Big Data Security an Overview

P.Joseph Charles[1], I.Carol[2], S.Mahalakshmi[3]

[1, 2] Assistant professor, Department of IT, St. Joseph's College, Trichy, India,
[3] II M.Sc., Computer Science, Department of IT, St. Joseph's College, Trichy

---

Abstract: *Data is currently one of the most important assets for companies in every field. The continuous growth in the importance and volume of data has created a new problem: It cannot be handled by traditional analysis techniques. This problem was, therefore, solved through the creation of a new paradigm: Big Data. However, Big Data originated new issues related not only to the volume or the variety of the data, but also to data security and privacy. In order to obtain a full perspective of the problem, we decided to carry out an investigation with the objective of highlighting the main issues regarding Big Data security, and also the solutions proposed by the scientific community to solve them. In this paper, we explain the results obtained after applying a systematic mapping study to security in the Big Data ecosystem. It is almost impossible to carry out detailed research into the entire topic of security, and the outcome of this research is, therefore, a big picture of the main problems related to security in a Big Data system, along with the principal solutions to them proposed by the research community.*

Keywords: Big Data; security; systematic mapping study

I. Introduction

Over the last few years, data has become one of the most important assets for companies in almost every field. Not only are they important for companies related to the computer science industry, but also for organisations, such as countries' governments, healthcare, education, or the engineering sector. Data are essential with respect to carrying out their daily activities, and also helping the businesses' management to achieve their goals and make the best decisions on the basis of the information extracted from them[1]. It is estimated that of all the data in recorded human history, 90 percent has been created in the last few years. In 2003, five exabytes of data were created by humans, and this amount of information is, at present, created within two days. This tendency towards increasing the volume and detail of the data that is collected by companies will not change in the near future, as the rise of social networks, multimedia, and the Internet of Things (IoT) is producing an overwhelming flow of data[2]. We are living in the era of Big Data. Furthermore, this data is mostly unstructured, signifying that traditional systems are not capable of analysing it. Organisations are willing to extract more beneficial information from this high volume and variety of data. A new analysis paradigm with which to analyse and better understand this data, therefore, emerged in order to obtain not only private, but also public, benefits, and this was Big Data[3].

Each new disruptive technology brings new issues with it. In the case of Big Data, these issues are related not only to the volume or the variety of data, but also to data quality, data privacy, and data security. This paper will focus on the subjects of Big Data privacy and security. Big Data not only increases the scale of the challenges related to privacy and security as they are addressed in traditional security management, but also create new ones that need to be approached in a new way[4]. As more data is stored and analysed by organisations or governments, more regulations are needed to address these concerns. Achieving security in Big Data has, therefore, become one of the most important barriers that could slow down the spread of technology; without adequate security guarantees, Big Data will not achieve the required level of trust. Big Data brings big responsibility.

According to the Big Data Working Group at the Cloud Security Alliance organisation there are, principally, four different aspects of Big Data security: infrastructure security, data privacy, data management, and integrity and reactive security. This division of Big Data security into four principal topics has also been used by the International Organisation for Standardisation in order to create a security standard for security in Big Data.

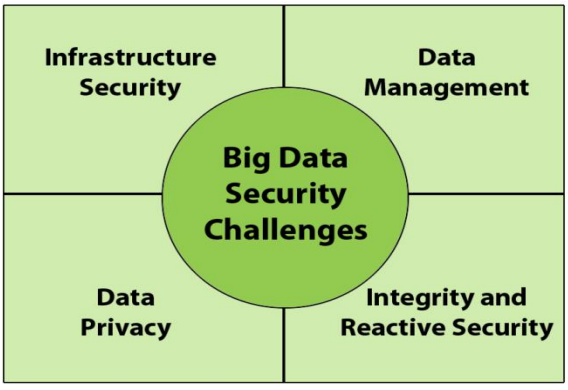

Figure 1. Main challenges as regards security in big data security, based on.

2 Infrastructure Security

When discussing infrastructure security, it is necessary to highlight the main technologies and frameworks found as regards securing the architecture of a Big Data system, and particularly those based on the Hadoop technology, since it is that most frequently used. In this section we shall also discuss certain other topics, such as communication security in Big Data, or how to achieve high-availability.

2.1 Security for Hadoop

The graphic shows that the main topic dealt with by those researching infrastructure security is security for Hadoop. As explained in previous sections, Hadoop can be considered as a de facto standard for implementing a Big Data environment in a company. The security problems related to this technology have, therefore, been widely discussed by researchers, who have also proposed various methods with which to improve the security of the Hadoop system. This category is probably the most transverse since, in order to protect it, the solutions use different security mechanisms such as authenticity or cryptography.

For example, there is a proposal for a security model for G-Hadoop (an extension of the MapReduce framework to run on multiple clusters) that simplifies users' authentication and some security mechanisms in order to protect the system from traditional attacks[5]. A few papers focus on protecting the data that is stored in the HDFS by proposing a new schema, a secure access system, or even the creation of an encryption scheme.

2.2 Availability

Researchers have also dealt with the subject of availability in Big Data systems. One of the main characteristics of Big Data environments, and by extension of a Hadoop implementation, is the availability attained by the use of hundreds of computers in which the data are not only stored, but are also replicated along the cluster. Finding an architecture that will ensure the full availability of the system is, therefore, a priority.

For instance[6], in the authors propose a solution with which to achieve high availability by having multiple active Name Nodes at the same time. Other solutions are based on creating a new infrastructure of the storage system so as to improve availability and fault tolerance.

2.3 Architecture Security

Another different approach is that of describing a new Big Data architecture, or modifying the typical one, in order to improve the security of the environment. The authors of propose a new architecture based on the Hadoop file system which, when combined with network coding and multi-node reading, makes it possible to improve the security of the system. Another solution focuses on secure group communications in large-scale networks managed by Big Data systems, and this is achieved by creating certain protocols and changing the infrastructure of the nodes.

2.4 Authentication

The value of the data obtained after executing a Big Data process can, to a great extent, be determined by its authenticity. A few papers deal with this problem by proposing solutions related to authentication. In, the authors suggest solving the problem of authentication by creating an identity-based signcryption scheme for Big Data.

2.5 Communication Security

The security as regards communications between different parts of the Big Data ecosystem is a topic that often is ignored, and only a small number of papers therefore deal with this problem. One paper approaches the topic by explaining the regular data life cycle in a Big Data system, following the different network protocols and applications that the data pass through. The authors also enumerate the main data transfer security techniques.

3 Data Privacy

Data privacy is probably the topic about which ordinary people are most concerned, but it should also be one of the greatest concerns for the organisations that use Big Data techniques. A Big Data system usually contains an enormous amount of personal information that organisations use in order to obtain a benefit from that data. However, we should ask ourselves where the limit regarding the use of that information is.

Organisations should not have total freedom to use that information without our knowledge, although they also need to gain some benefit from the use of that data. Several techniques and mechanisms with which to protect the privacy of the data, and also allow companies to still make a profit from it have therefore been developed, and attempt to solve this problem in various different ways.

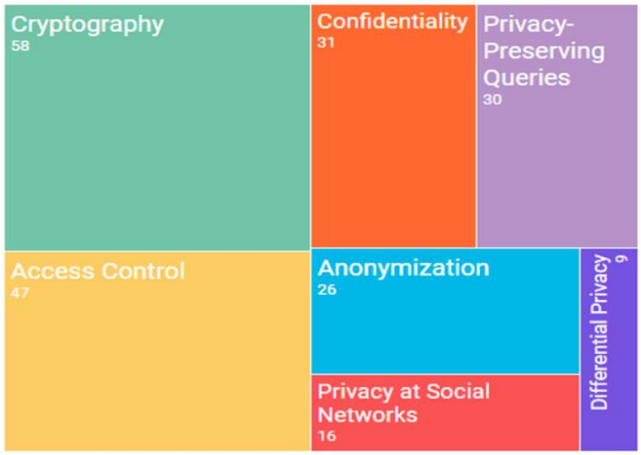

Figure 2 Main topics on data privacy.

3.1 Cryptography

The most frequently employed solution as regards securing data privacy in a Big Data system is cryptography. Cryptography has been used to protect data for a considerable amount of time. This tendency continues in the case of Big Data, but it has a few inherent characteristics that make the direct application of traditional cryptography techniques impossible.

One example of the use of cryptography can be found in, in which the authors propose a bitmap encryption scheme that guarantees users' privacy. Other authors' research is focused on how to process data that is already encrypted. One paper, for example, explains a technique with which to analyse and programme transformations with *PigLatin* in the case of encrypted data.

3.2 Access Control

Access control is one of the basic traditional techniques used to achieve the security of a system. -Its main objective is to restrict non-desirable users' access to the system. In the case of Big Data, the access control problem is related to the fact that there are only basic forms of access control. In order to solve this problem, some authors propose a framework that supports the integration of access control features. Other researchers focus their attention on the MapReduce process itself, and suggest a framework with which to enforce the security policies at the key-value level.

3.3 Confidentiality

Although privacy is traditionally treated as a part of confidentiality, we decided to change the order owing to the tremendous impact that privacy has on the general public's perception of Big Data technology.

The authors that approach this problem often propose new techniques such as computing on masked data (CMD), which improves data confidentiality and integrity by allowing direct computations to be made on masked data, or new schemes, such as Trusted Scheme for Hadoop Cluster (TSHC) which creates a new architecture framework for Hadoop in order to improve the confidentiality and security of the data .

3.4 Privacy-Preserving Queries

The main purpose of a Big Data system is to analyse the data in order to obtain valuable information. However, while we manipulate that data we should not forget its privacy. A few papers pay attention to the problem of how to make queries whilst simultaneously not violating the privacy of the data.

One way in which to achieve this protection is by encrypting the data, as discussed previously, but this adds a new problem: how do we analyse the encrypted data? Some authors propose that this problem can be solved by means of a secure keyword search mechanism over that encrypted data.

3.5 Anonymisation

One of the most extended ways in which to protect the privacy of data is by anonymising it. This consists of applying some kind of technique or mechanism to the data in order to remove the sensitive information from it or to hide it. Big Data usually implies a large amount of data, and this problem, therefore, increases in Big Data environments.

The authors of propose a hybrid method that combines the two most frequently used anonymisation schemes: top-down specialisation (TDS) and bottom-up generalisation (BUG).

3.6 Differential Privacy

The objective of differential privacy is to provide a method with which to maximise the value of analysis of a set of data while minimising the chances of identifying users' identities. A few papers focus on achieving privacy in Big Data by applying differential privacy techniques. For example, in the authors attempt to distort the data by adding noise.

4 Data Management

This section focuses on what to do once the data is contained in the Big Data environment. It not only shows how to secure the data that is stored in the Big Data

system, but also how to share that data. We shall also discuss the different policies and legislation that authors suggest in order to use Big Data techniques safely.

5 Integrity and Reactive Security

One of the bases on which Big Data is supported is the capacity to receive streams of data from many different origins and with distinct formats: either structural data or non-structural data. This increases the importance of checking that the data's integrity is good so that it can be used properly. This topic also covers the use case of applying Big Data in order to monitor security so as to detect whether a system is being attacked.

5.1 Integrity

Integrity has traditionally been defined as the maintenance of the consistency, accuracy, and trustworthiness of data. It protects data from unauthorised alterations during its lifecycle. Integrity is considered to be one of the three basic dimensions of security (along with confidentiality and availability). Ensuring integrity is critical in a Big Data environment, and authors agree as to the difficulty of achieving the proper integrity of data when attempting to manage this problem.

For example, they propose an external integrity verification of the data[7] or a framework to ensure it during a MapReduce process[8].

5.2 Attack Detection

As occurs with all systems, Big Data may be attacked by malicious users. Some authors, therefore, take advantage of the inherit characteristics of Big Data and suggest certain indicators that may be a sign that the Big Data environment is under attack.

For instance, in[9] the authors develop a computational system that captures the provenance data related to a MapReduce process. There are also researchers who propose an intrusion detection system especially intended for the specific characteristics of a Big Data environment[10].

5.3 Recovery

The main purpose of this topic is to create particular policies or controls in order to ensure that the system recovers as soon as possible when a disaster occurs. Many organisations currently store their data in Big Data systems, signifying that if a disaster occurs the entire company could be in danger. We have found only a few papers that cover this problem. For example, in[10] there are some recommendations regarding what can be done to recover from a desperate situation.

6. Summary and Conclusions

The infrastructure security, the main problem dealt with by researchers would appear to be security for Hadoop systems. This is not surprising since, as stated previously, Hadoop can be considered as a de facto standard in industry. The remaining problems addressed in this topic are usually solved by modifying the usual scheme of a Big Data system through the addition of new security layers.

The most frequently dealt with by researchers would appear to be privacy. There are a lot of different perspectives as regards ensuring privacy. Authors usually propose different means of encryption, based on traditional techniques but with a few changes in order to adapt these techniques to the inherit characteristics of a Big Data environment. Owing to the large amount of papers found on this topic in comparison to the others, we believe that it is advisable to split this category up into, on the one hand, data privacy itself, and on the other, cryptography and access control techniques.

This section includes almost the entire lifecycle of the data used in a Big Data system, from its collection to its sharing, and also includes how to properly govern the security of that data. With regard to collection and to its sharing, authors propose the creation of new schemas, frameworks, and protocols with which to secure data. Other authors also suggest toughening up the legislation concerning the privacy of the data used by companies. Furthermore, we have found a lack of papers dealing with the need to create a framework that covers security data governance in a Big Data system in its entire lifecycle.

In this section, the main topic discussed by researchers would appear to be the integrity of data. In order to secure that integrity, they propose various kinds of verification to ensure that the data has not been modified. This section also covers the possibility of detecting the attacks that a Big Data system may undergo. This is probably a consequence of the high availability that a Big Data system usually achieves, but this topic should not be overlooked.

This paper provides an explanation of the research carried out in order to discover the main problems and challenges related to security in Big Data, and how researchers are dealing with these problems. This objective was achieved by following the systematic mapping study methodology, which allowed us to find the papers related to our main goal.

Having done so, we discovered that the principal problems are related to the inherent characteristics of a Big Data system, and also to the fact that security issues were not contemplated when Big Data was initially conceived. Many authors, therefore, focus their research on creating means to protect data, particularly with respect to privacy, but privacy it is not the only security problem that can be found in a Big Data system; the traditional architecture itself and how to protect a Hadoop system is also a huge concern for the researchers.

We have, however, also detected a lack of investigations in the field of data management, especially with respect to government. We are of the considered opinion that this is not acceptable, since having a government security framework will allow the rapid spread of Big Data technology.

In conclusion, the Big Data technology seems to be reaching a mature stage, and that is the reason why there have been a number of studies created the last year. However, that does not mean that it is no longer necessary to study this paradigm, in fact, the studies created from now should focus on more specific problems. Furthermore, Big Data can be useful as a base for the development of the future technologies that will change the world as we see it, like the Internet of Things (IoT), or on-demand services, and that is the reason why Big Data is, after all, the future.

Reference :

1. Mayer-Schönberger, V.; Cukier, K. Big Data: A Revolution that Will Transform How We Live, Work, and Think;Houghton Mifflin Harcourt: Boston, MA, USA, 2013.

2.Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Ullah Khan, S. The rise of “big data” on cloud computing: Review and open research issues. Inf. Syst. 2015, 47, 98–115.

3. Eynon, R. The rise of Big Data: What does it mean for education, technology, and media research?Learn. Media Technol. 2013, 38, 237–240.

4. Wang, H.; Jiang, X.; Kambourakis, G. Special issue on Security, Privacy and Trust in network-based Big Data.Inf. Sci. Int. J. 2015, 318, 48–50.

5. Zhao, J.;Wang, L.; Tao, J.; Chen, J.; Sun,W.; Ranjan, R.; Kołodziej, J.; Streit, A.; Georgakopoulos, D. A securityframework in G-Hadoop for big data computing across distributed Cloud data centres. J. Comput. Syst. Sci.2014, 80, 994–1007.

6. Wang, Z.; Wang, D. NCluster: Using Multiple Active Name Nodes to Achieve High Availability for HDFS. In Proceedings of the 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), Zhangjiajie, China, 13–15 November 2013; pp. 2291–2297.

7. Liu, C.; Yang, C.; Zhang, X.; Chen, J. External integrity verification for outsourced big data in cloud and IoT. Future Gener. Comput. Syst. 2015, 49, 58–67.

8. Wang, Y.; Wei, J.; Srivatsa, M.; Duan, Y.; Du, W. IntegrityMR: Integrity assurance framework for big data analytics and management applications. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 33–40.

9. Liao, C.; Squicciarini, A. Towards provenance-based anomaly detection in MapReduce. In Proceedings of the 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid),
Shenzhen, China, 4–7 May 2015; pp. 647–656.

10.Tan, Z.; Nagar, U.T.; He, X.; Nanda, P.; Liu, R.P.; Wang, S.; Hu, J. Enhancing big data security with collaborative intrusion detection. IEEE Cloud Comput. 2014, 1, 27–33.