the magic bullet for PHM or the cure for all health system woes. But they do change the business of healthcare and our understanding of the ecology of health and healthcare systems. Big Data tools for PHM hold great potential for improving value in healthcare by shifting our perspective, demonstrating new evidence from disparate sources, and introducing transparency and accountability.

ENDNOTES

1GNS Healthcare (www.gnshealthcare.com/technology).
2Polypharmacy is the uncoordinated use of multiple drugs.
6Cerrato, Paul. “Beth Israel Deaconess Medical Center Embraces Analytics.” InformationWeek, 12 September 2012.
9“Pharmasim.” Health IQ (http://healthiq.co.uk/pharmasim).
10The International Classification of Diseases (ICD) is “a health care classification system that provides codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease. Under this system, every health condition can be assigned to a unique category and given a code, up to six characters long. Such categories can include a set of similar diseases” (Wikipedia).


Holly Korda is the Director for the Center for Policy and Research Translation in the Systems Research and Initiatives Group at Altarum Institute, where her experience includes design, implementation, and evaluation of new delivery models, payment methods, and incentives in healthcare. She received a PhD in health services research from Tufts University. She can be reached at Holly.Korda@altarum.org.

Ann V. Magee is a Senior Principal at Noblis, where her experience includes healthcare consulting, electronic health records, product recall management, and military healthcare facility design. She received a BS in nursing from Old Dominion University and holds certifications in evidence-based design and project management. She can be reached at AnnMagee@noblis.org.

Lori A. Damiano is a senior staff member at Noblis, where her experience includes data analysis, market analysis, research, geographic information systems, and project organization. She received an MBA from Regis University. She can be reached at Lori.Damiano@noblis.org.
Big Data Analytics: Outsourcing vs. Crowdsourcing
by Saeed Lajami, Anson Mok, Mario Wahyu Prabowo, and Sara Cullen

The analytics of Big Data promises to unlock vast amounts of information just waiting to be known and used. However, Big Data is characterized not only by its extraordinary volume, which grows exponentially every day, but also by its nature — very unstructured, very raw, and far too voluminous to be analyzed using traditional methods such as relational databases.

While many organizations have Big Data, not all of them have the resources and expertise to conduct Big Data analysis. If an organization wants to unlock the promise of Big Data, it may need to consider outsourcing. This article looks at the business case for both traditional outsourcing and crowdsourcing — a recent trend in outsourcing driven predominantly by Big Data. There are pros and cons to both approaches, and there is no one right or best answer. We hope that this article will help those who may be considering outsourcing — and particularly crowdsourcing — Big Data analytics by highlighting the benefits as well as the challenges. Crowdsourcing may have attractive economics and offer greater creativity, but once the confidentiality issues are considered, the “right” answer might well be a hybrid approach.

HARNESSING BIG DATA

The world generates millions upon trillions of pieces of raw data every day. This data is growing exponentially, with 90% of today’s Big Data having been created within the last two years alone.

Many organizations are looking to harness this vast amount of unprocessed and unstructured data in order to gain business value, such as more effective marketing and innovative product/service ideas in the commercial sector, better-informed research in the academic sector, and improved problem solving (such as crime reduction) in the government sector. The processing and examination of Big Data in order to discover the hidden patterns, unknown correlations, and other valuable information within is known as Big Data analytics (BDA).

Big Data is often expressed in petabytes (one thousand terabytes) or exabytes (one million terabytes). Examples of Big Data include blogs, data from social networks, Internet text and searches, call detail records, medical records, e-commerce records, data used by the sciences (astronomy, atmospheric science, biogeochemical, biological, medical, etc.), and media archives (photos, videos, etc.).

We believe that Big Data is best defined by the following:

Big Data is not a precise term; rather it’s a characterization of the never-ending accumulation of all kinds of data, most of it unstructured. It describes data sets that are growing exponentially and that are too large, too raw, or too unstructured for analysis using relational database techniques.

Outsourcing vs. Crowdsourcing

As this article focuses on an organization’s decision to choose to outsource or crowdsource BDA, it is important that we spend a few moments differentiating between traditional outsourcing and crowdsourcing.

Traditional outsourcing in the BDA context entails contracting with a third-party organization to conduct the data structuring and analytics. The provider is typically selected through a formal competitive bidding process. With crowdsourcing, the problem-solving task is outsourced to an undefined public (the crowd) through an open call via a Web-based business model. This approach has spawned a new breed of crowdsourcing technology platforms (e.g., Mechanical Turk, crowdSPRING), service providers (e.g., Kaggle, CrowdFlower, InnoCentive), and “crowdworkers” (individuals who participate in crowdsourcing initiatives for a living). The crowd conducts the BDA and the best solution receives a monetary reward, or “prize money.” In traditional outsourcing, the “winner” is chosen from the bids and then is paid to do the BDA work. In crowdsourcing, the crowd first does the BDA work and the individual or team with the best solution is awarded the prize money.

Not all crowdsourcing involves prize money, however. Nonprofit crowdsourcing is increasingly used as a problem-solving mechanism for government, academic, and nonprofit organizations. A famous example of this...
is the June 2012 public release of thousands of CCTV images (on a Facewatch app) by the UK police so as to engage the public in helping find those suspected of committing crimes.

Six years since the term was first coined by Jeff Howe in a 2006 *Wired* column, crowdsourcing has become an increasingly popular alternative and now presents a challenge to traditional outsourcing of BDA work. In the remainder of this article, we will compare and contrast the outsourcing and crowdsourcing approaches to BDA based on three key areas that organizations typically consider in the business case:

1. **Cost.** Both traditional outsourcing and crowdsourcing offer cost savings in different ways.

2. **Quality.** The manner in which the solutions occur, and are chosen by the client, is radically different.

3. **Risk.** BDA involves data — lots of it. This data is often private and often commercially valuable. Privacy and security are paramount.

### WHICH APPROACH IS CHEAPER?

One of the most influential considerations in any business case is cost, and even more so when it comes to outsourcing decisions. Vendors of traditional outsourcing, particularly in cases of offshoring to lower-wage countries such as India, can usually provide a lower cost to a client than if the client organization performed the work inhouse. This assumes, of course, that the specification is well written, the vendor is selected diligently, and the client manages the contract well.

Crowdworkers tend to be nonprofessional but competent workers who are not paid unless their solution wins. Individuals are the common recipients of the prize money, not organizations, thus no wages are actually paid, nor is there much in the way of overheads.

However, the overheads to the client can be just as high with crowdsourcing as with traditional outsourcing, even though they are dramatically different in nature. Crowdsourcing can initially be cheaper because there is not the need to write a detailed specification and contract, run a competitive tender, and manage the contract. Rather the client must only disseminate the data and the problem to be solved. The larger overhead cost of crowdsourcing occurs in the evaluation. Rather than evaluating a few bid responses to choose the winning provider, the client must evaluate many varied solutions. Thus, the cost of crowdsourcing is not necessarily lower; it just occurs later in the process. The evaluation of the solutions can actually be quite costly depending on the complexity and number of solutions submitted.

At this point in time, there has not been much in the way of research into the client-borne costs associated with crowdsourcing. We do know from a number of global studies (including one performed by Cutter in 2007) that, on average, it costs 6% of the value of an outsourcing contract to manage that contract. However, we have little equivalent data on the cost of crowdsourcing. Nonetheless, some academics have put forward the hypothesis that crowdsourcing’s net cost to a client is substantially lower than inhouse work and lower than traditional outsourcing as well. In crowdsourcing, only the winner is paid, not everyone who performed the BDA and lost. Also, it is entirely up to the client as to how much prize money it wishes to offer, not for the participants to say how much they wish to get paid to do the work (as in traditional outsourcing). For many crowdworkers, it’s not just the money, but also the prestige of winning the competition that drives them.

### WHICH APPROACH GIVES BETTER QUALITY?

The quality of BDA work sourced from either a vendor or the crowd is driven by two major factors: the expertise of the analysts and the degree of competitiveness of the possible providers.

With outsourcing, vendors can produce high-quality outcomes for their clients because the vendors provide access to a specialized, skilled workforce to carry out the work. In contrast, crowdsourcing’s quality is derived from the combined efforts of many and the diversity of skills of the crowd. Crowdworkers often lack formal training in the task to be carried out, and the approach introduces laypersons and hobbyists.

The competitive bidding process most common with outsourcing filters out companies that do not have the technical expertise to carry out BDA work, thus reducing the likelihood of contracting out to an incompetent, low-quality provider. On the other hand, the crowdsourcing approach is an “open call,” inviting anyone to participate who has an interest in developing solutions to the problem. It generally does not discriminate in
terms of the participation eligibility; the competitive element is in the solution, not the nature of the participants. It does not matter who comes up with the solution, as long as it is the best one.

In fact, both methods of contracting out BDA work involve the use of competition — the key difference between the two approaches is the very different ways in which they do so. In traditional outsourcing, the vendors invited to participate in the bidding process compete predominantly on the price offered (but also on input-based qualitative criteria such as experience and viability of the company, proposed approaches and outcomes, analyst expertise, etc.). In crowdsourcing, on the other hand, the competition is based on actually producing the best solution or outcome, not just proposing to do the work for a certain price. The prize money offered to the various participants does not differ, thus financial criteria are not applicable.

WHICH APPROACH IS HIGHER RISK?

With crowdsourcing, there is a possibility of project failure due to lack of participants or low-quality solutions. But this is no different to outsourcing BDA work if there is insufficient competition in the market and/or the specification is ill-prepared by the client organization (which it notoriously is for outsourcing in general). So assuming a reasonable degree of market maturity under either approach, what are the unique risks of each method?

Personal Data

In recent years, the privacy and security of confidential data have been a major focus for many organizations. Data protection constitutes a major determining factor in the decision to outsource or crowdsourc e. In outsourcing, involved parties are well defined in the outsourcing contract and can be held liable by law in case of a privacy breach. Crowdsourcing, in contrast, is by nature open and shared among participants and/or the public. Trying to protect private data has proven difficult in crowdsourcing efforts as well, as the crowd can do analysis unintended by the client, such as seeking redacted data.

This suggests that crowdsourcing may be a poor model for analyzing personal data, and indeed it has resulted in class action lawsuits. One such case involved an annual Netflix crowdsourcing contest to develop the best algorithm to predict user ratings for films. Although the data was constructed to protect individual privacy, two university researchers were able to identify individual users through data matching to the Internet Movie Database (IMDb). The contest was cancelled in 2010 after a class action settlement. AOL had a similar issue when it released data from 650,000 users and 20 million search queries for research. The New York Times showed how, despite AOL’s efforts to anonymize the data, one could still find the identities of individual users. This had the effect of temporarily shutting down AOL’s research division after key people were either fired or left.

While most courts have held that the unilateral click agreement prior to entering a website is a valid contract, crowdsourcing presents significant challenges in many legal areas that cannot be easily clicked away.

Commercially Sensitive Data

Another major challenge in performing BDA work arises when the analysis is to be carried out on potentially commercially sensitive information. In traditional outsourcing, the provider who wins the tender and the client organization execute a legally binding agreement. The outsourcing contract formally names the parties, has specific clauses that address the duty to protect and have required controls over personal and commercially sensitive data, and states the liability for failure to do so. Moreover, because the parties are named in the outsourcing contract, issues such as intellectual property ownership, patent applications, copyright, and similar legal issues can be easily solved and written into the agreement.

This is in contrast to the simple acceptance of website terms and conditions in the case of a crowdsourcing project. In crowdsourcing, once an organization makes a problem public, membership to the project is open to anyone, including competitors. Rivals may acquire access to valuable business knowledge and intelligence, with the potential to leverage such information in their own commercial interests.

While most courts have held that the unilateral click agreement prior to entering a website is a valid contract, crowdsourcing presents significant challenges in many legal areas that cannot be easily clicked away. These include employment law (crowdworkers as employees vs. independent contractors), patent applications (naming all contributors to the invention), and copyright
WEIGHING THE PROS AND CONS

On a purely financial basis, there appears to be a case for crowdsourcing BDA work as opposed to using traditional outsourcing or insourcing, since the client does not pay for the solution until after it has been created and been judged the best of all the solutions provided. Under traditional outsourcing, the client pays for the most capable provider to develop solutions that meet the specification (but are unlikely to be the best). Payment is commonly made progressively, with little option not to pay unless there is a substantial nonperformance of the specification and the client organization has the will and funds to prove it in court.

But BDA crowdsourcing comes with a very significant caveat concerning the security of the data to be analyzed. Crowdsourcing basically means the data is insecure, despite client attempts to omit or redact it. Thus, if the shared data contains personal data and/or commercially valuable information, then traditional outsourcing provides far better protection of and control over that data than crowdsourcing does.

"OUTCROWDING": A HYBRID BDA SOLUTION

It is clear from the foregoing discussion that crowdsourcing cannot be universally applied to all BDA work. The nature of crowdsourcing itself hinders the ability to pursue it as an option when confidentiality, privacy, and data security are concerns. Also, organizations with a strong corporate legal framework may prefer a formal, clear, and binding outsourcing arrangement rather than evidence of a looser contract via a click.

Because of the limitations of crowdsourcing, we propose “outcrowding,” a practical solution in which outsourcing and crowdsourcing coexist, effectively constituting a hybrid approach to the external sourcing of BDA work. In this scenario, the organization’s end-to-end spectrum of BDA work would involve both traditional outsourcing and crowdsourcing components as appropriate. For example, BDA work that requires innovation would be crowdsourced, while BDA work that involves confidential information would instead be outsourced through traditional means. This “best of both worlds” approach could harness the value for money of both approaches, the innovativeness of crowdsourcing, and the data protection of traditional outsourcing.

It’s true that this hybrid approach could add management complexity, given the need to coordinate both models of contracting BDA work. Outcrowding should therefore be seen as another possible tool in a Big Data manager’s toolbox. It is not necessarily an easy answer, but it could be a very worthy one.

ENDNOTES


6Some organizations do not run an open call, instead limiting the crowd to select members of the population; for example, inviting only universities to participate.


Saeed Lajami is a Business Analyst at Saudi Aramco, a leader in the production of petroleum-based energy. He specializes in business process reengineering and end-user training of SAP HR and CRM. He is a functional subject matter expert in the employee performance management process. Mr. Lajami has managed several project implementations of SAP CRM utilizing the service request architecture. He is currently pursuing a master’s degree in IT at the University of Melbourne (Australia) and a master’s degree in HR management at Mays Business School, Texas A&M University. He can be reached at lajami@me.com.

Anson Mok is the Managing Director of WebSlice, a provider of cloud hosting, IaaS, and cloud consultancy services based in Auckland, New Zealand. His role involves service delivery to small and medium-sized businesses in New Zealand and abroad. Mr. Mok is currently pursuing a master of information systems degree at the University of Melbourne and holds a bachelor of commerce degree with majors in management and information systems from the University of Auckland. He can be reached at anson@webslice.co.nz.

Mario Wahyu Prabowo is a Software Engineer at Redflex Traffic System Australia, a leader in traffic enforcement products and services designed to improve public safety. His work specializes in traffic incidents data processing, including adjudication processes and data analytics. He has worked in software development for several IT projects in Indonesia and Australia using Java and Microsoft .NET technologies. Mr. Prabowo is currently pursuing a master’s degree in information systems at the University of Melbourne and holds a bachelor degree of informatics engineering from Sepuluh Nopember Institute of Technology (Indonesia). He can be reached at mariowahyu@gmail.com.

Sara Cullen is a Senior Consultant with Cutter Consortium’s Business Technology Strategies practice and a member of the Government & Public Sector team. She is the Managing Director of The Cullen Group, a specialist organization offering consulting, training, and methodologies regarding commercial agreements. Dr. Cullen was a former national partner at Deloitte in Australia, where she ran the outsourcing consulting division. She is an Honorary Fellow at Melbourne University and an associate of the Outsourcing Unit at the London School of Economics. Dr. Cullen specializes in the design and management of outsourcing agreements. She has consulted to more than 135 private and public sector organizations, spanning 51 countries, in over 175 outsourcing projects worth US $16 billion. Dr. Cullen has written over 110 best practice and research publications and has trained over 2,500 people. She can be reached at scullen@cutter.com.
About Cutter Consortium

Cutter Consortium is a truly unique IT advisory firm, comprising a group of more than 100 internationally recognized experts who have come together to offer content, consulting, and training to our clients. These experts are committed to delivering top-level, critical, and objective advice. They have done, and are doing, groundbreaking work in organizations worldwide, helping companies deal with issues in the core areas of software development and agile project management, enterprise architecture, business technology trends and strategies, enterprise risk management, metrics, and sourcing.

Cutter offers a different value proposition than other IT research firms: We give you Access to the Experts. You get practitioners’ points of view, derived from hands-on experience with the same critical issues you are facing, not the perspective of a desk-bound analyst who can only make predictions and observations on what’s happening in the marketplace. With Cutter Consortium, you get the best practices and lessons learned from the world’s leading experts, experts who are implementing these techniques at companies like yours right now.

Cutter’s clients are able to tap into its expertise in a variety of formats, including content via online advisory services and journals, mentoring, workshops, training, and consulting. And by customizing our information products and training/consulting services, you get the solutions you need, while staying within your budget.

Cutter Consortium’s philosophy is that there is no single right solution for all enterprises, or all departments within one enterprise, or even all projects within a department. Cutter believes that the complexity of the business technology issues confronting corporations today demands multiple detailed perspectives from which a company can view its opportunities and risks in order to make the right strategic and tactical decisions. The simplistic pronouncements other analyst firms make do not take into account the unique situation of each organization. This is another reason to present the several sides to each issue: to enable clients to determine the course of action that best fits their unique situation.

For more information, contact Cutter Consortium at +1 781 648 8700 or sales@cutter.com.

The Cutter Business Technology Council

The Cutter Business Technology Council was established by Cutter Consortium to help spot emerging trends in IT, digital technology, and the marketplace. Its members are IT specialists whose ideas have become important building blocks of today’s wide-band, digitally connected, global economy. This brain trust includes:

- Rob Austin
- Ron Blitstein
- Tom DeMarco
- Lynne Ellyn
- Israel Gat
- Vince Kellen
- Tim Lister
- Lou Mazzucchelli
- Ken Orr
- Robert D. Scott