International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT239023

Security Process in Hadoop Using Diverse Approach

Dr. A. Antony Prakash

Department of Information Technology, St. Joseph's college, Trichy, Tamilnadu, India

ARTICLEINFO

Article History:

Accepted: 01 March 2023
Published: 14 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
66-71

ABSTRACT

Today, data is one of the most important assets for nearly every sector of industry. Advances in information technology and widespread growth in various fields such as business, engineering, medicine, and scientific research have resulted in an explosion of data. This data explosion presents new challenges that traditional methods cannot handle. One of the biggest concerns of our time concerns the security and protection of sensitive information. In today's Big Data era, where organizations collect, analyse, and make decisions based on analysis of massive datasets from various sources, security in this process becomes more and more important. At the same time, more and more organizations need to enforce access controls and privacy restrictions on these records to meet regulatory requirements such as HIPAA and other privacy laws. Network security breaches by internal and external attackers are on the rise, often taking months to discover, and those affected pay the price. Organizations that fail to properly control access to records face lawsuits, negative publicity, and fines.

Keywords: Encryption, Decryption, Knox, Ranger, Kerberos, Apache sentry

I. INTRODUCTION

Hadoop is a software framework for storing and processing large amounts of data. This article discusses Hadoop security. This article first explains the reasoning behind Hadoop security. We then discuss Hadoop security, the 3A's of Hadoop security, and how Hadoop enables security. This article covers Kerberos, Transparent Encryption in HDFS, and HDFS file and directory permissions that solve HDFS security issues. The article also lists some components of the Hadoop ecosystem for monitoring and managing Hadoop security.

Big data is the collection of vast amounts of historical and significant data that is the most valuable asset of any organization and, when used intelligently in business, can support decision-making based on facts rather than perceptions. . Charles Tilly in Oxford Dictionary coined the term big data in 1980 [1]. Hadoop is synonymous with big data. Big data

consists of the four V characteristics of volume, velocity, variety, and veracity [2], [3].

The massive growth of data is creating issues that affect not only data volume, velocity, diversity and accuracy, but also data security and privacy. Recently, another "V", that is Vulnerability to add (see Figure 1) [4].

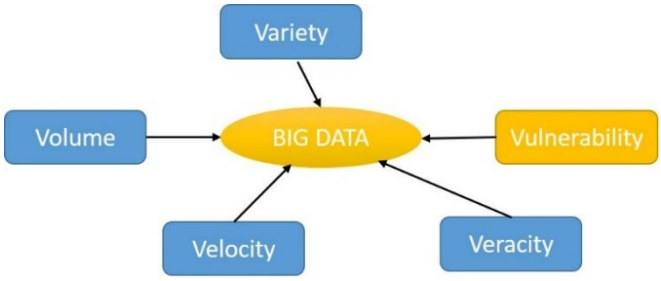

Figure 1 Vulnerability of big data

Initially there was no security model. Hadoop did not authenticate users or services and had no privacy. Hadoop was designed to run code across distributed clusters of machines, so anyone could submit code and run it. Previous distributions implemented audit and authorization controls (HDFS file permissions), but since users can impersonate other users with command line switches, such access controls are It was easy to avoid. Spoofing was so prevalent and performed by most users that the security controls in place were not effective.

At the time, security-conscious organizations segregated Hadoop clusters on private networks and restricted access to authorized users. However, Hadoop had few security controls, and many mishaps and security incidents occurred in such environments. Well-intentioned users can make mistakes (for example, using distributed deletion to delete large amounts of data in seconds). All users and programmers have the same level of access to all data in the cluster, all jobs have access to all data in the cluster, and all users can potentially read all records is ready. Map Reduce had no concept of authentication or authorization, so a malicious user could deprioritize his other Hadoop jobs to make them complete faster, or worse, kill other jobs.

The new technology is designed to ensure your privacy and security, unlike its predecessors. However, this technique can also be used for negative purposes. As more and more businesses and individuals use this technology to store and process personal data, it has become a prime target for data attacks

II. LITERATURE REVIEW

Vulnerabilities fall into three main categories: infrastructure security, privacy, and data management [5]. These are further divided into three dimensions: the architecture dimension, the data lifecycle dimension, and the data value chain dimension. According to the authors, infrastructure includes hardware and software vulnerabilities that exist in the architectural dimension. Data protection concerns data at rest and data in transit and affects the lifecycle of data.

Consider setting method Between NameNode and accept as true buyer. User authentication completed on NameNode uses hashing technology (SHA-256). User authentication by sending the hash function to the NameNode NameNode generates a hash function "observe". Hash token sent by the person. Use of RSA, Rijndael, AES, and RC6 encrypt randomly. Statistics for high quality nodes in the cluster. In Hadoop frame appearance, all information is split into chunks 64 MB covers recording encryption and decryption via his Map Reduce function on Hadoop devices [6].

It introduces the period of big data and big data analytics, highlights protection and privacy-related situations addressed using big data tools, and provides information on possible solutions and strategies for mitigating waivers. He discussed the following

possible answers to privacy protection of vast amounts of information: rules and legality, encryption, authentication, metadata and tagged data, unstructured distribution, anonymization. , interest tracking [7].

With more and more cloud packages Certainty of fact is becoming a key issue in the cloud Computing. Privacy is a key issue in the cloud-computing environment. All documents are stored in HDFS its plaintext and managed through the main server. Therefore, HDFS is unfamiliar with storage servers you might look at. Statistics Material Cloth. Additionally, Hadoop and HDFS has a weak security version. In particular, Conversations between data nodes and customer-to-customer Data nodes are not always encrypted. to ensure Measurement confidence in Hadoop-based cloud files Carport, a lonely triple cipher plan proposed, Updates combining HDFS reporting and encryption DEA (Data Encryption Algorithm) and Knowledge Scratch Encryption with RSA, then the person is encrypted IDEA (international data encryption algorithm) [8].

The author explored the security risks in cloud computing and smart advances businesses can take to bounce back and protect their resources. Key areas in cloud computing quality, vulnerability, and data opportunity management. In addition, they are aware of the association's safety hazards, hazards, and Precautions available before adopting this technology [9].

Lack of restraint cloud security data disasters, isolation, Privacy when accessing web applications Cloud. Intentions such as RSA, DES, AES, I tried comparing using blowfish a survey between them was also shown secure your data in the cloud. DES, AES, blowfish is a symmetric key computation, use a single key for both Message encryption/decryption [10].

III. SECURITY RISKS IN HDFS

Hadoop uses group whoami and bash –c utilities Unix with isolation for individual clients and assemblies, that's it Weaknesses for approval and record crowds client. There are three types of security breaches HDFS, Unauthorized Access, Tampering Disclosure and Waiver of Control or Assets. Following are the areas where threats have been identified in Hadoop.

3.1 Encryption

Big data encryption tools must protect data at rest and in transit across large data sets. Organizations should also encrypt both user-generated and machine-generated data. As a result, encryption tools must work with multiple large-scale data storage formats, such as NoSQL databases and distributed file systems such as Hadoop.

3.2 User Access Control

User Access Control is a basic network security tool. Lack of proper access control measures can have disastrous consequences for big data systems. A robust user control policy should be based on automated role-based settings and policies. Policy-driven access control protects big data platforms from insider threats by automatically managing complex layers of user control, including multiple administrator settings.

3.3 Intrusion Detection and Prevention

The distributed architecture of big data is the advantageous intrusion attempts. An intrusion prevention system (IPS) helps security teams protect big data platforms from exploitation of vulnerabilities by inspecting network traffic. IPSs are often placed right behind a firewall to isolate intrusions before any real damage is done [12].

The following section describes security threats that can affect the operation of Map Reduce and all framework components in the absence of a protected Map Reduce environment. Map Reduce distributed and replicated processing can open up a variety of attack opportunities.

3.3.1 Authentication and Authorization: Identity and authentication are central to any security measure. Without them, it is impossible to determine who should and should not access data. It does this at the user level with full Active Directory (AD) and Lightweight Directory Access Protocol (LDAP) integration with Kerberos, and at the service level with node-level role-based access control (RBAC).
Administrative data access should be ACLs, file permissions, and separation of administrative roles from OS and Hadoop administrators.

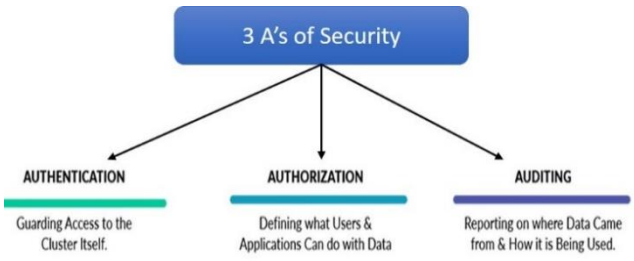

Figure 2 3 A'S of Security

The Hadoop stack has many different components that ship with no security by default. Configuration and patch management is required when different configurations and patch levels are running at the same time.

There are no built-in monitoring tools to detect exploits or block malicious queries. Logs should be configured to capture both the correct event types and enough information to determine user actions. Auditing and monitoring tools are important when data volumes and speeds are high.

The application is based on the web service model and is used especially in social networks such as Facebook, Yahoo and Snap Chat. Hadoop web applications can be vulnerable to known attacks due to insecure APIs. Big data cluster APIs need to be protected against common web service attacks, command injection, buffer overflow attacks, etc.

IV. TYPES FOR HADOOP SECURITY

Hadoop ecosystem includes several tools to support Hadoop security

4.1 Knox

Knox is a REST API-based perimeter security gateway that performs authentication and supports monitoring, auditing, authorization management, and policy enforcement on Hadoop clusters. Commonly authenticates user credentials against LDAP and Active Directory. Only successfully authenticated users can access her Hadoop cluster.

4.2 Ranger

It is an authorization system that grants or denies access to Hadoop cluster resources such as HDFS files, Hive tables, etc., based on defined policies. User requests are assumed to have already been authenticated while coming to Ranger. There are different authentication capabilities for different Hadoop components such as YARN, Hive, HBase.

4.3 Kerberos Security

Kerberos is one of the leading network authentication protocols designed to provide strong authentication services at both the server and client ends using secret key cryptography. It has proven to be very secure as it uses encrypted service tickets for the entire session.

4.4 Apache Sentry

Apache Sentry [13] is a fine-grained, role-based authorization and multi-tenancy management module for Hadoop. Sentry can manage access to data and metadata by applying precise levels of permissions to authenticated users and applications within your Hadoop cluster. It is highly standardized and may support acceptance of different data models in Hadoop. It is versatile and allows you to define authorization rules to validate a user's or application's request to access her Hadoop resources. Sentry aims to be a pluggable authentication engine for Hadoop elements such as Apache Hive, Apache Solr, Impala and HDFS.

4.5 HDFS Encryption

HDFS encryption is an impressive advancement that Hadoop has taken so far. Here data is fully encrypted from source to destination (HDFS). This method requires no changes to the original Hadoop application, so only the client can access the data.

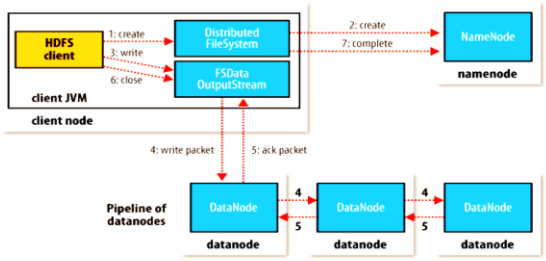

Figure 3 HDFS Encryption

Figure 3 shows the operation of saving each piece to HDFS. Customer divides each dataset into fixed size and square Encrypt before sending to the Hadoop document framework. It is considered possible to encrypt and decrypt [14]. It is basically updated using Java classes. Customer, Perform the encryption itself using AES calculations on the CPU Swap the encoded pieces to HDFS (Data Node). Include Point Collector Data Node (first Data Node where pieces are stored) Reproduce the obstacle on two different Data Nodes.

4.6 DECRYPTION IN HDFS

Customer assembles information on Data Node Sequentially, but during Map Reduce Occupancy. A number of squares are read (decoded) in parallel of task tracker [15]. Figure 4 shows anything the MapTask has read. Encodes blocked information when using Task Tracker AES encryption strategy.

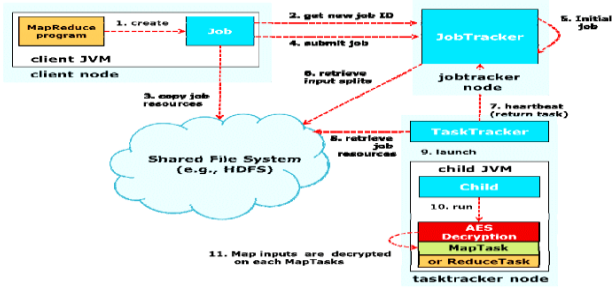

Figure 4 Decryption in HDFS

There are many considerations for this MapTask runs on a special Hadoop target. HDFS supports a read-once model. That's all concurrent decoding of HDFS squares explained it makes sense for some Map Reduce applications.

V. CONCLUSIONS

While traditional enterprise security products implement security controls, they are still insufficient to holistically address the security challenges posed by big data. To address the problems posed by data aggregation, organizations must find new ways to safeguard their critical tools, techniques, and procedures used to acquire, maintain, and analyses data of particular concern to improve the robustness of its security infrastructure, as most commercially available security software have access to all real and derived data available on the network. The add-on security features provided by the third party are not useful in the Big Data environment.

Furthermore, analysts and other users of information technology need more effective security training that will help them understand the specific threats they

face, that how their security choices will impact the outcomes, and how to reconcile security objectives with mission objectives. Finally, system designers need to consider the security implications affecting user-facing tools. Rather than relying on separate security tools for forensic verification, users benefit from the ability to provide security-related feedback throughout their workflow. Measuring the impact of these "small" safeguards can be difficult and problematic. However, field studies such as those cited in this white paper suggest that scaling security can have a significant impact in ways that scale with the amount of data you protect. Therefore, the entire exposed Hadoop system must be updated and all security features configured separately without installation.

VI. ACKNOWLEDGEMENT

The heading and image should be treated as a 3rd level heading and should not be assigned a number.

VII. REFERENCES

[1] R. Dontha, “The origins of big data “, February 2017 , Accessed 2nd Jan 2019.

[2] I.A.T. Hashem, et al. “The rise of "big data" on cloud computing”, review and open research issues Inf Syst (2015), pp. 98-115.

[3] J.S. Hurwitz, A. Nugent, F. Halper, M. Kaufman, “Big data for dummies wiley” , USA (2013).

[4] S. Sharma “Vulnerability introducing V of big data”,july 2017, Accessed 2nd Jan 2019.

[5] H. Ye, X. Cheng, M. Yuan, L. Xu, J. Gao, C. Cheng, “A survey of security and privacy in big data”. 2016

[6] Pradeep Adluru, Srikari Sindhoori Datla, Xiaowen Zhang, “Hadoop Eco System for Big Data Security and Privacy” 978-1- 4577-1343-9/12/$26.00 ©2015 IEEE.

[7] Youssef Gahi, Mouhcine Guennoun, and Hussein T. Mouftah “Big Data Analytics: Security and Privacy Challenges” IEEE Symposium on Computers and Communication (ISCC), 2016.

[8] Chao YANG, Weiwei LIN, Mingqi LIU “A Novel Triple Encryption Scheme for Hadoop-based Cloud Data Security” Fourth International Conference on Emerging Intelligent Data and Web Technologies, IEEE 2013 DOI 10.1109/EIDWT.2013.

[9] Owen O'Malley, Kan Zhang, Sanjay Radia, Ram Marti, and Christopher Harrell Hadoop Security Design‖, Yahoo, 2009.

[10] Rachna Arora, Anshu Parashar, "Secure User Data in Cloud Computing Using Encryption Algorithms" International Journal of Engineering Research and Applications (IJERA) Vol. 3, Issue 4, Jul- Aug 2013, pp.1922-1926.

[11] C.B. Raj Samani “ McAfee threats report 2018”, accessed 2nd Feb 2019.

[12] Duygu Sinanc Terzi; Ramazan Terzi; Seref Sagiroglu, “A survey on security and privacy issues in big data”, 10th International Conference for Internet Technology and Secured Transactions (ICITST), December 2015.

[13] Regha, S., Manimekalai M. “Approval of Data in Hadoop Using Apache Sentry”, International Journal of Computer Sciences and Engineering, Vol.-7, Issue-1, Jan 2019.

[14] Chao YANG, Weiwei LIN, Mingqi LIU, "A Novel Triple Encryption Scheme for Hadoop-based Cloud Data Security", International Conference on Emerging Intelligent Data and Web Technologies, 2013.

[15] GitHub, RJ97/Kuber, “A Framework for Large Scale Encryption in Hadoop Environment,” Mar. 2017.

Cite this article as :

Dr. A. Antony Prakash, "Security Process in Hadoop Using Diverse Approach", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.60-65, March-April-2023. Available at doi :
https://doi.org/10.32628/CSEIT239023
Journal URL : https://ijsrcseit.com/CSEIT239023