Big Data Analytics

Dylan Maltby
University of Texas at Austin
School of Information
1616 Guadeloupe, Austin, TX 78701
512-471-3821
dylan.maltby@gmail.com

INTRODUCTION

In recent years "big data" has become something of a buzzword in business, computer science, information studies, information systems, statistics, and many other fields. As technology continues to advance, we constantly generate an ever-increasing amount of data. This growth does not differentiate between individuals and businesses, private or public sectors, institutions of learning and commercial entities. It is nigh universal and therefore warrants further study.

This review aims to provide a brief overview of big data, and how it is used in analytics. After reviewing the methodology of research used in creating the review the investigation of big data will begin by attempting to craft a satisfying definition of the term. Many of the relevant technologies and techniques used in big data analytics will be covered briefly and the benefits of big data analytics across various sectors will be explored. The review will also present several of the challenges and barriers faced by purveyors of big data analytic tools and attempt to determine if the results of the analytics offset the costs of overcoming these challenges sufficiently to make them a wise investment. The review concludes with recommendations for the future.

METHODOLOGY

In a perfect world literature reviews would examine all published information addressing their topic thus giving a nearly flawless interpretation of the current climate and what led to that climate. In reality this literature review is limited by constraints of time and manpower. Therefore, the first problem is developing a methodology for determining which sources to use and how to find them.

This is the space reserved for copyright notices.

ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.
Copyright notice continues right here.

Finding Sources

To select from among the myriad of sources available on this subject I used the EBSCOhost searching tool available through the University of Texas at Austin's library website. EBSCOhost provides access to several databases, many relevant to this subject. The databases included in my search are: Academic Search Complete, Business Source Complete, Communication & Mass Media Complete, Computer Source, eBook Collection (EBSCOhost), ERIC, GeoRef, Information Science & Technology Abstracts, Inspec, Internet and Personal Computing Abstracts, Library Literature & Information Science Full Text (H.W. Wilson), Library Information Science &Technology Abstracts, Military & Government Collection, Regional Business News, and Science & Technology Collection.

These databases were searched using the terms "Big Data" and "Big Data Analytics." The resulting sources were narrowed down by filtering for scholarly (peer reviewed) sources. This not only helped narrow the scope of study, it also helped return accurate information by ensuring that each of these articles had at least been reviewed by someone knowledgeable about the field prior to publication.

At this point the abstracts for all the results were examined to determine if the article addressed one or more of the topics outlined in the introduction of this review (defining big data, relevant technologies and techniques, benefits of big data analytics across sectors, challenges and barriers to big data). From those that remained an attempt was made to select an even distribution of articles from academic sources, corporate sources, and popular publications.

In addition, one source was included upon the recommendation of the professor overseeing the course for which this literature review was written. The source is, *The Fourth Paradigm Data-Intensive Scientific Discover* edited by Tony Hey, Stewart Tansley, and Kristin Tolle.

DEFINITIONS

In almost all of the articles used in this review the terms "big data" and "big data analytics" are used interchangeably. This reflects the common opinion that "big data" does not just refer to the problem of information

overload but also refers to the analytical tools used to manage the flood of data and turn the flood into a source of productive and useable information.

This is demonstrated in the definitions of big data presented in some of the articles. Though they define big data in terms of its size, their measuring stick is whether the system is capable of performing analytics on the data. A couple of examples are Siemens and Long (2011), who define big data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze" and Chen, Chiang, and Storey (2012), who call it "data sets … that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique storage, management, analysis, and visualization technologies." These definitions show that the authors think of big data in terms of how it gets analyzed, not how many specific terabytes of space it fills.

Other definitions focus more directly on the data. Patrick Russom (2011) writes that for data to be classified as big data it must possess the three Vs: Volume, Variety, and Velocity. Many people assume that big data simply has volume, but Russom clarifies that the other two Vs are just as essential. Big data is not just large, but it is varied. It comes in many formats and can be organized in a structured or non-structured way. Velocity refers to the speed at which it is generated. One of the reasons we build larger and larger stores of data is that we can generate it much more quickly. And Russom explains that volume doesn't have to refer to terabytes or petabytes. He suggests that other ways to measure volume of data could be number of files, records, transactions, etc.

Echoing Russom's three Vs, Siemens and Long (2011) cite Google's Marissa Mayer defining data by three Ss: Speed, Scale, and Sensors (where 'sensors' refers to new types of data). It is easy to see how these characteristics line up with Russom's Vs.

TECHNIQUES AND TECHNOLOGIES

Since big data is not only large, but also varied and fast-growing many technologies and analytical techniques are needed in order to attempt extracting relevant information. Many of these are topics large enough to support an entire review on them alone. As such, this report is not designed to provide an in-depth knowledge of all these tools. Rather, it gives a broad overview of some of the most commonly used techniques and technologies to help the reader better understand what tools big data analytics is based on.

Techniques

There are a myriad of analytic techniques that could be employed when attacking a big data project. Which ones are used depends on the type of data being analyzed, the technology available to you, and the research questions you are trying to solve. Some of the tools that came up frequently in the reviewed material are summarized here.

- *Association rule learning:* A way of finding relationships among variables. It is often used in data mining and according to Chen, Chiang, and Storey (2012) it lends support to recommender systems like those employed by Netflix and Amazon.
- *Data mining:* Manyika et al. (2011) calls data mining "combining methods from **statistics** and **machine learning** with database management" in order to pinpoint patterns in large datasets. Picciano (2012) lists it as one of the most important terms related to data-driven decision making and describes it as "searching or 'digging into' a data file for information to understand better a particular phenomenon."
- *Cluster analysis:* Cluster analysis is a type of data mining that divides a large group into smaller groups of similar objects "whose characteristics of similarity are not known in advance." (Manyika et al. 2011) and attempts to discover what the similarities are among the smaller groups, and if they are new groups, what caused these qualities.
- *Crowdsourcing:* Crowdsourcing collects data from a large group of people through an open call, usually via a Web2.0 tool. This tool is used more for collecting data than for analyzing it.
- *Machine learning:* Traditionally computers only know what we tell them, but in machine learning, a subspecialty of computer science, we try to craft "algorithms that allow computers to evolve based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data" (Manyika et al. 2011). Miller (2011/2012) gives the example of the U.S. Department of Homeland Security, which uses machine learning to identify patterns in cell phone and email traffic, as well as credit card purchases and other sources surrounding security threats. They use these patterns to try to identify future threats so they can handle them before they become large problems.
- *Text analytics:* A large portion of generated data is in text form. Emails, internet searches, web page content, corporate documents, etc. are all largely text based and can be good sources of information. Text analysis can be used to extract information from large amounts of textual data. This can be done to model topics, mine opinions, answer questions, and other goals.

These are just a few of the many techniques used in big data analytics. For the interested reader, some additional analytical tools not discussed here include classification, data fusion, network analysis, optimization, predictive modeling, regression, special analysis, time series analysis, and others.

Technology

As with the analytical techniques, there are several software products and available technologies to facilitate big data analytics. Some of the most common will be discussed here.

- *EDWs:* Enterprise data warehouses are databases used in data analysis. Russom (2011) writes that for many businesses that are trying to start handling big data the big question is "Can the current or planned enterprise data warehouse (EDW) handle big data and advanced analytics without degrading performance of other workloads for reporting and online analytic processing?" Some institutions manage their analytic data in the EDW itself while others use a separate platform, which helps relieve some of the stress on the server resulting from managing your data on the EDW.

- *Visualization products*: One of the difficulties with big data analytics is finding ways to visually represent results. Many new visualization products aim to fill this need, devising methods for representing data points numbering up into the millions. Russom (2011) lists this field as one of those having the most potential, saying it is "poised for aggressive adoption." Beyond simple representation visualization can also help in the information search. Hansen, Johnson, Pascucci, and Silva wrote an article included in Hey, Tansley, and Tolle's collection *The Fourth Paradigm* (2009) discussing visualization in data-intensive science in which they explain that visualization products allow us to compare models and datasets and "enables quantitative and qualitative decision-making." Their article stresses scalability in visualization technologies and their ability to track provenance in real-time.

- *MapReduce & Hadoop*: MapReduce is a programming model used to handle a lot of data simultaneously and Hadoop is one of the more popular open-source implementations of that model. Szalay and Blakeley wrote an article in Hey, Tansley, and Tolle's *The Fourth Paradigm* (2009) in which they discuss this particular software. They explain that the principles MapReduce uses are similar to the "distributed grouping and aggregation capabilities that have existed in parallel relational database systems for some time" but they are able to scale very well to accommodate for exceptionally large data sets. They go on to explain that Hadoop implements a "data-crawling strategy over massively scaled-out, share-nothing data partitions" where various nodes in the system are able to perform different parts of a query on different parts of the data simultaneously. This works very well for big data, but for smaller projects they remind their readers that this product isn't as effective "when a good index might provide better performance by orders of magnitudes."

- *NoSQL databases*: These databases are designed specifically to deal with very large amounts of information that don't utilize a relational model. They scale very well and are often useful for tracking and analyzing real-time lists which grow quickly.

These cover some of the more common technologies used in big data analytics. But not everyone will use all these techniques and technologies for every project. Anyone involved in big data analytics must evaluate their needs and choose the tools that are most appropriate for their company or organization. These needs change, not only from business to business, but also from sector to sector. Now that some of the tools and techniques have been examined the application of big data in various sectors can be more closely examined.

BENEFITS ACROSS DIVERSE SECTORS

When it comes to big data analytics each sector has different needs and potential. Health care will use big data analytics differently than the private sector. Public administration will not utilize big data in the same way as higher education. Each area also will garner a different amount of return on their big data analytics investment. Some are poised for greater gains than others. This section of the review looks at how some of these fields are currently using big data analytics and how they could leverage it in their favor in the future.

Heath Care

This is a field with a lot of untapped potential. Miller (2011/2012) says that biomedicine has some catching up to do when it comes to big data (they still have work to do as far as collecting and structuring data to provide a large knowledge base) but there are really exciting projects already in the works. Miller talks about one company, Medco, which has used genotyping, gene expression, and next-generation sequence data to perform studies that have already yielded results. One of their projects searched for drug-drug interactions. They discovered that taking Plavix (a drug for preventing blood clots) at the same time as a proton-pump inhibitor lessens the effect of the Plavix. It also found that antidepressants can block the effects of tamoxifen, which reduces the risk of breast cancer recurrences. These and other discoveries allowed them to drop the co-use of known interacting drugs by one third among their pharmacists.

Research isn't the only area of health care looking to utilize the power of big data analytics. Siemens and Long (2011) talk about moving medical diagnosis away from clinical based medicine into evidence based medicine. By evaluating medical records, insurance records, pharmaceutical records, and more, analytics hopes to accurately identify medical problems in patients rather than relying on the experience of one single doctor. Miller (2011/2012) quotes Colin Hill, the CEO, president, chairman, and cofounder of GNS Healthcare (a healthcare analytics firm) on this subject as saying "When I go to my doctor for some treatment, he's kind of guessing as to what drug works … We need to make this system smarter and use data to better determine what interventions work."

The system is not perfect. Miller (2011/2012) explains that in biomedicine very small changes in people can mean very big differences in how they react to certain treatments. This makes it very challenging to group people together and increases the chances of getting false results – something that has extremely serious repercussions when dealing with health care.

Public Sector

Manyika et al. (2011) asserts that the public sector doesn't stand to gain as much from big data analytics because it does not keep as much data as other sectors so they do not have as much on which to perform analytics. Anyone that has filled out a tax return or applied for a passport may disagree with that assessment, but Manyika et al. also claims that the value of big data analytics to Europe's public sector could be as much as 250 billion euros. Regardless of the amount of available data everyone agrees that the government can benefit by utilizing the opportunities available in big data analytics.

Government has already started to tap the lucrative resource of big data. Chen, Chiang, and Storey (2012) list campaign advertising, voter-mobilization, policy discussion, donations, and more as areas where the public sector has already begun using web 2.0 and using web analytics and social media analytics to further their causes. They add that most of the work in this area is carried out by the governments themselves; little of it is being done in academia.

Science and Technology

In Hey, Tansley, and Tolle's (2009) collection of articles *The Fourth Paradigm* there is an article by Parastatidis that calls for a fourth paradigm in science – a new research methodology. Parastatidis goes on to outline this need by saying that current technologies commonly used in science and research are great for things like managing and indexing data, but they fall short when asked to "discover, acquire, organize, analyze, correlate, interpret, infer, and reason." He ends his article by mentioning the MapReduce computational pattern and calls on the research community to develop equivalent platforms for knowledge-related actions like reasoning, aggregation, inference, etc.

Parastatidis is not the only voice calling for change in this field. Chen, Chiang, and Storey (2012) tell that the National Science Foundation now has a requirement that every project must provide a data management plan. NSF has a 2012 BIGDATA program supported by US funding that

> aims to advance the core scientific and technological means of managing, analyzing, visualization, and extracting useful information from large, diverse, distributed and heterogeneous data sets … encourage the development of new data analytic tools and algorithms [and] facilitate scalable, accessible, and sustainable data infrastructure.

The call includes new systems in the scientific community capable of handling big data appears to have been answered.

Higher Education

Education is utilizing available technologies more and more each year. Picciano (2012) tells us that one third of the higher education population in 2010 enrolled in at least one fully online course and many more enrolled in blended courses (a mix of online and face-to-face teaching). Since teachers and students are trending towards increased use of technology for instruction there are a lot of recorded data generated. In fact, Siemens and Long (2011) claim that big data and analytics are going to be the biggest factors in what is going to shape the future of higher education.

Picciano (2012) gives at least four areas within higher education that can benefit from big data analytics. Siemens and Long (2011) list nine. Among these are recruitment and admissions processing, financial planning, student performance monitoring, administrative decision making, donor tracking, providing help to at-risk students, understanding an institutions successes and challenges, recognizing the hard and soft value of faculty activities, and others. Perhaps the most interesting of these is student performance monitoring.

Picciano (2012) cites a fascinating case of a school in Arizona that uses an analytics program to track student progress on the website of online courses they offer. They track all student activity: login/logout information, number of mouse clicks, number of page views, how long students viewed each page, student post content, etc. They use this to ascertain which students are struggling to understand the material. It is true that picking out which students are struggling does not necessarily require big data analytics. But through analytics this process can happen very quickly and without devoting faculty time to evaluating each student's individual progress. Picciano quotes an associate dean from the school identifying three main predictors of student success: frequency of a student logging into a course, site engagement (do students engage with materials on the site? Readings, practices exercises, etc.?), and how many points received on assignments. Looking at this and other information the associate dean says "We can predict, after the first week of a course, with 70 percent accuracy, whether any given student will complete the course successfully (with a grade of 'C' or better)." By catching struggling students early on instructors are able to reach out to students before the student falls far behind in the course.

Private Sector

The group using big data analytics more than any other is clearly the private sector. Part of the reason for this is that the private sector encompasses a wider variety of organizations and therefore there is a broader range of applications for analytics. Some general rules do apply across the various goods and services offered in the private sectors. Russom (2011) asserts "Anything involving customers could benefit from big data analytics." He specifically mentions targeted marketing and customer base segmentation as effective uses of analytics for most, if not all, customer-oriented businesses. Customers themselves

are often the ones enabling this behavior. Chen, Chiang, and Storey (2012) explain that with web 2.0 it is incredibly easy to analyze customer needs and track information with cookies and server logs. But not all of this occurs online. Manyika et al (2011) gives some examples of companies that are using big data effectively already and they point to Amazon's recommendation system, but also to Wal-Mart's vendor-managed inventory and Harrah's marketing which is targeted to increase customer loyalty.

CHALLENGES AND BARRIERS

After looking at some of the more impressive things big data analytics has done or is doing it is important to look at some of the more realistic sides of big data. Big data analytics is certainly not a panacea for all analytical problems, and it raises some concerns that will have to be addressed if it is going to reach its full potential.

Privacy

A lot of big data contains personal information about customers, clients, patients, and other types of users. People are concerned about how information relating to them is used, particularly how it is used to affect them. Many people are uncomfortable with information such as their health condition(s) or their current location being known by others, even if they receive a clear benefit from providing that information. According to Manyika et al. (2011) health care is one of the fields best poised to benefit from big data analytics, but because of privacy concerns it is held back. Chen, Chiang, and Storey (2012) mention the same problem, but add hopefully that people are developing privacy-preserving health data mining techniques that would allow us to use health information while keeping it anonymous.

People are not only concerned about their health care information. Picciano (2012) give the case study mentioned above where the Arizona school predicts student performance early in the term. Shortly after the study is discussed Picciano cautions that care is needed when storing student behavior information. Some would find the idea of having their entire online interaction recorded. We need to ensure that big data does not become Big Brother.

Security

Tied closely with privacy is security. If companies have massive amounts of data related to their customers or clients they must have the capacity to ensure that this information is secure. This becomes even more important in fields like education or health care where the stored data are not just purchases and orders but grades and medical test results. Chen, Chiang, and Storey (2012) point out that, at the time they wrote their article, large companies were expected to spend 32 billion dollars on computer security in 2012 and for medium companies security was projected to be the greatest expense. Small wonder, when the alternative is to risk violations of privacy and/or identity theft (Picciano, 2012).

Legal Rights

Who owns data? The person who generated the data? The people to whom the data refers? Who has copyrights to datasets? If I track my friends' activities on Facebook am I allowed to run analytics on that information or does it belong to Facebook? What constitutes 'fair use'? Who is accountable for inaccuracies that lead to negative consequences? Manyika et al. (2011) raises these questions and acknowledges that there are no good answers at today, but the industry definitely is concerned with them.

Human Capital

When thinking about the resources required for big data analytics it is easy to assume that because computer storage capacity and processor capabilities continue to grow and continue to decrease in cost we have everything we need to perform advanced analytics on massive data sets. This line of thinking ignores the human element of the equation. This is a new field and there are not enough professionals trained in the proper techniques to meet the current demand. Manyika et al. (2011) projects that by 2018 demand for analytical talent will be greater than supply by 50 to 60 percent. Russom (2011) reports than during a survey of a few hundred companies the top ranked barrier to big data analytics was inadequate staff and skills. Basic economics tells us that since data scientists are scarce resources they are valuable ones. Companies wanting to implement big data analytics must pay increasingly large sums to attract the employees with the expertise to perform the work requested.

DOES BIG DATA = VALUABLE INFORMATION?

Given these problems, does having big data mean being able to tap valuable information? Does having to invest the resources to overcome issues like privacy, security, and the cost of human talent cancel out the benefits that big data analytics provides?

An example of this conflict is found in Miller's article (2011/2012) when she describes how Medco Research Institute has a way of sequencing DNA and using it to generate "genome sequences for large number of people at low costs." The process takes four terabytes of data per person. A sampling of only a few thousand quickly enters the petabyte range. Medco not only has to pay for the systems that handle this large amount of information, they also must hire people with the needed skills, and they must invest in computer security as well. In addition they must guarantee the privacy of the individuals whose genetic information they are using. All of this together may seem like more than it is worth, but Miller quotes Colin Hill from GNS Healthcare as saying that if we're going to cure cancer we'll do it with big data analytics. Hill's comment would support Russom's opinion (2011) that big data used to be more of a problem, not it is more of an opportunity.

RECOMMENDATIONS

Looking over the materials covered in this review there are a few places where I would suggest we invest future resources.

First, I would have liked to seen some studies that showed a cost benefit analysis of existing big data projects. Some of the articles and reports I read for this review claim to be able to save certain industries so many millions or billions of dollars but those claims aren't followed up by much solid data. If we had a few examples to work from showing the incurred expense and the perceived benefit it would help determine where big data analytics can most beneficially be employed in the near future.

Second, to address the growing need for data scientists more schools need to put together programs with data science in mind. Chen, Chiang, and Storey (2012) outline some requirements such a program would need. According to them students should have some courses in Information Systems, Computer Science, Statistics, and others. Some schools are already implementing this suggestion and offering programs focused on developing analytical and IT skills pertinent to big data.

Finally, I want to see some best practices for dealing with sensitive data. What measures are needed to protect privacy and ensure security? This type of research ought to be done by a neutral party, probably within academia.

CONCLUSION

Today we see information overload almost everywhere. Big data analytics is trying to take advantage of the excess of information to use it productively. The benefits are many and varied, ranging from higher quality education to cutting-edge medical research, and while further research is needed for things like ensuring people's information is protected from exploitation, there are many exciting discoveries waiting to be uncovered through big data analytics.

REFERENCES

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm data-intensive scientific discovery*. Redmond, Wash.: Microsoft Research.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity* (pp. 1–143). The McKinsey Global Institute. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Miller, K. (n.d.). Big Data Analytics in Biomedical Research. *Biomedical Computation Review*, (Winter 2011/2012), 14–21.

Picciano, A. G. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education. *Journal of Asynchronous Learning Networks*, *16*(3), 9–20.

Russom, P. (2011). *TDWI Best Practices Report: Big Data Analytics* (Best Practices) (pp. 1–35). The Data Warehouse Institute (TDWI). Retrieved from http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx?tc=page0

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, *46*(5), 30–32,.