IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# HDFS AND THE CIA TRIAD: A NEED FOR COLLABORATION

[1]Aman Preet Singh Bhalla, [1]Post Graduate Student, [1]Master of Cyber Law and Information Security Department, [1]National Law Institute University, Bhopal, India.

***Abstract:*** The triad of confidentiality, integrity, and availability compromises in a significant proportion of data blocks in the HDFS system. The author has adopted a doctrinal approach and selected the diagnostic research methodology in order to ascertain the extent and significance of the combination of Kerberos, Blue-eye, and Name node approaches for protecting each of the three CIA triad of data sets in HDFS in order to protect the privacy of Big Data. The problem of security breach exists, but its solution lacks clarity. A huge volume of sensitive data is being generated by various end points due to technological advancements, and organizations are processing it for their businesses to run profitably. However, due to a lack of technological tools, flaws in policies, and incomplete knowledge of protective measures, the big data is susceptible to security breaches, compromising the business operations. As described in the paper, there are numerous variables that go into offering total security, and while several methods have been created to secure the HDFS in big data, they are not being used efficiently or, even worse, are not being properly integrated. As a result, one of the three key triangle principles of the CIA—confidentiality, integrity, and availability—is compromised. As a result, the author concluded that the combination of Kerberos, Name Node, and Blue Eye will undoubtedly protect each component of the CIA triad in HDFS and suggests combining the three HDFS established approaches to provide CIA triad security that is 360 degrees complete as each approach is analogous to one of the respective CIA triads. The right adjustments must be made to the current technologies and processes in addition to the combination for this mechanism to be sustainable in the future.

***Index Terms*** **– HDFS, CIA triad in HDFS, security issues in Hadoop.**

## I. INTRODUCTION

With the evolution of computing technology, large sums of data are being generated from various end points (sensors, social media, surveillance systems, etc.) and shared over computer networks. After which, it is processed and analysed with some algorithms and techniques (in order to find out useful information from it), which is now popularly known as BIG DATA.

Big data consists of huge datasets associated with 5 V's: *"high volume, velocity, variety, veracity, and value"* [1]. As per IDC, *"The quantity of data to be analysed is expected to double every 2 years."* [2]

With such large amounts of sensitive data, the concern for privacy always comes up, as these confidential datasets can be easily shared between business giants or can be breached by an attacker for monetary profits. So protection and privacy are big concerns with respect to big data, and as huge data grows by volume each day, each minute, every second, so are the concerns on the rise.

There are certain issues in big data with respect to privacy, such as the lack of management, processing, storage, and security techniques. So, based on our research, we discovered that different architectures were incorporated to protect big data from security breaches, such as Hadoop, PbD framework, and EA modelling, but our focus is on the Hadoop Architecture, or more specifically, the Hadoop Distributed File System (HDFS) in the Hadoop Architecture.

Hadoop was created by *"Goug Cutting & Mike Cafarella"*[3] in 2005. *"It is an open-source Java framework technology that helps to store, access, and gain large resources from big data with a low cost of distributed fashion and a high degree of fault tolerance with high scalability."*[4] *Hadoop is made up of HDFS and Map reduce*[5]*, where "HDFS provides data storage and Map Reduce provides data analysis in a clustered environment."* [6]

But there are security issues in HDFS too, which gave us the motivation for research, such as: *"risk of unwanted data access and theft when embedded a data in single Hadoop Environment."* [7] However certain different approaches were developed to handle data sets securely in HDFS as – Kerberos, Blue-eye and Name node. But again, using any one of them will not purely secure our digital assets in big data and there are chances for breaches.

---

[1]Dr. Fatma Mohammed Abdullah, 'Privacy, security and legal challenges in big data' (iaeme.com, 13 December 2019) <https://iaeme.com/MasterAdmin/Journal_uploads/IJCIET/VOLUME_9_ISSUE_13/IJCIET_09_13_167.pdf> accessed 1 August 2022.

[2] International Data Corporation, 'Data Creation and Replication Will Grow at a Faster Rate than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts' (idc.com, 24 March 2021) <https://www.idc.com/getdoc.jsp?containerId=prUS47560321> accessed 29 July 2022.

[3]B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvan, 'Big Data and Hadoop-A Study in Security Perspective' (*sciencedirect.com*, 8 May 2015) <https://www.sciencedirect.com/science/article/pii/S187705091500592X> accessed 20 August 2022.

[4]Rida Qayyum, 'A Roadmap Towards Big Data Opportunities, Emerging Issues and Hadoop as a Solution' (*researchgate.net*, 2 April 2020) <https://www.researchgate.net/publication/343484997_A_Roadmap_Towards_Big_Data_Opportunities_Emerging_Issues_and_Hadoop_as_a_Solution> accessed 16 August 2022.

[5]B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvan (n 3).

[6]ibid.

[7]ibid.

This problem motivated us to perform research by combining all 3 approaches in HDFS and defining it as a best security practise for large data sets and even for the base layer in Hadoop.

## II. Review Of Literature

Dr. Fatma Mohammed Abdullah, "*Privacy, Security And Legal Challenges In Big Data"*[8]-
The researcher introduced the 5V's of big data, its vulnerabilities, existing privacy policies, and concerns with a proposed suggestion to develop some security approaches in big data.
Nick Hajli ,Farid Shirazi , Mina Tajvidi,and Nurul Huda, "*Towards an Understanding of Privacy Management Architecture in Big Data: An Experimental Research"*[9]-
This paper discussed the technological vulnerabilities in the present architecture of big data, leading to security breaches.
José Moura, Carlos Serrão, "*Security and Privacy Issues of Big Data"*[10]-
The author of this research paper briefed us about the challenges in existing security approaches and issues occurring in big data due to the ineffectiveness of present security tools and hardware.
B. Saraladevia , N. Pazhanirajaa , P. Victer Paula , M.S. Saleem Bashab , P. Dhavachelvanc, "*Big Data and Hadoop-A Study in Security Perspective*[11] *"*-
The researcher's paper helped us understand the flaws in the Hadoop Architecture's storage, management, processing, and security of personal sensitive data. It presented various vulnerabilities in the HDFS architecture with some proposed solutions to mitigate them.
Rida Qayyum, "*A Roadmap Towards Big Data Opportunities, Emerging Issues and Hadoop as a Solution*[12] *"*-
The author of this paper made us aware of the working of HDFS, its architecture, and the problems in HDFS, leading to the need to develop some innovative solutions in order to solve these issues.

## III. Statement Of Problem

In the HDFS environment, the triad of confidentiality, integrity, and availability compromises in large number of data blocks.

## IV. Hypothesis

Combining Kerberos, Blue-eye, and Name node approaches to protect each of Hadoop Distributed File System's CIA triad of data sets

## V. Research Questions

Why only one of the CIA triad is secured now and not all 3 of them in HDFS?
Why all 3 of the approaches have to be combined and not any 2 of them for the security?
Will there be any loop-holes even after combination of these 3 approaches

## VI. Research Objectives

To determine the scope and importance of the combination of Kerberos, Blue-eye, and Name node approaches for protecting each of the three CIA triad of data sets in HDFS in order to protect the privacy of Big Data.

## VII. Research Methodology

To successfully achieve the research outcome within a limited time frame, we have adopted for doctrinal approach and chosen the diagnostic research methodology as the problem of security breach exists but its resolution has low clarity.

## VIII. Security of Present Architecture

### 8.1 Introduction

Confidentiality, Integrity, and Availability are represented by the three letters "CIA triad." The CIA triad is a prominent model that serves as the foundation for the creation of security systems. They are used to identify weaknesses and develop strategies for problem-solving. The CIA trinity divides these three concepts into different focal points because they are essential to the operation of a business: information must be confidential, trustworthy, and readily available. This distinction is useful because it directs security teams as they decide how to best address each concern. The organization's (for example, HDFS) security profile should be stronger and better able to respond to threat situations if all three criteria have been met.

### 8.2 Discussion

When referring to the concerns about privacy of big data as per the literature *"privacy, security, and legal challenges in big data"*[13] *—we* found and analyzed that—there may be a lack of fundamental knowledge about how to protect these enormous volumes of data, and because adequate education isn't always provided regarding how to provide security and privacy to those huge scale data, technology lacks sufficient security and privacy preservation functions.

---

[8]Dr. Fatma Mohammed Abdullah (n 2).
[9]Nick Hajli, Farid Shirazi, Mina Tajvidi and Nurul Huda, 'Towards an Understanding of Privacy Management Architecture in Big Data: An Experimental Research' (*researchgate.net*, 15 July 2020) <https://www.researchgate.net/publication/343642296_Towards_an_Understanding_of_Privacy_Management_Architecture_in_Big_Data_An_Experimental_Research> accessed 29 August 2022.
[10]José Moura, Carlos Serrão, 'Security and Privacy Issues of Big Data' (*arxiv.org*, 2 May 2019) <https://arxiv.org/ftp/arxiv/papers/1601/1601.06206.pdf> accessed 31 August 2022.
[11]B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvan (n 3).
[12]Rida Qayyum (n 4).
[13]Dr. Fatma Mohammed Abdullah (n 2).

Furthermore, modern technologies have poor privacy and protection features; therefore, they are constantly being violated both unintentionally and on purpose. Therefore, it is necessary to constantly review and update current procedures in order to prevent data leakage. Also, when identifying traditional point-to-point defenses, which are the foundation of traditional security mechanisms that focus on end-to-end communication assurance (e.g., TLS/SSL, IPSec, etc.)[14], these defenses are less effective against the new cyber advanced continuous threat assaults. Any weak spot in networks, applications, or large data processing can be the target of malevolent intent.

### 8.3 Conclusion

After examining the aforementioned factors, it is clear that big data frequently faces the threat of losing one of its security triads due to a lack of knowledge, skills, tools, and policies. In addition, the implemented security mechanisms are unable to handle the assaults from threats that are exponentially growing, which causes weak points in the big data architecture. All of these flaws result in breaches, making it difficult to secure all three members of the CIA triad.

## IX. Combination of Approaches

### 9.1 Introduction

When we study the three approaches in HDFS for security, we can see that across an un-trusted network, such as the internet, Kerberos, a computer network security protocol, authenticates service requests between two or more trusted hosts. It authenticates client-server applications and confirms users' identities using secret-key cryptography and a reliable third party. Therefore, it can be analogous to the Integrity triad. The second method for safeguarding sensitive information is the Blue Eye strategy. It is used to protect sensitive information from all angles (360°) to determine whether everything is maintained securely and enables the designated individual to properly preserve personal information. Hence, it works as a confidentiality triad. The Name Node technique, which provides security in data availability, is the third approach. It ensures that data is accessible in a secure manner by serving as an arbitrator and repository for all HDFS meta data. As a result, it is comparable to the CIA's Availability triad. Similarly, the introduction of chapter-2 discusses that by combining the CIA triad, security always strengthens and threat solving becomes an easy task.

### 9.2 Discussion

When we refer to the literature – *"Big Data and Hadoop- A Study in Security Perspective"*[15]we found that large data sets can be managed more easily, but ineffective tools, threats and weaknesses in public and private databases, unexpected and voluntarily data leaks, and a lack of public and private policy encourage hackers to gather resources whenever they are needed.

In addition to that, the finding of the researcher in the above mentioned literature is that-

*"We can improve security in big data by using any one of the approaches or by combining these three approaches in the Hadoop Distributed File System, which is the base layer in Hadoop, where it contains a large number of blocks." These approaches are introduced to overcome certain issues that occur in the name node and also in the data node. In the future, these approaches will be implemented in other layers of Hadoop Technology."* [16]

In addition to that, we analysed that Hadoop is a technology that the future depends on. Effort can be made to install Hadoop on a cloud server to manage big data and interact with new frameworks so that big data can be broadly accepted.

Further considering the literature *"A Roadmap Towards Big Data Opportunities, Emerging Issues and Hadoop as a Solution,"*[17] the researcher found that-"*The fundamental challenge mentioned in this paper is, data is not available in a structured mode and, simultaneously, it is colossal in size and volume and demands fast and efficient processing."* [18]

### 9.3 Conclusion

From the debate, we may draw the conclusion that, because of the many technical breakthroughs highlighted, the demand for data security is growing daily. Additionally, this vast amount of data is unstructured, which makes it more difficult to process. Additionally, because technical tools are poor, the data is subject to leakage, especially when it is stored on cloud platforms. Furthermore, according to the study, when all three ways—Kerberos, Blue-eye, and Name node—are used together in HDFS, the security is superior than when just one of them is used, or when any two of the approaches are combined.

Therefore, we can deduce that by combining all of the aforementioned methods in HDFS, namely – Kerberos, Blue-eye, and Name node, the system will be more resistant to intrusions, resulting in improved security.

## X. Flaws In Combination

### 10.1 Introduction And Discussion

The shortcomings that we can assume based on talking about the preceding research elements may be the inability to follow the measurements of all three approaches, which can be caused by a potential knowledge gap. Furthermore, while no system is ever completely safe, we may use layer security to our advantage by integrating all of the CIA triads. As attackers continue to find ways to penetrate systems, it will always be difficult to keep up with rapidly evolving technology and adhere to occasionally necessary regulations. To defend against these threats, we must use the most up-to-date technological instruments.

[14]ibid.
[15]B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvan (n 3).
[16]B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvan (n 3).
[17]Rida Qayyum (n 4).
[18]ibid.

### 10.2 Conclusion

Hence, we can conclude that in spite of all these shortcomings, using the CIA triad will be beneficial for the security of HDFS as these are generally occurring flaws that are present with each technological development but can be mitigated by preventive measures. Also, using these measures will help our system to move towards perfection.

## XI. Results and Conclusion

### 11.1 Conclusion

After viewing all the aspects and critically examining previous research, we come to the conclusion that due to advancements in technology, a huge volume of sensitive data is being generated by various end points and organizations are processing it for their businesses to run profitably, but due to a lack of technological tools, flaws in policies and incomplete knowledge of protective measures, the big data is vulnerable to security breaches, thereby compromising the CIA triad of Hadoop architecture in big data.

Offering complete security is a difficult task to work on because there are many factors to it, as discussed in previous chapters, and even though different approaches have been developed to secure the HDFS in big data, they are not being used effectively or, better yet, properly integrated. This results in the compromise of one of the CIA's three core institutions. As shown by our interpretation of the existing solutions, the combination of Kerberos, Name Node, and Blue Eye will undoubtedly protect each component of the CIA triad in HDFS, demonstrating the validity of our hypothesis.

### 11.2 Suggestions

The author after reaching the conclusion suggests that the three HDFS established approaches should be combined to offer CIA triad security that is 360 degrees complete as each approach is analogous to one of the respective CIA triad. To make this mechanism sustainable in the future, proper improvements in the existing set of technologies and techniques must be made in addition to the combination.

### 11.3 Scope Of Future Research

Appropriate technological measures and procedures should be developed to ensure the effective integration of Kerberos, Name Node, and Blue-eye approach in Hadoop HDFS.

### 11.4 Acknowledgment

I want to thank the professors who assisted by effectively guiding me, the students who aided me along the path of my study by answering my questions, the librarian, and the support staff for all of their help.

## XII. References

**[1]** Dr. Abdullah F, 'Privacy, security and legal challenges in big data' (*iaeme.com,* 13 December 2019) <https://iaeme.com/MasterAdmin/Journal_uploads/IJCIET/VOLUME_9_ISSUE_13/IJCIET_09_13_167.pdf> accessed 1 August 2022.

**[2]** Hajli N and others, 'Towards an Understanding of Privacy Management Architecture in Big Data: An Experimental Research' (*researchgate.net,* 15 July 2020) <https://www.researchgate.net/publication/343642296_Towards_an_Understanding_of_Privacy_Management_Architecture_in_Big_Data_An_Experimental_Research> accessed 29 August 2022.

**[3]** Moura J and others, 'Security and Privacy Issues of Big Data' (*arxiv.org,* 2 May 2019) <https://arxiv.org/ftp/arxiv/papers/1601/1601.06206.pdf> accessed 31 August 2022.

**[4]** Saraladevia B and others, 'Big Data and Hadoop-A Study in Security Perspective' (*sciencedirect.com,* 8 May 2015) <https://www.sciencedirect.com/science/article/pii/S187705091500592X> accessed 20 August 2022.

**[5]** Qayyum Rida, 'A Roadmap Towards Big Data Opportunities, Emerging Issues and Hadoop as a Solution' (*researchgate.net,* 2 April 2020) <https://www.researchgate.net/publication/343484997_A_Roadmap_Towards_Big_Data_Opportunities_Emerging_Issues_and_Hadoop_as_a_Solution> accessed 16 August 2022.

**[6]** Alam A and others, 'Hadoop Architecture and Its Issues' (*ieeexplore.ieee.org,* 29 May 2014) <https://ieeexplore.ieee.org/document/6822351> accessed 1 September 2022.

**[7]** Sharma M and others, 'Investigating the Inclinations of Research and Practices in Hadoop: A Systematic Review' (*ieeexplore.ieee.org,* 10 November 2014) <https://ieeexplore.ieee.org/document/6949381> accessed 20 August 2022.

**[8]** Kandrouch N and others, 'A Novel Security Architecture Based on Haystack System for HDFS Storage System: Extended Work' (*ijitee.org,* 4 February 2020) <https://www.ijitee.org/wp-content/uploads/papers/v9i4/C8892019320.pdf> accessed 21 August 2022.

**[9]** Namavaram V and others, 'Two Layered Privacy Architecture for Big Data Framework' (*researchgate.net,* 10 October 2017) <https://www.researchgate.net/publication/321375343_Two_Layered_Privacy_Architecture_for_Big_Data_Framework> accessed 26 August 2022.