International Journal of Advance Research in Science and Engineering
Vol. No.4, Issue 08, August 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

ENHANCED SECURITY MEASURES FOR HADOOP IN CLOUD COMPUTING

Suparna Gaur

Computer Science and Engineering, UIET, Kurukshetra University, (India)

ABSTRACT

As the cloud computing technology is prospering day by day, the hadoop framework of cloud computing has been widely adopted on a large scale for storage as well as processing of massive data of users arriving from various sources. Therefore, the security aspect of hadoop platform is the major issue of concern in today's scenario in order to protect user's data from any unauthorized access. The hadoop framework in its current form uses the OS authentication system to access its services. We propose here a methodology to enhance the authentication and authorization system of hadoop for its users and authorization between services. In its current form hadoop creates a delegation token on successful authentication and not the kerberos ticket. Now, if this job sends request to other services, then it does not have the kerberos ticket to send with the job for authorization. So we implement kerberos so that we use the JobConf object to pass the ticket to the job which can further be used for accessing other services within or outside hadoop. To add to the security, we convert the TGT into a json string before passing. Next we propose a new CP-ABE level based scheme to assign each user with a levelbased on his attributes. This level when compared with the file access key, defines the level of user's access to the file.

Keywords*: *Cloud Computing, Cp-Abe, Encryption, Hadoop, Kerberos, Security.

I. INTRODUCTION

Cloud Computing is one of the latest and emerging technologies in the field of Information Technology today. It is a network based computing paradigm which basically facilitate the storage and usage of services over the internet. Cloud Computing is a practical and new approach of computing in which dynamically scalable and virtualized resources are provided over the internet. . In simple words, it is a subscription based service which allows its users to store and access data and programs over the internet rather than the hard disk of our physical computer on a pay-for-use basis according to their needs. Therefore it is also termed as On-Demand Computing. Many experts expect that cloud computing will reshape the IT market place in the coming future. The cloud users can access storage and application development platforms over the internet with the help of variety of devices like laptops, smart phones, personal computers etc by the services offered by various cloud providers.

Hadoop is an open source java based programming framework that processes large datasets called Big Data in distributed environment efficiently. This is the project of Apache and managed by the same software company hence referred as "Apache Hadoop". Hadoop is the creation of Mike Caferella and Doug Cutting in the year 2005. At that time, Doug Cutting was working in Yahoo and he named this framework as Hadoop after the stuffed toy elephant of his son. Map reduce software framework by Google is the inspiration behind hadoop.The actual project that evolved to hadoop was a software called Nutch. It was based on java and other languages and

which was used for building a web search engine that is open source.The main feature of hadoop is its resiliency i.e processing is redirected to other nodes if any node fails, other features are Scalability i.e more nodes can be added to cluster whenever required, Cost effective as it runs on clusters of commodity hardware, Flexiblity i.e any variety of data can be stored and processed over hadoop and it is simple and easily accessible too. Its Contents are: Hadoop Distributed File System (HDFS), MapReduce, Yet Another Resource Navigator (YARN) and Hadoop Common.

HDFS: Hadoop Distributed File System, It uses master-slave architecture in which the master node called Namenode controls over all slave nodes called DataNodes. It continues processing even in case of any node failure as the work is redirected to some other location of data.

MapReduce: It includes distributed processing of data on clusters of commodity hardware. It has a master node called Job Tracker and some slave nodes called Task Tracker. Its primary goal is to split the large input datasets into small chunks called blocks or fragments that can be assigned to any node in the cluster and processed parallely.

Yarn: Yet Another Resource Navigator, It allows resources like CPU time, storage and memory to be assigned to the applications running on hadoop. Using this, multiple applications in hadoop can share a common resource management. It basically removes the dependency of hadoop on map reduce.Earlier only the map reduce applications could run on hadoop,Yarn allows other applications to run on hadoop as well.

Hadoop Common: It is the set of utilities and libraries whose work is to provide support to rest all hadoop modules.

Hadoop Security

Hadoop runs in non secure mode bydefault and it doens't require authentication in this mode. But when hadoop is configured to run in secure mode, each user or service accessing hadoop services are needed to be authenticated properly.Current hadoop system doesn't have a strong authentication mechanism or convincing method of security. If you are having the block location details, then there is no need to communicate with NameNode in order to get the blocks because blocks can be asked directly by communicating with the datanode. Hence the overall objective of study is to identify the loopholes in the system and suggesting measures to overcome them.

Rest of the paper is structured as follows. In Section 2 related work is discussed. Section 3 describes the problem formulated after going through the related work. Section 4 describes the proposed methodology for the problem formulated and finally section 5 concludes this paper.

II. RELATED WORK

B. Clifford Neuman *et al.* [1] stated that multiple users can use the services provided by modern computer systems which require the user's identity to be verified accurately. Authentication is the process of verifying user's identity that generated some data and this process is significant for security of computer systems. In traditional systems, password based authentication was used for verification but these passwords can be intercepted by eavesdroppers when sent across computer networks. Kerberos is a strong authentication method where authentication is based on cryptography and this method is suitable for insecure environments as it hides passwords efficiently.

John Bethencourt *et al.* [2**]** stated that in most of the distributed systems, if a user had some credentials only then he was allowed to access data. This policy could be applied by deploying a trusted server in order to store the data. This was the one possible method. But this method was not so effective because CIA of whole data would be at risk if server storing the data compromised. To overcome this issue, author proposed a CP-ABE scheme to be applied on encrypted data so as to provide an access control over the data stored on the server. Now, even if the server is not trusted or is compromised, still the stored is secure. This method secures the data from collusion attacks too. In the proposed system, credentials of users are based on some attributes and decryption of data is based on some policy which is decided by the one encrypting that data while the previous systems encrypted data was based on some attributes user key's were associated with policies. Proposed method is implemented by the authors and it is somewhat similar to Role-based access control. The limitation of the proposed scheme is less efficient and it is proved to secure under generic group heuristics.

Bao Rong Chang *et al.* [3**]** defined cloud computing as a popular and emerging paradigm in today's time, providing its users with on demand services through internet. In green energy point of view, four aspects of cloud computing are large data storage, low cost, reliability and efficiency. Building hadoop cloud computing is the motive of the paper hence authors introduced hadoop platform in cloud computing alongwith access security with the help of techniques like face recognition and finger print identification. Lots of mobile device are served either through wired or through wireless medium by hadoop cloud computing. Additionally lots of slave nodes are connected and controlled by a master node in the architecture defining hadoop system in which services like SaaS i.e software as a service, PaaS i.e platform as a service and IaaS i.e infrastructure as a service is provided by cloud computing. In this paper, cloud system's effectiveness and efficiency towards access security is successfully verified in 2.2 seconds using face recognition and finger print identification techniques in hadoop system. Linux platform connected to hadoop with the help of Ethernet or Wifi in order to maintain a connection between client and server.

Yanli Ren *et al.* [4**]** created a constant size CP-ABE scheme. The scheme achieved complete security without random oracles. The scheme admitted threshold decryption policies based on an identity-based encryption scheme. In a ciphertext-policy ABE (CP-ABE) scheme, an encryptor could express any access policy, and define what type of receivers would be able to decrypt the message in the encryption algorithm. Normally in CP-ABE schemes, the size of cipher texts was not constant, which depended linearly on the number of attributes involved in the policy for that cipher text. The only constant size CP-ABE scheme was selective secure without random oracles. The scheme only admitted (s, s) -threshold decryption policies.

Jason C. Cohen *et al.* [5] stated that for the organisations which are dealing with big data, Hadoop is an efficient framework for distributed computation and storage for large datasets on clusters of commodity servers. During Hadoop's development, authentication services were not of much concern and hence it is susceptible to threats and now HDFS requires better security framework to address the security issues. In order to get a more robust framework, Hadoop software layer is integrated with Trusted Computing Group (TCG) technology in the paper. Web interfaces of Hadoop can be protected by SSL with TPM storing private key. For data integrity HDFS files are checked by checksum and strong authentication can be provided by Kerberos and PKI.

Shuaishuai Zhu *et al.* [6] stated that as cloud computing technology is spreading, security of data stored is becoming a major requirement. Sharing of file is the most common service that cloud computing provides. But sharing and transferring of data suffers from some threats like unauthorised access, leakage of essential

information etc. Most of the times files in plaintext are being shared in cloud. Bandwidth and processing overhead required by traditional mechanisms for providing security is very high. To manage file sharing, Attribute based encryption (ABE) is one of the best possible ways with its attribute properties. This paper describes ABE scheme without pairing and on the basis of this scheme, a secure file sharing system with attribute support is designed in the paper which includes algorithms for example, system_init, storing, Attribute_Get and so on. CP-ABE-WP scheme uses four algorithms namely setup, encrypt, keygen, decrypt. Comparison and analysis among KP-ABE, CP-ABE and CP-ABE without pairing showed that SCFS system based on CP-ABE without pairing is more efficient and provides better security.

First A. Huang Jing *et al*. [7] stated that though cloud computing is receiving popularity, yet the major obstacle in its progress is the security issues associated with it. Some of the aspects in which cloud security is weak are transmission, access control, data storage, data verification. To address these issues, a cloud disk storage safety scheme based on Hadoop is proposed in the paper. This scheme takes help from the authorization process of Kerberos protocol. For encryption and authentication, algorithms like AES and RSA are utilized in the paper. For safe connection and storage among clients and servers, three times handshaking is described in the paper i.e First Hand shaking , Second hand shaking and Third hand shaking. The encryption scheme described in the paper ensured transmission and storage security effectively and efficiently but the system described above is not sophisticated enough. So, in the future work, more sophisticated versions of the current system for encryption and authentication is planned to be implemented.

Zhiqian Xu *et al*. [8] stated that one of biggest challenges in cloud storage system is protection of data stored on the cloud. If the data is outsourced to unreliable third party cloud storage then physical control on the data by data owners is lost. Traditional approaches for security are not adequate as one user's data can be intermixed with other users whether on same server or across different domains of security and moreover data in those approaches are protected by encryption. Who can access the key to decrypt the encrypted data is governed by key distribution. Key distribution is not outsourced to third party providers as it is having a critical role in securing encrypted data. As a result data owners themselves are responsible for managing key distribution. Attribute based encryption i.e ABE is an efficient tool for protecting cloud data and provides fine grained access control to encrypted data and allows automatic key distribution hence eliminating the responsibility of data owners. Key policy ABE i.e KP-ABE and ciphertext policy ABE i.e CP-ABE are the major classes to ABE. Key refreshing and revocation are some of the issues while deploying attribute based encryption. To address these issues, a generic framework has been proposed in the paper that can work with any ABE scheme and also incorporates features to any of the ABE scheme, either CP-ABE or KP-ABE without any need of modifying basic ABE as direct deployment of ABE schemes lead to performance, scalability etc issues.

Huixiang Zhou *et al.* [9] stated that in hadoop platform of cloud computing, protecting user's data legally and confidentially is the main issue of concern. Though information stored in hadoop is protected by Access control and encryption mechanisms yet enough security is not provided as hadoop applications require strong security measures. Traditional mechanisms like PKI and IBE have some defects like all the relevant information of user is needed by the resource provider which could damage the privacy of user and it requires more bandwidth and processing overhead too. In order to address the mentioned issues, CP-ABE based encryption scheme is proposed in the paper.

Masoumeh RezaeiJam *et al.* [10] stated that Hadoop which is an open source framework for cloud computing and big data has been widely used in business community today but the biggest obstacle in its development is its poor security mechanism. In the present security scenario, Hadoop considers the cluster, users or client and network as trusted, files stored in HDFS are in plain text, communication among hadoop and client host or clients and datanodes is not encrypted. Present security mechanisms for hadoop described in paper are, Apache knox gateway for perimeter level security, Kerberos for authentication but for password guessing attacks and multipart authentication, it is ineffective, file access permissions for authorization and data encryption for OS security and data protection. Security violations in hadoop are, malicious user can access the file, can modify the block, can submit/delete any job. Access control is one possible solution for this. Apace Sentry is also an option for providing authorization. Design of a trusted file system for hadoop which uses fully homomorphic encryption has been proposed in the paper. A novel triple encryption scheme in which DEA, RSA and IDEA is used to encrypt HDFS files, data key and RSA private key of user has been implemented in this paper. This paper concludes that at file system level, hadoop has strong security but granular support is lacking which is required for full secure access to data by clients and other business applications.

III. PROBLEM FORMULATED

Currently the Hadoop system is suffering from following vulnerabilities.

- Hadoop doesn't support fine grained access control: HDFS is not able to provide fine grained access control. We do have access control lists for HDFS but they are insufficient for many applications such as assured information sharing and there is a need to support more complex policies.
- Lack of strong authentication for its users: Users are not authenticated by hadoop at all if one has configured it by its defaults. If any user who got the access to Job Tracker, can submit any job with the privileges of the account used to set up the HDFS. In coming days we might see HDFS support protocols like Kerberos authentication protocol for user authentication and encryption of data or files being shared.

John Bethencourt [2] and Huixiang Zhou [9] specified CP-ABE scheme for securing HDFS data from unauthorized access but they didn't specify the access privileges or levels to which users can access the data. For providing authentication Kerberos protocol is used in hadoop. But when any job wants to access any non hadoop service in Kerberos environment itself, then it doesn't have the Kerberos ticket and hence can't authenticate itself.

IV. PROPOSED METHODOLOGY

After having gone through the above mentioned vast researches in the field of hadoop security, we here suggest the following amendments to improve on the security aspect.

- The authentication process in Hadoop among users and services can be induced by employing Kerberos authentication protocol. The Kerberos works in following steps:
  - The client doesn't send password to the server. Rather it requests a ticket from server called ticket granting ticket for which it send the username to the server.
  - The server, searches for username and if successful, sends the Ticket Granting Ticket (TGT) to the client whose encryption is done using the password of user itself.

- The client decrypts the TGT with the password he has. The TGT is valid for the current session (mainly 8 hrs). The TGT is contained with the expiry time.
- All subsequent requests to the server for any service are accompanied by the TGT from the client.

The Kerberos authentication in hadoop can be easily configured by setting the authentication type to Kerberos instead of simple. After successful authentication, a delegation token is created and passed to hadoop along with the job submission. Normally, the Kerberos ticket is not sent to the job. So, if any job requires requesting another service then it doesn't have service ticket. We can improve this by passing TGT as a string through JobConf map to hadoop. This string can be converted back to TGT by hadoop and this TGT can be used for authentication to other services. The security can further be improved by encrypting the TGT string and avoiding any issues. Alternatively, we can use a data interchange format like json to convert the TGT into a json string and later retrieving the TGT string back.

- Hadoop stores its data in HDFS on the cloud computing platform. On such platforms, the main concern is always user confidentiality and data security through proper authentication and authorization schemes. The traditional public encryption systems incur excessive processing load. So, many authors in past have discussed ABE schemes categorized as CP-ABE and KP-ABE. Huixiang Zhou *et. al* in [9] discussed a CP-ABE based scheme where they suggested to use multiple attributes for client identification rather than single key attribute for encryption. The authors experimentally demonstrated the effectiveness of scheme and compared results with traditional CP-ABE scheme.
  - Now another scheme LB-CP-ABE (Level based CP-ABE) is proposed here which uses the concept proposed in the base algorithm and further extends the concept by assigning level numbers to different groups of users. The level no. of each user depends on the attributes of each user. The data on the files on the data nodes also complies the functionality by assigning an access key to each data file, which is stored on the name node. The user's access privileges or the level no., based on user attributes define the limit of user's access to the file after comparison with the file access key. So, the security feature is enhanced by adding selective security and maintaining different user groups who can access different files depending on the level number.

V. CONCLUSION

This paper focuses on the security issues associated with the hadoop system. Current hadoop system doesn't authenticate its users strongly and no fine grained access control is provided as well. To enhance authentication system of hadoop for its users and authorization between services, we have proposed Kerberos authentication protocol so that we use the JobConf object to pass the Kerberos ticket to the job which can further be used for accessing other services within or outside hadoop. To add to the security, we convert the TGT into a json string before passing. Next we have proposed a new CP-ABE level based scheme which defines the level of user's access to the file. This proposed method is supposed to enhance security for hadoop system.

REFERENCES

[1] B. Clifford Neuman and Theodore Ts'o, "Kerberos: An Authentication Service for Computer Networks", IEEE, Vol. 32, No. 1, pp: 33-38, 1994.

[2] John Bethencourt, Amit Sahai, Brent Waters, "Ciphertext-Policy Attribute-Based Encryption", IEEE Symposium on Security and Privacy, pp: 321-334, May 2007.

[3] Bao Rong Chang Hsiu Fen Tsai Zih-Yao Lin Chi-Ming Chen, "Access Security on Cloud Computing Implemented in Hadoop System", 2011 Fifth International Conference on Genetic and Evolutionary Computing, IEEE, pp: 77-80, August-September 2011.

[4] Yanli Ren, Shuozhong Wang, Xinpeng Zhang, Zhenxing Qian, "Fully Secure Ciphertext-Policy Attribute-Based Encryption with Constant Size Ciphertext", 2011 Third International Conference on Multimedia Information Networking and Security, IEEE, pp: 380-384, November 2011.

[5] Jason C. Cohen, Dr. Subrata Acharya, "Incorporating Hardware Trust Mechanisms in Apache Hadoop", GC'12 Workshop: First International workshop on Management and Security technologies for Cloud Computing IEEE, pp: 769-774, 2012.

[6] Shuaishuai Zhu, Xiaoyuan Yang*, XuGuang Wu, "Secure Cloud File System with Attribute based Encryption", IEEE, pp: 99-102, 2013.

[7] First A. Huang Jing, Second B. LI Renfa, and Third C. Tang Zhuo, "The Research of the Data Security for Cloud Disk Based on the Hadoop Framework", IEEE, pp: 293-298, 2013.

[8] Zhiqian Xu, Keith M. Martin, "A Practical Deployment Framework for Use of Attribute Based Encryption in Data Protection", IEEE, pp: 1593-1598, 2013.

[9] Huixiang Zhou, Qiaoyan Wen, "A New Solution of Data Security Accessing for Hadoop Based on CP-ABE", IEEE, pp: 525-528, 2014.

[10] Masoumeh RezaeiJam, Leili Mohammad Khanli, Mohammad Kazem Akbari, Morteza Sargolzaei Javan, "A Survey on Security of Hadoop", IEEE, pp: 716-721, 2014.