DEVISE AND ACCOMPLISHMENT OF HADOOP DISTRIBUTED FILE SYSTEM BY USING ELLIPTIC CURVE CRYPTOGRAPHY FOR AMPLIFY THE REFUGE: A STUDY

Regonda.Nagaraju[1]
[1]Department of Computer Science and Engineering,
[1]Sreyas Institute of Engineering and Technology, Hyderabad,Telangana,India.

ABSTRACT:-

In recent years, a number of platforms for building Big Data applications, both open-source and proprietary, have been proposed. One of the most popular platforms is Apache Hadoop, an open-source software framework for Big Data processing used by leading companies like Yahoo and Facebook. Historically, earlier versions of Hadoop did not prioritize security, so Hadoop has continued to make security modifications. In particular, the Hadoop Distributed File System (HDFS) upon which Hadoop modules are built didn't provide robust security for user authentication. This paper proposes a token-based authentication scheme that protects sensitive data stored in HDFS against replay and impersonation attacks. The proposed scheme allows HDFS clients to be authenticated by the data node via the block access token. Unlike most HDFS authentication protocols adopting public key exchange approaches, the proposed scheme uses the hash chain of keys. The proposed scheme has the performance (communication power, computing power and area efficiency) as good as that of existing HDFS systems.

KEYWORDS: -HDFS(Hadoop Distributed File System), BIG DATA,HASH CHAIN.

1.INTRODUCTION

Hadoop Distributed File System (HDFS) has master/slave architecture. An HDFS cluster consists of a single name node, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of Data Nodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files.

Internally, a file is split into one or more blocks and these blocks are stored in a set of Data Nodes. The Name Node executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to Data Nodes.

The Data Nodes are responsible for serving read and write requests from the file system's clients. The Data Nodes also perform block creation, deletion, and replication upon instruction from the Name Node.

The Name Node and Data Node are pieces of software designed to run on commodity machines. These machines typically run a GNU/Linux operating system (OS). HDFS[1] is built using the Java language; any machine that supports Java can run the Name Node or the Data Node software. A typical deployment has a dedicated machine that runs only the Name Node software. Each of the other machines in the cluster runs one instance of the Data Node software.

The existence of a single Name Node in a cluster greatly simplifies the architecture of the system. The Name Node is the arbitrator and repository for all HDFS metadata. The system is designed in such a way that user data never flows through the Name Node

1.1 EXISTING SYSTEM

Data is currently maximum crucial belongings of any organization in every field with which many operations may be completed. The continuous growth within the importance and volume of records has created a brand new hassle: it cannot be dealt with by means of conventional evaluation strategies. This hassle was, therefore, solved through the advent of a new paradigm: Big Data. Big Data originated new troubles associated with records safety and privativeness. This survey paper consists of a few security issues in addition to privativeness issues in Big statistics. Big Data is often heterogeneous, noisy, dynamic & untrustworthy. Nevertheless, even noisy Big Data could be extra treasured than tiny samples because standard statistics obtained from common styles and correlation evaluation normally overpower person fluctuations and regularly disclose greater dependable hidden styles and know-how. Further, interconnected Big Data bureaucracy massive heterogeneous facts networks, with which information redundancy can be explored to atone for

missing facts, to crosscheck conflicting cases, to validate straightforward relationships, to disclose inherent clusters, and to discover hidden relationships and models. Big data analysis is the technique of applying advanced analytics and visualization strategies to large information sets to discover hidden styles and unknown correlations for powerful decision making. The evaluation of Big Data entails more than one wonderful levels which consist of records acquisition and recording, facts extraction and cleaning, information integration, aggregation and representation, question processing, statistical modeling and analysis and Interpretation. Each of those stages introduces challenges.

2.PROPOSED SYSTEM

Earlier versions of Hadoop did not prioritize security, so Hadoop has continued to make security modifications. In particular, the Hadoop Distributed File System (HDFS) upon which Hadoop modules are built didn't provide robust security for user authentication. This paper proposes a token-based authentication scheme that protects sensitive data stored in HDFS against replay and impersonation attacks. The proposed scheme allows HDFS clients to be authenticated by the data node via the block access token. Unlike most HDFS authentication protocols adopting public key exchange approaches, the proposed scheme uses the hash chain of keys. The proposed scheme has the performance (communication power, computing power and area efficiency) as good as that of existing HDFS systems.

A token-based authentication scheme is proposed that protects sensitive data stored in HDFS against replay and impersonation attacks. The clients of HDFC could be authenticated using the block access token by data node.

Unlike most HDFS authentication protocols adopting public key exchange approaches, the proposed scheme uses the hash chain of keys. The proposed scheme has the performance (communication power, computing power and area efficiency) as good as that of existing HDFS[2] systems. Apache Hadoop provides strong authentication for HDFS data.

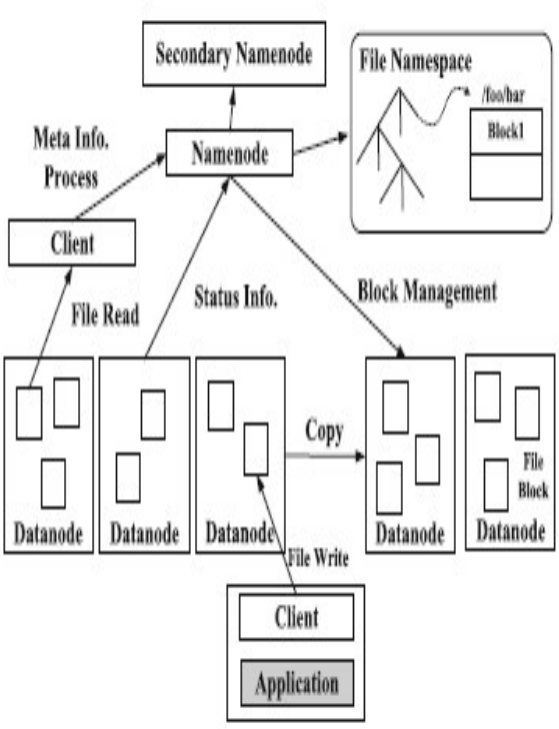

Fig:i- HDFS Architecture

All HDFS accesses must be authenticated:

1. Access from users logged in on cluster gateways
2. Access from any other service or daemon (e.g. Catalog server)
3. Access from Map Reduce tasks

In this approach **ECC**[3] is an approach to public-key cryptography based on the algebraic Structure of elliptic curves over finite fields.

ECC requires smaller keys compared to non-ECC cryptography[4] to provide equivalent security.

Elliptic curves are applicable for encryption, digital signatures, pseudo-random generators and other tasks.

They are also used in several integer factorization algorithms that have applications in cryptography.

2.1 HASH CHAIN

A hash chain is the successive application of a cryptographic hash function to a piece of data. A hash chain is a method to produce many one-time keys from a single key or password.

A server which needs to provide authentication may store a hash chain rather than a plain text password and prevent theft of the password in transmission or theft from the server.

A server begins by storing $h^{1000}(password)$ which is provided by the user. When the user wishes to authenticate, he supplies $h^{999}(password)$ to the server. The server computes h $(h^{999}(password)) = h^{1000}(password)$ and verifies this matches the hash chain it has stored. It then stores $h^{999}(password)$ for the next time the user wishes to authenticate[7].

An eavesdropper seeing h999 *(password)* communicated to the server will be unable to re-transmit the same hash chain to the server for authentication since the server now expects h^{998} *(password).* Due to the one-way property of cryptographically secure hash functions, it is infeasible for the eavesdropper to reverse the hash function and obtain an earlier piece of the hash chain.

In this example, the user could authenticate 1000 times before the hash chain is exhausted. Each time the hash value is different, and thus cannot be duplicated by an attacker.

Big statistics describes huge information sets which have greater various and complex shapes like weblogs, social media, e-mail, sensors, and snapshots. This plenty less set up information and forte trends from traditional databases normally associated with greater headaches in storing, reading and applying similar strategies or extracting consequences. Big records analytics is the method of examining incredible portions of complex information on the manner to find out unseen patterns or spotting furtive correlations. Since conventional databases systems cannot be used to the method the massive information, it poses several demanding conditions to the studies community. Security and privacy[6] are the crucial problems with facts. However, there exists incongruity some of the Big facts safety and privacy and the massive use of huge information. This paper offers insights on evaluating of huge facts, related worrying conditions, privatives and safety troubles and the differentiation among privatives& protection[5] requirements in huge records. Also, we cantered on numerous privacy fashions which may be stretched to massive statistics area, reading the blessings and downsides of Data anonymity privacy models.

3.REFERENCES

[1] D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, pp. 2127-2130,Dec-2001.

[2] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, ''Information safety in large information: Privacy and information mining,'' in IEEE Access, vol. 2, pp. 1149–1176, Oct. 2014.

[3] H. Hu, Y. Wen, T.-S. Chua, and X. Li, ''toward scalable structures for huge information analytics: A generation

educational,'' IEEE Access, vol. 2, pp. 652–687, Jul. 2014.

[4] Dr. Regonda. Nagaraju, Reddygari Aishwarya Reddy, S. Yamini, P. Mythry, ''Hybrid Encryption with Verifiable Delegation using Cloud Computing,'' International Journal of Research in Electronics and Computer Engineering, vol. 7,issue 1, no. 2, pp. 2448–2450, January-March 2019.

[5] C. Hongbing, R. Chunming, H. Kai, W. Weihong, and L. Yanyan, ''Secure huge information garage and sharing scheme for cloud tenants,'' China Commun., vol. 12, no. 6, pp. 106–115, Jun. 2015.

[6] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, ''Privacy-keeping multi- key-word ranked search over encrypted cloud records,'' IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 1, pp. 222–233, Jan. 2014.

[7] O. M. Soundararajan, Y. Jenifer, S. Dhivya, and T. K. P. Rajagopal, ''Data safety and privacy in cloud the usage of RC6 and SHA algorithms,'' Netw. Commun. Eng., vol. 6, no. Five, pp. 202–205, Jun. 2014.