How Big Data is Revolutionizing Business

I S 470 - Decision Support Systems

This report is prepared for Professor Hongyu Chen. It is an in-depth analysis of how big data is revolutionizing business and management. It includes case studies and information on how big data will impact students' careers.

Prepared by:

Group 2
Daniel Becerra
Tommy Pham
Trevor Williams

November 26, 2012

"Data-driven decisions tend to be better decisions. Leaders will either embrace this fact or be replaced by others who do. In sector after sector, companies that figure out how to combine domain expertise with data science will pull away from their rivals."

— Andrew McAffee, MIT Researcher

"Big data could transform the way companies do business, delivering the kind of performance gains last seen in the 1990s, when organizations redesigned their core processes. As data-driven strategies take hold, they will become an increasingly important point of competitive differentiation."

— Dominic Barton and David Court, Analysts at McKinsey & Company

Table of Contents

Table of Figures

Executive Summary

The term "Big Data" has been buzzing around in the IT industry for the last few months. Top universities in the nation are rushing to offer courses on it. Although data-rich companies such as Amazon, Google, and Yahoo! started this process, traditional companies are now exploring it as well. It is a prominent field that could transform management as we know it. In this paper, the authors look at the implications Big Data has for the overall management and direction of a company, specifically in terms of business strategy.

To be sure, this paper is not about learning Big Data, but rather how it integrates with the overall strategy of a company. We begin by defining what Big Data is, and what it is not. We then look at what motivated us to write about big data and its management implications. We also look at applications of Big Data, specifically Hadoop and Google BigQuery. Finally we explore career opportunities in the field of big data, and offer conclusions for this document.

Take one word of encouragement and caution regarding Big Data. The encouragement is that the field is relatively new so there is a lot to explore and learn, and plenty of time to catch up. The caution is that Big Data can change information systems and business strategy as we know it; and so while one can choose to not become a data miner, we cannot choose to ignore it completely. Students and current business employees should learn enough to work side-by-side with data scientists. Gone are the days where marketers, financers, even political aides can ignore the disrupting power of data, Big Data.

Big Data and data scientists played a big role in the latest presidential elections. The Obama campaign, through a team of data scientists, predicted and targeted the groups of voters that would make the biggest difference in the elections. The campaign made some odd moves, such as putting the President in a Reddit (a social network) chat for one hour. He appeared in a late-night comedy show, made campaigns, and raised funds in unique ways, against conventional wisdom. However, the results were clear: he won almost all of the swing states (Scherer). Big Data in many ways determined the direction the country is heading. Could it also direct the direction many companies will go? We believe so.

Project Motivation and Introduction

Our motivation for this topic is twofold: First, we want to gain a greater insight into this field that could change business as we know it, and second, we want to offer the insights gained to our classmates and workplaces. Let us begin by getting a grasp of what big data really is.

Definition

In order to define Big Data, it can be useful to clarify what Big Data is not:

1. It's not just analytics.
2. It's not database management.

What differentiates Big Data from past business intelligence can be narrowed to three areas:

1. **Volume**: As of 2012, 2.5 exabytes (an exabyte is one billion gigabytes) of data is created each day. That is almost 5 million times the space of an average computer today, daily. Moreover, this data not only comes from the Internet.

2. **Velocity**: Data analytics cannot provide real-time information while Big Data can. And when it comes to prediction, this is major. The nearer to real-time the information can be obtained, the sooner the insights can be made.

3. **Variety**: Whereas database managers and data analytics can only look at data within the company or data that has been stored, Big Data's main sources or raw data are relatively new. Social networks, GPS signals from phones, Internet searches, and so on are also part of big data. Big Data has access to thoughts, feelings, and patterns that executives a decade ago could only dream of.

How Big Data Affects Companies' Performance

MIT researcher Andrew McAffee and MIT professor Erik Brynjolfsson, working in partnership with McKinsey & Company, conducted a study that found that the data-driven companies performed better on objective measures of financial and operational results. On average, they are 5% more productive and 6% more profitable than their competitors (McAfee and Brynjolfsson). Let's look at some business areas and how Big Data can gain competitive advantage in those areas:

Customer Relationship Management

We live in a time where customers have seen the power of mass customization and social media. Large companies are increasingly expected to provide customized service. In order for interactions to feel individualized and human, they must be well informed, which makes data very useful to keep customer relationships very useful. What's the flip side? What if companies don't use Big Data adequately to improve customer experience? Perhaps this quote explains it best:

> *"Synchronized data is worth the expense because it's a hallmark of human interactions. If I talk to a friend and they keep asking me for information I know they already have, I have a right to get irritated. In the age of Big Data, I hold brands to the same standards." – Sean Madden*

Marketing Directives

Every piece of marketing material counts, even down to an impression. When it first began popular to "market" a product, advertising it everywhere helped the most. As years passed, companies became more strategic as to where to put their advertisement, for example, what do 18-24 year olds read and watch? They would invest there. Today, data-driven measurements give an advantage to use your marketing efforts effectively down to the penny. Big Data allows your customers to be heard through data measurements while giving your marketing team focused tracking data to restructure marketing pushes with. That's how Verizon was available to reduce the churn rate, Big Data scientists and marketers teaming up (Klawinski). This quote may explain it well:

> *"Starting mass marketing campaigns means money will fly out of your business bank account. The greater part of those monies may currently reside in someone's spam box or garbage can. With useful methods in collecting big data, businesses can maintain cost control simply by having clarity in marketing campaigns necessary to propagate brand awareness which is a realistic project management expectation."*
>
> *– Dave Klawinski*

Finance and Investing

Financial institutions are, perhaps, the ones looking to big data the most. Most immediately, they want to identify potential frauds. Their systems, as well as those of big card issuers such as Capital One, Visa, and Master Card look at vast numbers of transactions for unusual patterns or connections. This has taken fraud detection to a whole different level. If the two different transactions from two locations very far from each other showed in a short amount of time, then that would be easy to trace, but legitimate looking payments that add up to a major fraud could be harder to detect. That's where complex algorithms and Big Data comes in. Another area of use is investing. Many investors use data analytics to make investing decisions. It's every investor's dream to have access to the most accurate data from the most sources in real time to make the most precise decisions – big data allows that.

Human Resources

Some companies are using Big Data to make decisions on how to hire and how much to pay. Interviewers may have their intuition and observation to rely upon, but companies like Xerox have found that data tells them otherwise. For example, Xerox's call center used to hire based on experience, but the data suggested personality is more important. The result was the attrition rate being cut by half. Another area of human resources where Big Data can be helpful is pay and benefits. Rachel Silverman from Wall Street Journal stated "Some of these data-driven findings seem counterintuitive: Rotely paying workers more may not be enough to prevent defections. The key may be a flexible work schedule, or simply a nicer boss." Data is gathered from different sources, then a model is built – controlling factors as employee tenure and pay relative to peers – to figure out reasons of quitting, and what action would make a difference. Again, this is data-driven decisions versus intuition and experience. One final and more influential application of big data is building predictive models to attract the right candidates to the firm. How this would impact a company's future!

Supply Chain Management

Apple recently experienced a supply chain problem. More iPhones were in demand than they were able to meet. And it's safe to say that Apple was <u>not</u> careless with their forecast. The environment is rapidly changing, in more ways that even giant Apple cannot fully control. Big Data can do something for supply chain. Is there a supplier at risk of not meeting a company's needs? Could Big Data be used to predict future supply problems, environmental problems, and even better, could it be used to pose an alternative when problems show up? We believe it can. By accessing information in real time from vast sources, a better strategy for supply chain – one aligned in one solid plan for the company – can be created. Big data can help cut costs, eradicate weak links, and thrive in a world divided by companies that are partners and competitors at once.

Other Industries

In the executive summary, we mentioned that Big Data played a decisive role in the latest elections. Beyond businesses and politics, other industries are about to be or currently being transformed by Big Data; students would do well to become familiar with it. For one, Big Data in Health Care is inevitable. It can help organizations personalize care, engage patients, reduce variability and costs, and improve quality. Dr. Richard Tayrien, chief health information office for Hospital Corporation of America – a leading voice in the industry – stated ""Big Data solutions can result in personalized medicine that makes a dramatic difference by redirecting the care of a patient toward the most favorable outcome before predictably sustaining an adverse clinical event." These are lives that are at stake! Other areas are education, national security and even the food industry.

Applications

Apache Hadoop

What Is Hadoop?

Hadoop, by Cloudera, is an open-source software framework for distributed processing of extremely large amounts (petabytes) of data. In addition to size of data, Hadoop can also handle almost any data type – whether structured or unstructured -- such as log files, pictures, videos, audio files, email, and many more. The software itself is written in Java. It uses what is known as Hadoop Common, which provides access to compatible file systems. Data sets stored in HDFS (Hadoop Distributed File System) are organized across clusters of computers. Advantages are scalability, reliability, and portability. In terms of scalability, which is a huge consideration for enterprise, Hadoop is able to handle anywhere from a single server to thousands of machines. It also provides reliability by creating redundant data. With Hadoop, you do not need to build a schema prior to storing information in the file system. Overtime, patterns would be revealed that would never have been known with traditional relational databases containing only structured data (file systems can better handle unstructured data, whereas traditional databases typically contain structured data). Being able to use Hadoop to process a wide array of data allow for more closer and more accurate representation of queries. For better security, SSH is used for data propagation between nodes in the cluster. (Howson)

Structure

Servers are grouped into clusters for the purpose of efficiency and are provided with nodes. There is a node on the network switch that gathers location information and also a node on the server where data is stored. Based on location information, the Distributed File System replicates data obtained from one server node and stores it in another location for redundancy, in case of a power outage or server failure. Hadoop's MapReduce engine collects a large amount of data in smaller portions and then processes these portions to create an output, as shown in Figure 1.
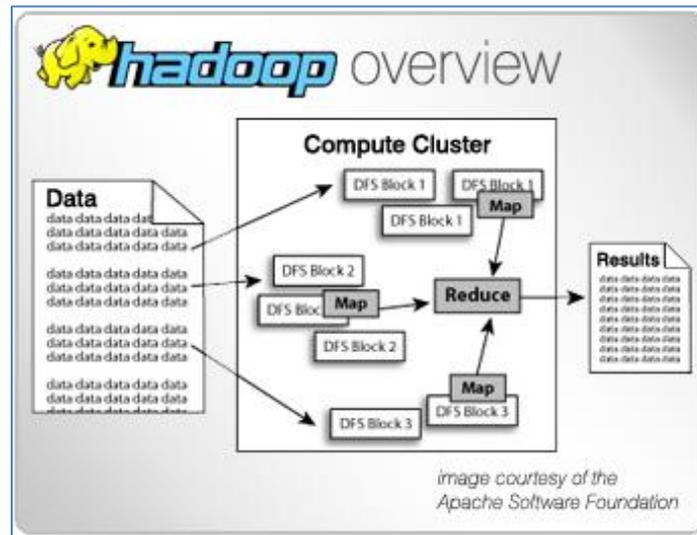

Figure 1 – Hadoop Overview

Source: IBM's website at http://www-01.ibm.com/software/ebusiness/jstart/hadoop/.

The Hadoop software can be found at http://hadoop.apache.org/ (Welcome to Apache Hadoop).

Case Studies

Many BI vendors are now starting to support Hadoop. However, there are other applications that can be used for the same purpose, such as analytic appliances or spreadsheets. What sets Hadoop apart from other methods is its structure and reliability.

There are many large companies that use Hadoop to handle their vast amount of company data (King):

1) **AOL**
   AOL Advertising provides online advertising for clients at a large scale, reaching millions of people each month. Their main challenges were figuring out how to analyze user data to gather demographic, psychographic, and behavioral characteristics into "cookie profiles", how to make those profiles quickly accessible to their advertising platform, and how to keep the profiles updated.

AOL's solution was to use a combination of Apache Hadoop and Membase for rapid data analysis. User data goes into Hadoop to create user profiles, then based on the current advertisements running, profiles are loaded into Membase to run queries to make decisions for ad placement (Zuanich).
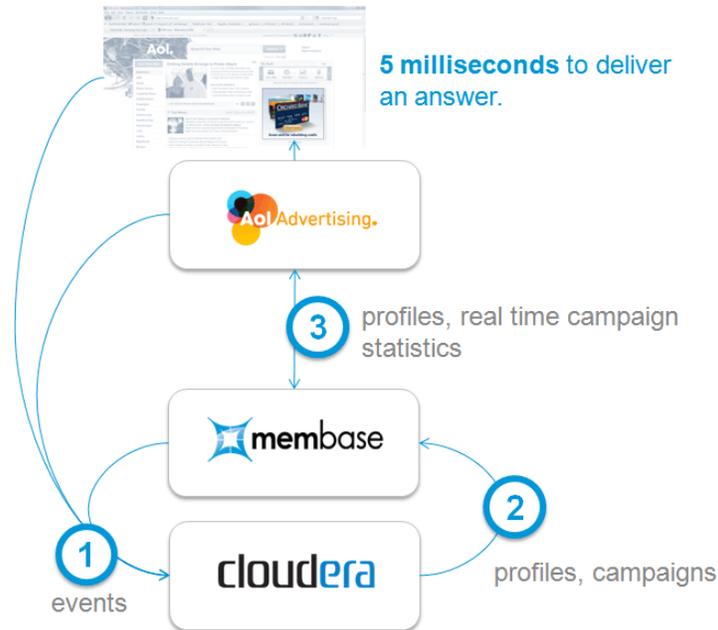

Figure 2 – AOL Advertising Model

Source: Cloudera at http://blog.cloudera.com/blog/2011/02/an-emerging-data-management-architectural-pattern-behind-interactive-web-application/

2) **Bank of America**
   Bank of America uses its data to analyze billions of records to gain a better understanding of the impact of new and existing financial products. Factors about users to consider when analyzing a product are credit, risk, insurance, use of online banking, etc. This information can be found with Hadoop.

3) **General Electric**
   General Electric's marketing teams needed to evaluate perceptions of the company through its campaigns. This would be accomplished through sentiment analysis in Hadoop text mining. The sentiment analysis process involves retrieving text from posts on GE (title, message), classifying it as a subjective or objective post, and if subjective, determining whether it is positive or negative. Hadoop improved accuracy to 80% and set new standards for data mining projects (Six Super-Scale Hadoop Deployments).
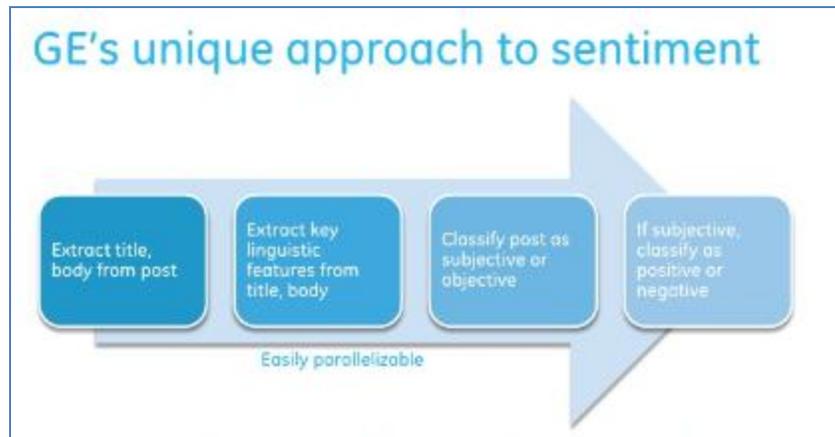
Figure 3 – GE's Unique Approach to Sentiment

Source: GE Sentiment Analysis Presentation from Datanami at http://www.datanami.com/datanami/2012-04-26/six_super-scale_hadoop_deployments.html?page=3

4) **Disney**

Disney uses data from its different business units such as theme parks, hotels, and cable television channel to analyze its impact on the company. Some statistics that can be found are theme-park attendance, reservations at hotels, and number of Disney Channel viewers.

5) **Facebook**

Facebook's large user-base and content provided by these many users provide a flood of data to Facebook's data centers. Using Hadoop, they are able to store over 30 PB of data (as of March 2011), and at the same time, be able to analyze all of this data to provide users with the best experience and most relevant information to them.

Google BigQuery

What Is BigQuery?

BigQuery is a cloud-based query service for large datasets powered by Dremel. Dremel is a Google system for executing SQL-like queries against data sets up to terabytes in size; according to Google, "Dremel can scan 35 billion rows without an index in tens of seconds" (Sato). This is because Dremel handles records by columns, placing each record's column value into a different storage volume as shown in Figure 2. BigQuery lets the public use this power in Google's cloud platform.
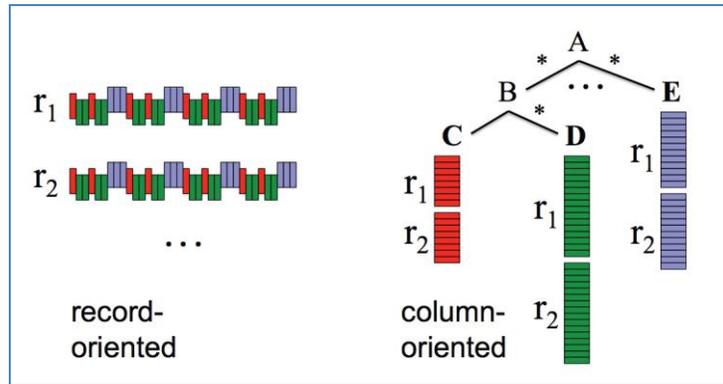
Figure 4 – Dremel's Columnar Storage

Source: "An Inside Look at Google BigQuery" at
https://cloud.google.com/files/BigQueryTechnicalWP.pdf.

BigQuery is a good example of an application that handles Big Data because it is available for anyone with a Google account. Moreover, users who already understand SQL should be able to execute queries easily, as shown below in Figure 3.

Figure 5 – BigQuery Query Example

Source: Google Developers BigQuery home page at https://developers.google.com/bigquery/.

BigQuery is not a database as it does not support data manipulation commands such as insert and delete; its main purpose is to perform fast queries of large data sets (up to multi-terabytes) for data analysis (Henschen). BigQuery is meant to handle structured data from imported files. Users try to find records matching certain criteria, gather statistics, and run ad-hoc quries. Therefore, it is suited for online analytical processing and business intelligence usage for simpler queries.

Case Studies

Analytics companies like Claritics use BigQuery to handle large amounts of data quickly. Claritics originally used Hadoop servers to process data on social and mobile game usage but ran

into problems maintaining the data. According to the CEO, "Once the data was in, running queries – asking simple questions and getting responses – took a lot of time. Complex queries on large datasets could take more than 30 minutes." (Analytics Firm Provides Game-Changing Capabilities). Claritics needed to process data quickly in a scalable service and deliver products more quickly.

The solution was to replace Hadoop with the Google BigQuery service. The SQL language was easier to use for staff and the system did not require extensive maintenance, saving 40% of their time. Running complex queries was more efficient and enabled the team to look up more trends in near real-time. The queries that took more than 30 minutes with Hadoop were completed in 20 seconds with Google BigQuery.

Management Challenges

For all the technical challenges Big Data presents, perhaps the biggest challenges rest within top management. What will top executives do? What questions will they ask? Are there discrepancies between top level executives and IT executives? Recent research tells us one of the first struggles companies must overcome comes from within the company. Let's look at some of those challenges:

- **Leadership** – We will still need leaders who can ask the right questions. If the wrong questions are asked, the Big Data will only find the right answers to questions that lead to nowhere. Yes, there is a need for technical skills in mining data, but there is as much need for business-oriented thinking. Seasoned executives must work side by side with front-level employees, all for the improvement of the company.

- **Talent management** – Here comes the flip side of executives working with front-level employees: The best data scientists are also comfortable speaking the language of business and helping leaders reformulate their challenges in ways that big data can tackle. Not surprisingly, people with these skills are hard to find and in great demand. Could it be that well-grounded students in both data mining and business objectives will lead companies in the near future? The data seems to tell us so.

- **Company culture** – The first question a data-driven organization asks itself is not "What do we think?" but "What do we know?" This requires a move away from acting solely on intuition or conventional wisdom. This requires breaking a bad habit in many organizations: pretending to be more data-driven than they actually are. Even more dangerous, assuming some things are true simply because of experience or because someone else said so. In the era of Big Data, mistaken assumptions can destroy a company. (i.e. Borders executives assumed no one would buy books online. Amazon did not assume that.)

In a study done by Paul Barth and Randy Bean, the authors found a discrepancy between the understanding of top executives and IT managers – as noted in the following information:

High Expectations: When asked what they thought the major benefits would be, C-suite executives named improvements in "fact-based decision making" and "customer experience" as #1 and #2. Many of the initiatives they had in mind were still in the early stages:

- 85% of organizations reported that they have Big Data initiatives planned or in progress.
- 70% report that these initiatives are enterprise-driven.
- 85% of the initiatives are sponsored by a C-level executive or the head of a line of business.
- 75% expect an impact across multiple lines of business.
- 80% believe that initiatives will cross multiple lines of business or functions.

Capabilities Gap: In spite of the strong organizational interest in Big Data, respondents painted a less rosy picture of their current capabilities:

Rank your access of data today as adequate or world-class	15%
Rank your analytical capabilities as adequate or world-class	21%
Rank your ability to use data and analytics to transform your business as more than adequate or world-class	17%

Figure 6 – Capabilities Gap

Problem with Alignments: When the researchers began to probe beneath the surface of these responses, IT executives and line-of-business executives had quite different perceptions of their companies' capabilities. Some examples:

	IT Managers	Top Executives
How would you rate the access to relevant, accurate and timely data in your company today? World-class or more than adequate	13%	27%
How would you rate the analytic capabilities in your company today? World class	13%	0%
How would you rate your company on leaders' ability to use data and analytics to improve or transform the business? Less than adequate	57%	18%

Figure 7 – Alignments Gap

There is clearly discrepancy between the judgments of capabilities of one another. IT personnel think they are more capable than what their top executives perceive them to be, and vice versa. Job #1 is to get the organization aligned. Without that ground work, Big Data can't live up to its promise.

Career Opportunities

According to Gartner, an increase of 4.4 million jobs to support big data is projected by 2015, with 1.9 million of these jobs in the United States (Hall). An issue, however, is the thin labor pool of people who have knowledge in analytics. Consequently, data experts are scarce, but increasing. Companies must find a way to bring in new fresh talent into the industry and to their companies.

Big data analytics is a growing field because, as Gartner puts it, "Your economic advantage depends on the data you have plus your ability to transform that data into meaning insights." (Hall) . The right people need to be put into place to ensure that the data being gathered is used with

maximum efficiency. According to a McKinsey report, "The United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data."

A salary survey conducted by Robert Half International predicts that data will be one of the hottest areas of employment in the coming year. Big Data roles and projected salary growth for 2013 are as follows:

Job Title	Projected Salary Growth	Projected Salary Range
Data Analyst/report writer	5.4%	$64,250 - $96,000
Data Architect	6.9%	$104,250 - $143,500
Data Modeler	7.6%	$92,000 - $126,750
Data Warehouse Manager	7.4%	$108,750 - $145,750
Data Warehouse Analyst	6.3%	$93,500 - $126,500
Business Intelligence Analyst	7.3%	$94,250 - $132,500

Figure 8 – Big Data Roles and Projected Salary

Conclusion

The evidence is clear: Data-driven decisions tend to be better decisions. Business leaders will either embrace this fact or be replaced by others who do. We wouldn't be surprised if some giant companies choose to trust their instincts, or what has been working so far, instead of the power of Big Data. In sector after sector, companies that figure out how to combine domain expertise with data science will pull away from their rivals. We can't say that all the winners will be harnessing big data to transform decision making. But, almost ironically, the data tell us that's the surest bet.

Not everyone is cut out to become a data scientist, or to enter the world of Big Data. However, leaders across most, if not all, industries must grow comfortable with Big Data. Directors across marketing, finance, supply chain, human resources, and especially top management, must know enough to create a partner-like environment across departments. As we saw, one of the biggest challenges now is the disconnection between top management and front-level employees when it comes to Big Data. This must be addressed, or else Big Data will make little to no progress.

Finally, students across all business concentrations would greatly benefit from learning about Big Data. Much like IT, Big Data is now ubiquitous. It's everywhere and we gain much by acknowledging it, and can potentially lose it all by ignoring it. It's a blunt statement, but the authors recognize and agree that companies are increasingly becoming data-driven decisions, and as stated before, these data-driven decisions are better decisions, and better decisions means a better future.

References

"Analytics Firm Provides Game-Changing Capabilities with Google BigQuery Service." *Google*. 2012. Web. 21 Nov 2012. <https://cloud.google.com/files/Claritics.pdf>.

"Apache Hadoop." 15 Nov 2012. *Wikipedia*. Web. <http://en.wikipedia.org/wiki/Apache_Hadoop>.

Hall, Susan. "Gartner: Big Data Will Generate 6 Million U.S. Jobs by 2015." *ITBusinessEdge* 23 Jan 2012. Web. 20 Nov 2012. <http://www.itbusinessedge.com/blogs/charting-your-it-career/gartner-big-data-will-generate-6-million-u.s.-jobs-by-2015.html>.

Henschen, Doug. "Google Launches BigQuery Analytics Service." *InformationWeek* 2 May 2012. Web. 21 Nov 2012. <http://www.informationweek.com/software/business-intelligence/google-launches-bigquery-analytics-servi/232901316>.

Howson, Cindi. "Big Data Meets BI: Beyond the Hype." *InformationWeek* 1 Jan 2012. Web. 20 Nov 2012. <http://www.informationweek.com/software/business-intelligence/big-data-meets-bi-beyond-the-hype/240012412>.

King, Rachel. "How Large Companies Tackle Big Data With Hadoop." *Bloomberg Businessweek* 11 July 2011. Web. 20 November 2012. <http://images.businessweek.com/slideshows/20110906/how-large-companies-tackle-big-data-with-hadoop>.

Klawinski, Dave. "Project Managers Can Leverage Big Data For Marketing Directives." *Business 2 Community* 18 November 2012. Web. 20 Nov 2012. <http://www.business2community.com/marketing/project-managers-can-leverage-big-data-for-marketing-directives-0335303#8dUBhM4l4e47GmxX.99>.

McAfee, Andrew and Erik Brynjolfsson. "Big Data: The Management Revolution." *Harvard Business Review* 26 Sept 2012. Web. 22 October 2012. <http://hbr.org/2012/10/big-data-the-management-revolution/ar/1>.

Sato, Kazunori. "An Inside Look at Google BigQuery." 2012. *Google*. Web. 21 Nov 2012. <https://cloud.google.com/files/BigQueryTechnicalWP.pdf>.

Scherer, Michael. "Inside the Secret World of the Data Crunchers Who Helped Obama Win." *Time Swampland* 7 Nov 2012. Web. 19 Nov 2012. <http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/#ixzz2CiSkLqnZ>.

"Six Super-Scale Hadoop Deployments." *Datanami*. 26 Apr 2012. Web. 20 Nov 2012. <http://www.datanami.com/datanami/2012-04-26/six_super-scale_hadoop_deployments.html?page=3>.

"Welcome to Apache Hadoop." 17 Oct 2012. *Hadoop*. Web. 20 Nov 2012. <http://hadoop.apache.org/>.

Zuanich, Jon. "An emerging data management architectural pattern behind interactive web applications." *Cloudera* 16 Feb 2011. Web. 22 Nov 2012. <http://blog.cloudera.com/blog/2011/02/an-emerging-data-management-architectural-pattern-behind-interactive-web-application/>.