ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

A Survey on Data Security System for Cloud Using Hadoop

Dharmik H. Patel, Dr. S. N. Gujar

M.E Student, Dept, of I.T, SKN College of Engineering, Pune, India

Assistant Professor, Dept. of I.T., SKN College of Engineering, Pune, India

ABSTRACT: Now a day there is an upcoming challenge to Store, Manage, and Distribute Big data across several server nodes, for this purpose, the new platform has been developed which is called as HADOOP. This survey gives the information about the big data issues and it mainly deals more on security issue that comes across in Hadoop. Hadoop Distributed File System (HDFS) is architecture base layer. The HDFS security is improved by using approaches like Kerberos Algorithm and Name node. The complex requirements of big data are handled by Hadoop file system. The previously developed model effectively prevents the intruder prone operations such as data alteration, data deletion, and insertion. The cluster nodes which are prone to severe threat are secured using the HDFS functionality which uses Apache Sentry for authorization of Clients and Kerberos for authenticating access to the nodes. Hadoop projects treat Security as a top work item which in turn represents which is again classified as a critical item. It is financial applications that deemed sensitive, to healthcare initiatives; Hadoop is traversing new territories which demand security environments. The growing acceptance of Hadoop, there is an increasing trend to incorporate more enterprise security features. System summarizes (authentication, auditing, authorization, and encryption) within a cluster. There is no some of the production environments that support Hadoop Clusters.

KEYWORD: Security in Hadoop, Kerberos security, Hadoop security, big data security, Authentication security.

I.INTRODUCATION

The prominent security concerns for Hadoop Clusters are data security and access control because of its Internet-based data storage. Cloud computing is an expansion of grid computing and distributed computing, which is a software concept in reality, it works through a mixture of technologies such as software technologies, management, integration and the use of various hardware sources. Cloud computing is realize mainly through the virtual technology. The virtual technology can be separated into single virtualization and multiple virtualizations. The single virtualization is apply the virtual technology on a machine to work on numerous machines working commonly such as VMware, while multiple machine virtualization ties the machines through the control center and makes them work like the single machine. Hadoop is the representative of related technology. The distributed storage system stores the data in similar devices which are independent of another.

HDFS developed by Apache Foundation, not only takes full advantages of the power it high-speed computing clusters and storage but also demonstrates high performance in big data storage. The main intension of network security is to ensure that irrelevant individuals who have no right to use but attempt to obtain the remote service cannot read or modify the information which will be passed to other acceptor. Most of the emergence of the network security problems is because malicious people try to intercept or modify the information which does not belong to him originally, in order to obtain some kind of benefit or harm many intentionally. It is obviously that to guarantee the network security not only needs to make the program without programming errors, but also to guard against stalkers, hackers especiallythat have abundant time and money to brute force attack the system.

Kerberos is a network authentication protocol. It is designed to provide strong authentication for client and server applications by using secret-key cryptography. The cryptographic algorithm typically employee in Kerberos is Data Encryption Standard (DES).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

II. LITERATURE SURVEY

B. Saraladevi and et al. The security issue is pointed more in order to increase the security in big data. We can improve security in big data by using any one of the approach or by combining these three approacheswhich is the base layer in Hadoop Distributed File System, where it contains large number of blocks. These approaches are introduced to overcome certain issues occurs in the name node as well as in Data node. In Future these approaches are also implemented in other layers of Hadoop Technology.

S. Saranya and et al. With the growth of business and data in really unpredictable amounts and enterprises are currently movingtowards tools to manage and perform computing on these data. To handle such large amounts of data Cloud Service Providers are soughed for storage purpose. In order to handle queries and process these data for dynamic and real time applications, Hadoop is being preferred. Since these data are really sensitive to determine even the future growth of an organization or a commodity, they are in a possibility of being in the demand for hackers. But neither Cloud nor Hadoop provides a promising security for the data or processed.

Yannan Ma and et. al. It system improved the ability of security and storage efficiency of the system through specialdecoding mode and multi-nodes reading. Meanwhile, we can store part of the metadata in Datanode to reduce the memory usage of the system, so that the storage efficiency of the system may improve obviously through the decrease of Namenode's workload because work of data encoding is done on the Namenode.

Bao Rong, Chang and et al. In this study Hadoop cloud computing together with access security by applying the rapid identification on fingerprint and face has been realized so that cloudcomputing initiates the services like SaaS, PaaS, and/or IaaS. The connection between client and server has employed a way of low-capacity Linux embedded platform linked to Hadoop via Ethernet, 3G or Wi-Fi. At client side, JamVM virtual machine is utilized to form the J2ME environment, and GNU Class path is viewed as Java Class Libraries. To verify the cloud system effectiveness and efficiency in access security, the rapid identification on fingerprint and face in Hadoop has been done successfully within 2.2 seconds to exactly cross-examine the subject identity.

Venkata Narasimha Inukollu and et al. Cloud environment is widely used in industry and research aspects; therefore security is animportant aspect for organizations running on these cloud environments. Using proposed approaches, cloud environments can be secured for complex business operations.

Dhananjay M Dakhane and et al. This is technology challenge for the future of Cloud Computing. There are many areas to improve in the various security aspects of Hadoop clusters and new technologies are proposed to ensure the security in terms of reliability and flexibility. Still lot of work is remaining to make Hadoop Clusters as Full-fledged Database system in terms of user accountability and dynamic data updating. These two fields open new avenues for research and interesting work can be proposed for third party auditor for existing and new security models in Hadoop clusters.

Karthik D Aiming and et. al. The existing popular cloud disc security weakness, we put advance a security encryption methods based on Hadoop which assures the data transmission and storage security and satisfy the server executes digital signature for client data at the same time. It is a distributed encryption system that could reduce the load on the server, and lastly achieve security, stability, efficient and successful storage.

Priya P. Sharma and et. al. In Big Data Era, where data is accumulated from various sources, security is a major concern (critical requirement) as there is no fixed source of data. With the Hadoop gaining larger acceptance within the industry, a natural concern over the security has spread. A growing need to accept and assimilate these security solution and commercial security features has surfaced. In this paper we have tried to cover all the security solutionto secure the Hadoop ecosystem.

Xianqing Yu and et. al. implemented SEHadoop model that consists of SEHadoop runtime model, SEHadoopBlock Token and SEHadoop Delegation Token to improve compromise resilience of Hadoop in a public cloud. SEHadoop model enhances isolation level among Hadoop components and enforces least access privilege on Hadoop processes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Our experimental results exhibit how enhanced isolation and least access privilege of SEHadoop prevent attackers from using compromised Hadoop processes to compromise the rest of components of Hadoop. SEHadoop Block Token does not appear to inflict overhead, and SEHadoop Delegation Token has very limited performance impact. The experiment result also showed migrating Hadoop jobs to SEHadoop is straightforward.

III.PROPOSED SYSTEM

This topic focuses on a specific security model for HDFS to secure the data on various operations in its nodes. The proposal model effectively precludes the intruder prone areas operations such as data deletion, insertion, and replacement which pose a severe threat to the HDFS functionality. The Hadoop file system has been designed to meet the complex requirements of big data. The high value associated with big data sets has also rendered big data storage system attractive target for cyber attackers. In this work, we focus on a process specific security model for HDFS in big data to secure the data on various operations in nodes of HDFS. The proposed model effectively precludes the intruder prone operations such as data deletion-insertion and replacement in the cluster nodes which pose a severe threat to the HDFS functionality using Apache Sentry for authorization of Clients and Kerberos for authenticating access to the nodes. This paper deals with the various issues that Hadoop faces in terms of security and proceeds towards proposing an efficient technique. The technique proposed here for Hadoop security is based on the interacted working of the tools like Apache Sentry, Network Authentication Protocol- Kerberos and Non- Relational Database- Mongo DB. The Technique proposed is named as DPE Technique, which symbolizes Hadoop distributed file system that yearns for security, uses a creamed layering of Sentry and Kerberos for a protocol for authorizing the user access and authenticating access to the data nodes respectively.

This proposed architecture is a combination of the components like the database Mongo DB, Apache sentry that assigns roles for the user, Kerberos which is a protocol for authenticating the user access to the Hadoop Distributed File System. Figure 1 depicts the architecture of the proposed system with all the components that act to provide a highly refined secure system. [2]

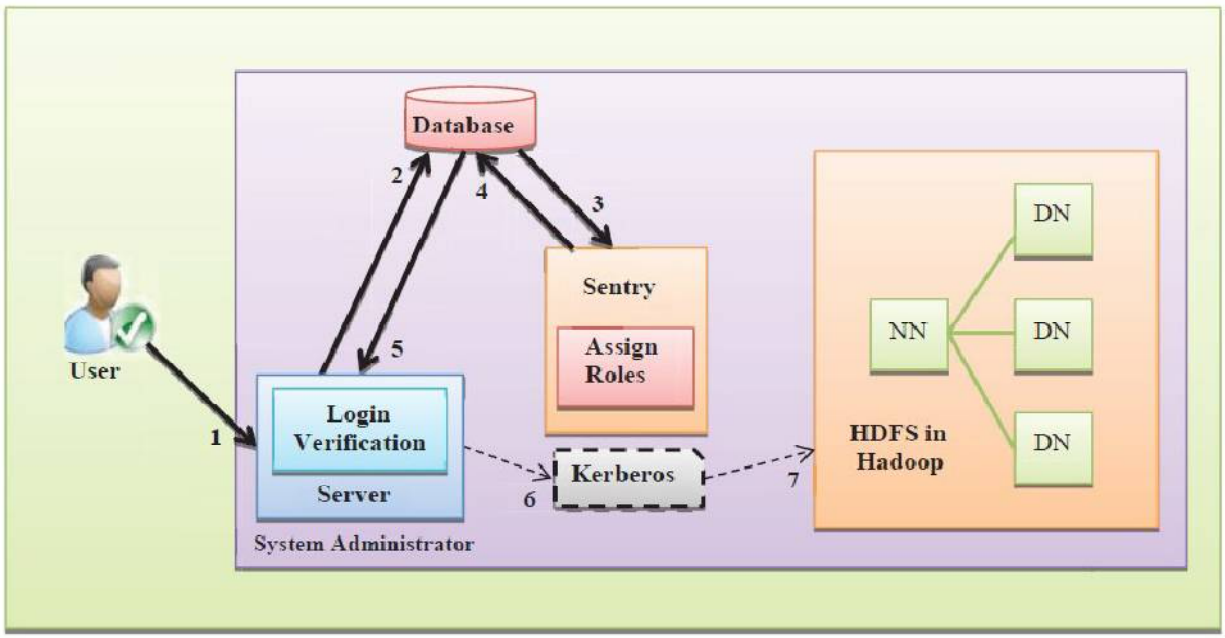

Fig.1.1: Architecture of Hadoop Kerberos security. [2]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

The next step of the working of Kerberos authentication is where the decryption of the TGT plays a major role. If the client successfully decrypts the TGT (ticket granting ticket) (i.e., if the client has entered the correct password), then it keeps the decrypted TGT. This process which involves the saving of the decrypted TGT indicates the proof of the client's identity. The TGT has a feature such that it has the ability to expire at a specified time, it permits the client to obtain additional tickets, which gives permission for a set of specific services. The requesting and granting of any of the additional tickets by the services or the system are user-transparent.

IV. PSEUDO CODE

- Step 1: Kerberos authentication is based on symmetric key cryptography.
- Step 2: The Kerberos KDC provides scalability.
- Step 3: A Kerberos ticket provides secure transport of a session key.
- Step 4: The Kerberos KDC distributes the session key by sending it to the client.
- Step 5: The Kerberos Ticket Granting Ticket limits the use of the entities' master keys.

V. CONCLUSION

Hadoop is a secure system and offers key features for securely processing enterprise data. But the security work never ends. It is working on several projects to enhance Hadoop security from the inside, shore up defenses from the outside with Apache Knox and to keep up with evolving requirements by providing more flexible authentication and authorization and by improving data protection. We are also working to improve integration with enterprise Identity Management and security systems. We can make the enterprise identity management more secure by additional security technologies like watermarking, one time password, and public key cryptography as these technologies have not been used yet.

VI. ACKNOWLEDGMENTS

For everything we get, the credit goes to all those who had helped us to complete this survey successfully. I am thankful to "Dr. S. N. Gujar for guidance and review of this paper. I would also like to thanks, all faculty members of SKN College of Engineering".

REFERENCES

1. B. Saraladevi, N. Pazhaniraja, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvan, 'Big Data and Hadoop-A Study in Security Perspective', vol 12, 2nd International Symposium on Big Data and Cloud Computing,2015.
2. Hong-ryeol Gil1, Joon Yoo1 and Jong-won Lee2,'An On-demand Energy-efficient Routing Algorithm for Wireless Ad hoc Networks', Proceedings of the 2nd International Conference on Human. Society and Internet HSI'03, pp. 302-311,2003.
3. S. Saranya, M. Sarumathi, B. Swathi, P. Victer Paul, S. Sampath Kumar, T. Vengattaraman, 'Dynamic Preclusion of Encroachment in Hadoop Distributed File System'2nd International Symposium on Big Data and Cloud Computing 2015.
4. Yannan Ma, Yu Zhou, Yao Yu, Chenglei Peng, Ziqiang Wang, Sidan Du, 'A Novel Approach for Improving Security and Storage Efficiency on HDFS', The 6th International Conference on Ambient Systems, Networks and Technologies, 2015.
5. GangChen,, SaiWu, YuanWang, 'The Evolvement of Big Data Systems: From the Perspective ofanInformation Security Application',
6. Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri, 'SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING', International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014
7. Rajesh Laxman Gaikwad, Prof. Dhananjay M Dakhane and Prof. Ravindra L Pardhi, 'Network Security Enhancement in Hadoop Clusters', International Journal of Application or Innovation in Engineering & Management (IJAIEM)
8. Karthik D, Manjunath T N, Srinivas K,'A View on Data Security System for Cloud on Hadoop Framework', International Journal of Computer Applications, page no. 4.
9. Priya P. Sharma, Chandrakant P. Navdeti, 'Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution', International Journal of Computer Science and Information Technologies, Vol. 5 (2)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

10. Xianqing Yu, Peng Ning, Mladen A. Vouk, 'Enhancing Security of Hadoop in a Public Cloud', 2015 6th International Conference on Information and Communication Systems.
11. Y. R. Mukund, Sunil S. Nayak, 'Improving false alarm rate in intrusion detection systems using Hadoop', 21-24 Sept. 2016 International Conference. Vol.3 (3).
12. Young-Ae jung, Si-jee woo, sang-soo yeo, 'A Study on Hash Chain-Based Hadoop Security Scheme', 10-14 Aug. 2015 International conference, (1).