RESEARCH ARTICLE OPEN ACCESS

An Adaptive Authentication Based on Blockchain for Bigdata Hadoop Framework

Mithun Kankal[1], Pramod Patil[2]
1(Department of computer engineering, Dr. D. Y. Patil Institute of TechnologyPimpri, Pune-41118, India)
2(Department of computer engineering, Dr. D. Y. Patil Institute of TechnologyPimpri, Pune-41118, India)

Abstract:

The authentication protocols for giant information system like Apache Hadoop relies on Kerberos. In the Kerberos protocol, there are variety of security problems that have remained unsolved like single purpose of failure, replay attacks, DDoS and are some examples. These indicate potential security vulnerabilities and massive information risks in victimization Hadoop. Here we intended to presents drawbacks of Kerberos implementations and identifies authentication needs that may enhance the security of huge information in distributed environments. The enhancement planned relies on the rising technology of block chain that overcomes shortcomings of Kerberos.

Keywords — **Big Data, Distributed Network, Authentication, Hadoop, Security, Blockchain, DecentralizedAuthentication**.

I. INTRODUCTION

Security of Big Data become important because of sensitive data exchange increasingcontinuously. Big Data is the collection large amount of structured (relational databases) and unstructured data (document files, images, video) complex data. Data are being collected fromindependent multitude independent, where data been used and analyzed to generate knowledge. Enterprises and organizations use that knowledge to make corporate decision making processes optimal, predict future trends and more. Hence, the data and its analyzed outcome are a valuable asset in today's economy. The people needs benefits of Big Data, privacy and security of Big Data, stored on distributed or cloud storage, has become an important issue. The Concerns has focused on security and protection of sensitive data or information, where it's related to new threats to information security and adopting existing traditional security measures is not adequate. Cloud Security Alliance (CSA) published a document that lists the top ten challenges to protecting Big Data systems and one of the ten challenges stated and most critical is granular access control.

The current authentication system of Apache Hadoop exposes the entire Big Data solution to a security issue due to Kerberos' system vulnerabilities. Limitations of Kerberos are evident in version 4 and early drafts of version 5; replay attacks, key exposure and time synchronization are vulnerabilities identified. It present the structural drawbacks and identify authentication requirements that can enhance security of BigData in distributed environments. Generally, the requirement of authentication process are policies that determines how a user must authenticate before access is granted to a protected service. It improves policies by enabling disintermediation of a third-party entity, empowered user control, durability and reliability of an authentication protocol. The intent is to take advantage of blockchain technology such as password-less/keyless authentication, data encryption and a lack of need for a third party or a central database. Blockchain authentication was first introduced in Bitcoin for verification of transaction.

II. LITERATURE SURVEY

A. Typical Hadoop Cluster:

Hadoop Cluster are normally any set of tightly connected or loosely connected computers that work together as a single system is called Hadoop Cluster. In simple words, a cluster of computer used to deploy Hadoop is called Hadoop Cluster.

Hadoop cluster is a type of computational cluster designed for storing and analyzing large amount of

structure and unstructured data in a distributed computing environment. These clusters run on commodity computers which can available at low cost.

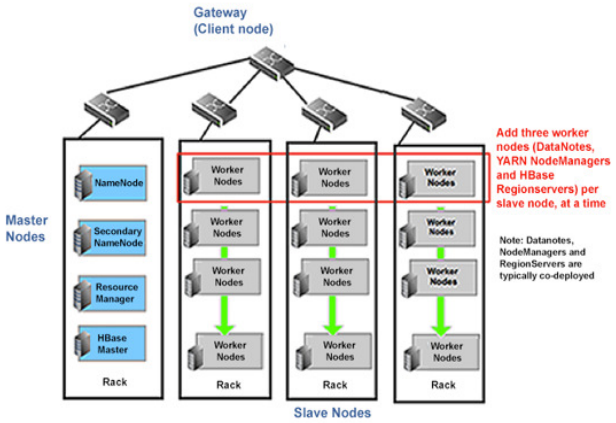

Fig. 1Typical Hadoop cluster.

Below are the Hadoop components that together form a Hadoop ecosystem

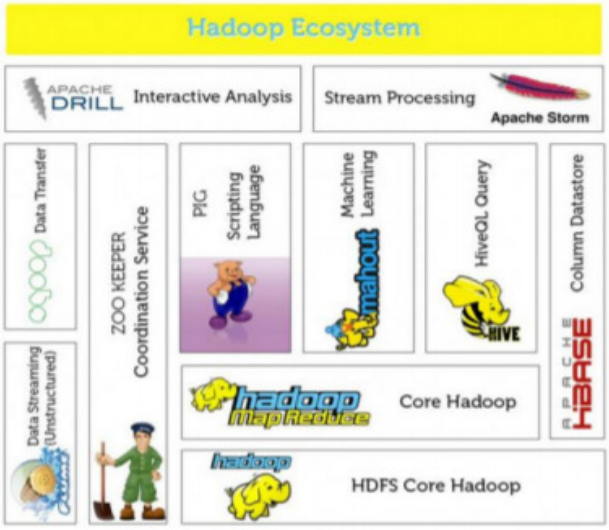

Fig. 2 Hadoop Ecosystem.

B. Kerberos:

Kerberos is an authentication system developed by MIT as part of athena project. Kerberos uses a trusted third party server or a middle man server, for authentication process and kerberos is based upon needham-schroeder-protocol.

Kerberos, a client which can be either a service or a user which sends a request for a ticket to the Key Distribution Center (KDC). The KDC creates a TGT (ticket-granting ticket) for the client, encrypts it using key which is client's password, and sends the encrypted TGT sent back to the client. The client then decrypt the TGT, using its password. If the client successfully decrypts the TGT then, it keeps the decrypted TGT, which indicates proof of the client's identity. Following figure shows the working and components involved in Kerberos.

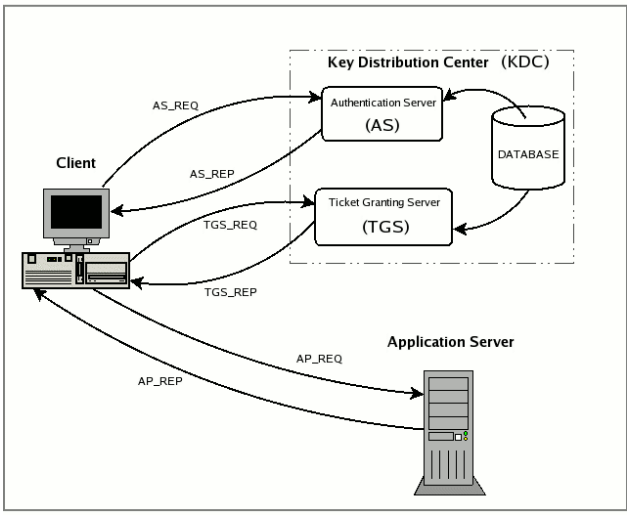

Fig. 3Working of Kerberos.

C. Kerberos and Hadoop Cluster

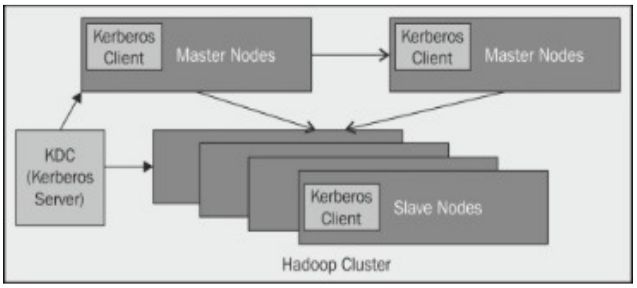

Fig 4. Hadoop Cluster and Kerberos.

D. Challenges of Current Authentication mechanism

- Keys Exposure
- Single Point of Failure
- Time Synchronization
- Denial-of-Service (DoS) Attacks
- Session Key

III. PROPOSED METHODOLOGY

Big Data is a complex distributed system where the main challenge is the complexity of managing a large implementation, new approaches to security are required. Authentication and data access control

should be managed by a strong authentication, flexible, scalable and decentralized that deny any malicious user from getting access to Big Data servers. Hence, new methodology need to overcome the shortcomings of security flaws in existing implementation. This section briefly discusses what is block chain and new approaches that can enhance authentication of Big Data.

A blockchain, originally block chain, is a continuously growing list of records which are linked and secured using cryptography. It is also called as blocks. Each block contains a cryptographic hash of the previous block, transaction data and timestamp of transaction.

Following fig shows where the block chain would be fit along with Hadoop ecosystem.

A. *Architecture:*

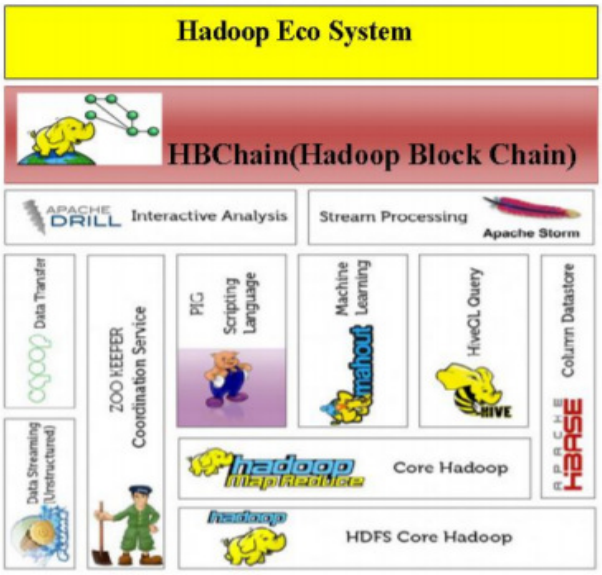

Fig 5. Blockchain layer on Hadoop ecosystem

B. *Approaches:*

The Methodology based on creation of new HDFS client/gateway interface using existing API of Hadoop ecosystem and new custom API based on Blockchain decentralized distributed database. The integration of Hadoop ecosystem and blockchain is major challenge for this implementation as blockchain is still evolving around the identification and authentication areas. The HDFS Client interface could be build using python/java along with new solidity languages of blockchain or blockchain using python and the user authentication data would be stored in blockchain decentralized distributed database. Following block chain approaches would be used to overcome the shortcoming of Kerberos systems.

- ***Decentralized Authentication***

Decentralized authentication replaces authentication mechanism which is based on username/password generated keys and the client-side SSL certificate with elliptic curve cryptographic generated keys; this is the same method used in a blockchain protocol. It removes central databases where user information is stored and managed centrally, which is vulnerable to hackers who compromise entire credentials. In this authentication the user password is only used in the user's own machine to access the private key.

The private key is never transferred/revealed through network or server and cannot be exchanged over a side channel between the server and client. This authentication protocol is based digital signature which uses irrefutable identity verification based on public keys. A user is verified when transaction or message was assigned by an approved private key. It is inferred that the exact identity of the owner is irrelevant if whoever has access to the private key is the owner

- ***Unbreakable Record***

Blockchain technology is a new type of database, which can be shared and used by a group of non-trust parties withoutrequiring a central administration, unlike SQL or Oracle or NoSQL databases. Blockchain is a type of distributed database that maintains a chain of record which can grow irreversibly and each record in ordered chain list is called blocks. Each block contains a timestamp, transaction data and a link to a previous block.

The records are impossible and impractical to computationally alter or reverse; it would be possible to protect transactions from theft and fraudulent actions. By using the principle of hash and block, the data cannot be changed once data has been written to blockchain. The administrator or users of data would not able to change or delete the existing chain of blocks. Every copy of block in blockchain in the network, has to be the same throughout the network. Then, consensus is achieved by using a proof-of-work protocol in the mining process. A proof-of-work protocol is a

piece of data, which is difficult to produce but easy to verify by others user. This makes blockchain distributed database suitable for recording sensitive information, such as identity, medical and monetary information.

- ***No Session Keys***

Using SIN protocol is considered as safer than session key sharing across network in existing authentication protocol like Kerberos. The SIN can be shared with everyone openly, as its corresponding private key is secured/stored on the client-side and it never transmitted in network over the wire, and not shared with any user or entity. During authentication mechanism, server verifies or validate a user by user-shared public key against their digital signature and the SIN user shared previously. It confirms that, SIN previous nonce in blockchain block record to prevent against replay attacks and subsequently authenticate the user request. The benefits of using SIN in identification mechanism is its portability, where the same identification method can be used on multiple devices without exposing users session key and credentials across network.

- ***Zero Single Point of Failure System***

Blockchain is a decentralized and distributed database or data storage technology that maintains a chain of block or records which continuously grow in ordered manner. It removes the risks with data which stores centrally and reduces the vulnerability of single point failure or vulnerability of network hackers. Every blockchain server or node which connected in blockchain network contains the copy of the blockchain. Quality of data is maintained by massive replication of database and is cryptographically trusted. Blockchain used for user authentication in a system creates an un-hackable and tamper-proof digital identity. It potentially reduces effectiveness of phishing attacks.

The decentralized and distributed nature of the blockchain network would make it impossible for the infrastructure to fail under an excess of requests. Hence, the authentication method based on blockchain technology, it would be immune to DDoS attacks and un-hackable.

- ***Prevent Data Theft***

The increase in the amount of data theft and hacking incidents that has caused consternation concerning the accessing of personal sensitive information, in particular financial data such as bank accounts details, credit cards andhealth records or medical records. Petland, who are the Professor in MIT has explored blockchain to build Enigma, which could potentially allow blockchain distributed databases to retain sensitive information and process it without risking exposure to hacker or malicious parties. Enigma is described as a peer-to-peer network, enabling different user or organizations to jointly store and run computations on data while keeping the data completely private. The blockchain technology makes it harder to break into a system rather than the technology that does not completely hinder theft. The complete implementation and infrastructure of blockchain enhances privacy, freedom and security of conveyance of data.

- ***Blockchain Algorithms****:*

Consensus Algorithm: Blockchain uses a proof-of-work algorithm for reaching a consensus. The value must be smaller for cryptographic hash function of each block in order to be considered value. For this a nonce is included in the block. By using the proof-of-work method, in order to change the data in one block, a huge amount of calculation is necessary and all successors of that block must be re-written. In addition, the shorter ones would be discarded at the situation of branches of the chain whereas the longest chain would be accepted by the network. This method or process makes the data in blocks practically unmodifiable or un-hackable, and moreover the harder the processing of overwriting the data where more blocks built upon the block in which the data is contained.

However, the blockchain can also use other methods or process for consensus. For example, a blockchain may useother method than hash function such as Scrypt for proof-of-work algorithm. In addition, the blockchain could also be extended for scientific computation where a correct solution to a certain problem could act a validation method. In this way, the computation power may be used in scientific researches to help solving scientific problems.

Secure Identity Number: is the unique record identifier by which this identity will be known. A SIN record is a series of key value pair or hashes, validated by MPK digital signature. Each SIN

record each record has stable hash value due to user of stable binary encoding. SIN generates cryptographic key pair according to the following calculation:

Base58check(0x0F + 0x02 + ripemnd160(sha256(k1))

k1 is user public key from an ECDSA secp256k1 key pair,0x0F is the special byte for SINs and 0x02 is the type ofSIN.

IV. RESULT AND DISCUSSIONS

- ***Kerberos Implementation:***

The need and application of Big Data has presented a wide variety of security challenges because of Kerberos. Below challenges are indicated for Kerberos before a blockchain solution is presented.

A. Password-based Authentication

The Kerberos commonly uses the password-based authentication protocol. During the initial communication phase with Key Distribution Center (KDC) the user session key is formed by data encryption using the user's password which is entered by user during authentication process. Unfortunately, in case of Kerberos the password could be easier to break as it does not uses any password policy and as has been shown in many cases, passwords are relatively easy to break. For example, breaches have occurred due all other passwords in an affected database where they were encrypted with the same key. The problem with the encryption algorithm is that it does not handle identical plaintexts, which results in similar passwords being encrypted into similar ciphers. Disclosure of Key Distribution Center (KDC) passwords allows attackers to capture all user's credentials, which in turn, makes Hadoop's security to be useless.

B. Keys Exposure

Kerberos makes use of random session and secret keys for authentication of transactions. These keys are commonly stored in the user's workstation and used to prove the identity and verify the authenticity of messages. The secrecy of keys is critical to the integrity of the Kerberos system. If one of the keys is disclosed, then anything authenticated by that key cannot be trusted. This allows the attacker to read and modify any data passed over a connection.

C. Single Point of Failure

Centralized server and services of Kerberos relies on having a single point of failure. When the KDC is out of service, the entire big data system would suffers from Kerberos unavailability. If any attacker has access to the Key Distribution Center (KDC), the entire Kerberos authentication infrastructure is compromised and the attacker can gain root access to the database of encrypted passwords. The attacker also has access to the Kerberos software and configuration files, both of which the attacker can modify to make the system perform authentication differently. This issue highlights threats in Kerberos, including protection against an administrator who has the privilege to install hardware/software key loggers or malware to steal a user's credentials and other sensitive data (e.g., password, session keys and data).

- ***An Authentication using Blockchain Implementation:***

A creation of new HDFS client/gateway interface using existing API of Hadoop ecosystem and new custom API based on Blockchain decentralized distributed database. The integration of Hadoop ecosystem and blockchain is major challenge for this implementation as blockchain is still evolving around the identification and authentication areas. The HDFS Client interface could be build using python/java along with new solidity languages of blockchain or blockchain using python and the user authentication data would be stored in blockchain decentralized distributed database.

- ***comparison:***

	Kerberos	**Blockchain**
Authentication Type	Centralized	Decentralized
Authentication Mechanism	Password based	Password less
Session Key	Time based session Key authentication	No Session Key
Failure	Single Point Failure	Decentralized
Exposure to attacks	Brute Force, DDoS etc.	Unbreakable/un-hackable

V. CONCLUSIONS

Here, it represents the common security problems associated with Kerberos. The Kerberos

uses in large network such as the internet and increasingly used in variety of systems such as Big Data environment where security vulnerability is common and it is highly vulnerable due to shortcoming of Kerberos. Here the Kerberos limitations have been addressed. New solutions would be needed to big data environment in an era where greater security requirements needed as integration of data from different system with big data system increase rapidly. Blockchain technology, first introduced by Bitcoin, have provided a scalable solution tomany common security issues faced as Big Data become common.

Existing authentication mechanism using Kerberos would position Big Data systems to depend on many security risks and vulnerabilities. Several enhancements of authentication protocols using blockchain in the distributed environment has decentralized infrastructure that is scalable and reliable and has no single point failure.

Therefore, the utilizing the advantages of blockchain technology could be leveraged to harden security systems, including distributed authentication and no single point failure of Big Data system due to centralized servers as mechanism is based on distributed technique. It is needed as new identity system and authentication framework based on blockchain technology.

REFERENCES

1. *"CSA Releases the Expanded Top Ten Big Data Security & Privacy Challenges: Cloud Security Alliance." [Online]. Available: https://cloudsecurityalliance.org/media/news/csa-releases-theexpanded- Top-ten-big-data-security-privacy-challenges/. [Accessed: 19-Jan-2016].*

2. *"Welcome to ApacheTM Hadoop®!" [Online]. Available: https://hadoop.apache.org/. [Accessed: 12-Jan-2016].*

3. *S. M. Bellovin and M. Merritt, "Limitations of the Kerberos Authentication System," SIGCOMM Compute Common Rev, vol. 20, no. 5, pp. 119–132, Oct. 1990.*

4. *D. Davis and D. E. Geer, "Kerberos Security with Clocks Adrift." in USENIX Security, 1995.*

5. *R. M. Needham and M. D. Schroeder, "Using Encryption for Authentication in Large Networks of Computers," Commun ACM, vol. 21, no. 12, pp. 993–999, Dec. 1978.*

6. *D. E. Denning and G. M. Sacco, "Timestamps in Key Distribution Protocols," Commun ACM, vol. 24, no. 8, pp. 533–536, Aug. 1981.*

7. *S. Nakamoto, "Bit coin: A peer-to-peer electronic cash system," Consulted, vol. 1, no. 2012, p. 28, 2008.*

8. *B. C. Neuman and T. Ts'o, "Kerberos: an authentication service for computer networks," IEEE Commun. Mag., vol. 32, no. 9, pp. 33–38, Sep. 1994.*

9. *"Intel-hadoop/project-rhino," GitHub. [Online]. Available: https://github.com/intel-hadoop/project-rhino. [Accessed: 23-Mar- 2016].*

10. *"Lightweight Directory Access Protocol," Wikipedia, the free encyclopedia. 20-Mar-2016.*

11. *https://101blockchains.com/consensus-algorithms-blockchain/*