# BIG DATA PROBLEMS AND POSSIBLE SOLUTIONS

Young B. Choi, Regent University, USA (ychoi@regent.edu)

Augustina Hunter, Regent University, USA (auguhun@mail.regent.edu)

Alaina Thomas, Regent University, USA (alaitho@mail.regent.edu)

**ABSTRACT**

*Big Data problems are rapidly growing in many parts of daily lives. Security, privacy, discrimination, and being able to cheat the system are all problems which are arising. It is no longer enough to separate Big Data into volume, velocity, variety, and veracity then solve the issues as many of the issues follow into more than one sector. Especially with fast data on the rise which will overload already overloaded systems. However, with statistical software, time, training, and funding, many companies can get a better hold on Big Data problems and be proactive about preventing an overload rather than being reactive. With companies having Big Data under control people can have more peace of mind about their privacy, and security will also be better equipped to process the data they receive.*

**Keywords:** Big Data, Security, Privacy, Discrimination, Cheating the System, Volume, Veracity, Variety, Velocity, Statistical Software

## INTRODUCTION

The research on Big Data analysis is starting a phase called Fast Data where Big Data systems receive many gigabytes of data every second. Current data systems collect complex data streams due to the volume, velocity, variety, veracity, etc. in the collected data. The smaller more related data streams are, the more useful it is than collecting redundant and inconsistent data (Bi & Dorng, 2016). So, what is Big Data? Big Data is "…data sets, typically consisting of billions or trillions of records, that are so vast and complex that they require new and powerful computational resources to process" (Dictionary.com, 2017). With Fast data being collected from many different sources quickly, large quantities of data (Big Data) have become a problem. Because many companies are being bombarded with too much data, the 4v split does not always work, and the problems must be faced head on. Some problems with Big Data include: an abundance of security data which cannot be easily processed and Big Data tools can be cheated which will lead to wrong information given. However, Big Data is even causing issues on a personal stand point as a person's privacy, security, and discrimination are effected.

## VOLUME

Big Data Volume is the most related with Big Data because it can be very big as far as quantities of data that reaches almost incomprehensible proportions. Today, our society and the people are more connected than ever before which leads to more and more data sources, resulting in an amount of data that is larger than we could ever imagine. The increased volume of data requires ever increasing computing power in order to derive value from the data. With that said, the traditional computing methods would simply not go to work on the volume of data accumulating today. Let's take a look at one of the social media, Facebook. Facebook stores photographs and probably has more users than any country's population. Each user expects to store a whole lot of photos estimating Facebook storing if not more than 250 plus billion images.

When we are talking about volume in the world of Big Data, we are referencing the enormously large amounts of data. Every type of organizations ranging from healthcare, financial institutions, national security, energy industry, manufacturing, etc. are no doubt benefiting with the technology and Big Data. But how are these

organizations keeping up with the technology? It is undeniable that the twenty first century has been flooded with too much data and moving much too quickly for anyone to collect, verify, and understand Big Data.

For analytics and business leaders, the massive and growing volume of data available for analysis is just one challenge. As data volume increases, the value of various data records will decrease in proportion to age, type, richness, and quantity among other factors. The social networking sites alone are producing data in order of terabytes every day, and this amount of data is definitely difficult to be handled using the existing traditional systems.

## VELOCITY

Big Data Velocity in Big Data measures the speed of the data coming from various channels or sources such as business processes, networks, machines, mobile devices, social media, etc. These characteristics are not being limited to the speed of incoming data, but also speed at which the data flows and aggregated. For the researchers, business leaders and IT leaders, the real-time data helps make valuable strategic decisions in a competitive world.

A packet analysis for cybersecurity is an example of velocity. So, the Internet sends a massive amount of data or information across the world every second of the day. A portion of the massive data has to travel through the firewalls into the organization's network. Because of the rise in cyberattacks, cybercrime, and cyberespionage, sinister payloads can be hidden in that flow of data passing through the firewall. In order to prevent being compromise, the flow of data has to be investigated and analyzed for anomalies, patterns of behavior that are red flags. This could be challenging for the IT because more and more data is protected using encryption. The attackers are getting smarter in hiding their malware payloads inside encrypted packets or taking sensor data.

## VARIETY

Big Data Variety refers to the various type and sources of data both structured and unstructured. We are very familiar how we used to store data from like spreadsheets and databases. This evolved data comes in the form of emails, audios, photos, videos, PDFs, etc. These various unstructured data create challenges as far as storage, mining, and analyzing data. The examples mentioned above like photographs, encrypted packets and sensor data are all very different from each other, and it was mentioned intentionally to give us a variety of unstructured and structured data to compare. How are these different from the traditional? Well, these data are not the typical or traditional old rows and columns and database joins. It does not fit into the fields of spreadsheets nor databases, but rather different from application to application, and much of it is unstructured.

Let's use the email messages an example. A business transaction process might require sifting through probably thousands to millions email messages and none of those messages is going to be exactly like another. Every email consists of a sender's email address, a destination, plus a time stamp. And, for every email message it will have human-written text and possibly attachments.

## VERACITY

Big Data Veracity commonly refers to "trustworthiness" of the data. In data analysis, the veracity is one of the biggest challenge compares to volume and velocity. The reason being it requires team and business partners to keep data clean and processes as well as preventing the accumulation of dirty data in the systems.

A combination of volume and variety of data with fast access are not enough. The data quality and credibility allows the right action when it comes to business strategy and decision making. The checks and balances and complicated algorithms keep the gears turning to the right operation direction.

## ABUNDANCE OF SECURITY DATA

The Go Big Security research, carried out by public-private partnership MeriTalk and underwritten by Splunk Inc., also revealed that 86 percent of the IT leaders surveyed believe that Big Data analytics has the potential to help make cyber security risk management more effective and proactive. (Wray, 2015).

In theory. "…government cyber security professionals say they could: better detect a breach that is in process; monitor streams of data in real time; and conduct a conclusive root-cause analysis following a breach" (Wray, 2015). This will help the government and many companies in a tremendous way, however, it is not a simple task. Less than 30% percent of cyber security professionals can fully leverage the Big Data, while 90% supposedly are not able to have a full timeline. Thus, making over 70% of cyber security professionals working "reactively" instead of "proactively" (Wray, 2015).

The cyber security team is continually working on a solution for the Big Data overload. Some are working on improving cyber security as a whole, others are upgrading existing technologies within the security sector, taking time to train more, and utilizing network analysis/ visibility solutions. The government believes all of this plus more management support and funding will fix the problem at hand (Wray, 2015).

## CHEATING THE SYSTEM

Big Data programs for grading student essays often rely on measures like sentence length and word sophistication, which are found to correlate well with scores given by human graders. But once students figure out how such a program works, they start writing long sentences and using obscure words, rather than learning how to actually formulate and write clear, coherent text (Marcus & Davis, 2014).

This can cause problems for many schools and even work places. Of course, while a work place does not have a grading scale, it could have a proficient/accurate scale in which it reviews the research and data an employee gives. However, cheating the system can also be unintentional.

Amazon used an automatic algorithm system to set prices on books being sold. Somehow, a textbook was put at over one million dollars, and this caused not only issues for Amazon but also for Barnes and Noble. Barnes and Noble benchmarks their competitors and tries to be as competitive in pricing as possible. However, because of this the Amazon book reached over twenty-million dollars before it was caught, and Amazon manually changed the price to one-hundred dollars (Hoerl, Snee, & De Veaux, 2014, p. 225). While this may have been an algorithm error, whichever company caught on first got the advantage, and it also gave both companies the information on how to cheat their competitors' systems and come out on top.

## BIG DATA EFFECTING US ALL

The issues and challenges Big Data raises for people is data privacy, data security, and data discrimination.

We're now at the point where even a total technology boycott may no longer fully protect us. Unless, of course, you choose to walk everywhere you go, wear a different mask every day (to foil face-recognition technology) and use only cash (that you never deposit in a financial institution) … As Big Data increases in size and the web of connected devices explodes it exposes more of our data to potential security breaches (Marr, 2017).

It is obvious every organization and/or any type of business can benefit of Big Data, especially where discrimination is concerned. People already check credit scores which can affect renting, getting a loan, getting insurance, etc. As Big Data continues to increase, so do the laws which protect people from discrimination otherwise people may one day not have a livelihood (Marr, 2017).

## SOLVING BIG DATA ISSUES

The first step in fixing the problems and challenges is by the management of Big Data. This is done through network topology and working through communication and security challenges. "The security mechanism in cloud technology is generally weak. Hence tampering of data at the public cloud is inevitable and it is a big concern" (Suthaharan, 2014, p. 71).

Next, is learning of Big Data because many use machine learning (ML). However, ML has three major problem and the biggest is number three: "… [a]n ML technique is developed based on a single learning task, and thus they are not suitable for today's multiple learning tasks and knowledge transfer requirements of Big Data analytics…" (Suthaharan, 2014, p. 72). With Machine Learning and Big Data Problems continuous

lifelong learning must take place because what works short-term for Bog Data Problems will not necessarily work long term.

The greatest solution for Big Data problems lies in integration of modern technologies. "…Hadoop Distributed File Systems and Cloud Technologies, with the latest representation-learning technique and support vector machine to predict network intrusions through Big Data classification strategy" (Suthaharan, 2014, p. 73).

Some research has also been done on statistical software which can be used to help with Big Data. There is excel add-ins, SPSS, SAS, and R. Each have a cost and training must take place for uses, but each software can provide help. However, determining which one to use depends on the project because there is not one universal software to help with all Big Data issues (Ozgur, Kleckner, & Li, 2015). On the other hand, with the Big Data analytics being researched through mathematical and statistical techniques many of the problems and challenges can be better understood and solved; it will just take more research and trials.

## CONCLUSION

Big Data is never going away, it is only going to increase which only will enhance the challenges being presented. While Big Data can be split up into four sectors, volume, variety, velocity, and veracity, this will not always help the issues which arise. The challenges such as security, cheating the system, privacy and discrimination fit in more than one of the sectors of Big Data. The easiest route is to face the problems directly and come up with a solution for the individual situation rather than trying to formulate an algorithm for each sector.

Solving these issues all start with managing Big Data and training. Whether the training is for filtering through Big Data better or learning which software works best for a company or personal use, it all will help towards fixing the issues at hand. However, the government and companies need time and funding to manage the Big Data better, yet it keeps growing at a constant rate. Big Data is a constant issue which will have many trial and errors before being solved, but it will give companies a better hold on data which gives people a peace of mind about their privacy, and security will be able to process data better.

## REFERENCES

Acharjya, D. P., & P, K. A. (2016). A Survery on Big Data Analytics: Challenges, Open Research Issues and Tools. *International Journal of Advanced Computer Science and Applications, 7*(2), 512-518. Retrieved from http://thesai.org/Downloads/Volume7No2/Paper_67-A_Survey_on_Big_Data_Analytics_Challenges.pdf

Bi, J., & Dorng, D. (2016). *Improve telco big data efficiency and utilization*. Retrieved from TMFORUM: https://inform.tmforum.org/sponsored-feature/2016/02/improve-telco-big-data-efficiency-and-utilization/

Davis, E. (2017). *The problems of big data, and what to do about them*. Retrieved from World Economic Forum: https://www.weforum.org/agenda/2017/02/big-data-how-we-can-manage-the-risks

Dictionary.com. (2017). *Big Data*. Retrieved from Dictionary.com: http://www.dictionary.com/browse/big-data?s=t

Gewirtz, D. (2016). *Volume, velocity, and variety: Understanding the three V's of big data*. Retrieved from ZD Net: http://www.zdnet.com/article/hpe-nasa-to-launch-a-supercomputer-into-space/

Hoerl, R., Snee, R., & De Veaux, R. D. (2014). Applying statistcial thinking to 'Big Data' problems. *WIREs Computational Statistics, 6*, 222-232. doi:10.1002/wics.1306

Marcus, G., & Davis, E. (2014). *Eight (No, Nine!) Problems With Big Data*. Retrieved from The New York Times: https://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

Marr, B. (2017). *3 Massive Big Data Problems Everyone Should Know About*. Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2017/06/15/3-massive-big-data-problems-everyone-should-know-about/#66782b356186

Ozgur, C., Kleckner, M., & Li, Y. (2015). Selection of Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities. *SAGE*, 1-12. doi:10.1177/2158244015584379

Soubra, D. (2012). *The 3Vs that define Big Data*. Retrieved from Data Science Central: http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

Suthaharan, S. (2014). Bug Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning. *Performance Evaluation Review, 41*(4), 70-73. doi:10.1145/2627534.2627557

Wray, S. (2015). *Is big data the answer to Government cyber secrity problems?* Retrieved from TMForum: https://inform.tmforum.org/news/2015/04/is-big-data-the-answer-to-government-cyber-security-problems/#prettyPhoto