Original Research Paper Engineering

A STUDY ON AUTHENTICATION CHALLENGES FOR BIG DATA IN PUBLIC CLOUD

Ashok Kumar J Bharathidasan University, Trichy, TN India.

Gopinath Ganapathy Bharathidasan University, Trichy, TN India.

ABSTRACT Big data applications are of great benefit to Organizations, Governments, companies, business, small and large scale industries. It is the De-facto standard for analyzing both streaming and static large data sets. It provides the predictive and historical results for critical decisions making. However, data security and privacy is the biggest concern in Big Data as it collects sensitive information from trusted or non trusted consumers. Data for this research are collected from a case study of Big data application platforms and Cloud Security Alliance (CSA). This paper, first discusses about the highlights of Big data security issues in Cloud Computing and improving further collaborative research for mitigating both security and privacy challenges relating to big data. Secondly, this paper elucidates the Kerberos authentication mechanism under HDFS and brings out the security issues and the computational performance that Kerberos identity authentication mechanism faced in HDFS cluster environment. Finally this paper elucidates the implementation of GnuPG for encryption or decryption of data for enhancing the data security, users privacy and access control in Hadoop Cluster.

KEYWORDS : Big data, cloud computing, Kerberos and GnuPG.

INTRODUCTION

Security in big data is magnified by the three V's, Volume, Variety and Velocity. Several authors [1] [2] have discovered a plethora of challenges for big data in public cloud environment which includes data storage and privacy. In the event of a security breach to big data it brings about both legal and reputational damage. Industries use this big data technology for storing and analyzing huge amount of data that relates to their day to day business and customers. Thus the information classification becomes more critical in Cloud Environment. Techniques such as honeypot detection and logging need to be used for securing Big Data applications in cloud environment [3].

The cloud consumer can outsource their sensitive data and personal information to cloud provider's servers which is not within the same trusted domain of data-owner. Thus the most challenging issues arises in cloud are data security, users privacy and access control. The author proposes a methodology of fine grained security with combined approach of Kerberos and PGP/GnuPG in cloud computing [4]. Kerberos is a trusted third party (KDC) and it authenticates each user over network. But the author comes out with Kerberos protocol issue which does not provide the non-repudiation feature and then suggests to implement digital signature mechanism in Kerberos authentication.

Data security is a methodology for encryption of the available data and the ensuring of enforcing appropriate policies connected with that data, are carried out securely for data sharing. Now a days Industries, Federal organizations etc are facing so many problems in securing "Security and data privacy" challenges. The widespread usage of big data in business, causes most companies in facing privacy issues. In order to meet this challenge there should be a good balance between security and data privacy.

KERBEROS

Kerberos [5] was designed and developed at Massachusetts Institute of Technology (MIT) in 1980. It is a Single-Sign-On (SSO) mechanism to authenticate various client systems. Kerberos is used in most of the colleges, universities and financial institutions as it is a reliable, secure and a well-established authentication protocol and a trusted third party authentication mechanism designed for TCP/IP networks.

Pretty Good Privacy (PGP)

PGP was developed by Philip R. Zimmermann [6]. GnuPG (GNU Privacy Guard) is an open source compatible encryption system based on OpenPGP. PGP/GnuPG encryption uses combination of public key cryptography, data compression, hashing and symmetric-key cryptography. It is used in several security constraints such as confidentiality, integrity and authentication for electronic mail and file storage applications etc., [7]. GnuPG creates the digital signature for the given data to verify the authenticity of the sender. Sender sends the hash digest along with the given data to the receiver. Then receiver uses the sender's public key to verify the digital signature. If it matches the digital signature, it will be confirmed that it is from the expected sender.

LITERATURE SURVEY

Cloud Security Alliance (CSA) identifies the top ten security and privacy problems which are addressed to make Big Data processing and computing infrastructure more secure(Fig. 1) [1] [2]. In top ten challenges, the following security issues are analyzed for providing better solutions.

- Distributed programming frameworks utilize parallel computation and storage to process massive amounts of data. An input file is splited into multiple chunks by Mapreduce framework and it is used in two phases. In the first phase, mapper performs computation after reading the data and output a list of key pairs. Then, a reducer combines all list of key pairs and produce the result. To overcome two major attacks, mappers should be secured with Kerberos and the data present in the untrusted mapper need to be secured with GnuPG.
- The visibility of the data can be controlled by limiting access in the operating system and the cryptography is used to encapsulate the data.
- The security properties designed in such a way that access to data is prevented from unprivileged user. The problem that exists with grained access data could be shared and swept into a more restrictive access for security purpose.

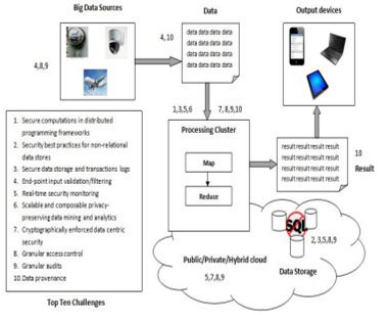

Fig. 1: Top Ten Security and Privacy Challenges

In open source implementation of Kerberos [5] [8] for Distributed Environment, Hadoop by itself does not provide secure authentication. so Hadoop user is authenticated with Kerberos. If a user needs to access Hadoop cluster then the user contacts the KDC and request access. On providing a valid credential the KDC provides the requested access.

The Author brings out the problems such as KDC security, time synchronization, denial mechanism and dictionary attacks for Kerberos protocol [8] and proposed the modification to use data signature mechanism and public key encryption for Kerberos. Replay Attack could still was possible because of using time synchronization mechanism. Then they came out with an issue of KDC bottleneck for larger network. This issue can be analyzed with the computation efficiency and security performance for Kerberos authentication.

Kerberos protocol suffers from limitation problems and they are found to be Replay Attack, Dictionary Attack, Key Storage Problem, Malware Attack, Authentication Forward Problem, Unauthorized Database Access, Single Point of Failure, Clock Synchronization and Digital Signature [9]. It could be seen that there is no single solution to address the above limitation. The following are some solutions to overcome the above limitations.

a. E. El-Emam et al., had proposed modifications for Kerberos database for possible password guessing attack. In this user password is not dependent on the private key. The algorithms used are SHA-256 hashing algorithm and 3DES encryption algorithm [10] in this method.

b. Dynamic password is used to overcome dictionary based password attack and Diffie-Hellman algorithm is used to exchange the password [11]. Security of KDC server and replay attack problem is not deeply analyzed.

c. The public key cryptography is used for initial authentication between KDC and the client [12]. It is based on PKINIT, PKCROSS and PKTAPP.

d. Ik Rae Jeong et al. , suggested a strong Diffie-Hellman-DSA mechanism for Kerberos and overcomes some security against key independence, forward secrecy and session state reveal attacks [13].

e. Z. Hu et al. , uses the Diffie-Hellman-DSA mechanism to improve the Kerberos protocol. Table 1 provides the comparison chart based on Traditional Kerberos V5, PKINIT implementation of Kerberos and Optimized DH-DSA improved protocol [14] .

Table 1 : Time Comparison (in ms)

Version	request time for Average TGT	Extra Time	Percent
Kerberos Version 5	8	0	25.8%
Optimized DH-DSA scheme	20	12	64.5%
PKINIT	31	23	100%

Naman S. Khandelwal et al. , proposed to use the concept of digital envelope and Visual cryptography for Kerberos. It solves the problem of password guessing attack and key distribution with the AES algorithm for symmetric key encryption and ECC algorithm for asymmetric key encryption [9]. Mazhar Islam et al , proposed a new symmetric key scheme "Image as Secret key" of high security and its key size is very large. It provides the comparison value based on throughput of packet size with various symmetric keys algorithm and they are AES as 5.3 , 3DES as 4.5, DES as 5.2 and the Image as Secret key as 136.06 [15] (Refer Table 2).

Table 2 . Execution Time (ms) of Symmetric Algorithms with different packet size

Packet Size (KB)	DES	3DES	AES	Image as Secret key
98	79	107	119	7.18
118	75	99	96	8.62
200	106	138	150	14.77
494	119	188	188	36.52
642	156	254	313	47.6
1388	264	373	352	102.6
1798	392	470	429	132.3
1926	407	460	372	142.6
10690.56	2079	2301	1892	792.8
14620.672	2648	2887	2248	1084.1
Throughput (MB/Sec)	5.2	4.5	5.3	136.06

Kerberos Delegation

In the distributed network, kerberos has many client-server interactions to authenticate several applications in the network. It impacts a high load on the KDC. Hence the Hadoop uses delegation tokens [16] to allow later authenticated access for the kerberos. These tokens are created with delegation rights for Hadoop and transparently authenticated without contacting the kerberos to sign in once again.

Data Encryption/Decryption with PGP

Natasa prohic had proposed the differences in certificate for PGP certificates and PKI certificates [17]. PGP public certificate has multiple signatures values whereas PKI certificates has a single signature. PKI certificate allows single name for each owner but PGP uses the different labels to identify several users. Danish Shehzad et al. , had proposed a hybrid encryption scheme. This scheme is using both symmetric key algorithm as "Images as Secret key" and asymmetric algorithm RSA to ensure proper Hadoop based cloud data security [18]. The results had shown an appreciable increase in throughput for this Images as Secret key versus symmetric key encryption schemes like AES,DES and 3DES (Table 3).

With Hadoop codecs, the new framework [19] implements the transparent encryption with OpenPGP. It is used for map reduce jobs to encrypt or decrypt data to protect the job secrets in the cluster. This transparent encryption uses more computational power for encryption and decryption. Thus it is not effective methodology for implementation with GnuPG, when compared with the hadoop normal computational power for computing the data.

Table 3: Comparison of execution time(ms): The Images as Secret key technique

Packet Size(KB)	DES	3DES	AES	Images as Secret key
102	122	112	83	7.38
124	102	108	72	8.64
200	154	142	104	14.8
500	189	178	122	37.48
640	298	246	152	50.72
1392	344	362	257	106.34
1788	431	466	384	133.6
1922	381	458	398	152.4
10698	1884	2289	2058	802.44
14628	2302	2996	2644	1094.3
Throughput (MB/Sec)	5.154503	4.348783	5.099458	13.28599311

GAP ANALYSIS

In Distributed nodes, the data can be computed in any set of nodes and can not be ensured the security of that computed node. Internode Communication for Hadoop does not store their data with encryption and so there is no protection for data present in those machines. The Administrator has full permission to access the data present in nodes and so the data can be manipulated with the help of rogue administrator. Hadoop does not authenticate the nodes. For parallel computation, nodes can join with other nodes by means of some third party authentication services like Kerberos.

Kerberos validates the authorized nodes in the cluster environment. For mitigating the Kerberos Limitations [9], the following necessary actions need to be addressed against attacks for the open source implementation of Kerberos.

- Replay Attack: Time stamp mechanism useless
- Dictionary Attack: Use of user's password as a secret key is not feasible for encryption in non-secure channel.
- Key Storage Problem: Sharing secret key and maintenance those key is not feasible in large networks.
- Malware Attack: It relates to the Kerberos client software. The implemented version of software checksums must be matched with the original one.
- Authentication Forward Problem: It is the feature available in Kerberos V. It leads to springbord attack.
- Unauthorized Database Access: Storing of user's password in database can be compromised by an attacker.
- Single Point of Failure: Implementing multiple Kerberos server solve this issue.
- Clock Synchronization: Use of timestamp mechanism is useless
- Digital Signature: The non-repudiation feature is missing and it should overcome from impersonation of user

The computation efficiency for security performance of Kerberos with public key implementation provides 64.5% whereas Kerberos traditional provides 25.8% [14]. It can be even reduced to some extend by replacing the AES algorithm [9] with Image as Secret key [15] symmetric algorithm for Digital Envolope technology in Kerberos authentication and thus it will overcome the limitations of Kerberos and improves the computation efficiency.

PKI (public key infrastructure) is a framework to provide public key certificates. OpenPGP supports both public-key cryptography and symmetric key cryptography. PKI is based on certificate authorities (CA) whereas OpenPGP depends on a web of trust model [20]. The users in the OpenPGP can choose whom they trust, whereas users in a PKI system has to depend on trusted CA. So the hybrid encryption scheme needs to be implemented with transparent encryption or decryption for proper Hadoop based cloud data security. In-built gnupg codec for Hadoop provides the transparent encryption in HDFS Cluster environment and consume more computational power than the normal computation of data.

Kerberos is used to authenticate the Hadoop user. Hadoop provides the access control policy instead from the Kerberos protocol. The external application should have the capability to manage the Kerberos delegated tokens on behalf of users and should provide the delegation tokens. Thus the later authenticated access should not contact the KDC again. The external program can be named as GTS (GnuPG Tracking System).

SOME SUGGESTED SOLUTIONS

The following proposed solutions ensured to mitigate some security issues mentioned in previous sections and to improve the security in cloud computing.

- Cloud service supports Kerberos as a trusted solution (Fig. 2) for authentication in Hadoop and accessing user. So it helps to identify the user and allow to attest only the trusted node and not allow the malicious nodes or impersonation nodes or rogue data nodes.

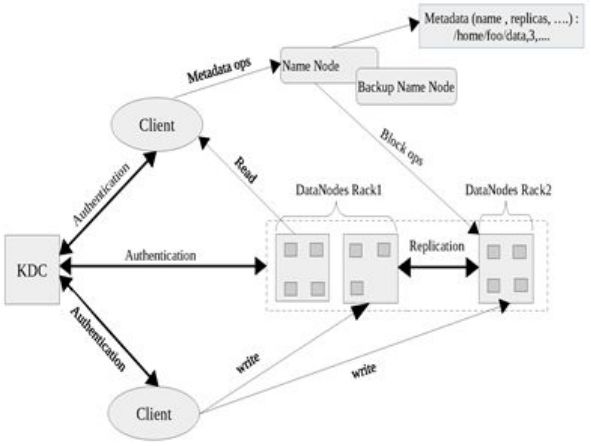

Fig. 2: Kerberos Authentication for Hadoop

- The Kerberos delegated external program is named as GTS (GnuPG Tracking System). It provides the delegation token to Hadoop Applications without having to contact the KDC. Mapper nodes computes the given data and produce the output during Map-Reduce framework in Hadoop. The reducer nodes collects all the output data and produce the overall result. This transaction information of data is getting logged into the GTS database for future reference.
- Normal user and Hadoop system administrator can not access the stored data in the Data Nodes and will be preventing to unauthorized access the data in the Data Nodes. Thus the data must be encrypted with PGP and could be accessible from the authorized user only.
- Mitigate the security attacks and improve the Kerberos authentication mechanism by implementing the symmetric key algorithm of Image as Secret key with Digital Envelope technology. Thus it supports the non-repudiation feature in Kerberos identity authentication and eliminates the most security attacks.
- In Cluster, GnuPG can be used to encrypt or decrypt the data in a secure manner. The GTS itself generates a public key and private key with associated unique ID for the delegated username. It can hold the private key and the private key password by itself. The public key can be shared securely with other nodes in the cloud. So it does not need any administrator or user intervention of creating these public and private key pair. So the random password can be generated by an application and set as password to the private key. The administrator should not have any privilege to access these key pairs.
- The Hadoop uses RPC procedure calls to communicate with other nodes. This communication should happen over TLS so that hacker cannot extract useful information or manipulate packets.
- SELinux [21] is used to prevent information leak in distributed environment.
- Real time access control can be implied in cloud environment to modify the Map Reduce framework and the Java Virtual Machine.
- With the use of Kerberos client application named GTS (GnuPG Tracking System), the malicious actions performed by the user with help of rogue administer can easily be monitored and get logged for their actions. It helps the organizations to identify the malicious administrator and malicious nodes present in the cloud environment.
- The label security method protects sensitive data by assigning data label as public, sensitive and confidential modes. The access is provided if the user label is matched with data label. Auditing should be done to detect the problems then and there to avoid them. This Auditing must be done in a secure manner and should help for forensic data analysis.

CONCLUSION

In this paper, BIG Data security issues identified by the Cloud Security Alliance such as Secure Computations in Distributed Programming Framework, Secure Communication, Cryptographically Enforced Access Control and Granular Access Control are analyzed. The work also has proposed solution for mitigating the same to enhance security in Big Data in cloud environment.

The new identity authentication mechanism is suggested under HDFS to modify the Kerberos protocol with less computational power and high throughput value of efficient symmetric key cryptographic algorithm image as secret key. Thus the Kerberos authentication mechanism under HDFS is enhanced with good security practices and overcomes the Kerberos Limitations. This modification suggested for Kerberos Authentication is applicable for all open source implementation of Operating Systems like Debian and BOSS (Bharat Operating System Solutions). The encryption and decryption of data at rest is the good practice for ensuring the security and apart from that the GnuPG supports the data security, users privacy and access control for Hadoop Cluster. Thus the data at rest and data at transit are secured with encryption and decryption process which keep the data safe from the malicious user and administrator.

References

1. Cloud security alliance members. (2013). Expanded Top Ten Big Data Security and Privacy Challenge. Retrieved 21 August, 2017, from https://downloads. cloudsecurityalliance . org/initiatives/bdwg /Expanded_Top_Ten_Big_ Data_ Security_ and _Privacy_ Challenges.pdf
2. Eweka raphael osawaru, & Riyaz ahamed, .A. .H. (2014). A Highlight of Security Challenges in Big Data. International Journal of Information Systems and Engineering, 2(1), 1-10.
3. Venkata narasimha inukollu, Sailaja arsi & Srinivasa rao ravuri. (2014). SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING. International Journal of Network Security & Its Applications, 6(3), 45-56.
4. Subhash chandra patel , Ravi shankar singh & Sumit jaiswal. (2015). Secure and Privacy Enhanced Authentication Framework for Cloud Computing. IEEE SPONSORED SECOND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEMS, 1631-1634.
5. William stallings (2006). Cryptography and network security principles and practices. (4th ed ed.). Pearson Prentice Hall.
6. Kamarudin shafinah & Mohammad mohd ikram (2011). File Security based on Pretty Good Privacy (PGP) Conce. Computer and Information Science, 4(4), 10-28.
7. Michael louie loria. (2014). Pretty Good Privacy. Retrieved 21 August, 2017, from http://slidedeck.io/michaellouieloria/pgp
8. Daming Hu, Deyun Chen, Yuanxu Zhang, & Shujun Pei. (2015). Research on Hadoop Identity Authentication Based on Improved Kerberos Protocol. International Journal of Security and Its Applications, 9(11), 429-438.
9. Naman khandelwal , & Pariza kamboj (2015). Two Factor Authentication Using Visual Cryptography and Digital Envelope in Kerberos. Electrical, Electronics, Signals, Communication and Optimization (EESCO), International Conference.
10. Eman El-Emam, Magdy Koutb, Hamdy Kelash, & Osama Farag Allah. (2009). An Optimized Kerberos Authentication Protocol. IEEE International Conference on Computer Engineering & Systems (ICCES), 508-513.
11. Chundong wang & Chaoran feng. (2013). Security Analysis and Improvement for Kerberos Based on Dynamic Password and Diffile-Hellman Algorithm. Fourth International Conference on Emerging Intelligent Data and Web Technologies, 6(3), 256-260.
12. Sufyan t faraj al-janabi & Mayada abdul-salam rasheed. (2011). Public-Key Cryptography Enabled Kerberos Authentication. IEEE Developments in E-systems Engineering, 209-214.
13. Ik rae jeong, Jeong ok kwon & Dong hoon lee. (2007). Strong Diffie-Hellman-DSA Key Exchange. IEEE COMMUNICATIONS LETTERS, 11(5), 400-404.
14. Zhao hu, Yuesheng zhu & Limin ma (2012). An Improved Kerberos Protocol based on Diffie-Hellman-DSA Key Exchange. IEEE International Conference on Natural Language Processing, 400-404.
15. Mazhar Islam, Mohsin Shah, Zakir Khan, Toqeer Mahmood, & Muhammad Jamil Khan. (2015). A New Symmetric Key Encryption Algorithm using Images as Secret Keys. International Conference on Frontiers of Information Technology, 1-5.
16. Rajesh laxman gaikwad, Dhananjay m dakhane & Ravindra pardhi. (2013). Network Security Enhancement in Hadoop Clusters. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2(3), 151-157.
17. Natasa prohic. (2005). Public Key Infrastructures – PGP vs X509. Institute of Communication Networks and Computer Engineering.
18. Danish Shehzad, Zakir Khan, Hasan Dag, & Zeki Bozkus. (2016). A Novel Hybrid Encryption Scheme to Ensure Hadoop Based Cloud Data Security. International Journal of Computer Science and Information Security, 14(4), 480-484.
19. Benoy antony. (2013). Map Reduse Encryption and Key Protection - Add support for PGP Encryption. Retrieved 21 August, 2017, from https://issues.apache.org/jira/browse/MAPREDUCE-4552
20. Nikos mavrogiannopoulos (2017). GnuTLS OpenPGP key support. Retrieved 21 August, 2017, from https://www.gnutls.org/openpgp.html
21. Murray mcallister,scott radvan,daniel walsh, dominick grift, eric paris & james morris. (2010). Security-Enhanced Linux. Retrieved 21 August, 2017, from https://docs.fedoraproject.org/en-US/Fedora/13/pdf/Security-Enhanced_Linux/Fedora-13-Security-Enhanced_Linux-en-US.pdf