Smart SIEM: From Big Data Logs and Events To Smart Data Alerts

Mohammed EL ARASS, Nissrine SOUISSI

Abstract: Cyber defense has become an increasingly recurrent and even a mandatory requirement for any type of organization that holds an Information System. In this context, a SIEM is the most suitable system for malicious activities detection. However, current classical SIEMs neglect Big Data issues. To fill the gap, this paper proposes a new generation open source SIEM composed of a Big Data platform ELK integrated with other intrusion detection and load-balancing tools named Smart SIEM. The features of the proposed system have been tested in a virtual environment composed of several Windows and Linux devices to see how it behaves against some of the most well-known attack scenarios in the literature, and the results were challenging. The proposed prototype was also compared to the most advanced SIEM QRadar and another new generation SIEM from scientific research.

Index Terms: Big Data, Cybersecurity, Data LifeCycle (DLC), Smart DLC, Elastic Stack Log and Kibana (ELK), Security Information Event Management (SIEM).

I. INTRODUCTION

Nowadays, companies no longer regard cyberattacks as a possibility but as a reality that will, undoubtedly, occur. It will be necessary to be aware of it as much as possible in order to deploy all available means to face it [1].The leading network security provider of network security solutions for defense. Corero confirmed in [2] that Distributed Denial of Service (DDoS) increase each year by 40% with more than 400,000 attacks monthly in 2018. Hence the need for a centralized system that monitors all the Information System (IS) activities to infer events that could turn into security incidents that require special care. A Security Information Event Management (SIEM) has been defined in [3], [4] as a software solution that provides security information of an Information System. However, [5] defines the SIEM as a set of security software tools including the following systems: log management, log security, event management, security information management, and security event correlation.
The main functions of the SIEM can be summarized as follows:

- Log collection and analysis
- Log Transfer in a standard format
- Security threats notification
- Security incident detection
- Incident response workflow

Commercial SIEMs that rely on relational databases struggle against the strangulation of their databases for companies with consistent IS [6]. The purpose of a SIEM is to align the heterogeneous information and event logs from various devices in a network to provide a common interface for visualizing and analyzing that data. The case study conducted by Zions Bancorporation in [7] revealed that it would take 20 minutes to an hour to query a month's security data from their traditional SIEMs. However, the same query takes only one minute using Big Data platforms such as Hadoop. As a result, Big Data technologies are now essential in next generation SIEMs to optimize their performance against the huge amount of data to manage. We identified two issues that we are trying to address in this paper:

- The majority of high-performance SIEMs are very expensive and built in a black box. Few organizations are able to acquire them. Designing and implementing an open source SIEM with the same functionalities as commercial ones is a scientific challenge;
- The current classic SIEMs do not manage Big Data. Analyzing security data from heterogeneous sources can be difficult for intrusion detection when homogeneous sources already have problems with huge amounts of this data.

The contribution of this paper is to design and implement an open source prototype of a new generation SIEM named Smart SIEM adapted to the Big Data context for monitoring IT devices security. This prototype is an integration of several existing open source cybersecurity and load balancing tools with a Big Data platform.

In this article, we design a new architecture of the SIEMs adapted to the Big Data context and we develop this architecture through an implementation of an open source new generation SIEM research prototype that consists of collection, storage and processing security information and events based on a Big Data platform. We took advantage of a Big Data platform ELK capability to identify the most appropriate architecture for this context and also to monitor malicious activity. We also used an open source tool for load balancing management to optimize our prototype in the real-time. The rest of this paper is as follows: Section 2 compares the proposed SIEM to the market-leading SIEM and another most relevant of the scientific literature. Section 3 describes our method for the design and the implementation of Smart SIEM. Section 4 presents and tests the proposed prototype. Section 5 discusses the results of the conducted tests. Section 6 concludes the article and gives some future directions.

Revised Manuscript Received on June 05, 2019

Mohammed EL ARASS, Mohammed V University in Rabat, EMI-SIWEB Team, Rabat Morocco

Nissrine SOUISSI, Mohammed V University in Rabat, EMI-SIWEB Team, Rabat Morocco

II. RELATED WORKS

To identify the commercial SIEM to compare with our prototype, we used the latest Gartner magic quadrant for SIEMs published in October 2018 presented in Figure 1.

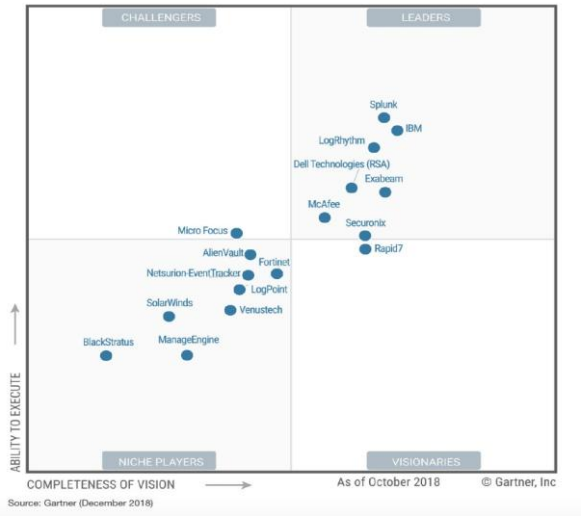

Fig. 1. The latest Gartner magic quadrant for SIEMs

According to Gartner's magic quadrant for SIEMs in October 2018, IBM's SIEM QRadar is the market leader. For this, we retain the QRadar architecture as defined in [8] as a basis for the Smart SIEM architecture.

The QRadar architecture is composed of 4 modules:

- **QRadar Console**: allows the results visualization
- **QRadar Event Collector**: collects, normalizes, and restructures the received logs according to the recommended format.
- **QRadar Event Processor**: processes logs received by QRadar Event Collector using a predefined rules engine to generate alerts.
- **QRadar QFlow Collector**: performs the same role as QRadar Event Collector but for real-time data such as network traffic captures.
- **QRadar Flow Processor**: performs the same role as QRadar Event Processor but processes the data received from QRadar QFlow Collector.
- **QRadar Data Node**: adds additional storage and analysis capabilities as needed.

We believe that all QRadar modules must be available in Smart SIEM. We think that other modules are important to better manage the received data, including integration, filtering and enrichment module, must also be supported.

In addition, QRadar distinguishes between log data and stream data. For this purpose, it defines a specific collection and analysis module for log data (QRadar Event Collector and QRadar Event Processor) and another module for the same roles for other data types (QRadar QFlow Collector and QRadar Flow Processor). We do not want to make any distinction as to data format type, as this will optimize the processing of the data from their reception by a single module (collection) in order to analyze them by a single module (analysis). QRadar also manages the Big Data aspect through the QRadar Data Node module, which allows adding as many nodes as needed to manage the number of received logs. Our SIEM will also benefit from this advantage as well by its distributed architecture composed of a master server and several slave servers as by the adopted Big Data platform. A load sharing tool in RAM will also be used to avoid losing data in case of congestion, unlike QRadar which only has a queue of logs that are waiting to be processed.

QRadar is a commercial SIEM billed based on the number of events received per second (EPS), when this number is exceeded the data is stored in the queue of the QRadar Event Collector module until the rate is decreased. However, if this rate is exceeded and the waiting queue is filled, the system deletes the events and QRadar notifies this. QRadar is intended for large companies because its acquisition cost is too expensive. But our Smart SIEM is free and could concern both small companies with standalone architecture and large companies with distributed architecture.

In the literature, there is little new generation SIEM that manages the Big Data aspect. We have identified only three new generation SIEMs in the literature: [9]–[11]. We have excluded [10] because it is a SIEM designed for monitoring only IoTs, and we are looking for a next generation SIEM that can monitor any IT equipment. [11] is an extension of [9], for this, we retain this SIEM to study its architecture.

[11] defines an architecture composed of eight modules:

- **The log collection module:** ensures the collection of log files.
- **Enrichment module**: adds additional data depending on the monitoring context.
- **The normalization module:** transforms the received data from the enrichment module to present them to the correlation and analysis module according to a predefined format.
- **The correlation and analysis module:** is the main element of this SIEM and provides results using historical results.
- **The storage module**: stores the logs after their collection and the normalized logs, it also stores the results provided by the correlation and analysis module for a possible comparison by this one.
- **The reporting module:** generates reports following the results found.
- **The monitoring module**: ensures results visualization provided by the correlation and analysis module.
- **The incident alert and response module:** manages the sending of alerts to the various stakeholders to react or provides automatic responses if an incident occurs.

This SIEM enriches all data collected, we believe that data enrichment should take place after their integration and filtering to optimize the processing. In addition, it stores only the collected logs at the beginning and the results are provided by the analysis and correlation module. If this SIEM manages only log data the one we propose will handle all types of data including log files. Finally, the SIEM proposed by [11] has not been implemented to validate its designed model.

III. METHOD

Based on the research works proposed in the literature, we have designed a new generation SIEM in a distributed architecture with a master server and several slave servers to face the cybersecurity and Big Data demands, which classic SIEMs cannot satisfy.

To do this, after conducting a data lifecycles (DLC) analysis through a literature review in [12], we used a DLC adapted in the Big Data context called Smart DLC [13] to identify our SIEM modules. This is justified by the fact that a new generation SIEM is a data lifecycle that manages log and flows data in a Big Data context. Our SIEM has been designed to be a new generation that takes into account the traditional SIEMs limitations in terms of Big Data management. This requirement has led us to choose a Big Data platform for managing the huge amount of logs that a classic SIEM cannot perform. In order to achieve this, we have defined several requirements that a new generation SIEM must fulfill:

- Collect data from several and different sources and verify the 7Vs of Big Data (Volume, Velocity, Variety and Veracity, Value, Variability, and Visualization) defined in [14];
- Ensure a system elasticity that is a rational and optimal resources allocation [15];
- Ensure detection and rapid reaction as this has a considerable financial impact [16];
- Ensure easy use of the system by analysts and experts. The effort to integrate and configure new devices must be the minimum possible;
- Ensure a high degree of automation to minimize analysts' interventions so that they can focus on the most timely cases;
- Ensure the integration of any type of IT device. Indeed, the system must handle several types of log formats;
- Ensure reusability of the system that is to say that a SIEM composed of several modules to facilitate its integration with other similar systems. [11];
- Provide IT monitoring and protection at a low cost, this includes the cost of acquiring the SIEM and the costs of installation, configuration, and maintenance.

IV. SMART SIEM

In this section, we present our proposed SIEM by defining its components as well as its architecture.

A. Smart SIEM modules

Given the new generation SIEM requirements, we propose a model consisting of seven modules as shown in Figure 2 that meets all these requirements. These modules were derived from Smart DLC [13].

Smart DLC is a process cartography organized into three types of processes according to the standard 9001: 2015 [17] and the CIGREF standard [18]: management process, production process and support process [19]. We selected from this cycle 6 realization processes and one support process (storage process).

We first looked at the Smart DLC realization processes explained in [13] because it is this type of process that manipulates data and has a direct impact on them. The selected modules correspond to the Smart DLC realization processes and which are: **collection module, integration module, filtering module, enrichment module, analysis module, and visualization module**.

To be able to perform the storage functions for our SIEM, we also retained the storage process that belongs to the Smart DLC support processes.

The first **collection module** is responsible for logs collecting and supplying these data to the following modules. The **integration module** allows the normalization of heterogeneous logs in a unified format, it also handles duplicate and erroneous logs. The **filtering module** allows selecting the information we want to monitor, which is usually JSON objects. The **enrichment module** uses the context data to add information that is not included in the raw logs. This is the case, for example, for the IP addresses geographic coordinates, in order to improve their processing thereafter. The **analysis module** represents the main module of our prototype because the alerts are recognized based on the incoming standardized data. The **visualization module** displays the results of the analysis module in a smart manner as it facilitates the detection of possible anomalies or attacks. All data manipulated in the system are stored in the storage module.

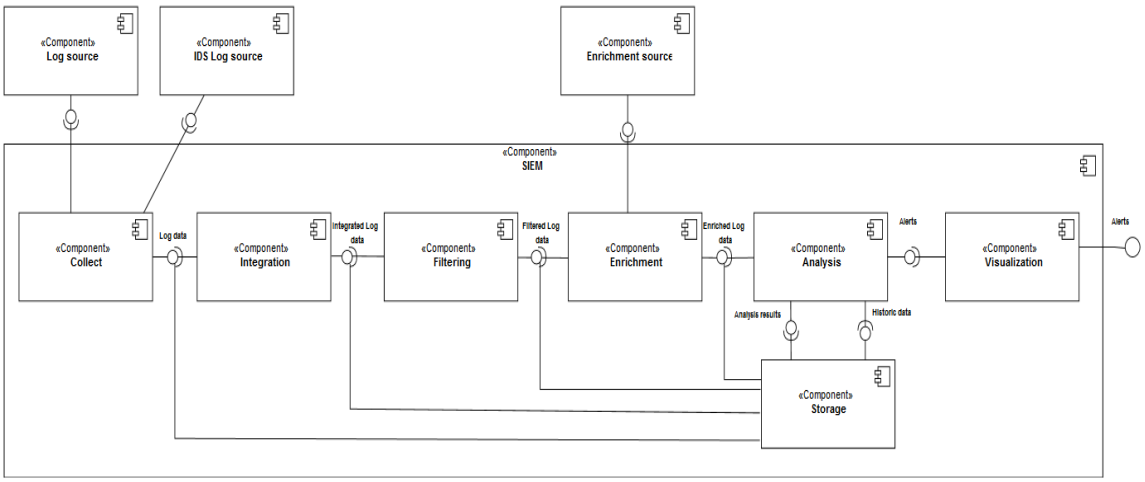

Fig. 2. The proposed SIEM component diagram

To implement the proposed model, we used the Big Data platform ELK as a base for our prototype [20]. ELK is a Big Data platform designed for centralized log management composed of several open source components. In itself, ELK is not a SIEM solution, but simply a solution for managing and aggregating log files. However, we have combined it with other security tools to enhance the ELK functionality to meet the new generation SIEM system requirements that we have defined previously. ELK is made up of Beats, Logstash, Elasticsearch and Kibana.

- **Beats** is an open source agent platform installed on remote machines to send log data to Logstash.
- **Logstash** is a dynamic open source pipeline that collects and integrates data from multiple sources simultaneously to transform and send them to the storage system, generally elaticsearch. [20], [21].
- **Elasticsearch** is free distributed, RESTful search and analytics. Its database is of type NoSql.
- **Kibana** is a powerful and intuitive visualization tool for data found in elasticsearch.

  ELK alone cannot ensure the security incidents detection, for this reason, we have found it useful to integrate other tools for incident detection, integrity checking, registry monitoring and detection, and intrusion prevention. For each feature we looked for, we chose the most powerful open source tool on the market. These tools are as follows:
- **Snort** is an open source network intrusion detection and prevention system (NIDS and NIPS) maintained by CISCO [22].
- **Sguil** is an open source heavyweight application that provides an intuitive graphical interface for visualizing events, session data, and raw packet captures. It was created by Network Security Analysts [23].
- **Zeek** is a very powerful network analysis framework that is much different from the typical IDS. It provides a complete platform for more network analysis [24].
- **OSSEC** is an intrusion detection system that analyzes logs from remote machines to check the integrity of Windows registers and generates real-time alerts [25].

In order to provide real-time monitoring, a load balancing and queue management tool are essential to not lose JSON objects in case of congestion. To do this, we used Redis which is an open source, in-memory data structure store, used as a database, cache and message broker [26].

We have implemented all the aforementioned tools in a distributed Linux environment with the following technical specifications.

Table 1. Technical specifications

Hardware specifications	**Software specifications**
Master server with 8 CPU cores, 16 GB RAM with 1 TB Hard Drive	VirtualBox 6.0.4
Storage Node with 4 CPU cores, 8 GB RAM and 100 TB Hard Drive	Linux Ubuntu 16.04
Forward node with 2 CPU, 2 GB RAM and 500 Go Hard Drive	Elasticsearch-6.6.1.tar.gz
	Logstash-6.6.1.tar.gz
	Kibana-6.6.1-linux-x86_64.tar.gz
	Java SE 11.0.2 (LTS)
	Redis 5.0.3
	filebeat-6.6.1-linux-x86_64.tar.gz
	packetbeat-6.6.1-linux-x86_64.tar.gz
	winlogbeat-6.6.1-windows-x86_64.tar.gz
	metricbeat-6.6.1-linux-x86_64.tar.gz
	Sguil 0.9.0
	snort-2.9.12.tar.gz
	Kali Linux 1.0.8

To ensure system elasticity, that means a rational and optimal resources allocation, in order to manage the Big Data aspect, we opted for a distributed architecture illustrated in Figure 3 to implement the proposed SIEM.

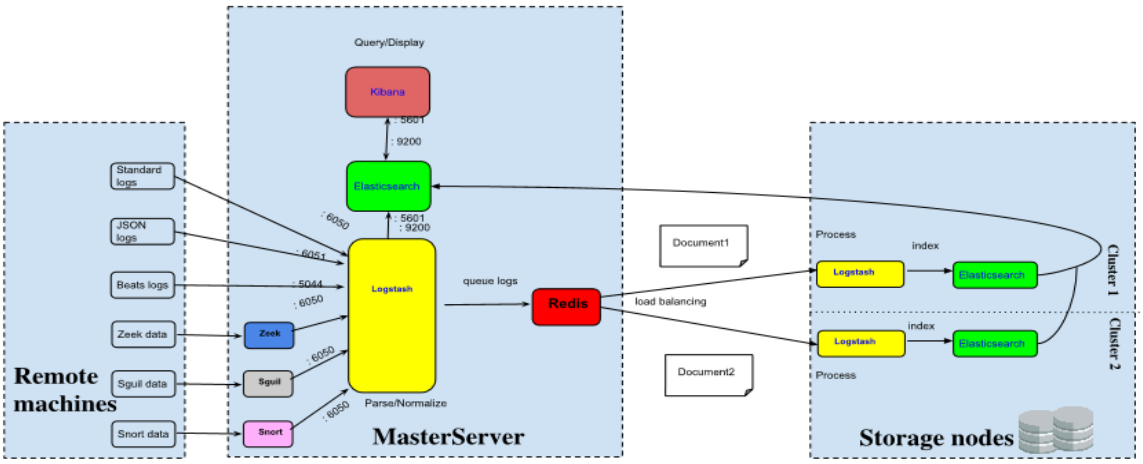

Fig. 3. Proposed SIEM architecture

B. Smart SIEM tests

To test the proposed SIEM functionality, a virtual lab combining a Web server and Windows and Linux workstations was created to provide a technical environment for the validation of our prototype. Using this network laboratory, a series of malicious cyber actions were performed to assess how well our SIEM was detecting and reporting them. To do this, we used a Linux Kali machine [27] which is the next generation of the industry-leading BackTrack Linux distribution for audit, security, and penetration testing.

Figure 4 illustrates the architecture of the proposed prototype test platform.

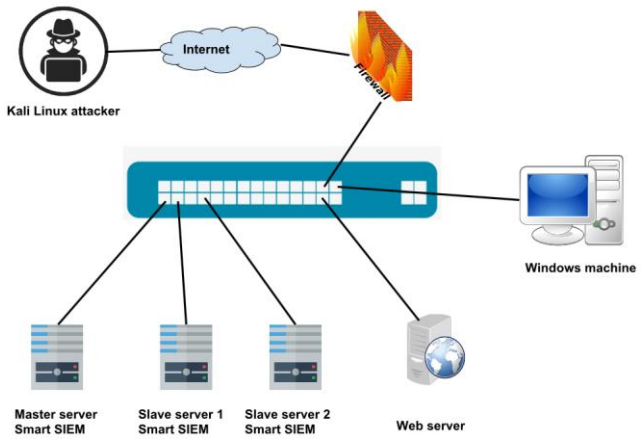

Fig. 4. Smart SIEM test platform

V. RESULTS AND DISCUSSION

In this section, we discuss the results of the tests performed. The first test was to check if all the tools we have integrated on the Big Data platform ELK are working properly. To do this, we checked the service status of all Smart SIEM components on the Master server. Figure 5 shows that all services are operational.

```
================================================================================
Service Status
================================================================================
Status:
  * sguil server[  OK  ]
Status: HIDS
  * ossec_agent (sguil)[  OK  ]
Status: Bro
Name         Type       Host          Status    Pid    Started
bro          standalone localhost     running   3201   25 Mar 15:50:56
Status: sct-virtual-machine-ens34
  * netsniff-ng (full packet data)[  OK  ]
  * pcap_agent (sguil)[  OK  ]
  * snort_agent-1 (sguil)[  OK  ]
  * snort-1 (alert data)[  OK  ]
  * barnyard2-1 (spooler, unified2 format)[  OK  ]
Status: Elastic stack
  * so-elasticsearch[  OK  ]
  * so-logstash[  OK  ]
  * so-kibana[  OK  ]
  * so-freqserver[  OK  ]
  * so-domainstats[  OK  ]
  * so-curator[  OK  ]
```

Fig. 5. Service status of Smart SIEM components

Then, we tested the Smart SIEM ability to detect multiple attacks that are programmed into the Kali machine as scripts. We have chosen a most common attack in cybersecurity which is port scan because attackers often start with this type of attack to enumerate the network resources victim including IP addresses and open ports. An unauthorized scan may indicate the first steps of a cyber-attack. After launching this attack, we found on Kibana as shown in Figure 6 that the number of alerts that is only the sum of NIDS alerts has skyrocketed from 14 alerts at 15:05 to 68 alerts at 15:15. This behavior should be significant enough to prompt SIEM analysts to investigate a possible attack on their network.

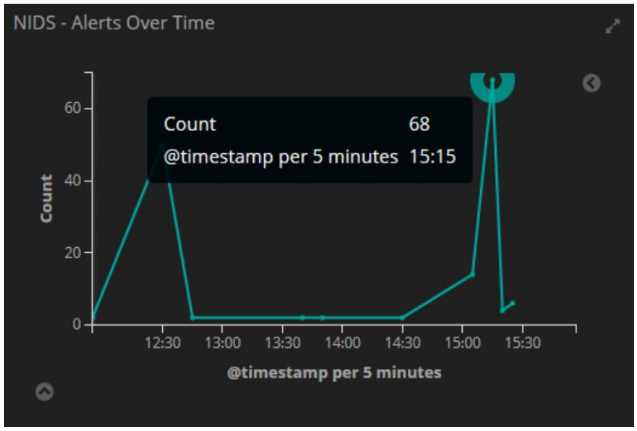

Fig. 6. Alerts after a scan attack

Subsequently, we launched several attacks at the same time to see the behavior of Smart SIEM. Figure 7 illustrates the alerts generated by the proposed SIEM and displayed by Kibana after having configured it to show the attacks type, the IP address of the attacker, and the address of the victim, and finally the number of alerts generated.

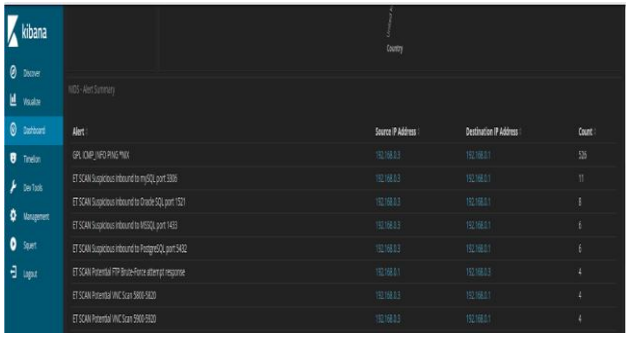

Fig. 7. Alerts generated as a result of multiple attacks

Our SIEM responds perfectly to the attacks made by the Kali machine. Indeed, it has detected SQL command injection attacks for MySQL, PostgreSQL, and Oracle database management systems. Also, it has raised alerts for brute force attacks and finally, DDos attacks by sending the web server a huge amount (528) of ICMP packets (ping). Finally, we were able, through Smart SIEM, to identify the target and the attacker machines.

VI. CONCLUSION AND FUTURE WORKS

In this article, we have proposed a new generation SIEM architecture that takes into account the traditional SIEMs limitations in terms of Big Data management. The designed prototype named Smart SIEM and made up of a Big Data platform ELK with other intrusion detection and load balancing tools could manage large heterogeneous networks composed of several devices with a log files centralization in order to make detection and analysis faster and efficient.

The proposed SIEM provides aggregation and normalization functions for log files collected from multiple sources to provide network analysts with rich information alerts via an efficient and easy-to-use interface.

The proposed prototype has been tested in a virtual environment composed of linux and windows devices to validate its operation first and then check its behavior against the most common attack scenarios. The results were challenging. Thanks to Smart SIEM, an analyst can not only monitor the cyber activity of his network but also reduce the lifecycle of the detection until the resolution of the detected incidents which increases the efficiency of monitoring level and the malicious activities detection.

Future works will be done to integrate the proposed prototype into a real production environment of a government defense agency.

REFERENCES

1. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, 'An evaluation framework for network security visualizations', *Computers & Security*, vol. 84, pp. 70–92, Jul. 2019.
2. Corero, 'DDoS Attacks Increase 40% Year on Year Confirms Corero Networks | Business Wire', 2019. [Online]. Available: https://www.businesswire.com/news/home/20180912005272/en/DDoS-Attacks-Increase-40-Year-Year-Confirms. [Accessed: 09-Apr-2019].
3. E. Al-Shaer, J. Wei, K. W. Hamlen, and C. Wang, 'Towards Intelligent Cyber Deception Systems', in *Autonomous Cyber Deception*, Cham: Springer International Publishing, 2019, pp. 21–33.
4. B. Geluvaraj, P. M. Satwik, and T. A. Ashok Kumar, 'The Future of Cybersecurity: Major Role of Artificial Intelligence, Machine Learning, and Deep Learning in Cyberspace', in *International Conference on Computer Networks and Communication Technologies*, vol. 15, S. Smys, R. Bestak, J. I.-Z. Chen, and I. Kotuliak, Eds. Singapore: Springer Singapore, 2019, pp. 739–747.
5. I. Alsmadi, 'Incident Response', in *The NICE Cyber Security Framework*, Cham: Springer International Publishing, 2019, pp. 331–346.
6. R. Zuech, T. M. Khoshgoftaar, and R. Wald, 'Intrusion detection and Big Heterogeneous Data: a Survey', *Journal of Big Data*, vol. 2, no. 1, Dec. 2015.
7. 'A Case Study In Security Big Data Analysis', 2019. [Online]. Available: https://www.darkreading.com/analytics/security-monitoring/a-case-study-in-security-big-data-analysis/d/d-id/1137299. [Accessed: 23-Mar-2019].
8. IBM, 'QRadar components', 2019. [Online]. Available: https://www.ibm.com/support/knowledgecenter/SS42VS_7.3.1/com.ibm.qradar.doc/c_qradar_comps2_deployment_guide.html. [Accessed: 24-Apr-2019].
9. F. Menges *et al.*, 'Introducing DINGfest: An architecture for next generation SIEM systems', 2018.
10. N. Miloslavskaya and A. Tolstoy, 'New SIEM System for the Internet of Things', in *Dynamic Programming for Impulse Feedback and Fast Controls*, vol. 468, London: Springer London, 2019, pp. 317–327.
11. M. Vielberth and P. Gunther, 'A Security Information and Event Management Pattern', Nov-2018.
12. M. El arass, I. Tikito, and N. Souissi, 'Data lifecycles analysis: towards intelligent cycle', in *Proceeding of The second International Conference on Intelligent Systems and Computer Vision, ISCV'2017, Fès17-19 April*, Fez, Morocco, 2017. https://doi.org/10.1109/ISACV.2017.8054938
13. M. El arass and N. Souissi, 'Data Lifecycle: From Big Data to SmartData', in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, Marrakech, 2018, pp. 80–87. https://doi.org/10.1109/CIST.2018.8596547
14. I. Tikito and N. Souissi, 'Data Collect Requirements Model', in *BDCA'2017*, 2017. https://doi.org/10.1145/3090354.3090358
15. I. Kotenko, A. Kuleshov, and I. Ushakov, 'Aggregation of elastic stack instruments for collecting, storing and processing of security information and events', in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, 2017, pp. 1–8.
16. Kaspersky Lab, 'Report: Measuring the Financial Impact of IT Security on Businesses | Kaspersky Lab official blog', 2019.
17. ISO, 'ISO 9001 Quality management', 2015. [Online]. Available: https://www.iso.org/iso-9001-quality-management.html. [Accessed: 10-Jul-2018].
18. CIGREF, 'CIGREF, "Les référentiels de la DSI : Etat de l'art usage et bonnes pratiques"', 2009. .
19. M. El arass, I. Tikito, and N. Souissi, 'An Audit Framework for Data Lifecycles in a Big Data context', in *2018 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*, Tangier, 2018, pp. 1–5. https://doi.org/10.1109/MoWNet.2018.8428883
20. Elasticsearch, 'Logstash', 2019. [Online]. Available: https://www.elastic.co/fr/products/logstash. [Accessed: 08-Mar-2019].
21. S. J. Son and Y. Kwon, 'Performance of ELK stack and commercial system in security log analysis', in *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*, Johor Bahru, 2017, pp. 187–190.
22. CISCO, 'Snort website', 2019. [Online]. Available: https://snort.org/documents. [Accessed: 08-Mar-2019].
23. Network Security Analyst, 'Sguil - Open Source Network Security Monitoring', 2019. [Online]. Available: http://bammv.github.io/sguil/index.html. [Accessed: 08-Mar-2019].
24. Zeek, 'The Zeek Network Security Monitor', 2019. [Online]. Available: https://www.zeek.org/. [Accessed: 09-Mar-2019].
25. OSSEC Foundation, 'Home — OSSEC', 2019. [Online]. Available: https://www.ossec.net/. [Accessed: 11-Apr-2019].
26. Redislabs, 'Redis', 2019. [Online]. Available: https://redis.io/. [Accessed: 08-Mar-2019].
27. Offensive Security, 'Kali Linux | Penetration Testing and Ethical Hacking Linux Distribution', 2019. [Online]. Available: https://www.kali.org/. [Accessed: 12-Apr-2019].

AUTHORS PROFILE

Mohammed EL ARASS is a Cyber Security Project Manager in Moroccan defense agency and member of Information System and WEB (SIWEB) in EMI School. He received his engineering degree from INSA Lyon in 2017. His research interests include Data lifecycle, Big Data management, System of Systems, and Cybersecurity.

Nissrine SOUISSI is a fulltime professor at the MINES-RABAT School, Morocco. She obtained a Ph.D. in computer science from the UPEC University in 2006, France and an Engineer degree from Mohammadia School of Engineers in 2001, Morocco. Her research interests include process engineering, business process management, databases, data lifecycle, smart data, hospital information system, and information system.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication