International Journal of Pure and Applied Mathematics
Volume 120 No. 5 2018, 1127-1132
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue
http://www.acadpubl.eu/hub/

Security improving and performance enhancing of big data by Anonymous Proxy Re-Encryption and Verfiable Hash Convergent Group deduplication

Dr. K. Meenakshisundaram
Associate Professor,
Department of Computer Science,
Erode Arts and Science College, Erode.
lecturerkms@yahoo.com

*M. Menaka
Ph.D-Research Scholar,
Department of Computer Science,
Erode Arts and Science College, Erode.
menarameac@gmail.com

Abstract- **Big data is among one of the emerging technologies because of increasing volume of data in internet. Big data also brings many security risks and privacy preserving issues. Security and privacy issues are magnified by variety, volume and velocity of big data. Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) is a privacy preserving technique which combines the benefits of proxy re-encryption with anonymous technique. In which a ciphertext can be securely and conditionally shared multiple times without disclosing knowledge of underlying message and both the identity information of ciphertext recipients and senders. The duplicated data may leads to problems like high computational complexity and high storage space. So in this paper, a de-duplication technique is proposed to resolve the high computational complexity and high storage capacity problem. A de-duplication technique called as Verfiable Hash Convergent Group Signcryption (VHCGS) is introduced for de-duplication of encrypted data by AMH-IBCPRE. VHCGS uses upload protocol and download protocol for de-duplication of encrypted data. The upload protocol stores a new ciphertext at the storage server and demonstrates the download protocol read by which the client can restore a ciphertext by verifying ownership. Thus this method provides better privacy preserving policy for big data with less computational complexity and less storage capacity. The experimental results are conducted to prove the effectiveness of the proposed method over existing method in terms of storage cost, retrieval time and search time.**

Keywords- Big data; Privacy preserving policy; Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption; Verfiable Hash Convergent Group Signcryption.

I. INTRODUCTION

The term big data refers to the massive amount of data which have more varied and complex structure. The size of big data may be in zeta bytes which cannot be able to handle using traditional data processing applications. Big data analytics (Mehta, B. B., & Rao, U. P. 2016) is very helpful in different fields like social media, medical science, semantic web, national security, etc. On the other hand, it creates a privacy threat to store and process huge volume of data very accurately and quickly. Privacy and security in big data is an important issue. Big data security model is not suggested in the event of complex applications due to which it gets disabled by default. However, in its absence, data can always be compromised easily. So there are various research has been carried out for big data security.

Information privacy (Jain, P., et. al., 2016) is the privilege to have some control over how the personal information is used and collected. It is the capacity of individual or group to stop information about them from becoming known to people other than those they give the information to. The approaches to privacy protection in data storage are chiefly based on encryption procedures. Encryption based techniques can be further classified as Identity Based Encryption (IBE) (Ye, F., et. al., 2015), Attribute Based Encryption (ABE) (Yang, K., et. al., 2017) and storage path encryption. Another approach is cryptography that refers to set of techniques and algorithms are employed to convert the plaintext into cipher text using encryption techniques. There are different methods based on this approach like digital signature, public key cryptography etc. The complex nature of big data makes traditional cryptography and encryption techniques not scalable up to the privacy needs of the big data.

Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) (Liang, K., Susilo, W., & Liu, J. K. 2015) is a privacy preserving technique with the properties of anonymity, multiple receiver update and conditional sharing to provide privacy preserving in big data. But it has major problem of high computational complexity and high storage capacity. Hence in this paper to reduce the computational complexity problem and storage capacity problem de-duplication process is introduced. Verifiable hash convergent group signcryption (VHCGS) framework with two protocols is proposed for de-duplication of encrypted data. By de-duplication process, the duplicated encrypted data is removed which minimize the high computational complexity and high storage capacity problem.

II. LITERATURE SURVEY

A Deterministic Finite Automata-based Functional Proxy Re-Encryption (DFA-based FPRE) (Liang, K., et. al., 2014) was proposed for secure data sharing. Here, a text was encrypted in a ciphertext related with an arbitrary length index string, and a decryptor was legitimate if and only if a DFA related with their secret key accepts the string. In addition to that, this encryption was allowed to be converted to another ciphertext related with a new string to which a re-encryption key was given. However, the proxy cannot gain access to the underlying plaintext. The flexibility of the user was increased by using this new primitive that delegates their decryption to others.

A Bi directional Multi-Hop Identity Based Proxy Re-Encryption (BiMH-IBPRE) (Liang, K.. et. al., 2014) was proposed to maintain the constant cipher text size and computational complexity during the encryption process. BiMH-IBPRE was similar to Multi-Hop Identity Based Proxy Re-Encryption (MH-IBPRE) expect the bidirectional property. BiMH-IBPRE captured conditional re-encryption, bi-directional, constant ciphertext and constant decryption complexity. In order to achieve a secure scheme, the subtleties combined the conditional re-encryption and conditional property with MH-IBPRE. Therefore, this type of construction is not trivial. Here, an extension of the new scheme was presented that can support conditional re-encryption with a set of conditions. But this method is failed to construct CCA secure BiMH-IBPRE system in an adaptive model.

An Attribute based encryption (ABE) was proposed (Wang, Z., et. al., 2016) with improved auxiliary input for big data security. This method tackled the problem of access control and privacy preserving in big data security. This is a security model of leakage resilient ABE with improved auxiliary input that allows the attacker to make more leakage queries seeing the challenge ciphertext. A concrete ciphertext policy- ABE (CP-ABE) and Key Policy-ABE (KP-ABE) with improved auxiliary input based on existing construction model. In addition to that, an improved strong extractor was proposed from the modified Goldreich–Levin theorem for the security proof. From the analysis it is proved that the proposed schemes were Chosen Plaintext Attack (CPA) secure under the security of the existing construction. The efficiency of the encryption algorithm is still low.

A conversion from non anonymous hierarchical identity based encryption (NaHIBE) (Shao, J., & Cao, Z. 2012) was presented for an important security requirement for uni-direction proxy re-encryption. In addition to the proposed conversion, the first concrete Chosen-Ciphertext Attacks (CCA) and collusion resistant multi-use unidirectional IDbased proxy re-encryption (MUIBPRE) scheme in the standard model based on Water09 scheme. This scheme resolved the various security problems. However, this method has different drawbacks like collusion-resistant MUIBPRE scheme increased the growth of size and high decryption time.

SecDep scheme (Zhou, Y. et. al., 2015) was proposed which is a secure de-duplication scheme. This scheme utilized two approaches named as Multi-Level Key Management (MLK) and User Aware Convergent Encryption (UACE) approaches for de-duplication process of big data. The key space overhead was reduced by MLK approach which utilizes file level keys for chunk level keys encryption. The computation overhead was reduced by UACE approach by combining the cross-user file level and inside-user chunk level de-duplication. The major disadvantage of this is high computational complexity.

A architecture was proposed (Bellare, M. et.al., 2013) which provided secure de-duplicated storage resisting brute force attacks and realized it in a system named as DupLESS. In DupLESS system, an obvious pseudorandom function (PRF) protocol is utilized to obtain clients encrypted under message based keys from a key server. It enabled the clients to store encrypted data with an existing service, have the de-duplication service on their behalf and it achieved strong confidentiality.

For secure data de-duplication a novel method called as PerfectDedup was proposed (Puzio, P. et. al., 2015). This method takes into account the popularity of data segments and leveraged the properties of Perfect Hashing to assume data confidentiality and block level de-duplication simultaneously. The main intend of this method was detect the popularity of data segments in a list of popular segments stored by Cloud Storage Provider (CSP) based on data segment identifiers calculated with a Perfect Hash Function (PHF). This scheme achieved de-duplication of encrypted data at block level in perfectly secure manner. But this scheme faces the overhead problem due to Perfect Hash Function (PFH) generation and transmission.

III. PROPOSED METHODOLOGY

In this section the proposed a privacy preserving ciphertext multi sharing mechanism with de-duplication of redundant encrypted big data is described in detail. Initially, the security of big data is considered by a ciphertext sharing mechanism called as Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) with the properties of anonymity, multiple receiver update and conditional sharing in big data environment. In order to reduce the computational complexity and storage capacity of big data, a scheme is designed. This scheme supports secure de-duplication where several groups are sharing data by using Verifiable hash convergent group signcryption (VHCGS) framework with two protocols. It provides better privacy preserving policy with less computational complexity and less storage capacity for big data environment.

A. *Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption based privacy preserving*

Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) is a ciphertext sharing mechanism it achieves multi cipher text receiver update, conditional data sharing and anonymity simultaneously in asymmetric bilinear group. The multiple receiver-update is defined as given a ciphertext the receiver of the ciphertext can be updated in multiple times. In this proposed work it is termed as Multi-Hop. The conditional sharing is defined as a ciphertext can be fine grained shared with others if the pre-specified conditions are satisfied. Anonymity is defined as given a ciphertext no one knows the identity information about the sender and receiver. In the proposed security model, when the challenge identity at the outset of security game of big data let the corrupted users to be adaptively selected by an adversary. Furthermore, there are four different models for security of big data is described for different practical purposes. This can be defined as follows:

- One of the basic security models is Multi-Hop Identity based Conditional Proxy Re-Encryption (MH-IBCPRE). In this model, Chosen-Ciphertext Attacks (CCA) was launched to the original ciphertext by a challenger. It can be achieved by playing the game with adversary by a challenger and then ciphertext was re-encrypted to solve a complex problem.
- The case when a proxy colludes with delegate is considered to compromise the secret key of delegator and the underlying message. In this case, the protection of message is very difficult to achieve as the delegate can always decrypt the corresponding ciphertext for the proxy. The secret key of the delegator, however, is possible to be secured.
- An adversary is allowed to acquire all re-encryption keys for the description of the collusion attack model. If the output of adversary game is a valid secret key of an uncorrupted user, then it is considered as the adversary wins the game. It is in the selective model where the adversary has to return a target identity at the outset of the game.
- Since to the security model of anonymity, it is complex in the sense that the game is classified into two sub games as anonymity for delegator and anonymity of re-encryption key.

The AMH-IBCPRE consists of following algorithm to privacy preserving policy for big data.

1. $(\mathrm{x},\mathrm{y}) \leftarrow \mathrm{Setup}(1^{\mathrm{m}})$: On input a security parameter m, output a master secret key x and master public key y.
2. $K_I \leftarrow KeyGen(x, I)$: on input x and an identity $I \in \{0,1\}^*$, output a secret key K_I.
3. $rk_c, I_i \rightarrow I_{i'} \leftarrow ReKeyGen(I_i, K_{I_i}, I_{i'}, c)$: on input a delegators identity I_i and the corresponding secret key K_{I_i}, a delegate's identity $I_{i'}$ and a condition $c \in \{0,1\}^*$, output a re-encryption key $rk_c, I_i \rightarrow I_{i'}$ from I_i to $I_{i'}$ under condition c.
4. $C_{1,I_i,c} \leftarrow Enc(I_i, c, s)$: on output an identity I_i, a condition c and a message s, a output a 1-level ciphertext $C_{1,I_i,c}$ under identity I_i and c.
5. $C_{l+1,I_{i'},c} \leftarrow ReEnc\left(rk_c, I_i \rightarrow I_{i'}, C_{l,I_i,c}\right)$:on input $rk_c, I_i \rightarrow I_{i'}$, and an l-level ciphertext $C_{l,I_i,c}$ under identity I_i and c, output an (l+1)-level ciphertext $C_{1,I_{i+1},c}$under identity $I_{i'}$and c or ⊥ for failure, where $l \geq 1, l \in N$.
6. $m \leftarrow Dec\left(K_{I_i}, C_{l,I_i,c}\right)$:on input K_{I_i}, and an l-level ciphertext $C_{l,I_i,c}$ under identity I_i and c, output a message s or ⊥ for failure, where $l \geq 1, l \in N$.

The above algorithms are used to encrypt the data and provide privacy preserving for big data. But the encrypted data may also contain some duplicate data which leads to high computational complexity and high storage capacity problem. In order to avoid these problems in big data, verifiable hash convergent group signcryption (VHCGS) is introduced along with the proposed encryption technique.

B. *Verifiable hash convergent group signcryption based de-duplication*

A Verifiable hash convergent group signcryption (VHCGS) (Cho, E. M., & Koshiba, T. 2017) is introduced with two sub protocols for the encrypted de-duplication process. Initially an upload protocol is designed to store a new ciphertext at the storage server. Then a download protocol is demonstrated red by which the client can restore a ciphertext by verifying ownership.

1) *Upload protocol*

In the upload protocol, the following process are done for de-duplication of encrypted data in big data environment.

1. The user generates the signature of a file F_1
2. The user encrypts the file n to F_2.
3. The user generates a tag by utilizing the ciphertext F_2
4. The user signcrypts the ciphertext of the file and ciphertext of the signature into F_3
5. The user uploads the cipher texts and the tag pair (F_1, T) to the server.
6. The server asks to the verifier for verification of the ciphertext.
7. The verifier can partially unsigncrypt for checking the validity of the file and return invalid or valid.
8. Then the server checks whether there exists the same ciphettext on the server.
9. If there is no duplicate file exists on the server, then the server keeps δ as signature of the message, F_3as the

ciphertext of the file and T to check for fake de-duplication respectively.

10. Else, the server stores only δ as the new signature of the existing ciphertext.

After the above algorithm, the de-duplicated encrypted message is stored in the storage server.

2) Download protocol

The following procedure is performed, when a user tries to download their own file from the storage server.

1. The user sends a download requests to the server through signature and file name of the file.
2. The server asks the verifier to verify the ciphertext.
3. To check the validation of the file, the verifier can partially unsigncrypt and it return invalid or valid.
4. The server checks the ownership of the ciphertext.
5. When the server finishes the validation of messages and signature, it gives the ciphertext to the client.
6. At last, the client unsigncrypts the ciphertext.

The above protocols de-duplicate the encrypted data which reduce the computational complexity and high storage capacity problem in big data environment.

III. Results and Discussion

In this section the experiment is conducted in healthcare data which contain health event of patients. For the experimental purpose, the existing AMH-IBCPRE method is compared with proposed AMH-IBCPRE-VHCGS in terms of storage cost, retrieval time and search time.

I. Storage Cost

Storage cost is the amount of space required to store the data. Thus for k data ranges from 1000 to 7000 the amount of space required to store the ciphertext for different methods such as Identity Based Encryption (IBE), Attribute Based Encryption (ABE), AMH-IBCPRE and AMH-IBCPRE-VHCGS. The AMH-IBCPRE-VHCGS requires less storage space because it does not store any duplicate encrypted data.

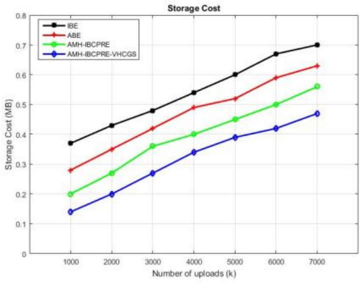

Fig. 1. Comparison of storage cost

Fig. 1, shows the comparison of storage cost between existing IBE, ABE, Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) and proposed AMH-IBCPRE with Verifiable hash convergent group signcryption (AMH-IBCPRE-VHCGS) methods. X axis represents number of uploads and Y axis represents the storage cost in MB. From the Fig.1, it is proved that the proposed AMH-IBCPRE-VHCGS method has less storage cost than the existing methods.

II. Retrieval time

Retrieval time is defined as the amount time taken to retrieve a data for user query. It is calculated by enter a query and find the amount of time taken to retrieve result for entered query. The retrieval time is calculated for different methods are IBE, ABE, AMH-IBCPRE and AMH-IBCPRE-VHCGS.

Fig. 2, shows the comparison of retrieval time between existing IBE, ABE, Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) and proposed AMH-IBCPRE with Verifiable hash convergent group signcryption (AMH-IBCPRE-VHCGS) methods. X axis represents methods and Y axis represents the retrieval time in seconds. From the Fig. 2, it is proved that the proposed AMH-IBCPRE-VHCGS method has less retrieval time than the existing methods.

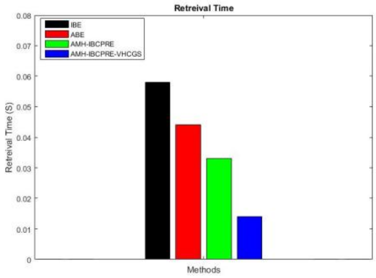

Fig. 2. Comparison of retrieval time

III. Search time

Search time is the amount of time taken to search the data in big data for user query. It is calculated by enter a query and find the amount of time taken to search result for entered query. The search time is calculated for different methods are IBE, ABE, AMH-IBCPRE and AMH-IBCPRE-VHCGS.

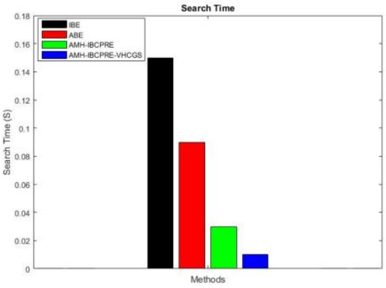

Fig. 3. Comparison of search time

Fig. 3, shows the comparison of search time between existing IBE, ABE, Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) and proposed AMH-IBCPRE with Verifiable hash convergent group signcryption (AMH-IBCPRE-VHCGS) methods. X axis represents methods and Y axis represents the search time in seconds. From the Fig. 3, it is proved that the proposed AMH-IBCPRE-VHCGS method has less search time than the existing methods.

IV. Conclusion

In this paper, an efficient privacy preserving technique for big data is proposed. The data are encrypted by Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) which is a ciphertext sharing mechanism it achieves multi cipher text receiver update, conditional data sharing and anonymity simultaneously in asymmetric bilinear group. Though it is an efficient privacy preserving technique, it faces some problems like computational complexity problem and storage capacity problem. In this paper, the above problems are resolved by proposed de-duplication technique called as verifiable hash convergent group signcryption (VHCGS) with upload and download protocols. This method de-duplicates the encrypted data and resolves the computational complexity problem and storage capacity problem. The experimental results show that the proposed method has less storage cost, less retrieval time and less search time in big data.

References

1. Bellare, M., Keelveedhi, S., and Ristenpart, T., "DupLESS: Server-Aided Encryption for Deduplicated Storage", IACR Cryptology ePrint Archive, vol. 429, 2013.
2. Cho, E. M., and Koshiba, T., "Big Data Cloud Deduplication based on Verifiable Hash Convergent Group Signcryption", IEEE Third Int. Conf. Big Data Comput. Serv. Appl., pp. 265-270, 2017.
3. Jain, P., Gyanchandani, M., and Khare, N., "Big data privacy: a technological perspective and review", J. Big Data, vol. 3, no. 1, pp. 25, 2016.
4. Liang, K., Au, M. H., Liu, J. K., Susilo, W., Wong, D. S., Yang, G., and Xie, Q., "Privacy-preserving ciphertext multi-sharing control for big data storage", IEEE trans. inf. forensics secur., vol. 10, no. 8, pp. 1578-1589, 2015.
5. Liang, K., Au, M. H., Liu, J. K., Susilo, W., Wong, D. S., Yang, G., and Xie, Q., "A DFA-based functional proxy re-encryption scheme for secure public cloud data sharing", IEEE Trans. Inf. Forensics Secur., vol. 9, no. 10, pp. 1667-1680, 2014.
6. Liang, K., Chu, C. K., Tan, X., Wong, D. S., Tang, C., and Zhou, J., "Chosen-ciphertext secure multi-hop identity-based conditional proxy re-encryption with constant-size ciphertexts", Theor. Computer Scienc., vol.539, pp. 87-105, 2014.
7. Mehta, B. B., and Rao, U. P., "Privacy preserving unstructured big data analytics: Issues and challenges", Procedia Comput Sci., vol. 78, pp. 120-124, 2016.
8. Puzio, P., Molva, R., Önen, M., and Loureiro, S., "PerfectDedup: Secure data deduplication", Int. Workshop Data Priv. Manag. Springer Int. Publ., pp. 150-166, 2015.
9. Shao, J., and Cao, Z., "Multi-use unidirectional identity-based proxy re-encryption from hierarchical identity-based encryption", Inf. Sci., vol. 206, pp. 83-95, 2012.
10. Wang, Z., Cao, C., Yang, N., and Chang, V., "ABE with improved auxiliary input for big data security", J. Computer Syst. Sci., vol. 89, pp. 41-50, 2016.
11. Yang, K., Han, Q., Li, H., Zheng, K., Su, Z., and Shen, X., "An efficient and fine-grained big data access control scheme with privacy-preserving policy", IEEE Internet of Things J., vol. 4, no. 2, pp. 563-571, 2017.
12. Ye, F., Qian, Y., and Hu, R. Q., "An identity-based security scheme for a big data driven cloud computing framework in smart grid", Glob. Commun. Conf. (GLOBECOM), IEEE, pp. 1-6, 2015.
13. Zhou, Y., Feng, D., Xia, W., Fu, M., Huang, F., Zhang, Y., and Li, C., "SecDep: A user-aware efficient fine-grained secure deduplication scheme with multi-level key management", Mass Storage Syst. Technol. (MSST) 2015 31st Symp. IEEE, pp. 1-14, 2015.