International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

A Novel Security Architecture Based on Haystack System for HDFS Storage System: Extended Work

I. Kandrouch, N. Hmina, H. Chaoui

Abstract: ***Over the past few years, the digitization of our different daily transactions and our activities have helped to create the new Big Data era. This phenomenon has become more widely used in different fields for many reasons, such as information retrieval, decision making, etc. But keeping this data safe from any attack attempts constitutes a real challenge. Indeed, Big Data security has become very important recently, since all technologies and large organizations have become dependent on these Data. For this reason, we deal with the Big Data security problems, especially those encountered during their management and treatment, and more precisely, those appeared during the processing with Hadoop system. The paper aims to highlight the Big Data system processing situation in terms of security, also it shed light on some existing security solutions used for overcoming these security issues, we will reveal its strengths and its weaknesses according to the security pillars. Subsequently, we will present our new security architecture, based on Haystack intrusion detection system, which serves to improve the inactive data security contained in HDFS, while analyzing its feasibility and its parameters***

Keywords: Big Data, Hadoop, Haystack ,HDFS, Security.

I. INTRODUCTION

Computer technologies use has today produced a technological revolution, which produced accordingly a large amount of data [1]. Indeed, data generation recently exceeds the limits of the digital universe. Following this technological revolution, working on Big Data [2] and its processing have become increasingly important for new sciences and companies[3] ... etc. In fact, Big Data is the collection of data that is difficult to collect, store, and control it securely during its processing[4] [5] . The emergence of this new Big Data phenomenon is accompanied by the emergence and / or evolution of systems that able to process these data, but unfortunately, the treatment is not done in a completely secure way, and that because several vulnerabilities that hampers the proper functioning of these systems attacks it over the years.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Ibtissame Kandrouch*, Department of Computer Science, National School of Applied Sciences, Ibn Tofail University, Morocco.
Nabil Hmina, Department of Physics, Mohammed V University, Morocco.
Habiba Chaoui, Nabil Hmina, Department of Physics, Mohammed V University, Morocco.

Apache Hadoop is now the leader in Big Data processing systems [6], but also, this system is not powerful enough in terms of data security. Stopping at these problems, led us to an efficient use of Big Data via Hadoop. For this reason, we will study in this article first, the security degree of this data, with the various concerns 'especially security problems' can arise during its entire life cycle

Then, we will present a statistical study result, showing the number of detected vulnerabilities that attack data management systems, in order to reveal the critical situation of these data during their processing.

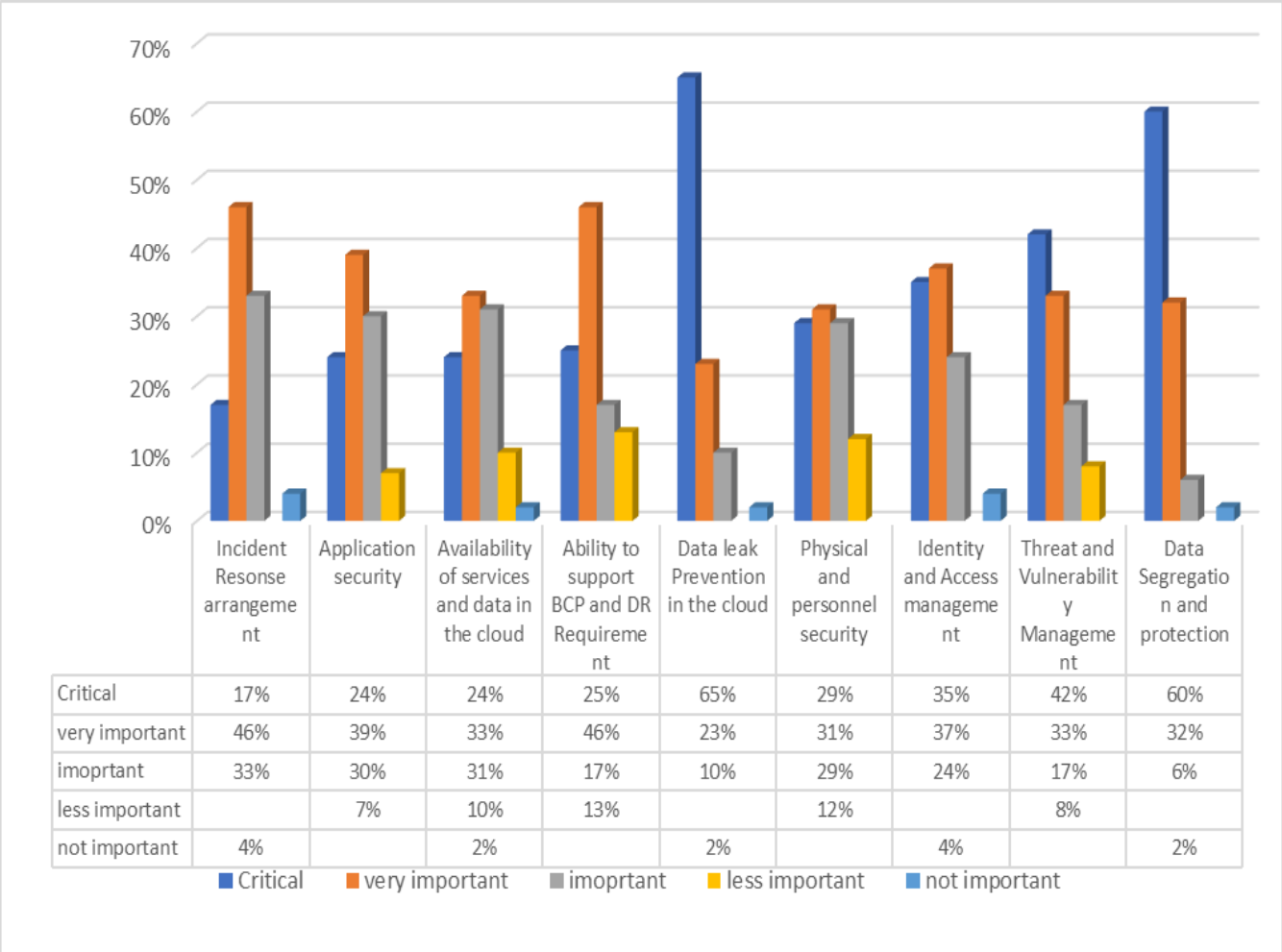
The third part is devoted to studying Big Data management system, Hadoop, including its distributed file storage system HDFS. This study allows us to understand how Hadoop stores and processes a large dataset in a distributed and scalable way. We will then identify the various security problems encountered when processing data by this system and highlight some of existing security technics with their strengths and weaknesses, also the problems encountered when adopting these solutions.

The presentation of our new architecture based on the haystack intrusion detector, the study of its feasibility, its characteristics and its requirements will be the objective of the fourth part.

II. BIG DATA SECURITY CHALLENGES

(This part develops the previous study presented in [11]) It is very remarkable that the growth of this type of data does not stop increasing over time, which pushes to focus on the security and confidentiality of this data. Big Data can be considered both an opportunity and a challenge. Indeed, Big Data technologies does not only offer unprecedented potential in detecting and predicting threats, but also a major challenge.[7]. The real challenge appears when it comes to loss, flight, and the various attacks, which can have a very negative impact on the parties involved.

The following figure shows that the most important factor with critical and very important challenges is the prevention of data leaks with a percentage of 88%. Also, 92% of security problems comes down to segregation and data protection. [8] [9].



	Incident Resonse arrangement	Application security	Availability of services and data in the cloud	Ability to support BCP and DR Requirement	Data leak Prevention in the cloud	Physical and personnel security	Identity and Access management	Threat and Vulnerability Management	Data Segregation and protection
Critical	17%	24%	24%	25%	65%	29%	35%	42%	60%
very important	46%	39%	33%	46%	23%	31%	37%	33%	32%
imoprtant	33%	30%	31%	17%	10%	29%	24%	17%	6%
less important		7%	10%	13%		12%		8%	
not important	4%		2%		2%		4%		2%

Fig. 1: Big Data Security Challenges

Figure 1 shows that a good security measure will be more effective if it implements challenges and solutions. So it's very obvious that the important and critical things to consider are the data leakage and also the security solutions.[9]

A. Security challenges during the Big Data life cycle

(This part develops the previous study presented in [9]) Securing Big Data involves the establishment of security mechanisms throughout the life cycle of Big Data.[10].
In the rest of the paper, we will raise the different problems found at each stage of the Big Data life cycle.
Figure 2 showing the life cycle of a typical Big Data platform

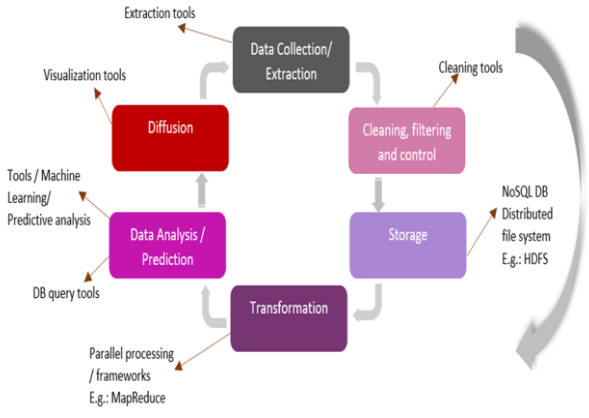


Fig. 2: Big Data life cycle

B. Security challenges during collection

Big Data processing is provided by non-traditional data management systems, designed especially for storing, cleaning, filtering, and for performing other operations on data, while ensuring data quality, persistence, and confidentiality as well.

The risks of compromising data and computer systems based on availability, integrity and confidentiality are the security challenges of these systems. In fact, their evolution over the years is followed at the same time by many security problems, a simple attack can scan such a system for searching vulnerabilities, and it became infected one they are exploited, disrupting all Big Data processing phases.
Figure 3 is a graph showing the number of vulnerabilities exploited in big data management systems over the years.

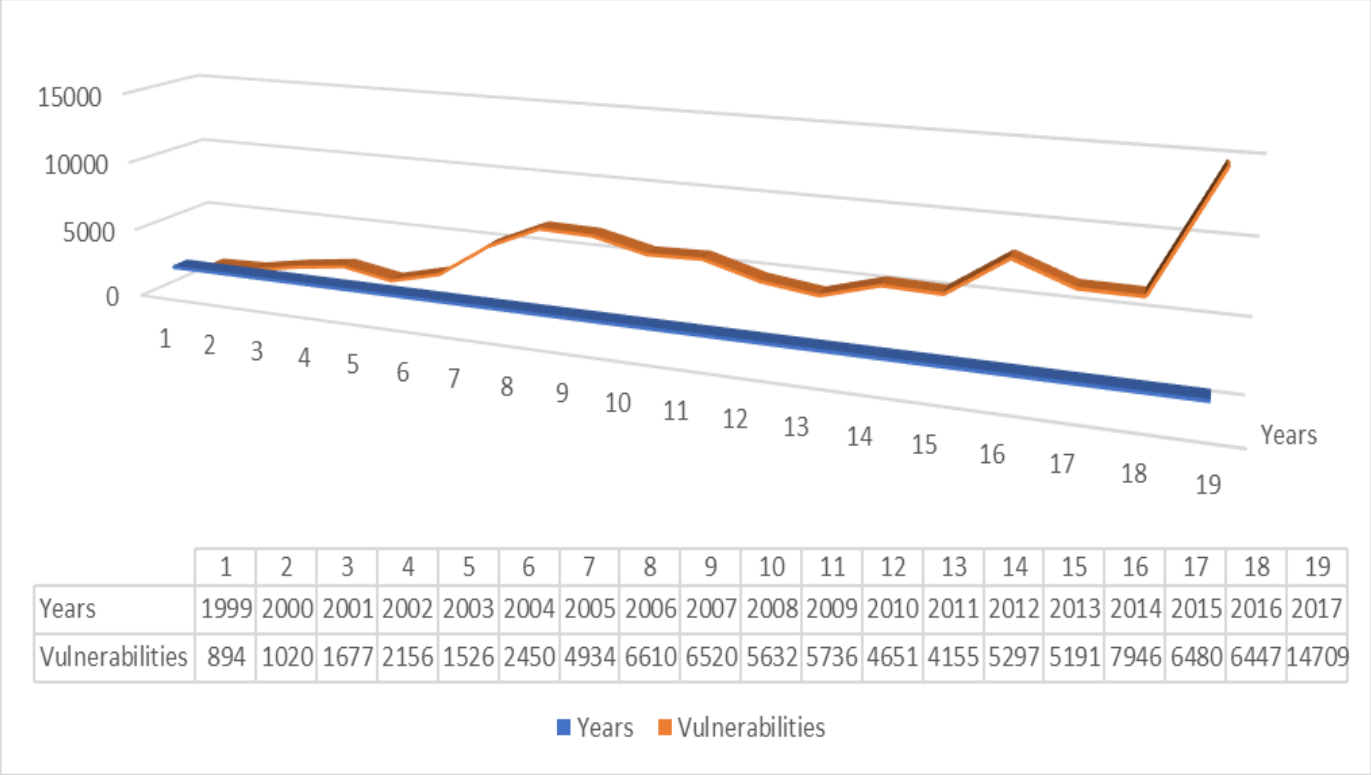
C. Big Data processing systems vulnerabilities over the years

vulnerability of such a data management system designates the degree to which the owners of the data (whether physical or organizations) are likely to suffer leaks and damage due to the exploitation of the detected vulnerability

To carry out an assessment of a vulnerability and the capabilities of a group of threats, it will be necessary first to define the number of vulnerabilities detected over the years by

Big Data management systems[9].

According to studies interested in data security, such as the annual report of cisco and cyber Security [12] , OWASP [13], and other studies work on the situation of Big Data, as well as to the security situation of systems allowing Big Data treatment, the annual Big Data guides as example, we have been able to make a graph showing the evolution of Big Data management systems vulnerabilities over the years, from 1999, until 2017[9].



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Years	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vulnerabilities	894	1020	1677	2156	1526	2450	4934	6610	6520	5632	5736	4651	4155	5297	5191	7946	6480	6447	14709

Fig. 3: Big Data processing systems vulnerabilities over the years

(This part develops the previous study presented in [9])According to the results obtained previously, vulnerabilities number remains high, and their exploitation, as described by the first 10 OWASP [13] aims to undermine the integrity or the availability of the information system with the possibility of viral contamination.

The number of vulnerabilities progresses in an exponential way until the years 2006 and 2017.

With 6447 vulnerability detected during 2016, we can say that the situation of data management systems still critical. This may be related to several factors, such as the lack of an integrated protection system, or the unavailability if the staff members. Detection, Analysis and Security Alerts Resolution in Data Management Systems[9].

In the rest, we will use the Hadoop Big Data Management System, as a system of study and implementation of our solution, not only because it is the most used system for managing massive data, but also because it is invented without any security layer [2], even if it considers the leader compared to other systems of management and Big Data treatment[9]

III. HADOOP BIG DATA MANAGEMENT SYSTEM

Hadoop (Highly Archived Distributed Object-Oriented Programming), a free and open source Java-based programming framework. It's a reference platform supporting data analysis, storage and Big Data manipulation.

Hadoop is based on Computing grids principal, which distributes the execution on several nodes or clusters of servers. This operation principle consists in dividing data into several parts, each piece of data must be stored in a cluster of different servers [2], this distribution makes it possible to distribute tasks on several clusters on which all data are distributed, instead of backing all the processing on a single cluster, actually it's the load balancing principle. Hadoop framework consists of several components; the main ones are the master nodes, and the slave nodes.

A.Hadoop components

Hadoop consists mainly of the following modules; these modules will be presented in table 1 below:

Table- I: Hadoop modules

Modules	Meaning
Hadoop Distributed File System (HDFS)	Distributed File Management System allows data storage on machines.
Hadoop Common	Refers to the collection of utilities and libraries supporting other Hadoop modules.
Hadoop Yarn (Yet Another Negotiator)	Cluster Management Technology
Hadoop MapReduce	Programming Model for Large-Scale Data Processing

Hadoop Distributed File System (HDFS) is a portable Java file system. It has many similarities with existing distributed file systems, suitable for applications with large datasets, and provides redundant storage for these amounts of data with low latency where performs read or write operations [15] [2]. It operates by fragmenting data into blocks with default sizes, and stored it in data nodes, the metadata of this data is stored in the name node.

Unauthorized access	Hadoop makes sure to authenticate any user or service (the unauthorized may become owners of the services via RPC)
Unauthorized data modification	Data Nodes have no access control mechanisms for protecting data blocks. And thereafter, possibility to modify Data Nodes Data
Denial of service or resource	An attacker can be present as a Hadoop service. A run of MapReduce can use the host Operating system interfaces.

bellow illustrate the HDFS architecture.

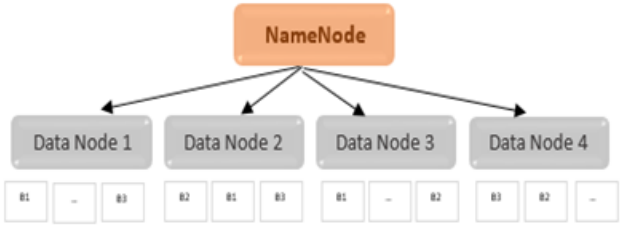


Fig. 4: HDFS file storage system architecture

B. 3.2 Hadoop security issues

Hadoop encounters several security problems, firstly at the permissions level:
Hadoop uses the UNIX 'whichi' and 'bash -c groups' utilities for individual users and groups respectively, and this is a failure because the permissions and the quota file are clients [16]. Regarding, HDFS confronts many security breaches can be defined in three types:

- Hadoop is based on data fragmentation, and writing on Data Nodes principle, which can be cascaded by copying into other Data Nodes. The copying of this data over several nodes aimed ensuring the data redundancy, and ensuring the information availability, but in return, more complexity will be added to these fragments of data, which presents several security issues.
- Data processing done where data resources are available, this process creates a new complexes environment that are vulnerable to attack [1].
- Communication between nodes is done through protocols providing remote communication, which doesn't implement a secure communication.
- Data Access control in Hadoop Framework based on the Role-Based Access Control (RBAC), this type of access control is essential for most databases, but implicitly assumes roles definition, assignment of role permissions, and user's role distribution are performed by a central authority [5] .

C. Hadoop security existing solutions

At the beginning Hadoop did not developed with a security layer, which means that there was no security or authentication model for users or services [2]. While Big Data security is a major concern for large companies, since every piece of technology today generates an impressive amount of data to be processed every minute [18][19].It is for this reason that many efforts have been made to protect environments and improve their security. So, we did an overview on several techniques, open source, commercial and other technologies, to meet this need [14].

- **Kerberos**

The purpose of authentication is identifying users. Hadoop provides Kerberos system as a first authentication. Kerberos [20]is a network authentication protocol, which allows node to transfer data via secure channel, with a tool called "ticket" to prove their unique identification. When authenticating via Kerberos, A simple Authentication Security Layer (SASL) can be added. During operation, SASL uses two ID to identify a user: the first authentication ID is used to give users access to the system. The second one is ID authorization, that allows checking which option (role or action) the user must access the system.

- **Bull Eye Approach**

The "Bull Eye Approach" approach which mainly aims at securing sensitive information is also used in the HDFS part in order to guarantee security in a node.
this approach is used in the rack node 1, in order to verify whether the sensitive data is securely stored.[2].

- **Name Node Approach**

In the existence of such a problem in the HDFS part, the name node becomes unavailable, which makes the group of services and the stored data also unavailable, therefore, it will be difficult to access the data of securely. therefore, it is necessary to use two name nodes. Both Node Name servers can run successfully in the same cluster. Both are provided by a redundant Name Node Security Enhance (NNSE), which holds Bull Eye Algorithm. It allows administrator to execute Hadoop options for both nodes [2].

- **SDFS**

SDFS, Secure Distributed File System for Data-at-Rest employs a combination of techniques to satisfy the aforementioned requirements and provide both solid security and high information accessibility at a lower stockpiling cost contrasted with what is given by HDFS with its standard replication and encryption systems [21].

- **ACLs**

Authorization is a process allowing authorize or deny information access for a user or a task. Hadoop implements an authorization type like the model used in UNIX, that of ACL, the HDFS has this type of permission (dynamic and static ACL). It has the advantage of extending permissions to several group of users [22]. MapReduce have permission checked by an ACL, this allows to filter users or groups who will have the right to submit a request in the queue or even modify tasks properties. There are also permissions to access HBase data through ACLs.

- **Randomized Encryption on Hadoop ecosystem**

To make Hadoop a secure system, the randomized encryption mechanism is implemented. The processing part in Hadoop system, MapReduce, used randomized encryption techniques. MapReduce will break the decryption / encryption process and speed up the process to ensure that system performance or scalability isn't affected. Several encryption techniques are implemented assuming that if a hacker succeeds in compromising some data, he will not be able to access sensitive data.

- **Anonymization**

Collecting data for analysis causes is a big privacy issue. The protection of Personal Identifiable Information (PII) is increasingly difficult because data is shared too quickly. Personal data must be anonymized and transferred into secure channels. However, person's identities can be discovered, if we have based on Artificial intelligence algorithms and analysis. The prediction made by this analysis can lead to ethical problems. Several anonymization tools are available, such as Comell Anonymization Toolkit [23]and ARX Toolkit [24], which work with Big Data tools like Hadoop system.

- **Apache Sentry**

Apache sentry is a Cloudera open source project; it is produced in the form of an authorization module for the Hadoop system. This project offers granular authorization and role-based authorization [25].

- **Apache Knox**

Apache Knox framework is a system for providing authentication and access for multiple services in Hadoop cluster [26]. It is a perimeter security solution, supporting multiple authentication verification scenarios, and cluster security management.

- **Project Rhino**

Project Rhino is a commercial project that provides an integrated security solution for Hadoop [26]. It offers an authentication, an SSO solution [20], and a cryptographic encoding for data stored as a block in Hadoop. Project Rhino also improves security provided by the HBase module by providing cell-level authentication, and encryption of stored tables [26] [25].

The table 2 below summarize our studies and present the different solutions in terms of the main security pillars.

Table- II: Existing security solutions across security pillars

Studies	Security policy	Confidentiality	Integrity	Availability	Authentication	Non-repudiation	Authorization	Scalability	Extensibility
[26] [21]	Randomized Encryption	V	--	V	--	V	--	--	--
[27] [28]	Anonymization	--	V	--	--	--	--	V	--
[26] [22]	Kerberos	--	V	--	V	--	--	--	V
[26]	ACL	--	--	--	--	V	V	--	--
[27]	Name Node Approach	--	--	V	V	--	--	--	--
[21]	SDFS	--	V	V	--	--	--	V	V
[26] [25]	Apache Sentry	--	--	--	--	--	V	--	--
[26]	Apache Knox	--	--	--	V	--	--	--	--
[25]	Project Rhino	--	--	--	--	--	--	--	--

The fundamental principles of security (Confidentiality, integrity and availability) apply to most systems. The security architecture must match security controls and security requirements that are sometimes dictated by the need to ensure the other three basic elements (scalability, non-repudiation and reliability), more to the other two inseparables: authentication and authorization.

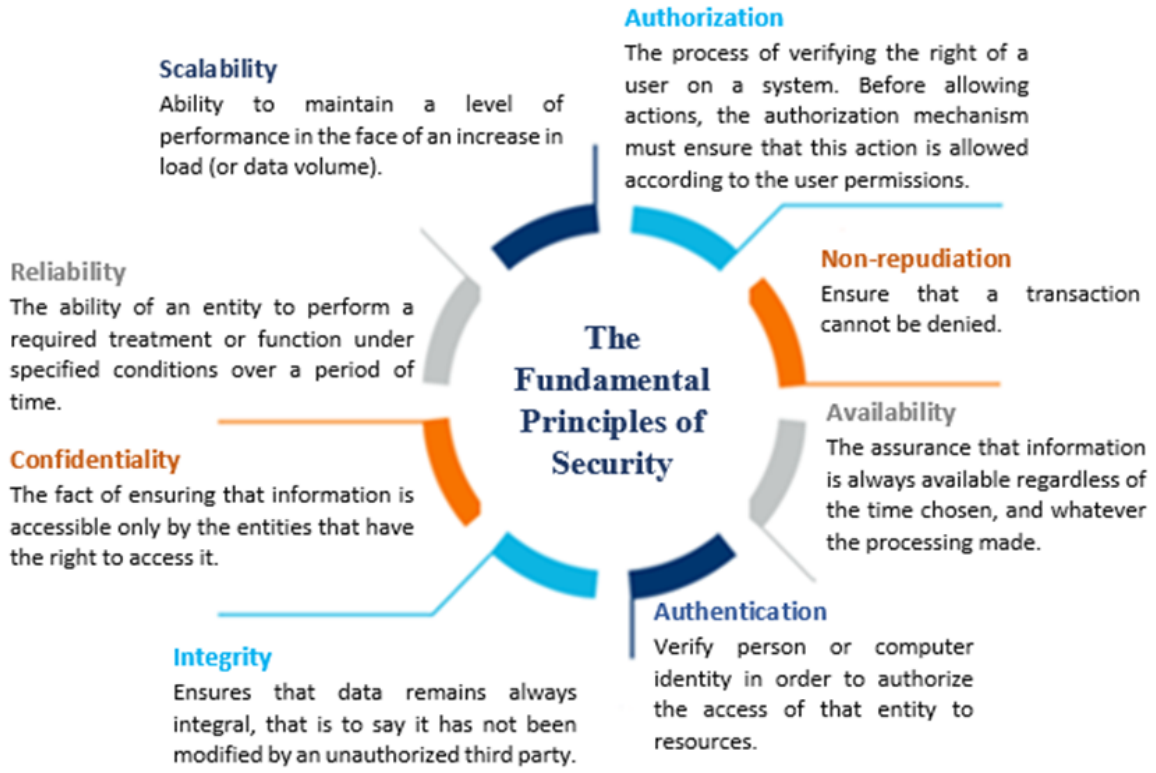


Fig. 5: The fundamental principles of security[9]

The various solutions recently cited are today the most adapted when handling Big Data. Through the table above, we realized a detailed analysis of mechanisms and security solutions for Big Data, mainly those used in the HDFS storage system, and we saw that Hadoop security is not a very strong, the reason why we proposed a new security architecture, aiming respond to the shortcomings of the solutions mentioned above.

IV. THE PROPOSED METHODOLOGY: THE NEW SECURITY ARCHITECTURE

Users interact directly with Hadoop, and therefore, in order to detect any abnormal behavior, it is more desirable to use intrusion detector systems. These Systems are one of the effective solutions to tackle the problems of system disruption.

There are many types of intrusion detection systems. An Intrusion Detection System can be classified as Host-Based Intrusion Detection System (HIDS), Network Based Intrusion Detection System (NIDS), and Distributed Intrusion Detection System (DIDS) [29].

A. Intrusion detector motivation

Data security is an approach that aimed firstly creating a completely safe and secure system against various violations, which is in the most cases cannot be possible. Indeed, several challenges do not make a secure information system because firstly the system software complexity, either in terms of its internal design, or the way in which it performs and executes its data processing

Thus, a system often contains security vulnerabilities that can cause major security problems, that can lead sometimes to the non-upgrading systems, which may introduce new problems.

As a response to these security problems aiming hinder the proper functioning of the systems, a new approach to detection, analysis, and reaction against intrusions has appeared. Intrusion detection systems are the appropriate systems that monitor systems looking for malicious activities [27], to prevent and detect security breaches, it allows automating the analysis process of events in order to detect security problems intended to compromise the proper functioning of the system.

B. IDS and Big Data

Intrusion detection often involves analysing data, especially large or Big Data. This database type always defined as a source of problems, especially when traditional data management systems are no longer able to handle these huge amounts of data.

Detecting and analysing the security of heterogeneous, voluminous, and diverse data can be very difficult. Indeed, the integration of a tool in order to automate data analysis process, and detecting malicious behaviour can increase the complexity of systems in terms of volume, in the fact of having to store a large amount of data [31], and in particular their variety and velocity.

Reinforcing the security level offered by big data management system using an intrusion detector means choosing the system first in order to be able to store and process data reliably, especially in front of the 5V Big Data characteristics. The intrusion detector also must handle data heterogeneity and diversity for being able to provide better results.

Figure 7 below shows the software architecture of an intrusion detector, which describes in a symbolic and schematic way the different elements of the system.

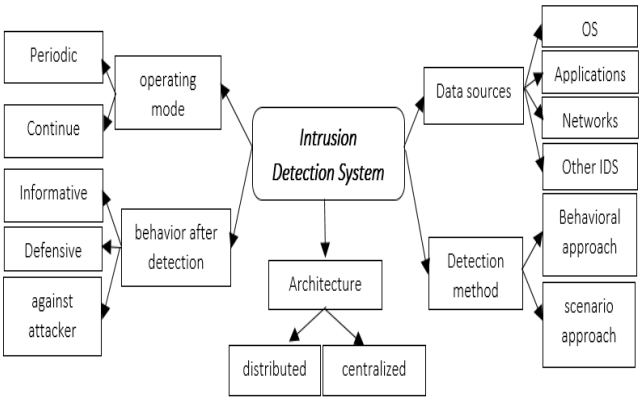


Fig. 6: Intrusion Detection System Architecture

Table-III: Intrusion Detection System Types

	HIDS	NIDS	DIDS
Characteristics	Have sensors, which focuses only on single host for intrusion detection. A HIDS monitors the incoming and outgoing Packets from the host and alerts the user or administrator of suspicious activity if detected any.	Have sensors which detect the Intrusions over the network. NIDS are placed at a strategic. Point or points within the network to monitor incoming and outgoing traffic of all devices on the network.	Integrates both types of sensors. DIDS consists of many IDS scattered over the large network. These IDS are arranged in the network in such a way that they can communicate with each, or are connected to the central server.
Fact	The operation of this IDS depends on information collected from logs; It's dependent on operating system of machines; It can operate even in encrypted environment.	NIDS is a dedicated hardware or software; Network analyzing network traffic; The operation depends on information collected through various sensors; It consists of single purpose sensors; NIDS have a very less impact over the performance of network; It does not have any kind of dependency on operating System.	The process of implementing DIDS is lengthy; It's very difficult to maintain liaison between large number monitors; The process of DIDS is hierarchal.

The figure 8 below shows the IDSs classification, thus the characteristics of each category.

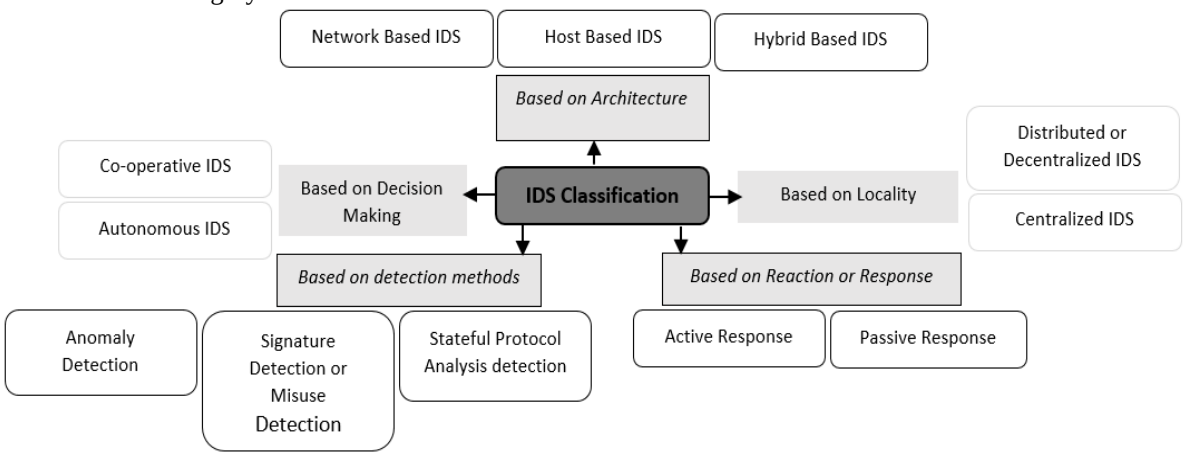


Fig. 7: Intrusion Detection System Architecture

Several factors can create differences between these systems, and classify them in terms of analysis performance, protection, ease of implementation and others ..., the table 4 bellow compares these three types of systems based on these factors.

Table- IV: Comparison between IDS

	HIDS	NIDS	DIDS
Analysis	Analysis logs & alerts host machine only	Analysis a network traffic directly	Interactive querying of data for analysis using aggregation
Protection	Protects even when turned off	Do not protect when turned off	Provides complete protection
Versatility	Highly versatile	Comparatively less versatile	Least versatile
Affordability	Low cost	Average Cost	High Cost
Ease of implementation	Easy	Easy	Difficult
Training	Requires minimum training	Requires certain training	Requires intense training
Bandwidth Requirement	No requirements of bandwidth	Utilizes LAN bandwidth	NIDS components utilizes Bandwidth
Pros	-High success rate; -Real time detection attacks; -No requirement of additional hardware.	-Real time detection; -Detects attacks remain undetected by HIDS;	-Works over extensively large network; -Can be implemented over any network types.
Cons	-Works for single system only; -Network level threats are not resolved.	-High ownership cost; Cannot detect encrypted attacks.	-Very High Cost; -Complex implementation.

Table- V: Interpretation and results

conclusion of the study carried out on ids (Table- IV)	• Intrusion detection systems are the solutions based on either network or host. • Network-based intrusion detection systems (NIDS) generally have sensors located at several points along network, for analyzing data packets entering and leaving network. • Host-based intrusion detection systems allow analysis of traffic to and from the machine on which the IDS is installed. This category of IDS provide functionality that cannot be obtained with NIDS. • The HIDS can monitor changes made in a system file, so it is possible to monitor the various attempts to replace these files, those also aimed at the trojans installation, or backdoors are also supervised and can also be stopped. • All these operations are provided by the HIDS, not the NIDS. • The comparison also made show that the HIDS are very easy for implementation compared to other systems types, is also characterized by a minimal cost, high versatility, and requires minimal training. • And even the disadvantages presented does not affect our purpose of use, (they are usable just if it is a single system, which perfectly our case. Also, the proposed architecture is not used at network level).
It is obvious to use an intrusion detector of HIDS category considering these characteristics. We have chosen Haystack system; this intrusion detector is designed to detect mainly the following anomalies [24].	
The proposed architecture will react:	1. When an unauthorized user attempts to enter the system, or a user without any role; 2. against many attacks such as those aimed at erasing files; 3. When such a user tries to access the system by the identity or role of another user; 4. When such a user wants to block access to system resources to other users; 5. When someone tries to change the system security configuration.

The following table shows the different characteristics of the two systems, Hadoop, and Haystack IDS.

Through this table, it's clear that there are several commonalities between the two systems, specifically the scalability of Hadoop system, and the possibility of adding new frameworks. Also, the access type for both systems, and their categories. These points are the bases that justify our choice of the haystack intrusion detection system.

Bellow, the table 5 present a comparative study between Hadoop and Haystack system.

Table- VI: Comparative study between Hadoop and Haystack

Hadoop	Hadoop is a multi-user data processing system [29] [2]	Configured to support RBAC access control type [29] [28].	Lack of a security standard, that imposes customization (i.e. there is not a single security management standard, the security management options are numerous with Hadoop.) [29] [2].	Hadoop also makes it possible to deploy a multi-tenancy architecture (multitenant architecture i.e. an architecture allowing software to be used by several departments of the company from a totally secure single installation)
Haystack	Haystack is a prototype system for intrusion detection in a multi- user computer system So, it is well suited for a multi-user system like Hadoop.	Haystack works when an authorized user tries to take the identity or role of another user [2]. Check on the role of each user.	there is no security standard for Hadoop, and so, adding Haystack intrusion detector does not pose any problem.	----------

4.3 Security architecture based on Haystack system

In the following, we will present our new architecture intended for stored data at rest, we will thereafter describe its operation and give the necessary arguments that prove both its feasibility and its reliability.

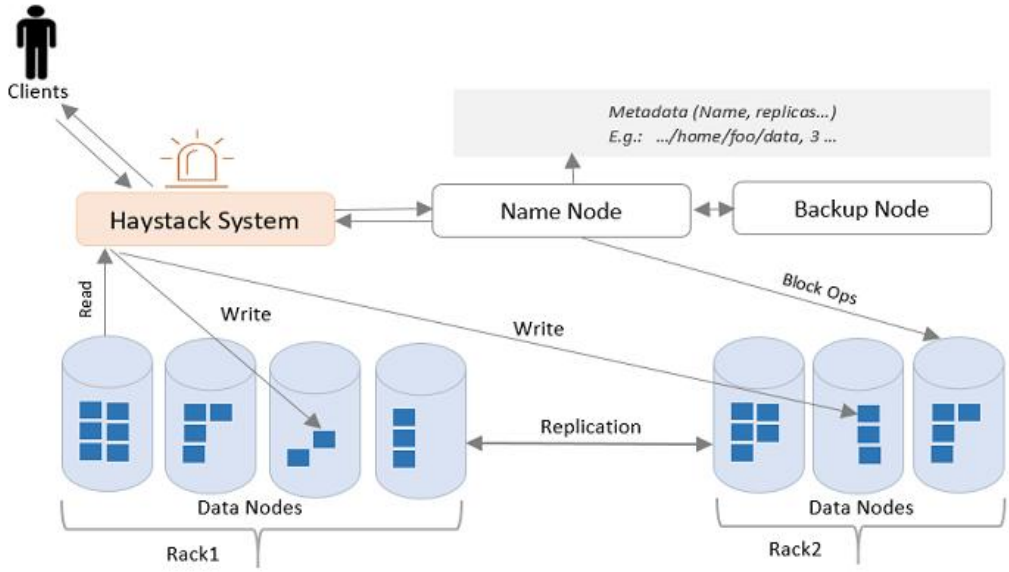


Fig. 8: Security architecture based on Haystack system.

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

1. Haystack IDS;
2. Name Node;
3. Data Node.

When a user solicits Hadoop to recover files:

Haystack Intrusion Detector checks first if the user is allowed access to the system.

If this is the case, it allows the user to continue his request after a second check on the requested service according the user's role. The services requested later will be retrieved via the name Node. This Name Node will tell the user what Data nodes contain blocks requested. The Name Node regularly receives a "heartbeat" and a Block report of all 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 manage the data traffic of data nodes.

4.4 Read and write an HDFS file

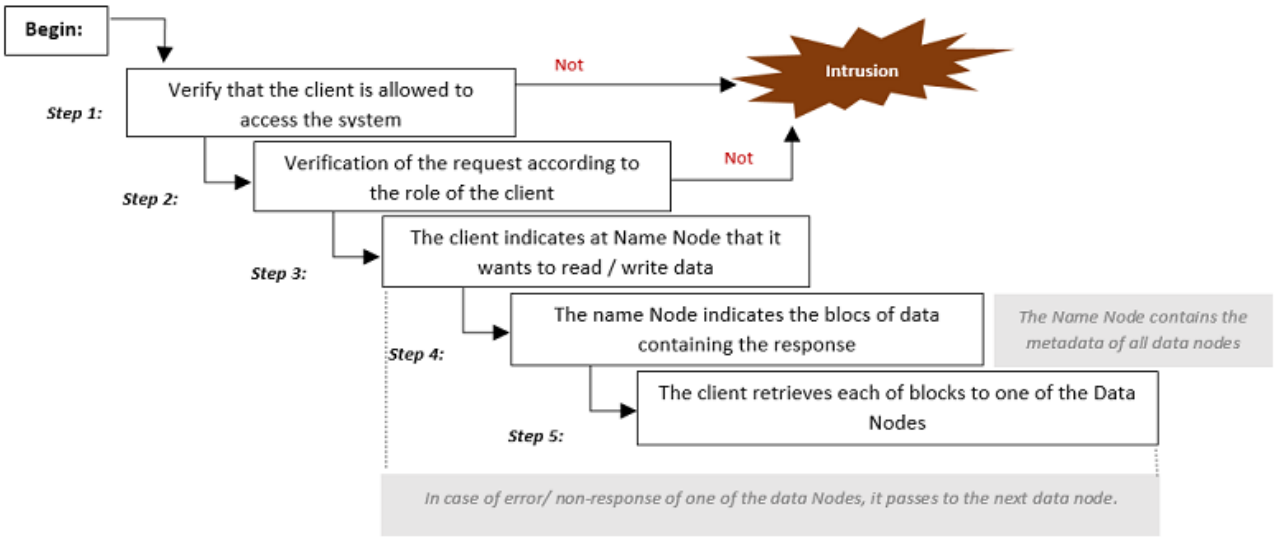


Fig. 9: Read/ write an HDFS file.

4.5 Result Analysis

Based on the analytical study above on power rating security of the new proposed architecture, according to the characteristics of the two systems which compose it, HDFS storage system, and the Haystack IDS. We can deduce the following results as table VII below shows.

Table- VII: Comparative study between Hadoop and Haystack

	Operation in Encrypted Environments	Monitoring	Real time detection attacks	Versatility	Ease of use	Investigation	recovery
HDFS with the use of haystack IDS	Yes	yes	Yes	Yes (High)	Yes	Yes	yes
Using HDFS alone	No	No	No	Yes (Low)	Yes	No	No

With this new architecture proposed, we notice that the security incidents response time will be reduced considerably, with the possibility of working on encrypted environments, while ensuring high versatility, recovery and monitoring.

4.6 The new architecture requirements

The proposed architecture must meet the specified requirements; it must also meet other needs. Indeed, several elements will motivate these requirements:

- **Costs and resources**: the cost and resources are required for design, and the architecture construction.
- **Reliability**: how much you can count on the new architecture, and the quality that defines how much the architecture can be relied upon to deliver the defined services, this quality is measured as ensuring that the technology can deliver services. It is the guarantee that technologies can provide services.
- security (privacy, integrity and availability) apply to most systems. The security architecture must match the security controls and security requirements that are sometimes dictated by the need to provide other three basic elements (reliability, performances and cost).
- **Performance**: the usefulness and usability of the system.
- **Legal and regulatory constraints**: related to the technical security controls, access policies and data retention.

V. DISCUSSION

In this paper, we have reviewed apache Hadoop security and its methods, its threats and some methods enhancing its security level. Each solution cited offers advantages, and contains weaknesses, either at the encryption level, the implementation cost, the access control ... etc.

For all these reasons, the new haystack-based security architecture has been put in place. Indeed, this architecture makes it possible to face the problems encountered with the solutions cited above, more to other functionalities like the detection of misuses, and the real time attacks.

In contrast, this architecture is characterized by RBAC access control, one of the major problems with this model is that all users associated with the same role necessarily have the same privileges. This reduces the flexibility of the security policies thus modelled.

VI. CONCLUSIONS

Big Data security needs are in the process of collection, storage or during their transfer.

For this reason, we have tried through this paper to study the data storing process, and the various problems encountered during their storage and manipulation. The answer to these problems is the objective of our proposed new architecture.

In our upcoming work, we will begin by dealing with the problems can be produced by the access control of the architecture,

We will do simulations to prove the feasibility and power of our architecture compared to other solutions. And we will extend the study not only for data at rest, but also for those on treatment.

REFERENCES

1. B. Matturdi, X. Zhou, S. Li, and F. Lin, "Big Data security and privacy: A review," China Communications, vol. 11, no. 14, pp. 135–145, 2014, doi: 10.1109/CC.2014.7085614.
2. B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, "Big Data and Hadoop-a Study in Security Perspective," Procedia Computer Science, vol. 50, pp. 596–601, 2015, doi: 10.1016/j.procs.2015.04.091.
3. A. Atmani, I. Kandrouch, N. Hmina, and H. Chaoui, "Big Data for Internet of Things: A Survey on IoT Frameworks and Platforms," in International Conference on Artificial Intelligence and Symbolic Computation, 2019, pp. 59–67.
4. C. L. Philip Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information Sciences, vol. 275, pp. 314–347, Aug. 2014, doi: 10.1016/j.ins.2014.01.015.
5. "Expanded Top Ten Big Data Security and Privacy Challenges - Cloud Security Alliance : Cloud Security Alliance." [Online]. Available: https://cloudsecurityalliance.org/download/expanded-top-ten-big-data-security-and-privacy-challenges/. [Accessed: 21-Mar-2018].
6. E. Burns, "Quand utiliser Hadoop... et quand s'en passer ?" [Online]. Available: http://www.lemagit.fr/conseil/Quand-utiliser-Hadoop-et-quand-sen-passer. [Accessed: 21-Mar-2018].
7. "Big Data et sécurité : enjeux et opportunités pour les entreprises – Perspectives IT." [Online]. Available: http://perspectives-it.fr/les-entreprises-peinent-a-lutter-contre-les-vols-de-donnees/. [Accessed: 06-Jan-2019].
8. R. V. Rao and K. Selvamani, "Data Security Challenges and Its Solutions in Cloud Computing," Procedia Computer Science, vol. 48, pp. 204–209, Jan. 2015, doi: 10.1016/j.procs.2015.04.171.
9. I. KANDROUCH, C. SAADI, N. HMINA, and H. CHAOUI, "Security Measures Assessment for Big Data Management Systems," in 2019 5th International Conference on Optimization and Applications (ICOA), 2019, pp. 1–5, doi: 10.1109/ICOA.2019.8727699.
10. Z. SABIR and A. AMINE, "Connected Vehicles using NDN for Intelligent Transportation Systems."
11. "La détection d'intrusion : une approche globale / MISC-072 / MISC / Connect - Edition Diamond." [Online]. Available: https://connect.ed-diamond.com/MISC/MISC-072/La-detection-d-intrusion-une-approche-globale. [Accessed: 06-Jan-2019].
12. "Cisco 2017 Midyear Cybersecurity Report," Cisco. [Online]. Available: https://engage2demand.cisco.com/LP=5897. [Accessed: 06-Jan-2019].
13. P. Adluru, S. S. Datla, and X. Zhang, "Hadoop eco system for big data security and privacy," 2015, pp. 1–6, doi: 10.1109/LISAT.2015.7160211.
14. "HDFS Architecture Guide." [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. [Accessed: 21-Mar-2018].
15. P. Vijaykumar Patil and N. Venkateshan, "Review on Big Data Security in Hadoop," International Journal Of Engineering And Computer Science, vol. 3, no. 12, 2014.
16. Q. Shen et al., "SAPSC: Security Architecture of Private Storage Cloud Based on HDFS," 2012, pp. 1292–1297, doi: 10.1109/WAINA.2012.80.
17. M. Weidner, J. Dees, and P. Sanders, "Fast OLAP query execution in main memory on large data in a cluster," 2013, pp. 518–524, doi: 10.1109/BigData.2013.6691616.
18. A. Cuzzocrea, R. Moussa, and G. Xu, "OLAP*: Effectively and Efficiently Supporting Parallel OLAP over Big Data," in Model and Data Engineering, vol. 8216, A. Cuzzocrea and S. Maabout, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 38–49.
19. S. T. F. Al-Janabi and M. A. Rasheed, "Public-Key Cryptography Enabled Kerberos Authentication," 2011, pp. 209–214, doi: 10.1109/DeSE.2011.16.
20. P. Zerfos, H. Yeo, B. D. Paulovicks, and V. Sheinin, "SDFS: Secure distributed file system for data-at-rest security for Hadoop-as-a-service," 2015, pp. 1262–1271, doi: 10.1109/BigData.2015.7363881.
21. V. Guillaume, "Big Data-Hadoop-Sécurité | SUPINFO, École Supérieure d'Informatique." [Online]. Available: http://www.supinfo.com/articles/single/3262-big-data-hadoop-securite. [Accessed: 21-Mar-2018].
22. X. Xiao, G. Wang, and J. Gehrke, "Interactive anonymization of sensitive data," in Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, 2009, pp. 1051–1054.
23. F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn, "Flash: efficient, stable and optimal k-anonymity," in Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), 2012, pp. 708–717.
24. N. Sudheesh, Securing Hadoop: Implement robust end-to-end security for your Hadoop ecosystem. Birmingham, UK.
25. P. P. Sharma and C. P. Navdeti, "Securing big data hadoop: a review of security issues, threats and solution," Int. J. Comput. Sci. Inf. Technol, vol. 5, no. 2, pp. 2126–2131, 2014.
26. D. S. Terzi, R. Terzi, and S. Sagiroglu, "A survey on security and privacy issues in big data," 2015, pp. 202–207, doi: 10.1109/ICITST.2015.7412089.
27. J. Sedayao, R. Bhardwaj, and N. Gorade, "Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues," 2014, pp. 601–607, doi: 10.1109/BigData.Congress.2014.92.
28. N. Kumar and S. Sharma, "Comparative Study of Intrusion Detection Systems in Cloud Computing."
29. C. SAADI and H. CHAOUI, "Proposed security by IDS-AM in Android system," in 2019 5th International Conference on Optimization and Applications (ICOA), 2019, pp. 1–7.

AUTHORS PROFILE

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, information systems, and also cloud environments.

Nabil Hmina, Received Degree in Physics, option Thermodynamics at Mohammed V University (Morocco) in 1989, DEA- Fluid dynamics and transfers, University and Central School of Nantes (France) in 1990. He was Director of the National School of Applied Sciences kenitra, since November 2011 to 2018, and, he's president of the university sultan moulay sliman, since November 2018 to date.

Habiba Chaoui, 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).