Big data privacy and security: A systematic analysis of current and future challenges

Nobubele Angel Shozi[1,3] and Jabu Mtsweni[2,3]
[1]Council of Scientific and Industrial Research (CSIR), Meraka Institute, Pretoria, South Africa
[2]Council of Scientific and Industrial Research (CSIR), Defence Peace Safety and Security, Pretoria, South Africa
[3]University of South Africa (UNISA), South Africa
ashozi@csir.co.za
jmtsweni@csir.co.za

Abstract: Big data is a term that describes data of huge volumes, variable speeds, and different structures. Even though the rise of big data can yield positives, the nature of big data poses challenges as capturing, processing and storing becomes difficult. One of the challenges introduced by big data relates to its privacy and security. Privacy and security of big data is considered one of the most prominent challenges as it directly impacts on individuals. Through big data, individuals lose control over how their data is used and are unable to protect it. An invasion of privacy occurs when one's data is used to infer aspects of one's life without our consent. The prospect of data breaches in big data is also expected and can result in millions of records containing personal information being leaked. This paper aims to understand the privacy and security challenges that relate to big data. In order to gain this understanding, a systematic literature review is conducted to firstly identify the general challenges of big data. Currently, a number of research papers are identifying the challenges of big data however these papers do not follow a sound methodological process in identifying these challenges. The systematic literature review process consists of sequenced steps that must be followed to ensure that your research produces the required results. The systematic literature review was chosen to ensure that the three questions posed in this research are answered. These questions are: What are the current big data related challenges, what challenges are related to privacy and security and what future challenges can be identified from the analysis of these challenges. The top challenges of big data are discussed briefly and narrowed down into the challenges that are related to privacy and security of big data. In conclusion, this paper will provide reflections on future big data challenges. This outcome of this research is firstly to identify the big data challenges, secondly to understand the privacy and security challenges that relate to big data and lastly to provide insight into the future challenges that can impact on big data.

Keywords: big data, privacy and security, challenges, systematic literature review

1. Background

Gartner (2013) defines big data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". This is but one definition of big data that is used by researchers as there is no universally accepted definition. This definition highlights the following characteristics of big data which are – Volume, Velocity, Variety, Value and Veracity. Volume refers to the size of the data. Velocity refers to the speed of the data (real time, streaming, batch or near time). Variety refers to the different formats in which data is found (structured, semi-structured mixed or unstructured). Value refers to the benefits that the use of big data could yield for an organisation. Veracity is sometimes added as a characteristic of big data and this refers to the integrity and trustworthiness of the data.

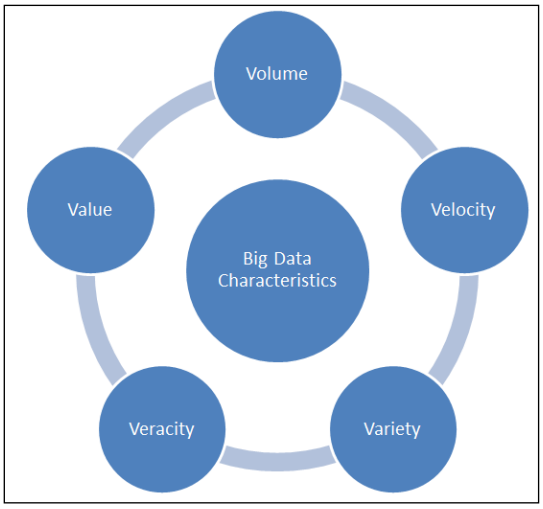

Figure 1: 5V's of Big Data

A number of papers identify the characteristics of big data as challenges (Chandio, Tziritas, & Xu, 2015; Yin & Kaynak, 2015):

- Volume: large amounts of data are produced daily from social media networking site and other sources. This is growing at an exponential rate as it was recorded in 2010 that the world's data has reached 1 Zetttabyte (Zb) (Zikopoulos et al., 2012). This poses challenges as traditional methods are unable to handle large amounts of data.
- Variety poses a challenge as the data varies in forms and can be found from different platforms. These bring about challenges to the hardware and software requirements of systems that will be processing big data.
- Velocity poses a challenge as this data comes in at different speeds and as it comes in, it needs to be processed. Velocity requires that the processing is done in a timely and accurate manner, which is a challenge.
- Value is also seen as a challenge as there needs to be interdisciplinary cooperation to ensure that value can be extracted from the big data.
- Veracity is the reliability and trustworthiness of the data. How this is measured and ensured is seen as a challenge.

Apart from the challenges posed by big data through its characteristics, big data gives rise to privacy and security challenges. Privacy and Security according to the researchers is the ability to ensure that data is only accessible to authorized people and only used for its intended purpose. The privacy landscape has changed drastically in the past couple of years. Before it was easy to protect one's privacy as our personal information was not digitized, but nowadays this has changed (Schadt, 2012). One of the latest and most popular big data privacy and security breach is the AshleyMadison.com breach. AshleyMadison.com is a dating website for individuals looking to have extramarital relationships. In this data breach, personal information belonging to 37million individuals as well as company financial records were breached (Information Is Beautiful, 2015).

Data breaches are not the only threat to privacy and security in big data. Big data analytics also pose a threat as has been shown by Target where they were able to identify potentially pregnant woman and send them market related products (Duhigg, 2012). This backfired when they sent these marketing products to a young woman who had not told her family about her pregnancy. The ability to combine datasets has also been seen a challenge to the privacy and security of big data as a lot of information can be inferred. This was demonstrated by researchers when they conducted a study in which a medical data set and a voters list were linked using shared attributes (Zip code, gender and date of birth). It was found that 87% of the US population could be identified based on these attributed (Sweeney, 2002).

The next section discusses the methodology that was used for data collection and analysis in the study presented in this paper.

2. Research Methodology

A literature review provides existing knowledge on a topic and does not have a prescribed methodology, while the systematic literature review has a clear purpose, question, defined search approach and produces a qualitative appraisal of articles (Jesson, Matheson, & Lacey, 2011). A systematic literature review was chosen for this research as it provides one with a methodologically sound way of performing a thorough literature review that is free from bias. The systematic literature review guides one to ensure that the research questions posed for the research are answered and that the selected literature adequately contributes towards answering the questions.

2.1 Systematic literature review approach

There are five (5) steps that can be followed when performing a systematic literature review (Khan, Kunz, Kleijnen, & Antes, 2003):

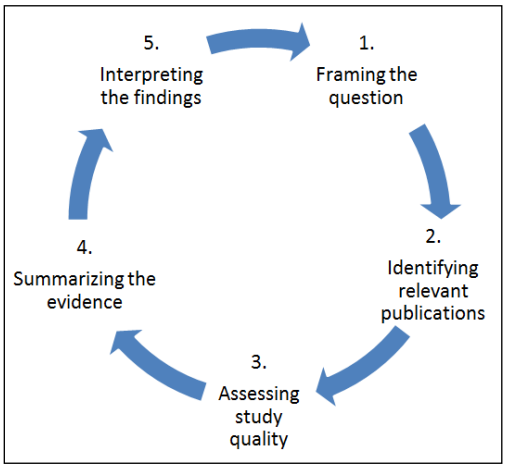

Figure 2: Five step process for performing systematic literature review

Step 1: Framing the question
The problem that is to be addressed must be specified in a clear, unambiguous and structured question format before beginning the review work. In this research, the problem that was being addressed pertained to understanding the challenges that relate specifically to big data privacy and security. In order to address this problem the following three (3) research questions are answered:

Question 1 - What are the current big data related challenges?
Question 2 - Which challenges are related to privacy and security?
Question 3 - What future privacy and security challenges can be identified from the analysis of these challenges?

Step 2: Identifying relevant publications
The literature review should be extensive and multiple resources without language restrictions should be searched. In this step, all the publications found are assessed to determine if they could be included or not. The reasons for exclusion and inclusion are recorded – an inclusion and exclusion criterion is created and discussed.

Searches for primary studies can be undertaken using digital libraries, however these are not sufficient for a full systematic literature review and sources such as journals, grey literature, conference proceedings and the internet must be used (Keele, 2007). According to Brereton, Kitchenham, Budgen, Turner and Khalil(2007), there are seven (7) electronic sources that are relevant to software engineers, these are: IEEExplore, ACM digital library, Google scholar, CiteSeerX library, Science Direct and EI Compendex. Web of Science is also added to this list. Therefore, for this research, these are the electronic resources that were investigated. Only electronic resources were used in this research as these are easily accessible and widely available.

This step aimed to identify publications that discuss big data challenges. The search string used was '"big data" and challenges'. These keywords needed to be found in the title of the paper to ensure that it is relevant to this study. The words do not however need to be in this order. Each database uses its own syntax for searching and therefore this search string might not be the same for each database that is searched. However, the essence of the search needs to be maintained in each database that is searched.

The years that are used for the search were from 2013 to 2015 to ensure that only the most relevant big data challenges are identified.

Table 1: Results from searches on databases

Database	**Results**
IEEExplore	73
ACM digital library	26
Google scholar	484
CiteSeerX library	60
Science Direct	19
El Compendex	0
Web of Science	93
Total	**755**

With these studies identified, their relevance needs to be assessed. In order to do this an inclusion/exclusion criterion is used to decide if an article will be included from the research or not:

- Does the article title include the words 'big data challenges'
- Is the article from 2013 till 2015?
- Is the article in English?
- Is it a scholarly peer reviewed article?
- Is it a full, openly available article?
- Does the article have a section listing/discussing the big data challenges?
- Does the article have more than one author?
- Does the article follow a strong methodology in order to identify the big data challenges?

Once the criterion was applied to the articles that were found in the various databases, the following were deemed to fit the criteria prescribed:

Table 2: Results after criteria applied to database articles

Database	**Results**
IEEExplore	9
ACM digital library	1
Google scholar	30
CiteSeerX library	2
Science Direct	4
El Compendex	0
Web of Science	6
Total	**52**

Step 3: Assessing study quality

The literature that is found is assessed to ensure that it meets the quality of the research. An in depth assessment for the risk of various biases is conducted to gauge the quality of the literature found. A quality hierarchy/assessment is developed to aid in the assessment of the literature.

Apart from the inclusion/exclusion criteria, a quality assessment must be conducted on the identified publications. This is to ensure the quality of the publications and also to ensure that no biases were involved

when selecting the publications. A checklist for qualitative studies was provided by (Keele, 2007) and this was adapted to suit this research:

- How appropriate/relevant is the paper for addressing the study questions?
- How generalizable are the findings to other research areas?
- To what extent can the study findings be trusted in answering the study questions?

Each of the identified papers was read and a score of 1-3 across the three dimensions listed above was given. Three (3) denotes high applicability, two (2) denotes medium applicability and one (1) denotes low applicability. These dimensions were used to assess the quality of each paper that was found.

Step 4: Summarizing the evidence
The findings of the literature review are summarized. Based on the identified literature, the challenges which were found are summarized in Table 3.

Step 5: Interpreting the findings
The interpretation of the results would be highlighted and discussed more in-depth in the results section (Section 3).

3. Results

The challenges that were found in the literature are summarized in Table 3. The challenges in Table 3 are listed as they are found in literature. In some cases, words which are seen as synonyms such as Cost and Budget are listed in the same challenge. Even though Security is commonly referred to as Confidentiality, Integrity and Availability, these are not integrated into one challenge as they are listed separately in literature and they each have their own meaning.

Table 3: Identified Big Data challenges

Challenge	IEEExplore	ACM	Google Scholar	CiteSeerX	Science Direct	Web of Science
Accuracy						1
Adoption			1			1
Analysis (Analytics)	5		5			1
Availability			2			2
Compliance	1		1			
Collaboration		1	1			1
Complexity			2		1	1
Compression			1			
Confidentiality						1
Connectivity			1			
Cost/Budget	1		5			1
Data access and sharing	3		4	1		2
Data input and output processes	1		1			
Data capture			1			1
Data ownership	1	1	4			1
Dealing with outliers			1			
Deployment			1			
Distributed mining			1			
Fault tolerance	2		2			
Governance	1		1			
Heterogeneity	3	1	13	1	2	3
Hidden big data			1			
Human collaboration			3			

Challenge	IEEExplore	ACM	Google Scholar	CiteSeerX	Science Direct	Web of Science
Incompleteness		1	8	1	2	2
Inconsistency		1	1			1
Interoperability	1		1			1
Integration			2			1
Integrity						1
Leadership			1			
Maintenance	1		2			1
Manageability (Management)	1		3	1		
Maturity	1					
Misuse of big data analysis results	1		1			
Organizational issues			2			1
Performance			1			
Privacy and Security	7	1	22	1	2	6
Processing	2		2			
Quality	3		5			
Scale (Scalability/Quantity/Capacity)	4	1	16	1	1	3
Skills/Talent Gap	3		9	1	1	1
Statistical significance			1			
Storage	4		6			2
Technology (Infrastructure/Architecture)	2		7	1	1	2
Timeliness	1	1	11	1	2	2
Time evolving data			1			
Transfer (Transport)	2		2			
Visualisation		1	3		1	2
Unstructured			2			

Table 3 shows the occurrences of each challenge in the papers and databases that were analysed. Table 3 is used to answer the first research question which is "What are the current big data related challenges". Through this analysis, the top five (5) challenges of big data are discussed. The top five challenges are selected as they are deemed the most prominent big data challenges based on the papers that were reviewed:

- Privacy and security: The privacy and security of big data can be seen as the most important and the challenge with the most impact. This is especially true when the privacy and security of big data is with regards to how an individual's privacy can be maintained and preserved (Garg & Somani, 2014).
- Scale: Managing large data is a challenge. In the past this used to be mitigated through CPUs getting faster, however nowadays data volume is increasing faster than CPU speeds and other computing resources (Jagadish et al., 2014).
- Heterogeneity: In order for data to be analysed, some form of structure is required. The analysis of these different types of data formats from structured, unstructured to semi-structured is complicated (Khan et al., 2014).
- Timeliness: As the data grows, real-time techniques are needed that can be able to deal with the data (Jagadish et al., 2014). This is to ensure that the data is dealt with in a timely manner and is processed and stored within a reasonable time frame.
- Skills: There is a shortage of professionals that are adequately skilled to work with big datasets. The United States alone faces a shortage of 140000 to 190000 professionals with analytical expertise in the big data field (Manyika et al., 2011).

The second question of this systematic literature review is "Which of the identified big data challenges relate to privacy and security". The biggest challenge of big data is preserving individual privacy (Michael & Miller, 2013). In all our daily activities, we are generating new data from updating our social networking profiles to

having surveillance cameras capture our movements. Data is constantly being generated about us. The digital footprints we create when combined can be used to uniquely identify us in cyberspace.

From the surveyed articles that define privacy and security as a challenge, the following challenges that specifically relate to privacy and security have been identified (Katal, Wazid, & Goudar, 2013; Singh, Srivastava, & Johri, 2014):

- Inference. The combination of personal information with other external datasets can lead to inference of new facts about the individual that were not meant to be revealed.
- Lack of consent. People's information is collected without their consent at times and used to add value and give an organisation a competitive advantage.
- Social stratification. Literate individuals can take advantage of big data predictive analysis to the detriment of illiterate and poor individuals.
- Multiple sources for information. The ability to access records of individuals such as health records from multiple sources poses a threat to privacy and security of an individual.
- Multiple uses for data. In some instances when data is collected there are no boundaries to its usage. It is sometimes used for other purposes other than the one it was collected for.
- Technology. Network infrastructure, big data sources and other aspects cause additional security risks as end-to-end security must be maintained to ensure that hackers are unable to access information.
- Privacy vs Utility. There is no clear balance between being able to share data while limiting disclosure yet ensuring sufficient data utility.

This section has provided an answer to the first and second research question that relates to challenges of big data. The various challenges of big data were discussed and the most prominent big data privacy and security challenges identified. The following section will answer the third question, "What future privacy and security challenges can be identified from the analysis of these challenges".

4. Future challenges

Although, governments are catching up in bringing forth policies and legislations dealing with the use, storage, and processing of personal information and big data, the nature of big data still gives rise to a number of future challenges:

- The rapid growth of big data will slowly erode and blur the lines of data ownership and usage. It will become more difficult to prevent one's data from being used in means which are considered obstructive and invasive.
- As data storage also becomes cheaper, organisations will be less compelled to delete data to make way for new data. This means that more data can be collected, processed and stored.
- Currently there is a lack of people who have the skills to work with big data. This can be seen for example in South African universities where there are not a lot of graduate courses that are focused on data science.
- Processing power needs to always be on par with the exponential growth of big data. Finding value in this data can be challenging especially if the processing is unable to handle the large volumes of data.
- The growth and variety of big data will continue to pose challenges with real-time techniques that are able to handle this type of data. These techniques would need to be able to handle the different forms of the data all within reasonable time frames.

Fan and Bifet (2013) also lists a number of other future challenges of big data:

- Until such a time where storage is cheaper, compression or sampling methods have to be considered in order to save space. The data can be compressed which takes more time but less space or it can be sampled which results in loose of information but saves space.
- Big data is constantly changing and data mining techniques need to be able to adapt to this change.
- As big data is large in size, visualizing this enormous datasets can be challenging and new methods are required to allow one to see data in more visually appealing ways.
- A large percentage of the world's available data is unstructured and untagged making it difficult to find any value in it. This is a challenge which will need to be dealt with in future.

5. Conclusion

The explosive growth of big data along, its structures, the speed of the data, the ability to trust data and wanting to find value in the data present challenges when one intends on using big data. The systematic literature review process followed in this research has provided one with an overview of the current big data challenges that are found in literature. These challenges go beyond the nature of big data to identify challenges that are outside of the characteristics of big data. The systematic literature review aids in giving one a diverse view into the various challenges facing big data. These include challenges such as the shortage of skills of individuals who are able to work with big data, the storage constraints of big data, the ability to process this data in a timely manner and many more other challenges. Privacy and security is highlighted as being the most challenging as it was the one challenge that was found in more literature than any of the other challenges. The privacy and security challenge is particularly challenging when it relates to preserving and maintaining an individual's privacy and security. This is especially difficult as we are constantly sharing our data on the internet and data is constantly being generated about us. This paper also discussed a number of future challenges that can be expected through the continuous use of big data.

6. References

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, *80*(4), 571–583. http://doi.org/10.1016/j.jss.2006.07.009

Chandio, A. A., Tziritas, N., & Xu, C.-Z. (2015). Big-Data Processing Techniques and Their Challenges in Transport Domain. *ZTE Communications*, *1*, 010.

Duhigg, C. (2012, February 16). How Companies Learn Your Secrets. *The New York Times*. Retrieved from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Fan, W., & Bifet, A. (2013). Mining Big Data: Current Status, and Forecast to the Future. *SIGKDD Explor. Newsl.*, *14*(2), 1–5. http://doi.org/10.1145/2481244.2481246

Garg, K., & Somani, S. (2014). Big Data Challenges: A Survey.

Gartner. (2013). What Is Big Data? - Gartner IT Glossary - Big Data. Retrieved July 14, 2015, from http://www.gartner.com/it-glossary/big-data

Information Is Beautiful. (2015). World's Biggest Data Breaches & Hacks. Retrieved from http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94. http://doi.org/10.1145/2611567

Jesson, J., Matheson, L., & Lacey, F. M. (2011). *Doing Your Literature Review: Traditional and Systematic Techniques*. SAGE.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and Good practices. In *2013 Sixth International Conference on Contemporary Computing (IC3)* (pp. 404–409). http://doi.org/10.1109/IC3.2013.6612229

Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*.

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, *96*(3), 118–121.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., … Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges, Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal, The Scientific World Journal*, *2014, 2014*, e712826. http://doi.org/10.1155/2014/712826, 10.1155/2014/712826

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., … Institute, M. G. (2011). Big data: The next frontier for innovation, competition, and productivity.

Michael, K., & Miller, K. W. (2013). Big Data: New Opportunities and New Challenges [Guest editors' introduction]. *Computer*, *46*(6), 22–24. http://doi.org/10.1109/MC.2013.196

Schadt, E. E. (2012). The changing privacy landscape in the era of big data. *Molecular Systems Biology*, *8*(1), 612. http://doi.org/10.1038/msb.2012.47

Singh, V., Srivastava, I., & Johri, V. (2014). Big Data and the Opportunities and Challenges for Government Agencies. *International Journal of Computer Science and Information Technologies*, *5*(4), 5821–5824.

Sweeney, L. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. http://doi.org/10.1142/S0218488502001648

Yin, S., & Kaynak, O. (2015). Big Data for Modern Industry: Challenges and Trends [Point of View]. *Proceedings of the IEEE*, *103*(2), 143–146. http://doi.org/10.1109/JPROC.2015.2388958

Zikopoulos, P., deRoos, D., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2012). *Harness the Power of Big Data The IBM Big Data Platform* (1 edition). New York ; Singapore: McGraw-Hill Osborne Media.