Introduction of an Authentication Method for Securing Data in Hadoop System

J. SRIVIDYA
Associate Professor
Dept of CSE
CMR Technical Campus
Hyderabad, T.S, India

VEERNALA SWAPNA
Assistant Professor
Dept of CSE
CMR Technical Campus
Hyderabad, T.S, India

Abstract: **The increasing popularity of cloud basis Hadoop system has led to the improvisation of security. It permits users for storing and processing purpose of huge data at exceptionally low costs however it lacks security measures to carry out satisfactory authentication as well as authorization of users and services. In our work we introduce a token-based approach which provides more secured protection of distributed file system data without burdening authentication functions. The introduced authentication approach protects sensitive distributed file system data against various attacks such as impersonation and replay attacks. The system will make use of hash chain of authentication keys, rather than public key-basis authentication keys which are found in existed file systems. The scheme allows clients to be verified by data node by the use of block access token while thwarting replay as well as impersonation attacks. The proposed system will include an additional layer of security towards the traditional symmetric key Hadoop's distributed file system authentication process. The technology of elliptic curve cryptography makes authentication keys unidentified, thus protecting them against various attacks. The scheme allows clients to be verified by data node by the use of block access token while thwarting replay as well as impersonation attacks.**

Keywords: **Hadoop System, Token-Based, Distributed File System, Hash Chain, Elliptic Curve Cryptography, Authentication Keys.**

I. INTRODUCTION

The Traditional systems of data processing as well as management systems are considered for processing of structured data, as a result they are not efficient in managing of unstructured and large-scale, data which is included in Big Data. The technology of Big Data needs novel techniques that make an analysis of huge volumes of data over network. Apache's Hadoop is a software support for processing of Big Data applications. It is used by several major online media companies and it mainly allows for distributed processing of huge data sets across several clusters of computers[1]. Hadoop's distributed file system is introduced for increase from single servers to several machines, where each of the machines offers local computation as well as storage. Hadoop software permits users for storing and processing purpose of huge data at exceptionally low costs. The data within Hadoop was not sensitive and accessible to cluster might be sufficiently restricted and it lacks security measures to carry out satisfactory authentication as well as authorization of users and services.

Hadoop security controls need name node as well as data node for sharing of private key to make use of block access token. When the key is known to the attacker, the data on the entire data nodes is exposed. Hadoop Distributed File System did not provide strong security for the purpose of user authentication which has made the system communication open to eavesdropping hence we introduce an authentication approach which is of token-based protecting sensitive Hadoop's distributed file system data against various attacks such as impersonation and replay attacks[2]. The proposed mechanism will permit the clients of Hadoop's distributed file system to be validated by data node by means of block access token.

II. METHODOLOGY

In the recent times, several platforms for structuring applications of Big Data of open-source and proprietary were proposed. Among them, one is Hadoop which is an open-source software approach for processing of Big Data used by most important companies. It is an open-source platform which is made up of stand-alone modules like distributed file system known as Hadoop's distributed file system, library for processing of huge distributed datasets known as MapReduce. It stores up data files across numerous machines and store up its metadata on name node. While Hadoop matured, additional data and further variety of data containing sensitive enterprise as well as personal data are moved to Hadoop's distributed file system. Hadoop's distributed file system lax authentication permits any user to impersonate any of the other user or else cluster services. For extra authentication purpose, Hadoop offers Kerberos that make use of symmetric key operations and when the entire components of Hadoop system are authenticated by means of Kerberos, then key distribution center of Kerberos might have a restricted access. To decrease Kerberos traffic, Hadoop system depends on tokens of delegation. The methods of delegation token make use of symmetric encryption and shared keys might be distributed to several hosts based on type of token. This makes Hadoop communication exposed to eavesdropping as well as modification, hence making replay as well as impersonation attacks more possible. Hence we introduce authentication approach of token-based for protecting file system data against various attacks. The proposed method will make use of hash chain of authentication keys, instead of public key-basis authentication keys which are generally found in existed Hadoop's distributed file system [3]. The projected scheme wills permit clients to be validated towards data node by block access token. The proposed system will show communication power, computing power as well as area efficiency equal to the existed Hadoop's distributed file system regarding performance.

III. AN OVERVIEW OF PROPOSED SYSTEM

Hadoop's distributed file system is a scalable file system that store up huge data files across numerous machines and store up its metadata on name node. It is based on a master–slave structural design as a result particular master node maintains the entire files within file system. The name node will manage the namespace and control access to files by means of clients. Name node will maintain track of which blocks must to be replicated and begin replication at any time necessary. Several copies of every file offer data protection as well as computational performance. The data nodes are accountable for storage of application data and provide read or write requests from clients. The data nodes moreover carry out block formation, removal as well as replication upon instruction from name node. Map Reduce is structure for processing of huge data sets across Hadoop cluster. Like Hadoop's distributed file system, it is on basis of master–slave representation. The master is particular node that coordinates activity among numerous worker nodes. The master obtains input data that needs to be processed. The input data is divided to minute chunks and these chunks are processed in parallel on numerous worker nodes which is known as Map phase. The workers will send their results back towards master node that gathers these results to generate the sum total which is known as Reduce phase. MapReduce functions on key-value pairs and specifically the entire input as well as output in MapReduce is within key-value pairs. Previous versions of Hadoop did not provide importance to security, hence this framework was continued to make the modification of security [4]. Particularly, Hadoop Distributed File System on which modules that is build did not present strong security for the purpose of user authentication which has made the system communication open to eavesdropping as well as modification, hence making replay as well as impersonation attacks more possible. For this reason we have introduced an authentication approach for protecting file system data against various attacks. Differing from most of the Hadoop's distributed file system authentication protocols implementing methods of public key exchange, the proposed system make use of hash chain of keys. In the proposed authentication scheme of Hadoop's distributed file system, delegation tokens are created by means of elliptic curve cryptography [5]. The proposed system develops protection system of authentication information that is exchanged between name node as well as data node. The scheme is considered to function in the design of master-slave which needs an operation environment in which single master node will coordinate activity among numerous slave nodes. The proposed system permits the clients of Hadoop's distributed file system to get verified by data node by means of block access token [6]. The system will show the performance in terms of communication power as well as area efficiency as good as existed Hadoop's distributed file system. The proposed system makes use of delegation tokens to append an additional layer of protection to existing Hadoop's distributed file system authentication.

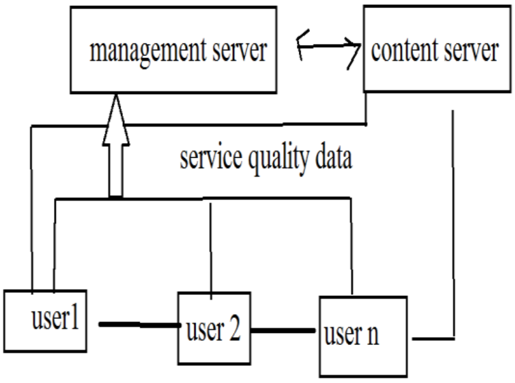

Fig1: proposed authentication scheme

IV. CONCLUSION

By the increased growth of social networks as well as smart devices, the usage of Big Data services has improved significantly. Hadoop is an open-source software approach for processing of Big Data used by most important companies but the system lacks security measures to carry out satisfactory authentication as well as authorization of users and services hence we introduce an authentication approach which is of token-based protecting distributed file system data against various attacks. The system will make use of hash chain of authentication keys, instead of public key-basis authentication keys. The proposed mechanism will authorize clients of distributed file system to be verified by data node by means of block access token. It includes an additional layer of security towards the traditional symmetric key distributed file system authentication process and makes authentication keys unidentified, thus protecting them against various attacks. In projected system, tokens are produced on the basis of elliptic curve cryptography and enhance security of verification messages communicated among name node and data node. The proposed system will show communication power, computing power as well as area efficiency equal to the existed systems regarding performance.

V. REFERENCES

[1]. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: Proceedings of Sixth Symposium on Operating System Design and Implementation (OSDI04), pp. 137–150 (2004).

[2]. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The Hadoop distributed file system. In: Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10 (2010).

[3]. Vatamanu, C.,Gavilut,D., Benchea, R.M.: Building a practical and reliable classifier for malware detection. J. Comput.Virol.Hacking Tech. 9(4), 205–214 (2013).

[4]. Lee, S.H., Lee,D.W.:Current status of BigData utilization. J. Digit. Converg. 11(2), 229–233 (2013).

[5]. White, T.: Hadoop The Definitive Guide, 2nd edn, pp. 41–47. O'Reilly Media, Sebastopol (2009).

[6]. Condie, T., Conway, N., Alvaro, P., Hellerstein, J.M., Elmeleegy,K., Sears, R.: MapReduce Online. In: Proceedings of NSDI'10 (2010).