Association for Information Systems
AIS Electronic Library (AISeL)

SAIS 2015 Proceedings | Southern (SAIS)

2015

Benefits and Risks of Big Data

Dana Cole
Georgia College and State University, dana.sires@bobcats.gcsu.edu

Jasmine Nelson
Georgia College and State University, jasmine.nelson@bobcats.gcsu.edu

Brian McDaniel
Georgia College and State University, brian.mcdaniel@bobcats.gcsu.edu

BENEFITS AND RISKS OF BIG DATA

Dana Cole
Georgia College and State University
Dana.sires@bobcats.gcsu.edu

Jasmine Nelson
Georgia College and State University
Jasmine.nelson@bobcats.gcsu.edu

Brian A. McDaniel
Georgia College and State University
Brian.mcdaniel@bobcats.gcsu.ed

ABSTRACT

Big data is one of the most prevalent topics in information systems today. The purpose of this paper is to explore big data, its past uses, legal history, current and potential security risks, and potential future uses. This literature review primarily focuses on the benefits of big data as well as the risks associated with big data. The research provides a general overview of big data and some of the technologies related to big data. The authors conclude with suggestions for future research of big data.

KEY WORDS

Big data; risk; security; data breach; open source; legal; future; information systems; business; NoSQL; medical; Epidemiology; the four Vs

INTRODUCTION

Big data is a relatively new term in the technology industry. Big data is defined by the four V's: volume, variety, velocity, and veracity (Goes). These V's describe the characteristics of big data. This new way of managing data allows companies to increase the amount of data which they are able to collect. This data, when analyzed, can become essential to making business decisions, answer research questions, and advancing medical science. Analyzing large amounts of data can identify trends that could help discover previously unrecognized relationships. Goes (2014) states, "For the last two or three years, the field of 'big data' has emerged as the new frontier in the wide spectrum of IT-enabled innovations and opportunities allowed by the information revolution."

Big data has many benefits but along with these benefits comes risk. Much of the risk with big data is incurred when cloud computing is involved. Most entities that participate in big data do not store their own data. The cost of software and hardware to store the amount of data which is available to companies is not economical. Outsourcing to cloud service providers allows companies to focus on their core business. When companies rely on others to store their data, some control over the data is lost. Confidentiality, system integrity, data integrity, and reliability may all become a greater risk when total control over data is lost (Neumann, 2014).

The intent of this literature review is to analyze the benefits and risks of big data. The benefits of big data will be explained using the health care industry setting. The risks of big data will cover security and data rights. The data rights will touch on companies control over their own data, service level agreements, and how much control individuals have over the data collected about them. Internal and external security in covered. This includes the issues with open source tools, NoSQL, and data breaches.

BENEFITS OF BIG DATA

Various companies use big data to analyze trends and thus gain a better understanding in different fields of study. For example, in the medical research field, scientists and medical professionals alike are capable of studying more patients at a

higher degree of speed and accuracy in an effort to increase their research efforts. One such example is in regards to global infectious disease. Mapping global infectious disease is an offensive strategy to combat the disease. There is much to learn about a disease by studying its natural habitat. For example, it is much easier to estimate and thus work to prevent an outbreak if we are aware of the geographical range of the disease. Until now, only about two percent of infectious diseases have actually been able to be mapped geographically. The reason for this slow rate of progression lies in the fact that science has previously attempted to establish the geographical range of diseases based on where they have been observed. These records are obtained from such sources as literature, web reports, and even GenBank. Next, scientists overlay these results with environmental factors such as rainfall and temperature. Statistical analysis is then performed on the data to obtain the relationship between outbreaks and environmental factors. These relationships are then used to predict when and where new outbreaks may occur. The problem with this approach is that it is extremely labor intensive, time consuming, and until now has been largely limited by human resources, not to mention delayed greatly. According to Hay, Moyes, and Brownstein (2013), the future of mapping global infectious disease is on the horizon. Big data has begun to accelerate this process by creating a feedback loop. The data is updated in real time every time a new occurrence of a disease is reported. This data then populates an evolving map of predicted risks. Furthermore, it is presumed that this data could be used to create entire databases on the occurrences of diseases (Hay, 2013).

Another rapidly expanding arena within the health data mining realm of big data is in regards to cancer. For many years, cancer research has been limited to the use of animals and the observation of diagnosed patients. Both methods are very time consuming, causing research to span many years, not to mention growing cost that accompanies the longevity. A few years ago, students from Penn State started a company on the simple premise of using big data to cure cancer. The research behind this premise involved interacting with various physicians and obtaining their view on the greatest hindrances they face when trying to treat their patients. The students learned the greatest portion of the problem was communication. Pertinent data for one cancer patient can come from many different sources such as oncologists, radiologists, surgeons, lab reports, pathology reports, and also in many other forms. For example, some physicians maintain audio recordings of their findings, some use hand-written notes, while still others may resolve to maintain their notes in an online format. If the multiple sources for this data are never combined, then the patient's treatment may be less effective.

The solution the entrepreneurs derived was to create an algorithm that pinpoints values in lab reports, use programs to read documents and extract pertinent data, and to search various databases to obtain information for patients diagnosed with similar conditions. Due to utilizing big data in the way, physicians are now able to search their database to find a patient who has similar genetic makeups, diagnoses, lab reports, and various other similarities. The physicians are then able to identify what treatment methods did or did not work for the similar patient and essentially treat their patient based off of another patient's treatment success. This approach greatly reduces the amount of guesswork in how physicians may treat patients in the future (Helft 2014).

Due to big data technology allowing the combining of cumbersome data, big data is becoming one of the most rapidly advancing tools in the field of medicine. As big data is fairly new and is especially new as far as its utilization in the field of medicine, the amount of properly documented research is sparse at present; however, the benefits are myriad, including assisting in cancer research and the prevention or more rapid containment of epidemic outbreaks, but these benefits are not without risks.

RISK ASSOCIATED WITH BIG DATA

Big data allows companies to gather large amounts of information about customers, possible customers, patients, gamers, and criminals; however, big data raises concerns when it comes to who owns the data, and who has rights to the data. Companies would like to think they own and control the data which they collect. This many not be entirely possible due to vendor lock-in and the composition of Service Level Agreements. Individuals would also like to know and be able to find out how much of their personal information is available on these mega databases. For individuals, big data does not only pose a data rights issue, but big data poses a privacy issue.

It is human nature to want to know what is being said about you and to make sure what is being said is as accurate as possible. The principle of "access/participation" has already been adopted by the Fair Credit Reporting Act. This principle gives persons the right to assess whether their personal information is correct. By the Fair Credit Reporting Act incorporating this principle, individuals are able to view their credit report and have incorrect information corrected (Navetta, 2014).

The principle of "access/participation" has not been adopted by many others. The United States does have a Fair Information Practice Principles. This states if information is collected about any persons, the person must be aware that information is being gathered; disclose who will have access to this information; and the purpose of the information being gathered (Navetta, 2014). This act allows individuals the ability to make informed decisions about their personal information. In the age where information is power and big data is able to help provide that power, data brokers are increasing in number and the amount of information they are selling is also increasing. When data brokers become involved, it is hard for individuals to find out which data broker has their information. In many cases, individuals do not have the financial resources to require the broker to allow access to personal information (Navetta, 2014).

The Fair Information Practice Principles makes it a requirement that individuals agree to having data collected about them. Individuals are not always handed a sheet of paper to sign. In the example of Facebook, individuals agree to have data collected about them when they sign up. In order to receive a Facebook account, an individual has to agree by using an informal consent button (Bechmann, 2014). A study conducted by Bechmann (2014) suggests that the informal consent button creates an imbalance between user and service provider in the favor of the service provider. The study also suggests that individuals using these services cannot comprehend the extent to which the service provider, Facebook, can use data including privacy inbox and news feeds (Bechmann, 2014). Using the button does not feel the same as physically signing personal information over to another; therefore, the user may not feel the need to read the fine print.

While individuals worry about having rights to the data about them, companies worry about having rights to the data which they collect. Service level agreements (SLA) can greatly increase a company's ability to retain full rights and ownership of its own data. A service level agreement is the agreement between the company who is providing the data storage and the company who is requesting the data storage service. A SLA needs to define the quality of service which will be provided (George et al., 2014). This quality of service needs to include parameters for when the data can be accessed, and how long the provider has to get the customer access to their data when the database goes down. The service level agreement should also include agreements to who can access the data, rights management, and data usage control (George et al., 2014). The service level agreement should also include in which format and when data is to be returned to a customer when services are terminated. The SLA should also have some reference to how the service provider is supposed to dispose of the data in their database and when confirmation has to be sent to the former customer.

SLA does not guarantee full control over data. One of the major concerns with big data is vendor lock-in. Vendor lock-in is due to the user having a close relationship to the product or service and the product being tailored to the purchaser (Greenstein, 1997). There are several ways which vendor lock-in can occur. The first of these ways happens when a vendor goes out of business or is bought out by another vendor. The second way for vendor lock-in to occur is the cost of moving vendors is too great or seems like too much of a struggle. A study performed by Greenstein (1997) suggests that switching cost will keep a purchaser from changing vendors. Some purchasers of new technology would like to leave the current vendor, but stay because of the switching cost. This restricts the purchaser from switching to a vendor that would better meet their needs (Greenstein, 1997). The third way is the cost of returning to using main frames, which are company owned, is too great. Vendor lock-in should be considered when companies are making decisions on how to manage their data.

Big Data causes concern when it comes to data rights; however, this area of concern is improving. Through the principle of "access/participation" and SLAs data rights have improved for companies and individuals. As the idea of big data increase and becomes more accepted and understood, data rights will come less of a concern.

Security and big data is currently a hot topic. The news seems to be riddled with data hacking stories. These stories normally end with individuals' private information being compromised. Due to these current events, security has become one of the scariest risks of big data.

Security, regarding big data, can be broken into two parts: internal security and external security. Internal security deals with people and inter-company issues, best practices and protocols. External security deals with the system itself, its vulnerabilities and structural risks.

The first area is external security. The realm of big data security is one in flux; part of the reason is that big data acts; therefore, it must be interacted with differently than past technology programs. Hamami (2014) states, "Traditional approaches to data security and resiliency don't apply to managing big data." The reason for the difference is simple, big data is a new concept with limited users. It takes time to figure out what works, where the flaws and holes are, and how to manage the risks. Many systems, including the increasingly popular NoSQL, are open source tools. These open source tools have

increased security risk because some fail to maintain a minimal level of security. Westermeier (2005) describes the risk in using open source tools, "open source licensee is not locked into a particular vendor and its particular schedule of bug fixes, modifications, and alterations." While proponents see this as a plus, the lack of consistent security is a large liability. Additionally, the newness of the big data systems creates gaps in understanding. "One cannot simply combine multiple databases, crunch the numbers, and magically uncover actionable correlations that can automatically and unthinkingly be implemented" (Bottles et al., 2014). While the massive amount of data and automation in sorting mitigates some risks, it increases other risk. Due to this tradeoff, knowledgeable people must be driving and interpreting the data in the right direction. If the data is not driven and interpreted in the right direction, risks to the information can lead to catastrophic and inaccurate results regarding sales projection, trending analysis, and cost drivers. While some companies are offering more secure versions of these open sourced systems, currently most companies who have larger budgets, such as Apple, Google and the like, can comfortably afford the more secure versions.

Conversely, some of the security risks are old problems in a new setting. Specifically, as Barki and Spears (2010) point out, "at least half of the breaches to information systems security are made by internal personnel". Internal security, whether it is by accident or maliciously, is statistically the greatest internal risk. Edward Snowden, the individual who leaked classified information from the National Security Agency, is the clearest picture of a malicious internal breach of security. He did not hack into a system but merely preyed on unsuspecting individuals and uploaded gigabytes of information onto a flash drive.

It remains to be seen how big data security will be handled. Further research needs to be performed to determine if NoSQL is the right, or even best, tool to utilize for the needs of companies, programs, and researchers. No matter the database structure used, the desire for end user to have all aspects of accurate and secure information is most important, especially for those desiring to use big data for medical purposes. Therefore, the potential benefits are incredible but the need for complete security is paramount.

CONCLUSION

Big data is a part of most of our everyday lives and is essential in the current information age we are living in. The ability to discover trends and analyze relationships has given big data a strong presence in the medical industry. From tracking diseases and patient care to increasing the abilities of cancer research, big data is now part of modern medicine. These benefits do not come without risks. Data rights and security are risks which users have to recognize and mitigate when using big data. Data rights have become a greater issue since big data has become prevalent in almost all parts of our lives and companies. As big data has advanced, so have ways to mitigate the issues with data rights. Even with the mitigation in place, there is plenty of room for improvement and adjustments. Internal and external security risks are important for users of big data to address. If security risks are not addressed, information could be comprised, misused, and rendered useless.

The world will continue to find ways to use big data to improve our lives. The risks with big data will always be present and ever changing, but as the technology matures, risk mitigation will advance. Huwe (2012), Director of Library and Information Resources at University of California-Berkeley says, "Big data is here to stay, we need to get used to it, and we would be well-advised to harness its potential".

There is a great deal of research left to do on the topic of big data. Suggestions for future research include determining the effects big data had on restricting the spread of the 2014 Ebola outbreak. Researchers could also analyze companies that are currently not pleased with their current data management provider and what is preventing them from changing providers. This would further help to understand vendor lock-in and find ways to prevent or lessen the effects of it. The last suggestion for research is to find out what is a correlation between data security and the cost of the cloud computing service.

REFERENCES

1. Bechmann, A. (2014). Non-Informed Consent Cultures: Privacy Policies and APP Contracts on Facebook. *Journal of Media Business Studies,* 11(1), 21-32.

2. Bottles, K., Begoli, E., & Worley, B. (2014). Understanding the Pros and Cons of Big Data Analytics. *Physician Executive,* 40(4), 6-12.

3. George, G., Hass, M.R., & Pentland, A. (2014, April). Big Data and Management. *Academy of Management Journal.* pp. 321-326. Doi:10.5465/amj.2014.4002.

4. Goes. P.B. (2014), Big Data and IS Research. *MIS Quarterly,* 38(3), iii-viii.

5. Greenstein, S.M. (1997). Lock-in and the Costs of Switching Mainframe Computer Vendors: What Do Buyers See? *Industrial & Corporate Change*, 6(2), 247-273.

6. Hamani, O. (2014). Big Data Security: Understanding the Risks. *Business Intelligence Journal*, 19(2), 20-26.

7. Hay, S., George, D., Moyes, C., & Brownstein, J. (2013). Big Data Opportunities for Global Infectious Disease Surveillance. *PLOS Medicine, 10*(4).

8. Helft, M. (2014). Can Big Data Cure Cancer? *Fortune, 170(2),* 70-76.

9. Huwe, T.K. (2012). Big Data, Big Future. *Computers in Libraries*, 32(5), 20-22.

10. Navetta, D. (2014). Legal Implications of Big Data. *Computer & Internet Lawyer, 31*(1), 1-5.

11. Neumann. P.G., (2014). Risks and Myths of Cloud Computing and Cloud Storage. *Communications Of The ACM*. October 2014; 57(10): 25-27. doi:10.1145/2661049.

12. Spears, J.L., & Barki, H. (2010). User Participation in Information Systems Security Risk Management. *MIS Quarterly*, 34(3), 503-A5.

13. Westermeier, J.T. (2005). Managing Open Source Software Risks in M&A Corporate Transactions. *Journal of Internet Law*, 9(5), 20-24.