ISSN 2278-3091

Volume 9, No.4, July – August 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse292942020.pdf

https://doi.org/10.30534/ijatcse/2020/292942020

Survey on Security Monitoring and Intrusion Detection in the Big Data Environment

Alaeddine BOUKHALFA[1], **Nabil HMINA**[2], **Habiba CHAOUI** [3]

[1, 3] System Engineering Laboratory, ADSI Team, National School of Applied Sciences, Ibn Tofail University, Morocco

[2] Sultan Moulay Slimane University, Beni Mellal, Morocco

[1] alaeddine.boukhalfa@gmail.com

[2] hmina5864@gmail.com

[3] habiba.chaoui@uit.ac.ma

ABSTRACT

Recently, new cities have appeared called smart cities (SC) that use electronic devices to manage and transfer information. These electronic devices are interconnected via a new generation of the internet called internet of things (IoT), and are generating and transferring continuously a large mass and variety of data between them, these data transfers can contain new intrusions. However, the current information security devices cannot identify these new intrusions, and have problems with the large amount and diversity of produced data. To solve these issues, new approaches have been proposed for security monitoring and intrusion detection in a Big Data environment. In this paper, we will give an overview on these proposals and analyze them, to identify their weaknesses in order to help researchers improve information security in the future.

Key words : Survey, Big Data, Internet of Things, Intrusion Detection, Security Monitoring.

1. INTRODUCTION

Currently the world is experiencing a great revolution in the field of information and communication technologies (ICT), new evolved urban areas have appeared recently called smart cities (SC), that use electronic devices to improve the quality, performance and interactivity of urban services, and also reduce costs and resource consumption, namely, the efficient management of traffic, transport, waste, supply of drinking water and electricity, improving the health system, and the development of the education system.

Several apparatus and electronic devices, like computers, smart phones, tablets, smart TVs, and others manipulated by citizens, or information systems and electronic sensors used by the various urban services of these smart cities, are interconnected through a large network called Internet of Things (IoT), in order to transfer and exchange information between them.

Unfortunately, transfers and exchanges between these various electronic devices are not always safe, they can sometimes hide many new attacks and intrusions created lately by hackers.

In addition, the exchanges and communications between the various electronic objects are at the origin of an exponential growth in the volume, and a variety of data generated over the network (IoT), such as videos, sounds, images, text messages, satellite geo-positioning (GPS) data, climate data, etc… which causes detection difficulties for current information security mechanisms and tools.

With this growth and diversity of data produced, traditional database management tools suffer from performance issues, this has pushed recently researchers to find new ways to solve these problems. As a consequence, a new concept was created called Big Data, to process the large amount and diversity of data.

To solve the issues cited above, many new solutions have been proposed for security monitoring and intrusion detection using Big Data techniques in environments of large mass and diversity of data. This paper exposes the state of the art of these propositions, it is organized as follows. Section 2 gives a brief summary of the background. Section 3 presents an analysis of studies in literature concerning security monitoring and intrusion detection in large data environments. Section 4 provides a synthesis of these studies to identify the weak points, in order to help research develop information security in the future. Section 5 gives the conclusion of this survey.

2. BACKGROUND

2.1 Smart Cities

Nowadays, the world is evolving by using new information and communication technologies (ICT), ancient cities have taken new forms, and transformed into evolved cities called smart cities (SC), as shown in Figure 1, these are cities that manipulate electronic devices and digital tools [1], for the purpose of managing efficiently urban services and systems, such as transport, security, health and education, etc...In order to solve problems, limit consumption of costs and resources, and achieve operational performance.

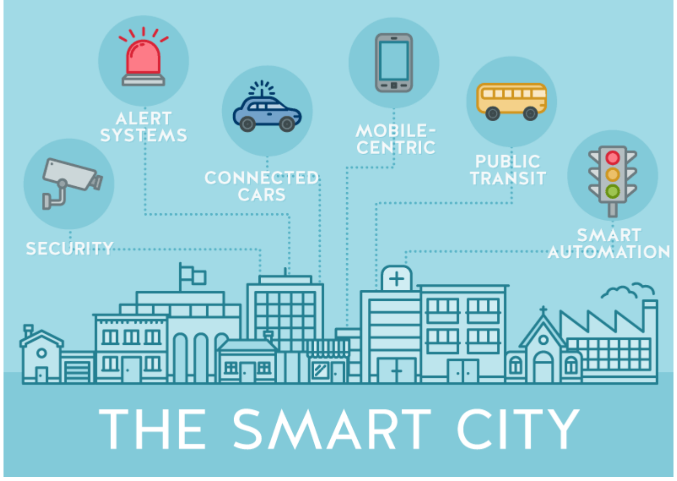

Figure 1: The Smart City

2.2 Internet of Things

To communicate and transfer information within and between these smart cities, information systems and digital systems of urban services, and also terminals like smartphones, digital terrestrial television, computers and laptops manipulated by users, are interconnected via a large network called Internet of Things (IoT). As illustrated in Figure 2, It is a new generation of the Internet that links electronic and digital objects of various users around the world [2].

Figure 2: Internet of Things Network

2.3 Big Data

The rapid and uncontrollable growth of connected objects, is responsible, without a doubt, for increasing the volume of information transmitted via the network (IoT), and the diversity of the data produced, namely, emails, scanned documents, images, geolocation data, etc…However, the traditional means and tools for managing databases and information, have really problems with organizing and processing the huge amount and variety of data generated. What recently launched research to find new techniques capable of processing quickly and easily data of different natures and in large quantities. Therefore, a new concept was created called Big Data to deal with these problems. These are new measures concerning the capture, research, sharing, storage, analysis and presentation of this data. As shown in Figure 3, Big Data is characterized by five characteristics called 5Vs [3]:

- Volume: it means a large amount of data to process.
- Variety: it means data of different natures from different sources.
- Velocity: it means the frequency of creation and processing of data.
- Value: it describes the gain from the use of this Big Data.
- Veracity: it designates the credibility of these data.

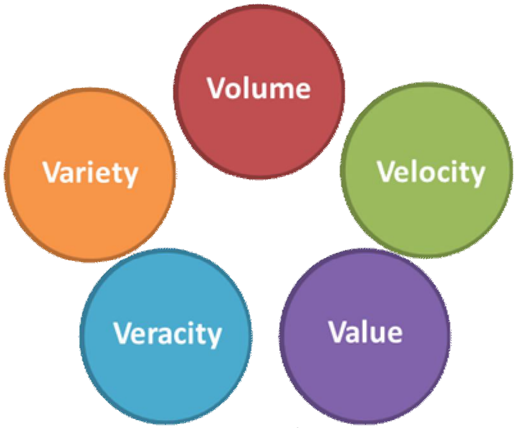

Figure 3: 5Vs of Big Data

3. SECUITY MONITORING AND INTRUSION DETECTION IN A BIG DATA ENVIRONMENT

It is true that communications and data transfers between electronic devices contain sometimes hidden new attacks and intrusions, and with the suffering of the existing means and tools of information security with this large quantity and variety of data exchanged, new models have been proposed for security monitoring and intrusion detection in large data environments, we will try to cite and analyze them in this section.

3.1 Security Monitoring

In a Big Data environment, network security monitoring was proposed in the manuscript [4], the authors affirmed that current security devices are not dedicated to Big Data environments. They illustrated, firstly, a system based on: the accumulation of the large amount of data, its integration that requires purification and classification operations, its analysis to get information, and exposing this information in order to make decisions. Secondly, they gave an overview on some correlation methods used to analyze data. The study was limited to the analysis of data and the presentation of the existing attacks without any action against these threats. It did not propose an improvement to predict new attacks. No Big Data management system has been proposed.

Another approach for security monitoring of Internet of Things (IoT) network was discussed in the paper [5], the authors confirmed that traditional methods and means of security suffer with the large amount and variety of data passing through the network. A new system has been suggested, based on collecting the large amount and variety of security log data from the electronic devices of public users, storing these data using Big Data management system Apache Hadoop [6], normalizing data to get a unique format of data, analyzing data by applying aggregation and correlation operations through Complex Event Processing (CEP) methods, and visualizing results employing advanced visualization tools. The principle of the article is based only on displaying the state of logs without any action against attacks.

In the article [7], the authors approve that the current solutions of deep analysis of the packets of the network are neither evolutionary nor adaptable with the large amount and diversity of network data. A new model for security monitoring and traffic analysis has been proposed, it groups two parts, the first part is dedicated to the storage of network traffic data in a distributed way using a Big Data management system, the second part is dedicated to the analysis of this network traffic using correlation algorithms in order to detect intrusions. The architecture has been evaluated using the best known Big Data management systems. The system represents an interesting idea for analyzing a large mass of network traffic data, except that the used correlation algorithms are not evolutionary to detect new threats.

One more approach for monitoring and analyzing the big traffic of the network was proposed in the paper [8], the authors mentioned that the old network traffic analysis tools are not efficient with a large amount of data. They proposed a new multi-layered traffic analysis system, that collects the data, stores it in the storage component of Apache Hadoop called Hadoop Distributed File System (HDFS) [9], analyzes it with the analysis program of Apache Hadoop called MapReduce [9], and presents the results via an interface. To verify its effectiveness, the system was evaluated with traffic from a mobile communications network, the results were promising. The proposed solution is remarkable, except that the analysis steps does not aim to detect attacks.

Another vision for monitoring Internet traffic was discussed in the article [10], the authors proved that the old traffic analysis tools have become obsolete with a large amount of data passing through the network. They proposed an online solution consisting of three stages, the first stage collects the packets that pass through the network via several collectors, the second stage receives the data in the form of messages via a distributed messaging system called Apache Kafka [11], the third step processes data via the Big Data management system Apache Spark [12] and displays the results to operators. The solution is not based on evolutionary analytical methods that can detect new attacks. The motivation for choosing the Big Data management system has not been presented.

3.2 Intrusion Detection

A concept for analyzing network traffic in order to detect anomalies has been suggested in the paper [13], the authors approved the necessity of new means to analyze the big traffic of the internet network in order to recognize attacks. They suggested a distributed manner of analysis established on the MapReduce analysis program. The method consists in dividing the network traffic into several parts, then the analysis is applied to each part of the data to detect attacks, finally, the results are displayed to network administrators. The proposal gave good results. The solution presented is interesting, but the analysis method adopted is not evolutionary to detect new intrusions.

Another idea for analyzing network traffic was presented in the article [14], the authors affirmed that network traffic constitutes a large amount of data that contains many attacks. They proposed a set of analysis algorithms employing the R language to deal with problems attached to volume, variety and veracity of big quantities of data. The suggested algorithms are interesting, except that they aim to improve the quality of the Big Data without any action against menaces. No Big Data management system has been proposed.

A new model for anomalies detection in a telecommunications domain was proposed in the paper [15], the authors admit that a large amount of data are generated currently by many devices around the world, and these data must be managed efficiently. They proposed a new concept for anomalies detection based on the collection of data from a telecommunications network, the preparation of data before processing, the analysis using an unsupervised clustering method, and the display of anomalies detected. The proposed model constitutes a new idea for anomalies detection but it is not dedicated to finding new intrusions. No Big Data management system has been employed.

The authors present, in paper [16], a new approach for intrusion detection in an environment of great quantity of data, they affirmed that intrusions increase as the quantity of generated data increases. They presented a new distributed architecture established on loading network traffic, and processing it employing Apache Spark Streaming [17]. The experiment was carried out and gave good results. The proposed model represents an interesting approach in a Big Data environment, but the authors did not describe the employed algorithms for intrusion detection. the choice of the used Big Data management System has not been justified.

Table 1: Synthesis and analysis of studies realized for security monitoring and intrusion detection

Study	Purpose	Proposition	Data environment	Data management system	Analysis method	Evolutionary analysis method
[4]	Security Monitoring	System for network security monitoring based on collection, normalization and analysis.	Big Data	Unspecified	Correlation algorithms	Not evolutionary
[5]		System for network security monitoring based on collection, normalization, analysis and visualization.	Big Data	Apache Hadoop	Complex Event Processing (CEP)	Not evolutionary
[7]		System for security monitoring based on: collection, storage and analysis	Big Data	Apache (Hadoop, Spark, Hive, Pig, Shark)	Correlation algorithms	Not evolutionary
[8]		System for monitoring based on: collection, storage, analysis and visualization	Big Data	Apache Hadoop	Apache Hadoop MapReduce	Not evolutionary
[10]		System for monitoring based on: collection, storage, analysis and visualization	Big Data	Apache Spark	Anlysis algorithms	Not evolutionary
[18]	Intrusion Detection	Architecture based on : collection, normalization and analysis	Big Data	Unspecified	Finite Dirichlet Mixture Model	Not evolutionary
[13]		Distributed approach for analysis using a big data management system	Big Data	Apache Hadoop	Apache Hadoop MapReduce	Not evolutionary
[14]		Analysis methods based on R language	Unspecified	Unspecified	R language	Not evolutionary
[15]		Architecture based on: collection, preparation, analysis and visualization	Unspecified	Unspecified	Unsupervised Clustering	Not evolutionary
[16]		Architecture based on: collection, preparation, integration and analysis	Big Data	Apache Spark	Apache Spark Streaming	Not evolutionary

A new system for intrusion detection established on an algorithm called Finite Dirichlet Mixture Model was suggested, in paper [18]. The authors claimed that a system that identifies no menaces per day is considered obsolete, they proposed a new model that captures and logs network data, prepares them for the next operation, and analyzes them using the Dirichlet Mixture Model algorithm in order to find intrusions. The system has been evaluated and the results were promising. The proposal represents a remarkable idea for intrusion detection, but it is not dedicated to identify new intrusions. No Big Data management system was proposed.

4. SYNTHESIS AND ANALYSIS

Table 1 presents the synthesis of the studies for security monitoring and intrusion detection in a Big Data environment mentioned above, it gives an overview on each study in terms of the objective to be reached, the proposal of the solution, the environment of data aimed by the solution, the data management system used, the analysis method adopted, and the evolution of the analysis method adopted.

As shown in Table 1, almost all proposed solutions for security monitoring and intrusion detection in a Big Data environment, describe essentially systems based on: firstly, the collection and accumulation of the large quantity and variety of data from electronic device logs and traffic networks. Secondly, the storage of these large data sets without using a Big Data Management system, or using Big Data management systems Apache Hadoop, Apache Spark and others, without justifying exactly the choice of the systems adopted. Thirdly, the analysis of these data to detect only the existing intrusions using non-evolutionary methods, namely, the distributed processing programs of Big Data management systems like MapReduce, Complex Event Processing (CEP) methods and correlation algorithms, without proposing new methods able to evolve automatically to predict and detect new intrusions. Fourthly, the display, presentation and visualization of the results of data analysis without any action against attacks, however, it is necessary to think about implementing mechanisms to block these attacks.

5. CONCLUSION

With the increasing amount of data generated over the network, new intrusions have appeared. Many new solutions have been proposed for monitoring security and intrusion detection in a Big Data environment. In this paper, we have tried to expose these studies and analyze them, for the purpose of identifying their failures, to be discussed by research in the future.

REFERENCES

1. V. Albino, U. Berardi and R. M. Dangelico, **Smart Cities: Definitions, Dimensions, Performance, and Initiatives**, Journal of Urban Technology, vol. 22, no 1, p. 3‑21, 2015, doi: 10.1080/10630732.2014.942092.
2. A. Atmani, I. Kandrouch, N. Hmina and H. Chaoui, **Big Data for Internet of Things: A Survey on IoT Frameworks and Platforms**, in Advanced Intelligent Systems for Sustainable Development (AI2SD'2019), Cham, 2020, p. 59 ‑ 67, doi: 10.1007/978-3-030-33103-0_7.
3. A. Boukhalfa, A. Abdellaoui, N. Hmina and H. Chaoui, **Network Traffic Analysis using Big Data and Deep Learning Techniques**, in 2020 IEEE 6th International Conference on Optimization and Applications (ICOA), Beni Mellal, Morocco, 2020, p. 1-4, doi: 10.1109/ICOA49421.2020.9094455.
4. L. Lan and L. Jun, **Some Special Issues of Network Security Monitoring on Big Data Environments**, in 2013 IEEE 11th International Conference on Dependable, Autonomic and Secure Computing, 2013, p. 10-15 doi: 10.1109/DASC.2013.30.
5. I. Saenko, I. Kotenko and A. Kushnerevich, **Parallel Processing of Big Heterogeneous Data for Security Monitoring of IoT Networks**, in 2017 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), 2017, p. 329-336
6. Ziyad R. Al Ashhab, Mohammed Anbar, Manmeet Mahinderjit Singh and Kamal Alieyan, **Detection of HTTP Flooding DDoS Attack using Hadoop with MapReduce: A Survey**, International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no 1, p. 71-77, 2019.
7. S. Marchal, X. Jiang, R. State and T. Engel, **A Big Data Architecture for Large Scale Security Monitoring** , in 2014 IEEE International Congress on Big Data, 2014, p. 56 ‑ 63, doi: 10.1109/BigData.Congress.2014.18.
8. J. Liu, F. Liu and N. Ansari, **Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop**, IEEE Network, vol. 28, no 4, p. 32‑ 39, 2014 doi: 10.1109/MNET.2014.6863129.
9. M. R. Ghazi and D. Gangodkar, **Hadoop, MapReduce and HDFS: A Developers Perspective** , Procedia Computer Science, vol. 48, p. 45‑ 50, 2015 doi: 10.1016/j.procs.2015.04.108.
10. B. Zhou et al., **Online Internet traffic monitoring system using spark streaming**, Big Data Mining and Analytics, vol. 1, no 1, p. 47 ‑ 56, 2018, doi: 10.26599/BDMA.2018.9020005.
11. « Apache Kafka », Apache Kafka. http://kafka.apache.org/
12. Mouad Banane and Abdessamad Belangour, **Querying massive RDF data using Spark**, International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no 4, p. 1481-1486, 2019 doi: 10.30534/ijatcse/2019/68842019.
13. R. Fontugne, J. Mazel, and K. Fukuda, **Hashdoop: A MapReduce framework for network anomaly detection**, in 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, avr. 2014, p. 494‑ 499 doi: 10.1109/INFCOMW.2014.6849281.
14. L. Wang and R. Jones, **Big Data Analytics in Cyber Security: Network Traffic and Attacks**, Journal of Computer Information Systems, p. 1‑ 8, 2020
15. V. K. Vasantham and V. Meka, **User-Anomaly Detection in Telecommunication Using Big Data Analytics**, vol. 7, no 5, p. 4, 2019.
16. M. T. Tun, D. E. Nyaung and M. P. Phyu, **Performance Evaluation of Intrusion Detection Streaming Transactions Using Apache Kafka and Spark Streaming**, in 2019 International Conference on Advanced Information Technologies (ICAIT), Yangon, Myanmar, 2019, p. 25‑ 30, doi: 10.1109/AITC.2019.8920960.
17. « Spark Streaming - Spark 2.4.5 Documentation ». https://spark.apache.org/docs/latest/streaming-programming-guide.html
18. N. Moustafa, G. Creech, et J. Slay, **Big Data Analytics for Intrusion Detection System: Statistical Decision-Making Using Finite Dirichlet Mixture Models**, in Data Analytics and Decision Support for Cybersecurity, I. Palomares Carrascosa, H. K. Kalutarage and Y. Huang, Éd. Cham: Springer International Publishing, 2017, p. 127-156.