Big Data Security Proposed Solution

Ibtissame KANDROUCH
ADSI Team, System Engineering Laboratory
National School of Applied Sciences
Ibn Tofail University, Kenitra, Morocco
Ibtissame.kandrouch@uit.ac.ma

Manale BOUGHANJA
ADSI Team, System Engineering Laboratory
National School of Applied Sciences
Ibn Tofail University, Kenitra, Morocco
boughanja.manale@gmail.com

Nabil HMINA
System Engineering Laboratory, ADSI Team
National School of Applied Sciences
Kenitra, Morocco
hmina@univ-ibntofail.ac.ma

Habiba CHAOUI
ADSI Team, System Engineering Laboratory
National School of Applied Sciences
Ibn Tofail University, Kenitra, Morocco
Mejhed90@gmail.ma

Abstract

Following the remarkable evolution of new technologies, and the computerization of different transactions and devices in several areas in recent years, the generation of data in its various forms has become increasingly important. Hence the emergence of the new era of Big Data. This new phenomenon of Big Data has appeared with several advantages, but in return, it confronts several challenges linked to security. This paper highlights the various security issues that may exist throughout the Big Data life cycle. And touches some needs necessary to reinforce the security of these data. Thus, the paper presents a solution allowing the improvement of the security level offered by systems allowing the management of these data during their real time treatment.

Keywords
Big Data, Security issues, real time.

1. Introduction

In the last twenty years, data has been increased. Every day, billions of bytes of digital information have been generated. Because of the use of different devices as well as social networks 90 % of all data available today in the world has been produced in the last two years (Mehmood et al. 2016). Hence the appearance of the new era of Big Data.

Big data is rapidly being turned into a hot spot that attract the attention of the world. It is a global concept that canbe defined as a set of voluminous and complex data, characterized by five dimensions: volume, that represents the data generated, velocity, or speed describes the frequency with which data is generated, captured and shared; variety,

indicate the heterogeneity of data sources; veracity, which mean the reliability of data and value, indicate the information inference. With such sophisticated technology the challenges appear, especially for information security and privacy.
The objective is to stop any spiteful from using this data to thread the private information related to the individual and identifying the sources of problems lies in the efficient use of big data. Therefore, we focused on the solutions to these issues.
For all these reasons, the second section aim to highlight Big Data security issues. In section III we will see the security needs. Section IV for presenting existing works that aimed securing Data in all its states, Last but not least, the paper presents as a solution a proposal to raise the level of security offered by some systems allowing the Big Data management, especially during its real time processing, while presenting the arguments that prove its validity.

2. Big Data Security issues

Big data is a recent technology that has been adopted by many industries to predict user behavior. Despite this, organizations are concerned about the benefits of big data and neglect the privacy and security issues of individuals. Several factors have shown that big data can pierce the privacy of individuals if it is not well managed, controlled and protected. There are a number of security and privacy issues that need to be addressed before building a data environment.

We highlight some important big data challenges that should be taken into account (figure.1), and we will present some security needs to deal with those problems.

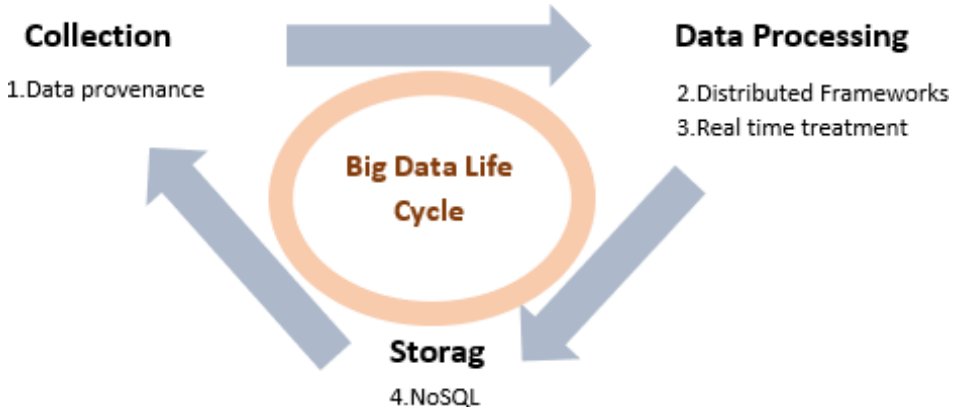

Figure 1: Big Data life cycle

- ***Data provenance***: Data provenance refers to the origin and creation of data. Usually the data is collected from a wide variety of sources; however, we don't know the origin of this data in order to determinate if it's came from unauthorized sources (Dev Mishra and Beer Singh 2016). Therefore, the data generated from different sources causes a great security challenge.
- ***Distributed framework:*** Big Data use distributed processing with many systems analyses (Dev Mishra and Beer Singh 2016). With the beginning of such a technology different framework have been created but they don't pay attention about the security one of this distributed framework is Hadoop Distributed File System it's considered as one of the technologies that are used for analysis that could create a breach of security.
- ***Real time security***: Real time security monitoring is intended to alert at the first sign of an attack (Dev Mishra and Beer Singh 2016). However, these alerts lead to many false positives, which are ignored. This problem is more critical with Big Data view at its great volume and velocity.

- ***NoSQL***: To solve the problem of storage to deal with unstructured data, many organizations has migrated from traditional database to a NoSQL (Not only structured query language) database. However, the main problem is security, due to the NoSQL database architecture (Dev Mishra and Beer Singh 2016).

Beside these issues, we should think about the privacy. The organizations that use big data can obtain a large quantity of personal information in order to obtain their own benefit from this data, and we should get this problem seriously.

3. Security needs

The Big Data evolution leads to many questions concerning security and data protection. Traditional security techniques such as encryption schemes, firewall can be broken by an attacker, even the anonymization can be re-identified. For this reason, it can be considered that the traditional technique is not efficient anymore (Matturdi et al. 2014), and thereafter, an increased need for mechanisms able of providing security to these large amounts of data.

The security and privacy issues should be considered in big data are shown in figure 2.

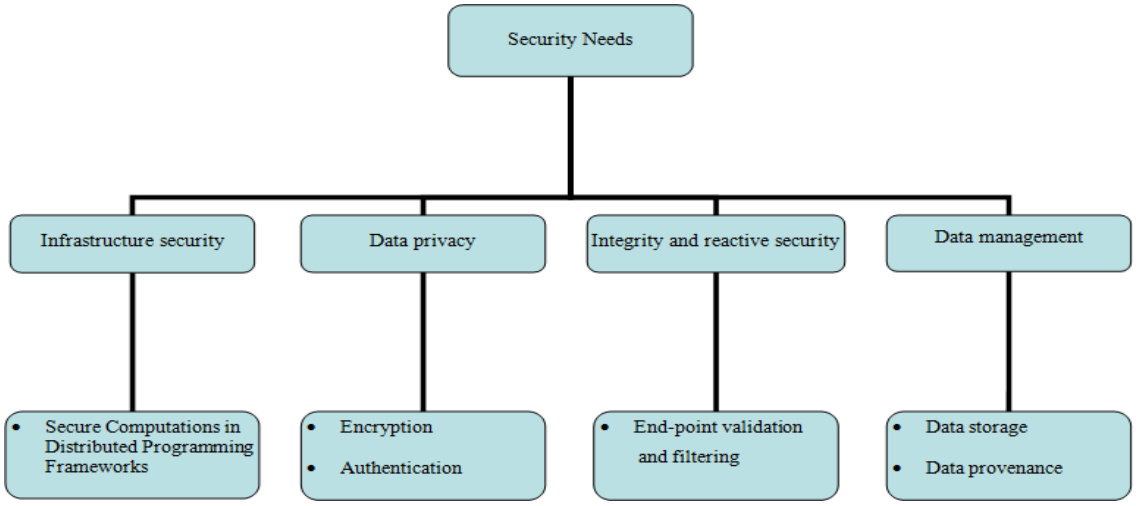

Figure 2: Big Data Security needs

- ***Secure computation in distributed programming frameworks***: in this case, a parallel programming is used to treat a large volume of data, the MapReduces is a framework of distributed programming. therefore we should secure the data from unauthorized mappers (Adluru, Datla, and Zhang 2015).

- ***Encryption:*** encryption is always a good way to protect our data from malicious person. For that, the encryption should be applied in different components such as storage and communication (Gahi, Guennoun, and Mouftah 2016).

- ***Authentication:*** is an efficient mechanism used to control access. In our case, we can use authentication for example in HDFS to have more security during the analysis and authenticate the workers(Adluru, Datla, and Zhang 2015).

- ***End point validation and filtering:*** the end point validation is required to identify the reliability of collected data, then a filter out of data to avoid any risk to the system. Also including low is important because many organizations use the collected data to gain benefits (Dev Mishra and Beer Singh 2016). However, there is no law or regulations that regulate the mining of this data(Ye et al. 2016).

Therefore, a set of regulation should be developed to protect the confidentiality and privacy for individual's life (Zhou et XU.2016). Another needs of security that we should considered as important is activity tracing which use

to use to detect if there were any malicious actions try to manipulate data and even the responsible of the action (Gahi, Guennoun, and Mouftah 2016).

4. Related work

The concept of big data refers to a large amount of data from various sources that are captured and recorded. The definition of the big data can be summarizing into the following 5 characteristics: volume, velocity, variety, veracity and value.

Traditional techniques have not become very effective to ensure the security of the big data, for this reason many researchers have been realized to protect, monitor and audit the big data process. Typically, this paper divides this research into six titles: Hadoop security, Cloud security, monitoring and auditing, Anonymization, Key management and real time processing.

Table 1 summarizes the studies in terms of method proposes and also what guarantees in the security.

Figure 3. Big data Security categorization

a. Hadoop security

Hadoop is an open source framework, it was not designed for data security but rather it was intended to operate in a trusted environment. For this reason, two techniques have been proposed:

The first method is done between the user and the Name node "HDFS component" in this mechanism the user must authenticate to have access to the Name node (Abouelmehdi et al. 2016). The user has to authenticate himself to access. The user sends a hash function then the name node produces a hash function. these two functions will be compare for purpose of granting access or not. The second method is using the Kerberos mechanism in order to increase the security of Hadoop (Adluru, Datla, and Zhang 2015), or using the bull eye algorithm to manage the relation between the original and duplicated data with this algorithm give access to the data for the authorized person only.

b. Cloud Security

The main problem on clouds is data storage. Therefore, the provider should take some measure to handle and share the data on the cloud platform (Saraladevi et al. 2015).the method that have been proposed are: authentication with email, encryption and decryption. A new scheme is proposed for encryption to strength the big data storage. This scheme uses cryptographic virtual mapping to create a path(Kumar, Lee, and Singh2012).

c. Monitoring and auditing

Monitoring is designed to verify the quality and safety of a network event,, and ensure that it is conducted, recorded and reported. Auditing is used to check the security policy. These two elements gain an important role in the security.

Detection and prevention of intrusion on network traffic is difficult. To solve this problem, a monitoring architecture has been developed (Cheng et al. 2015) in this method tree likelihood metrics are calculated: domain name, packet and flow to identify if there is a malicious action or not then send it to detection system. Another method that is use is a self assuring system it separate between the normal and abnormal behavior of users(Marchal et al. 2014).

New gaps for big data auditing appear such as availability, integrity, etc. Therefore, it essential to take all of these gaps in term of big data. To fix the problem of availability a multiple replicas should be present on big data environment. Data integrity is to ensure the whole data is recorded, the traditional way to ensure the integrity is by getting all data from the server and verified by the client but this solution is not working in our case of big data. Thus, to audit dynamic storage data some research has been done. In (Gupta et al. 2014) a proposed schema based on Merkele hash tree to solve the authentication problem.

d. Anonynization

The issue in big data analysis is keeping the privacy of individual's life. Protecting Personally Identifiable Information (PII) is difficult because the data are shared too quickly. To remove the privacy issue, a policy should be applied between the user and a company. Some research has been done; in (Liu et al. 2015) the proposed architecture makes anonymization in sensitive fields in log data with AES and it's measured by K- anonymity.

e. Key management

Because of the variety of big data the safety of unstructured data is more difficult than the structured data. Therefore a proposed approach has been developed for data node to secure the different type of this data. In (Mengke2016.Pdf n.d.) the security of unstructured data contains two stages, data analysis, and security. In data analysis a filtering and classification based on data sensitivity level is done, then assuring the security with a selecting algorithm.

f. Real time processing

The main issue with big data is to handle the real time data stream. To deal with such a problem several researches has been proposed. In (Twardowski and Ryzko 2014) multi agent architecture for processing big data based on lambda architecture has been proposed. This method provides capability for robust real time processing. In [17] the proposed solution is Strom architecture, because of it capabilities such as speed of computation and it's a simple model programming.

TABLE I. CATEGORIZATION OF BIG DATA SECURITY STUDIES

Studies	Purpose		Method	Security Guarantee
(Abouelmehdi et al. 2016)	Hadoop Security	Security and Privacy of hadoop	Encryption between name node and data node using the encryption techniques (AES)	Integrity, Authentication
(Adluru, Datla, and Zhang 2015)		HDFS Security	Based on Kerberos mechanism using the bull eye algorithm	Authentication
(Saraladevi et al. 2015)	Cloud Security	Secure Storage	Using the authentication and encryption/decryption	Authentication, Availability
(Cheng et al. 2015)	Monitoring	Intrusion Detection	Likelihood metrics to detect malicious flows	Tractability
(Marchal et al. 2014)		Detection of abnormal user behavior	Using the self assuring system	Non-repudiation
(Gupta et al. 2014)	Auditing	Audit dynamic data storage	Using Merkle Hash Tree (MHT)	Authentication
(Liu et al. 2015)	Anonymization	Ammonyzation of sensitive fields	K-anonymity	Confidentiality, Privacy
(Sedayao, Bhardwaj, and Gorade 2014)	Key Management	Secure unstructured data	Classification and filtering data	Integrity, Authentication
(Twardowski and Ryzko 2014)	Real Time Processing	Lambda architecture	Multi agent architecture	Integrity
(Yang et al. 2013)		Storm technology	Based on architecture of Strom	Integrity

Several improvements have been made to improve the security of the big data. The table above summarizes the research carried out to enrich the security of the data.

5. Proposed solution

As we already presented the main issues addressed for big data is security. Because of the number of data generated it's difficult to understand the system behavior and sort between the normal and abnormal behavior. For this reason the interest of using an application of naturally inspired algorithm has been increased, such as evolutionary algorithms.

Evolutionary algorithm (AE) is a method of optimization inspired by natural evolution theory. There are several categories of AE among them we can mention: Genetic algorithms, genetic programming, and evolution strategies. In this section we will explain the advantages of using an AE in big data.

In our case we will focused on using genetic algorithm because of their advantages. Their goal is to obtain an approximation of the solution to a complete problem by an optimization mechanism. Its use the notion of natural selection developed in the XIXe century by Darwin and applies it to a population of potential solutions to the given problem.

As we know since the appearance of big data, researchers are targeting their attention for the security of HDFS, but forget about the security for real time processing. For this reason the idea for the work to come is the realization of a solution that goes to hybrid between the treatment performed by mapreduce and the security that is achieved by the artificial immune system.

The use of this solution is come from its several benefits table 2 summarize the advantages of each algorithm:

TABLE II. ALGORITHM ADVANTAGES

MapReduce Algorithm	Immune System Algorithm
- Scalability - Flexibility - Speed - Security and Authentication - Parallel processing	- Adaptability - Distributed processing - Robust - Various - Fault tolerance - Autonomous

The table above show as the benefits of each algorithm. In our case we will try to hybrid between these two algorithms to be able to benefit more from their advantages the speed of execution with the help of the map reduce treatment and the reinforcement of the security with the algorithm of the immune system, especially in the case of real time processing.

6. Conclusion

The security is a major factor. Indeed, the key to success ensuring data privacy for organization is effective information management and governance. Data privacy applies to every internal and external network user within the global business community. However, many people are unaware that there are some measures to protect themselves against the most critical risk.

To conclude, we try to clarify and explain the concept of the big data. We mentioned the different challenges of security. Finally we proposed a solution to reinforce the processing and security of big data.

References

Abouelmehdi, Karim, Abderrahim Beni-Hssane, Hayat Khaloufi, and Mostafa Saadi
2016 Big Data Emerging Issues: Hadoop Security and Privacy. In Pp. 731–736. IEEE. http://ieeexplore.ieee.org/document/7905621/, accessed June 29, 2018.

Adluru, Pradeep, Srikari Sindhoori Datla, and Xiaowen Zhang
2015 Hadoop Eco System for Big Data Security and Privacy. In Pp. 1–6. IEEE. http://ieeexplore.ieee.org/document/7160211/, accessed June 21, 2018.

Cheng, Hongbing, Chunming Rong, Kai Hwang, Weihong Wang, and Yanyan Li
2015 Secure Big Data Storage and Sharing Scheme for Cloud Tenants. China Communications 12(6): 106–115.

Dev Mishra, Aditya, and Youddha Beer Singh
2016 Big Data Analytics for Security and Privacy Challenges. In Pp. 50–53. IEEE. http://ieeexplore.ieee.org/document/7813688/, accessed June 21, 2018.

Gahi, Youssef, Mouhcine Guennoun, and Hussein T. Mouftah
2016 Big Data Analytics: Security and Privacy Challenges. In Pp. 952–957. IEEE. http://ieeexplore.ieee.org/document/7543859/, accessed June 21, 2018.

Gupta, Archana, Ajita Verma, Parul Kalra, and Lokesh Kumar

2014 Big Data: A Security Compliance Model. In Pp. 1–5. IEEE. http://ieeexplore.ieee.org/document/7056963/, accessed June 29, 2018.

Kumar, A., HoonJae Lee, and R. P. Singh
2012 Efficient and Secure Cloud Storage for Handling Big Data. In 2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM2012) Pp. 162–166.

Liu, Chang, Rajiv Ranjan, Chi Yang, et al.
2015 MuR-DPA: Top-Down Levelled Multi-Replica Merkle Hash Tree Based Secure Public Auditing for Dynamic Big Data Storage on Cloud. IEEE Transactions on Computers 64(9): 2609–2622.

Marchal, S., X. Jiang, R. State, and T. Engel
2014 A Big Data Architecture for Large Scale Security Monitoring. In 2014 IEEE International Congress on Big Data Pp. 56–63.

Matturdi, Bardi, Xianwei Zhou, Shuai Li, and Fuhong Lin
2014 Big Data Security and Privacy: A Review. China Communications 11(14): 135–145.

Mehmood, Abid, Iynkaran Natgunanathan, Yong Xiang, Guang Hua, and Song Guo
2016 Protection of Big Data Privacy. IEEE Access 4: 1821–1834.

YANG, M., ZHOU, X., ZENG, J., XU, J., Challenges and solutions of information security issues in the age of big data, China Communication, pp.193-202, March, 2016

Saraladevi, B., N. Pazhaniraja, P. Victer Paul, M. S. Saleem Basha, and P. Dhavachelvan
2015 Big Data and Hadoop-a Study in Security Perspective. Procedia Computer Science 50. Big Data, Cloud and Computing Challenges: 596–601.

Sedayao., J., R. Bhardwaj, and N. Gorade
2014 Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues. In 2014 IEEE International Congress on Big Data Pp. 601–607.

Sedayao , J., Bhardwaj, R., and Gorade, N., Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues, Big Data (BigData Congress), pp. 601 – 607, Anchorage, AK, 2014.

Chingfang ,H., Bing, Z. and Maoyuan, Z., A novel group key transfer for big data security, Applied Mathematics and Computation, vol. 249, pp. 436–443, 2014.

Yang,W., Liu, X., Zhang ,L., and . T. Yang, L., Big Data Real-Time Processing Based on Storm , p. 1784-17, 2013

Twardowski., Bartlomiej., and Dominik., Ryzko., 2014 Multi-Agent Architecture for Real Time Big Data Processing. In Pp. 333–337. IEEE. http://ieeexplore.ieee.org/document/6928203/, accessed June 21, 2018.

Ye, Haina., Xinzhou., Cheng., Mingqiang., Yuan., et al.,
2016 A Survey of Security and Privacy in Big Data. In Pp. 268–272. IEEE.
http://ieeexplore.ieee.org/document/7751634/, accessed June 21, 2018.

Biographies

Ibtissame Kandrouch Received her Eng. Diploma in Computer Science, software engineering option from the National School of Applied Sciences, Ibn Tofail University (Morocco) in 2015;
She is currently preparing a doctorate in science and technology at the same university.
Research interests include the security of big data and information systems, also, the study of cloud environments.

Manale Boughanja Received DUT. Computer network administration in Higher School of Technology (EST) Salé in 2014.Received her LP. Diploma in network and telecommunication, Science University Rabat (Morocco) in 2015;
She is currently preparing a master in Security of Information Systems in National School of Applied Sciences, Kenitra
Research interests include the security of Big Data and information systems.

Nabil Hmina Received Degree in Physics, option Thermodynamics at Mohammed V University (Morocco) in 1989, DEA- Fluid dynamics and transfers, University and Ecole Centrale de Nantes (France) in 1990, University PhD - Engineering Sciences, University and Central School of Nantes in 1994, HDR (1st in Morocco) Ibn Tofail University, kenitra, 2002.
Actually, he’s Director of the National School of Applied Sciences kenitra, since November 2011 to date.

Habiba Chaoui Head of the Logistics and Mathematics Department (ILM), Responsible for the research master "Security of Information Systems" specialty "Security of Systems and Computer Networks", head of research team "Data analysis and information security», Also responsible for MUS "mobile technologies and security" at the national school of applied sciences, ibn tofail university (Morocco).