International Journal of Knowledge Based Computer Systems
11 (1) 2023, 01-07
http://www.publishingindia.com/ijkbcs/

Vulnerability Assessment and Mitigation Techniques on Hadoop Framework

Kimmi Kumari[1*], M. Mrunalini[2], Mamata Pandey[3], Sushma Verma[3] and Shubhojeet Paul[3]
[1]BIT Mesra, Ranchi, Jharkhand, India. Email: kimmikumari08@bitmesra.ac.in
[2]M S Ramaiah Institute of Technology, Bangalore, Karnataka, India.
[3]BIT Mesra, Ranchi, Jharkhand, India.
[*]Corresponding Author

Abstract: Hadoop has enormous data storing and large processing capabilities which makes it an efficient platform as compare to other platforms. In Hadoop much stress is given to deal with failures and coordination for complex distributed system. Early Hadoop Projects did not stance strong security measures. While new security challenges raised due to multi-occupant virtual environment of cloud computing. There is a phenomenal growth in cloud enabled services which also needs adoption of innovative security and privacy exploration. Due to flexible nature of Hadoop a strong mechanism is required to remove the vulnerabilities which can leads to security attacks. In this paper we discuss some vulnerabilities and effect of these vulnerabilities on the performance of Hadoop and which should be handled properly. Using vulnerability assessment and mitigation techniques, we have provided a security strategy for Hadoop in this study. This method is used because it can also reveal vulnerabilities which are unknown and helps in detecting the future threats while at the same time mitigating the present security threats.

Keywords: Assessment, Hadoop, Mitigation, Privacy, Security, Vulnerabilities.

I. Introduction

Information security model known for confidentiality, integrity and availability. While comparing to other platforms Hadoop is known for storing and processing enormous data efficiently. The early project of Hadoop stresses more on enhancing the technical aspects like developing the logic on how to deal with failures and coordination of such a large distributed system. It is considered that the distributed system's complete machine cluster is connected to a secure network, but it is not the reality. So Early Hadoop Project did not stand with strong security measures. Newer versions of Hadoop still evolving in terms of strong security measures and more stress is on protecting the tremendous data through encrypting the data and other data protection mechanism. Earlier it was assumed that Hadoop runs on trusted network and only authorized users were on the network but now encryption technique added for data transmission between the nodes. Hadoop is a distributed computing and parallel processing platform created by the Apache Software Foundation for batch processes. HDFS for huge storage, MapReduce for data processing, and YARN for cluster resource management are all parts of Hadoop. Big Data is handled by Hadoop, which makes use of various sorts of data that come from various sources. The Hadoop ecosystem is built on the Hadoop Distributed File System. Mass storage ability with the growing storage capacity is provided by the HDFS. Since HDFS filesystem logically spans in many servers so it is very important to understand the security perspective in HDFS or in all servers of the Hadoop cluster.

Transferring from the map phase to the reduced storage jobs is facilitated by aggregating the data volume as little as possible. MapReduce can be applied to transform data into executing business analytics. Lack of authentication and insecure communication between Hadoop daemons are two of the primary security issues with MapReduce. In Fig. 1, the main parts of Hadoop are shown. The component Job Tracker directs the MapReduce task to the nodes which having the data within the cluster. Task tracker accepts the task from MapReduce from the job tracker. Data node used to store the data in HDFS. Name node is used to make HDFS aware about each file location in the HDFS.

II. Related Work

Threats and risk are related to vulnerabilities. In a distributed system, vulnerabilities might take on a variety of different shapes [1]. Vulnerabilities are frequently found in the software itself. Every piece of software has weaknesses. Although it may seem harsh to say this, no piece of software is ever completely secure. An example of software vulnerability might be simply put; it is a piece of code that is vulnerable to a failure scenario or error that is not gracefully handled. Take the straightforward case of a piece of software that has a password screen that enables users to modify their passwords (we will assume that the intended logic for the software is to allow passwords up to 16 characters in length). What happens if the new password input form accidentally has a maximum length of 8 characters, truncating the selected password? Users could end up creating passwords that are shorter than they initially thought and, worse, less secure passwords that are easier for an attacker to guess.

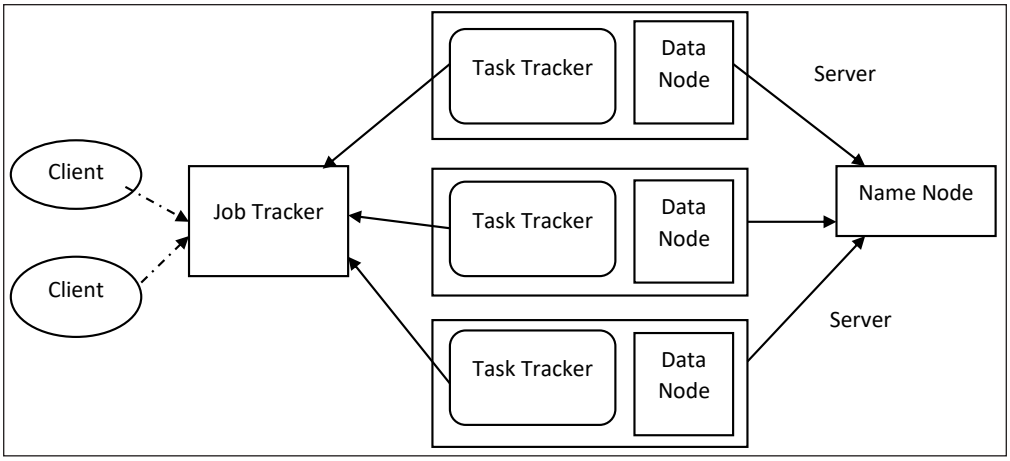

Fig. 1: Hadoop Architecture

Hadoop shares the inherent vulnerability [2] that any software possesses. Giving users access to master nodes may open the door for them to accidentally or intentionally use an unpatched Hadoop vulnerability. The more effective method of limiting access to the master nodes. Attackers may use software, a network, or a web interface to access all Hadoop components, like Sentry, Storm, and flink. Vulnerabilities are grouped into many categories because Hadoop is a combination of distributed computing software, hardware, application programme, and management rules. These categories could represent software flaws that allow hackers to target the Hadoop framework due to the language that it is written in [3] [4].

The "ping of doom" was the name of this assault. Although it has been mitigated, the fact remains that, prior to being fixed by network hardware vendors, this vulnerability had nothing to do with the software stack of the machines on the network, but an attacker could still use it to stop a specific machine on the network from performing its normal functions.

Every administrator's standard operating procedures should include regular schedules for installing patches to a distributed system's software stack since software patches are frequently issued to remedy vulnerabilities as they are found. The scope of patches should also cover the firmware for switches, routers, other networking hardware, disc controllers, and the server BIOS, as demonstrated by the ping-of-death example [4].

The security of big data can be seen from two aspects i.e., first while handling and storing information and the factors which affect the security of data and databases. Second aspect is while managing and maintaining the resources and dealing with operational safety issues with big data platform [5] [6].

It is very important to analyze the challenges while working with large data and management system when the Hadoop itself lacks the strong internal security management. In HDFS, cluster of servers may use different operating system platform and may leads to difficulty in level of patching [7] [8].

III. Proposed VAM Framework for Hadoop

While developing Vulnerability Assessment and Mitigation Framework for Hadoop, one of the most challenging tasks is to identify known and unknown vulnerabilities. Known vulnerabilities may come from Vulnerability database. After identification of these Vulnerabilities, evaluation should be done. While assessing or evaluating these vulnerabilities some deep learning methods or Artificial Intelligence may be useful. While evaluating a security or Vulnerability mitigation technique against a vulnerability some quantitative methods through deep learning can be utilized to present a relationship between individual vulnerability and its counter mitigation technique by a numeric value. This numeric value may be the outcome of these quantitative techniques of evaluator. The numeric value indicates whether a security technique can mitigate a vulnerability or may cause a vulnerability to incur. On the basis of these numeric values the relevance of the security technique whether it is primary or secondary against the vulnerability can be decided.

Detailed information of all the vulnerability mitigation techniques is analyzed and the study of appropriateness of a security technique among the available options must be presented in the comparative study. After organizing the risk mitigation techniques under primary or secondary importance a set of suggestions must be generated. The evaluation of risk mitigation techniques obtained through VAM methodology helps in refining the selection process of an effective risk mitigation model. After refinement of effective risk mitigation model one of the most effective vulnerability mitigation technique is selected according to priority.

In the proposed framework vulnerability assessment and mitigation for Hadoop can be distributed in various stages. Several challenging tasks need to be done to accomplish this task. These tasks can be performed in a phased manner.

- While adopting VAM framework for Hadoop, the most

challenging task is to identify known vulnerabilities. These known vulnerabilities may come from vulnerabilities database.

- If the threat is not recognized from the vulnerability database, it should be recognized as new novel threat.
- After identification of these Vulnerabilities, evaluation should be done. While assessing or evaluating these vulnerabilities some deep learning methods or Artificial Intelligence may be useful.
- For determining vulnerability and implementing countermeasures, we make use of recent developments in deep learning, which enable the development of systems that can learn from previous system events and predict a specific dangerous event that is likely to occur next. On the basis of the anticipated information, this system should recommend genuinely preemptive countermeasures.
- After assessing the specific vulnerability a matrix is prepared to show the relevance of specific security technique for each vulnerability.
- A comparative study of different vulnerability assessment and the counteraction is prepared.
- Finally the prioritization of vulnerability mitigation technique is decided.

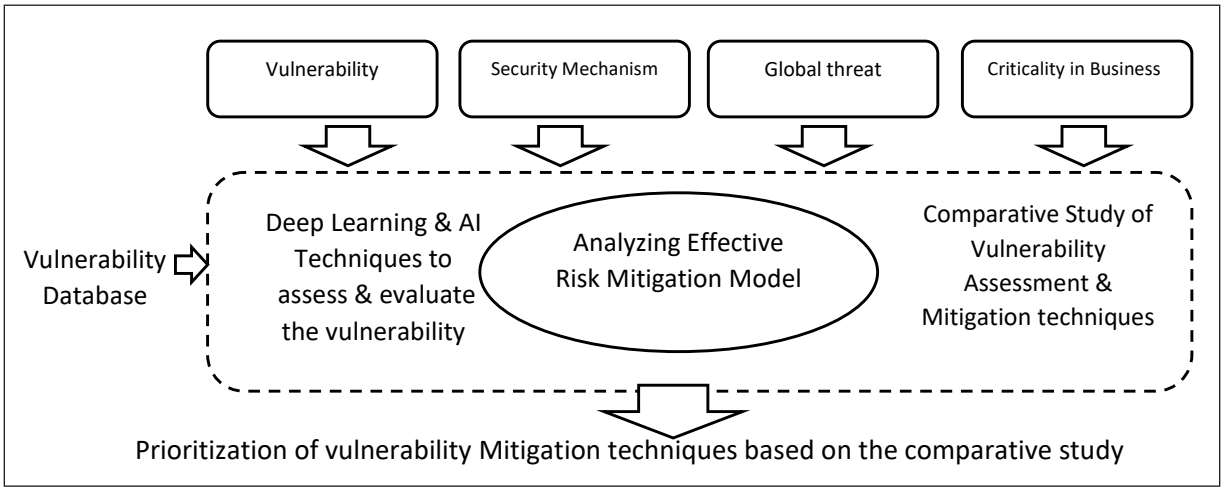

Fig. 2: Proposed Vulnerability Assessment and Mitigation Framework of Hadoop

Proposed Framework can be written in various steps:

Step 1: Identifying Vulnerability in Hadoop

While identifying vulnerabilities in Hadoop there are 3 types of vulnerabilities should be considered. First may be the technology related vulnerability i.e. the software platform in which Hadoop is developed and its possibilities to exploit by the cybercriminals. Secondly the configuration vulnerabilities for example there are many default settings need to be assigned to such a big platform. So these settings may be one of the targets by the attackers. The network security flaws, which reveal security breaches at both the service level and the data level, can also be taken into consideration.

Numerous vulnerability databases list numerous vulnerabilities, including a vulnerability management repository called the National Vulnerability Database (NVD) contains information on software problems that affect security. Another database that offers details on software vulnerabilities is the Computer Emergency Readiness Team (CERT). The Microsoft Security Response Center (MSRC) publishes Microsoft Security Bulletins, which are also connected to security flaws found in Microsoft products. Another database list of vulnerabilities with a unique ID is called Common Vulnerabilities and Exposures (CVE). The Open Source Vulnerability Database (OSVDB) offers truthful, unbiased data about security flaws. Numerous flaws in cyber-security software and hardware are listed in the CVE database.

Known vulnerabilities are those which can be handled by typical defense mechanism like DOS, script virus or worms' etc. unknown vulnerabilities may be an insider attack, network threat for which some special mechanism needed. Patch management is the process of retrieving the known and unknown vulnerabilities. Through vulnerability scanning the new vulnerabilities are discovered and resolved accordingly.

Step 2: Identifying Threats

Novel threats basically unknown threats which may be initiated by the insider or may be the network threats. These are harmful since threat protection mechanism is not well known.

For finding vulnerability or to identify insider threats conventional security approach are not sufficient. It is a challenging task due to user's behaviors. It can be done in various steps:

- To identify vulnerability or insider threat, log data is analyzed to detect threats.
- User behavior within the cluster is on surveillance to identify security incidents, monitoring violations and preparing baselines.

- Data mining methods are used to detect the abnormal behavior of the user and helps to undergone huge data to filtering out the true threats from the false positives.
- Vulnerabilities identified are shown through a basic metrics. Some systems may have dense vulnerabilities which may tend to gather in specific areas.
- Then the vulnerabilities shown with their counter action to fix and to repair. In this metric a clear picture will be depicted to how many vulnerabilities have been repaired.

Step 3: Vulnerability Assessment and Mitigation in Hadoop

The goal of vulnerability analysis is to identify as many security flaws as possible (a "breadth over depth" approach). It should be used frequently to keep a network safe, especially when there are network changes (e.g., new equipment installed, services added, ports opened). Additionally, it will work for firms that want to know about all potential security flaws but are not yet security mature.

After identifying vulnerabilities which is most challenging task we need to assess the vulnerability and to establish an effective vulnerability mitigation model.

- While assessing the vulnerability, its severity should be known. It can be assessed using threat likelihood which me be judged by exploitability, discoverability and reproducing capability.
- Vulnerability needs to analyze the impact on confidentiality and integrity.
- Each vulnerability should be ranked from critical to low relied on its severity and possible effect on Hadoop to create a vulnerability score system.
- Estimation of number of vulnerability which can be repaired should be done.

Refer Table II for the evaluation.

After identifying vulnerabilities which is most challenging task we need to assess the vulnerability and to establish an effective vulnerability mitigation model.

While assessing the vulnerability, its severity should be known. It can be assessed using threat likelihood which may be judged by exploitability, discoverability and reproducing capability.

- Vulnerability needs to analyze the impact on confidentiality and integrity.
- Each vulnerability should be ranked from critical to low relied on its severity and possible effect on Hadoop to create a vulnerability score system.
- Estimation of number of vulnerability which can be repaired should be done.
- Appropriate mitigation technique is adopted.
- Risk mitigation techniques for Hadoop must consider the following approaches to secure the system:
  - Ensure security of tools and techniques.
  - User account management and authentication.
  - Access control.
  - Securing transaction logs and large data sets.
  - Protection for software and hardware configuration.
- To protect Hadoop system from multiple novel attacks of hackers data analytics and AI techniques are used which can detect unknown vulnerabilities. Extended threat detection and false positive rate can helps analysts to detect unknown security breaches immediately.
- A comparative study of risk mitigation models can help in selection and prioritization of appropriate model.

Step 4: Identifying Mitigation Related to Vulnerability

For identifying mitigation techniques for specific vulnerability first we need to discover the vulnerability and then it can be exploited to find appropriate mitigation technique. Detailed process of mitigating vulnerabilities is discussed in Table I.

Table I: Vulnerability Assessment and Mitigation in Hadoop with Prioritization of Mitigation Technique

Attack (Vulnerability) Name	Evaluation/Assessment of Vulnerability	Effective Vulnerability Mitigation Technique	Priority of Vulnerability Mitigation Technique
Protect Session Privacy	• It is vulnerable to attack session privacy. • Session privacy is susceptible to hacking, whether it is used for communication between clients and data nodes or among nodes.	SSL (Secure Socket Layer)/TLS (Transport Layer Security)	• Transport Layer Security (TLS) is the first choice. • Transport encryption can protects all communication from access or modification by attackers. • But the TLS is difficult to implement and get certificate management right.
Data Usage	• System enforces to unapproved authorization to logical access of information and system resources.	Identity and authorization/ Masking/Application encryption	• Priority is given to Secure Shell (SSH) authentication within the Hadoop cluster.

Attack (Vulnerability) Name	Evaluation/Assessment of Vulnerability	Effective Vulnerability Mitigation Technique	Priority of Vulnerability Mitigation Technique
			• Identity and authentication can be used as a central security effort and considered as a second priority technique.
Data at Rest (External Threat)	• Vulnerable to attack on HDFS file system where data is stored.	Application/Object (HDFS) encryption	• In Hadoop, the file system contains embedded encryption. This indicates that data is encrypted transparently when it is placed into the file system, without modifying the programme running in the cluster.
Data at Rest (Credentialed Users)	• Vulnerable to tenant data privacy in multi-tenant clusters.	Application encryption/External key management	• The system can be linked with key management services from third parties or used with Hadoop's Key Management Service (KMS). Access Control Entries (ACE) or Access Control Lists (ACL), which are essentially file permission constructs, are utilized by some versions of Hadoop.
Node Authentication & Validation	• In Hadoop user application runs on a cluster of machine so vulnerable to attack any of the specific machine. • A node's identity can be forged.	PKI (Private Key Infrastructure)/Kerberos	• HDFS supports a variety of authentication schemes including PKI-based and Kerberos. • Although Kerberos credentials are accepted, those credentials are not passed along to the workers. • Therefore, jobs that need Kerberos keys to access resources.

Step 5: Vulnerability Mitigation through VAM Methodology

Security of Hadoop ecosystem based on proper functioning of three parts which are identity, authorization and authentication are discussed in detail in Table II. Due to distributed nature of Hadoop system nodes within the cluster are loosely coupled from authoritative identity sources. To ensure the security of data of all nodes various security protocols may be imposed. Below table shows the Vulnerability Assessment and Mitigation for Hadoop through VAM methodology.

TABLE II: VULNERABILITY ASSESSMENT AND MITIGATION FOR HADOOP THROUGH VAM METHODOLOGY

Vulnerability Causes	Effect of Vulnerability	Mitigation Technique
Identity/Data Breaches	• Files stored in HDFS may be unsecure. • Data in files can be maliciously accesses. • File transfer may be unsecure within and outside the cluster.	• Files within HDFS system can be encrypted. • SSL needs to be applied for MapReduce and web consoles. • Encryption should be adopted during HDFS file transfer.
Malicious Attack	• Lack of authentication between client and services. • Lack of authentication between Data node, Name node, Task tracker and Job tracker.	• With the use of single sign on Hadoop public key cryptography and other authentication mechanism can be applied. • Authentication of all the clients for all services can be performed through single server.

Vulnerability Causes	Effect of Vulnerability	Mitigation Technique
	• Unauthenticated delegation of tasks.	• For storing the encryption keys of authenticated users an effective credential management framework can be used.
Unauthorized Access Privileges	• Wrong file permission within HDFS system may be assigned. • Access privilege list may be controlled maliciously. • Organization job queues may be distorted in terms of priority sequence.	• Authorized HDFS file permission should be granted. • Access control list and job queues of various user groups should be verified before execution. • Fine grained and role based authorization should be imposed.

IV. Result and Discussions

Hadoop environment must be protected from a network security perspective. The vulnerability due to distributed nature of Hadoop can be addressed through various security mechanism viz. imposing firewalls, IDS and protecting the communication between the nodes. The VAM methodology helps in gaining the insights to employ specific security technique for handling vulnerability. This methodology also review the existing vulnerabilities and generate the security options to help in prioritization among these options. VAM methodology also identifies the future vulnerabilities and the level of mitigation techniques which may handle these unexplored vulnerabilities. The main focus of VAM methodology is to protect the information system but if we see in broader aspect it presents a detailed review of vulnerability and attacks which could be useful in understanding the nature of vulnerability and its mitigation technique. It also presents a comprehensive study to adopt any defending action during and after an attack. This methodology is very useful when the system is complex like Hadoop as well as when a new component is to be added to project, as it can look for new threats which may cause system failures or may be exploited by the malicious users. Since Hadoop is operated on cloud environment also so new components may be frequently added due to various reason viz. scalability, and efficiency etc. traditional security assessment system only look towards exploited vulnerabilities which is not the case in VAM methodology.

V. Conclusion

As Hadoop is used to store and process enormous data, big enterprises prone to use it. While using Hadoop for organization security administrators face the challenge to mitigate all kinds of threats and vulnerabilities to protect organizations data. It is also true that there is no single threat for such a big distributed system rather new threats may arise frequently. To protect the system from these threats mitigation techniques and many security control strategies must be adopted. More than one security control is applied to achieve a comfortable table off security. The idea of implementing multiple security control is called defense in depth.

In this paper we have discussed Hadoop framework, vulnerability, patch management, and security issues in big data/HDFS. We have also proposed vulnerability assessment and mitigation framework for Hadoop. This framework stresses on identifying vulnerability and to control these vulnerability through the counterpart mitigation techniques. In our framework we have used deep learning and AI methods to analyze the threats and effective mitigation techniques. The review of the vulnerability and their possible mitigation techniques helps in comparison and analysis of the effective models. VAM methodology helps in decision making for selection of appropriate vulnerability mitigation techniques against the threat likely to occur. VAM methodology helps in development of guidelines for security measures with recommended priority of mitigation techniques. It is also beneficial for risk analysis and cost effective options for mitigation of exploited vulnerability as well as future threats.

References

[1] R. Pandey, and C. Verma, "Big data representation for grade analysis through Hadoop framework," In *2022 6th International Conference-Cloud System and Big Data Engineering (Confluence),* IEEE, 2022, pp. 312-315.

[2] A. M. Martinez-Enriquez, M. Adnan, M. Afzal, M. Aslam, and R. Jan, "Minimizing big data problems using cloud computing based on Hadoop architecture," In *2014 11th Annual High Capacity Optical Networks and Emerging/Enabling Technologies (Photonics for Energy),* Charlotte, NC, USA, 2014, pp. 99-103.

[3] E. Fernández-Medina, J. Moreno, and M. A. Serrano, "Main issues in big data security," *Future Internet*, vol. 8, no. 3, p. 44, 2016.

[4] B. B. Rad, N. Akbarzadeh, P. Ataei, and Y. Khakbiz, "Security and privacy challenges in big data era," *International Journal of Control Theory and Applications*, vol. 9, no. 43, pp. 437-448, 2016.

[5] D. H. Manjaiah, and M. A. Shetty, "New security architecture for big data Hadoop," In N. Shetty, L. Patnaik, N. Prasad, and N. Nalini (Eds.), *Emerging Research in Computing, Information, Communication*

and Applications (ERCICA 2016). Singapore: Springer, 2019.

[6] R. Rapuzzi, and M. Repetto, "Building situational awareness for network threats in fog/edge computing: Emerging paradigms beyond the security perimeter model," *Future Generation Computer Systems*, vol. 85, 235-249, 2018.

[7] G. S. Bhathal, and A. Singh, "Big data: Hadoop framework vulnerabilities, security issues and attacks," *Array*, vol. 1-2, p. 100002, 2019.

[8] E. Spivey, *Hadoop Security: Protecting Your Big Data Platform*. O'Reilly Media Inc., 2018.