A Novel Evolutionary Model for Improving Big Data Security

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

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

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

Abstract

Big Data security is a critical mission which aims to keep the data and the various treatments made, safe from any attack attempt, especially in front of the number of attacks and vulnerabilities which increases exponentially. Several algorithmic solutions are put in place in order to reinforce the security of this massive data (in all these states), such as encryption, authorization, intrusion detection, and other techniques inspired by genetics. Indeed, an evolutionary method such as genetic algorithms is set up to increase the level of robustness and variation of the processed data. This paper gives an overview of some modern cryptography security algorithms, and other evolutionary genetic algorithms paramount in data security. He presents comparative and analytical studies of these algorithms in order to proof the effectiveness of one compared to others. To finally present as a proposal an evolutionary model inspired by the immune system of the human body, given its ability to defend against viruses and any unexpected attack, which will be intended for big data during their parallel processing in real time in the MapReduce part of the management system, Hadoop.

Keywords
Big Data Security; attack; algorithmic solutions; evolutionary genetic algorithms; immune system

1. Introduction

Several solutions and technologies have been introduced and / or have evolved in parallel with the exponential evolution of the use of big data, as traditional data management systems are no longer applicable to these huge amounts of data. Indeed, Big Data invites us to meet many challenges, including storage capacity, analysis, purification and processing. Specifically, the challenges regarding the security of the processed data. For all these reasons, this document highlights some algorithmic solutions, which aims improving Big Data Security and propose another solution that can help increase the level of security of this data. So, this document is organized as follows: the second part aims to present the needs of Big Data in terms of security, in order to reveal the increased importance of putting in place solutions that aims to improve the security level of these big quantities of data. The third section is for presenting some existing algorithmic solutions aimed at strengthening the security of data, knowing the way in which they operate, and highlighting their strengths, and especially their weakness. The last part is to present our proposed evolutionary solution for securing Big Data during their real-time processing by MapReduce.

2. Big Data Security needs

Big Data security was and is and will always be a pivotal for any organization handling this data. In fact, these large amounts of data address several security challenges, which are always related to its five characteristics (5V): Volume, Speed, Variety, Veracity, and Value.

1) Velocity

The first challenge is the non-secure computing of data, in fact, an unsecured program can access sensitive data (personal profile, age credit cards, etc.), can corrupt data leading to incorrect results and can perform a denial of service to a leading big data solution to financial loss.

2) Variety

The challenge that relates to the characteristic variety of data is the validation and filtering of endpoint input. There are two fundamental challenges in the data collection process: validation input and data filtering. The amount of data collection in big data makes it hard to validate and filter data. Likewise, there is another challenge that is about granular access control, this second challenge is due to the fact that existing Big Data solutions are designed for performance and scalability, keeping no security in mind.

4) Volume

These include data storage on various distributed data nodes, self-prioritization, real-time analysis and streaming, secure communication (between nodes, middleware, and end users) and transactional Big Data logs.

5) Variety, value and veracity

There are many concerns about monetization and sharing Big Data Analytics in terms of invasion of privacy, invading marketing and unintentional disclosure of information. It's actually the challenge of privacy preserving data extraction and analysis.

3. Big Data Existing Security solutions

Different researches have been conducted to achieve different security mechanisms. For that purpose, several cryptography algorithms have been developed and enhanced to meet this need. Cryptography is a concept to prevent unauthorized disclosure of information. For this reason, this concept is considered from its appearance, until the day among the most powerful techniques for securing data. Talking about cryptography, there are several types of this concept, such as modern cryptography techniques, and homomorphic encryption:

3.1 Modern Cryptography Algorithms

Modern cryptography designates all the principles, means methods of transforming data in order to encrypt their content, establish their authenticity, prevent their modification, and prevent their repudiation or unauthorized use. Among the most known algorithms of this type of encryption is the RSA(Shao et al., 2014), which is an encryption algorithm used in asymmetric cryptography, 99% of the certificates issued to date use the RSA method as an encryption algorithm(Minni et al., 2013). Similarly, there is the ECC, the most recent encryption method. It means Elliptic Curve Cryptography and offers a faster and more secure connection than RSA and DSA methods. ECC also offers shorter key lengths, which require less bandwidth and storage capacity. It is thus a method much more adapted to the mobile connections (smartphones, tablets ...). ECC is compatible with other algorithms. We can combine RSA, DSA and ECC to provide the most secure connection possible. Another method is invented in this way to not only electronically sign data, but it is also used both as a signature algorithm and encryption in SSL certificates(Ebrahim et al., 2014).

Table 1. Comparative study between security algorithms with keys

Factors	***RSA***	***ECC***	***DES***	***AES***
Key length / size (bits)	Based on no of bits	135	56	128, 198, 256
Block size (bits)	change (>1024)	change	64	128
Security rate	Good	Less	Not enough	Excellent
Execution time	Slowest	Fastest	Slow	More fast
Security against attack	Timing attack	----	Brute force attack	Choose n-plain, known n-plain text
cloud environment	AWS, Cipher cloud	AWS	Drop box, AWS	Google apps, eclipse, cipher cloud
Advantages	-Eliminate the security concerns -The minimum key length allowed allows RSA to remain the safest method for many years.	-Reduced storage requirements due to the reduced size of the key.	-Provide a trusted environment for store files -Resist to all linear differential or correlated key attacks carried out with reasonable financial and time resources - chiffre jusqu'à 1Go de données par seconde.	-Speed of Enc/ dec files -Conditions de stockage et de matériels relativement faibles
Weaknesses	Vulnerable to timing attack	Increasing the size of the encrypted message	Hacked by DPA, mim attack	Loss controller on BCC from the e.user, leaked for quantum attack

AES algorithm is better than DES, RSA and ECC. But disadvantage of AES algorithm is key sharing. There is no safe way to share the key. And there is also a loss of data when we are working on a large file. These algorithms had some security issues related to key length, block size, security rate and execution time(Ebrahim et al., 2014). DES is relatively easy to achieve physically and some chips encrypt up to 1 GB of data per second which is huge: it is more than what a normal hard disk can read. For the industrialists it is an important point especially in front of R.S.A. One of the main disadvantages of ECC is that it increases the size of the encrypted message much more than RSA encryption. In addition, the ECC algorithm is more complex and more difficult to implement than RSA, which increases the likelihood of implementation errors, thereby reducing the security of the algorithm.

3.2 Genetic Security Algorithms

AGs can be used to control a system evolving over time (production line, nuclear power station ...) because the population can adapt to changing conditions. AGs are also used to optimize networks, antennas ... They can also be used to improve the security of systems given their ability to self-adapt, and their instant response power against intruders. Below we mention the systems based on genetic algorithms that are essential for strengthening the security of information systems.

- FINGERPRINT, RETINA SCAN, And FACIAL RECOGNITION

Fingerprinting, retinal scanning and facial recognition systems are among the most advanced security systems in the world. Since the fingerprints of two people are not identical, not even the identical twins (who share exactly the same genes). Fingerprints are one of the best biometric credentials to use as a password for protected systems. Two other algorithms have proven their power with regard to the security of user accounts in cloud environments: MIST.

MIST is a security algorithm that allows an authorized person to access their account, facilitates the process of remembering their account login credentials, and limits the effectiveness of social engineering to bypass the system(LeJeune et al., 2016).

- MALACHI

Malachi takes a totally different approach to account security. Instead of using click-based interfaces like MIST, MALACHI relies entirely on typed user input(LeJeune et al., 2016).

Table 2. Data Security Algorithms characteristics

Security algorithm	***characteristics***
MIST	-protect passwords -Check the current IP addresses of users against previous IP addresses of these users -allows an authorized person to access their account, helps them in the memorization process to remember their account login credentials. - MIST has been developed and implemented for use in protected cloud systems.
MALACHI	Security of user accounts for cloud services.
FINGER PRINT	Biometric authentication for systems and ecosystems
RETINA SCAN	Biometric authentication for systems and ecosystems
FACIAL RECOGNITION	Biometric authentication for systems and ecosystems

4. Analysis

Although the security algorithms mentioned above contain weaknesses, but so far, they remain security references for several domains and applications. The problem with these algorithms is that they are no longer applicable for securing BIG DATA in a real time. (I.e. during their processing, their transfer via the network, or during their transition ...). For all these reasons, we thought about designing, developing, and programming an evolutionary algorithm whose operating principle is analogous to that of the human body's immune system, so that ultimately it can be integrated into MapReduce, the part that charge of distributed parallel processing of big data in real time. The algorithm must be necessarily evolutionary considering the capacity of this type of algorithms of the adaptation to the environments, and of the activation in an autonomous way following the 'Darwinian' principle of natural selection.

2. The proposed Solution

The use of artificial immune systems(Xiao and Zhang, 2017) in intrusion detection is an attractive concept for two reasons. First, the human immune system provides a high level of protection against pathogenic invasions in a robust, self-organized and distributed manner. Second, current computer security techniques are not able to cope with the dynamic and increasingly complex nature of computer systems and their security (Banković et al., 2007).
these systems would then have the same beneficial properties as the SIS such as error tolerance, adaptation and self-control.

Table 3. Data Security Algorithms characteristics: Pros and Cons

Security algorithm	***Finger print***	***Retina scan***	***Facial recognition***
Pros	-Simple and less intrusive tests. -May reduce innocent convictions -Can help solve crimes and identity issues	-Low presence of false positives -Extremely low false negatives (almost 0%) -Very reliable because no one has the same retinal profile -Fast results: the identity of the subject is checked very quickly	-Without contact - Automatically investigate people -Numerise several people at the same time
Cons	-May constitute a violation of his privacy -Sensifies concerns about third party access. -May be misused to convict innocent people.	-The accuracy of the measurement can be affected by a disease such as cataracts -The accuracy of the measurement can also be affected by severe astigmatism -The scanning procedure is perceived by some as invasive -Not very friendly -The subject being scanned must be close to the camera's optics High equipment cost	- Significant changes in weight. -Image quality requirements

Why the immune system?

The biological immune system is a weapon against intruders in a given body. It is complicated enough for an artificial simulation to be performed in a complete way. It is possible to identify the most important functions in a biological immune system.The application of the immune system for the defense against foreigners elements in an instantaneous manner in the MapReduce share, which operates by manipulation in a parallel and distributed way, autonomously, and this in an independent manner of any interactions with external antigens.

The proposed solution

After storing the data entered in the HDFS module, and during their parallel processing by MapReduce, they will be processed in addition to the new built-in algorithm.

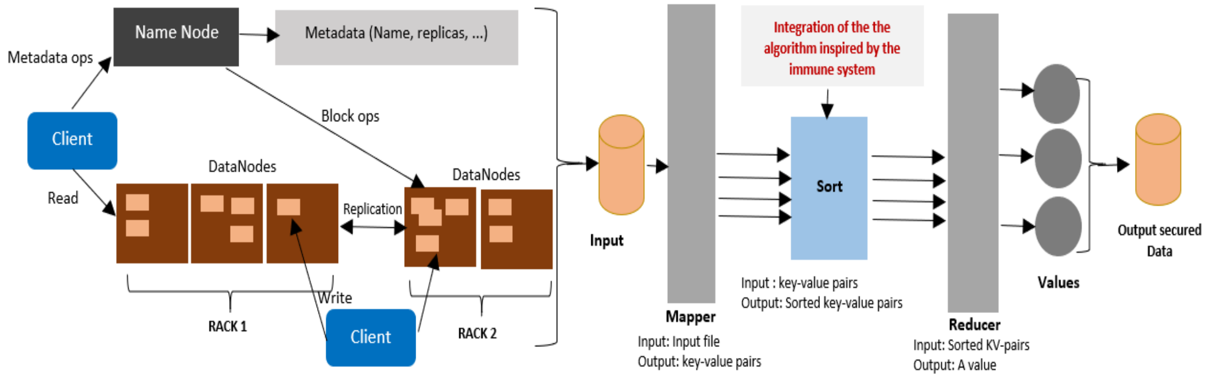

Figure 1. The new Hadoop architecture after integrating the genetic algorithm

How the architecture works?

The functioning of this new architecture is ensured by five daemons:

1. Name Node;
2. Data Node;
3. Mapper;
4. The integrated algorithm;
5. Reducer.

When a user solicits Hadoop to recover files:
The services requested will be retrieved via the name Node. This Name Node will tell the user which Data nodes contain the blocks. The Name Node regularly receives a "heartbeat" and a Block report of all the Data Nodes in the cluster to ensure that the data nodes are working properly. A Bloc report contains a list of all blocks in a Data Node. In the case of a data Node failure, the Name Node selects new data Nodes for new data block replications, balances the disk usage load and manages the data traffic of the data Nodes.

When the drive retrieves the data from the storage system, it must convert the data into <key, value> pair to continue the execution (this process is called data analysis). The analysis consists in decoding the data from their native storage format in order to transform them into a format that can be used by a programming language. During this transformation, the data will pass in addition by the new integrated algorithm whose purpose is to eliminate any intrusions or parasite encountered. The resulting data from this transformation will then go through the reducer of the goal to stand out from the value.

6. Conclusion

Although Big Data is the new digital revolution in the field of data, that presents a good number of benefits to the users, but in return, it faces a lot of security challenges. This paper aims to shed light on the security problems encountered, and especially on those that hinder the proper functioning of big data processing systems during their work. For these reasons, we tried to present through this paper as a solution to the problems mentioned, the application of an algorithm in the Map reduce, part of the Hadoop ecosystem, which is responsible for parallel processing in real time. Whose operation (of the algorithm), is analogeous to the work of the immune system of the human body.

Acknowledgements

The authors are grateful to the anonymous reviewers and the anonymous critics for their helpful comments and their remarks. This will improve the quality of the article, and will help us to advance in our future work

References

Banković, Z., Stepanović, D., Bojanić, S., Nieto-Taladriz, O., 2007. Improving network security using genetic algorithm approach. Computers & Electrical Engineering 33, 438–451.

Ebrahim, M., Khan, S., Khalid, U.B., 2014. Symmetric algorithm survey: a comparative analysis. arXiv preprint arXiv:1405.0398.

LeJeune, J., Tunstall, C., Yang, K., Alkadi, I., 2016. An algorithmic approach to improving cloud security: The MIST and Malachi algorithms, in: Aerospace Conference, 2016 IEEE. IEEE, pp. 1–7.

Minni, R., Sultania, K., Mishra, S., Vincent, D.R., 2013. An algorithm to enhance security in RSA, in: Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference On. IEEE, pp. 1–4.

Shao, C., Li, H., Zhang, X., 2014. Cryptographic implementation of RSA for ion fault injection attack, in: Consumer Communications and Networking Conference (CCNC), 2014 IEEE 11th. IEEE, pp. 791–796.

Xiao, X., Zhang, R., 2017. Study of Immune-Based Intrusion Detection Technology in Wireless Sensor Networks. Arabian Journal for Science and Engineering 42, 3159–3174.

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.

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).