The Performance Evaluation of Various Security Techniques in Data Aggregation Model for Big Data

K. Meenakshisundaram,M. Menaka

Abstract: ***Big Data is defined as a collection of huge size of data sets with different types. It brings some problems like privacy preserving problem and security risk. Event-Role-Attribute based Access Control mechanism was used to ensure the end to end security. It created a flexible boundary with the consideration of event, role and attribute to access the data. Then the data were encrypted by Anonymous Multi-Hop Identity Based Conditional Proxy Re-Encryption (AMH-IBCPRE) where a ciphertext was conditionally and securely shared multiple times without disclosing the identity information of ciphertext senders or recipients and knowledge of underlying message. The problem of high computational complexity and high storage capacity due to the duplicate cipher data was sorted out by a de-duplication technique called Verfiable Hash Convergent Group Signcryption (VHCGS). In this paper, Enhanced Conditional Proxy Re-encryption with Data Aggregation and Masking (ECPR-DAM) and Enhanced Conditional Proxy Re-encryption with Verifiable hash convergent group Signcryption, Data Aggregation and Masking (ECPRVS-DAM) are proposed for data aggregation and data security. The data aggregation model aggregate the data which is the response to maintain the ever improving demands of Big Data. In the process of data aggregation, Fibonacci search is used instead of dichotomic search. Because the Fibonacci search is reducing the average time required to access a location of data. Moreover, a watermarking construction is introduced to enhance the security of Big Data. The watermarking construction provides synchronization marks in the aggregated data and helps protect the data itself at the end points. It improves the big data security. The experimental results prove the effectiveness of the proposed ECPR-DAM and ECPRVS-DAM methods over existing method in terms of storage cost, retrieval time and search time.***

Keywords: Big data; Data Aggregation; Fibonacci Search; Watermarking Construction.

I. INTRODUCTION

Big data [7] and big data analytics plays a significant role in today's fast-paced data-driven businesses. Big data is a collective term referring to data that is so complex and large that it exceeds the processing capability of software techniques and conventional data management systems. Security and privacy concerns are growing as big data becomes more and more accessible.

The collection and aggregation of massive quantities of heterogeneous data are now possible. Large data sharing is becoming routine among clinicians, governmental agencies, scientists, citizens and businesses. However, the technologies and tools that are being developed to manage these massive datasets are often not designed to incorporate adequate security or privacy measures, in part because we lack sufficient training and a fundamental understanding of how to provide large-scale data security and privacy. Access control mechanisms are used to enable the end-users for controlling the access of their own data. Among various access control mechanisms Role Based Access Control (RBAC) and Attribute Based Access Control (ABAC) [6] are emerging as a promising paradigm. But these models provide static boundary for access control. An Event-Role-Attribute Based fine grained Access Control mechanism provides flexible boundaries for security policies. The security policies were developed based on the event, role and attribute. Privacy preserving is one of the techniques used to provide security for big data. Anonymous Multi-Hop Identity based Conditional Proxy Re-Encryption (AMH-IBCPRE) [5] is a privacy preserving technique with the properties of anonymity, multiple receiver update and conditional sharing to provide privacy preserving in big data. But this method has problem of high storage capacity and high computational complexity. A framework called as Verifiable hash convergent group signcryption (VHCGS) was introduced reduced the high storage capacity and high computational complexity problem by de-duplicate the encrypted data. In this paper, a data aggregation model is proposed to aggregate the de-duplicated encrypted data. The dichotomic search was used in data aggregation [9] model. Instead of dichotomic search, Fibonacci search is used in this paper in data aggregation model. Because the efficiency of Fibonacci search is ten times better than the dichotomic search. In addition to, the privacy of the big data is improved by introducing a watermarking construction which is a data masking technique. It provides highly efficient end to end security in big data.

II.LITERATURE SURVEY

A Scalable Privacy-preserving Big Data aggregation method (Sca-PBDA) [10] was proposed to provide privacy preserving big data aggregation. Initially sensor nodes in the network were splitted into clusters based on the pre established gradient topology structure. Then, based on privacy preserving configuration message received from the sink each node was modified. In order to reduce the energy consumption, inter and intra cluster data aggregation was employed during the big sensor data reporting phase. Finally aggregated results were recovered by the sink to complete the privacy preserving big data aggregation. However, the

Revised Manuscript Received on April 10, 2019
K.Meenakshisundaram,AssociateProfessor, Department of CS, Erode Arts and Science College
M. Menaka,Research Scholar, Department of CS, Erode Arts and Science College

scalability of uniform privacy preserving intra-cluster aggregation method was not maintained properly. A privacy preserving data aggregation scheme [3] was proposed based on secret sharing scheme. This scheme ensured that control center gets the integrated data without compromising the privacy of user's. A threshold of secret shares was set same with the number of group members to resist the differential attack by the conspiracy between Data Aggregation device (DA) and Control Center (CC). In addition, the user identity was masked by using the same group serial number. Then the malfunctioning Smart Meter (SM) was determined by adopting a hash table and achieved the fault tolerance for the normal aggregation by substitution. An efficient data aggregation method [4] was proposed for aggregation of big data. This technique was used for clustering based periodic wireless sensor network. For a local aggregation at sensor node level, the proposed data aggregation technique allowed cluster head to remove the redundant data which were created by neighboring nodes by using three data aggregation methods. The first method utilized similarity functions such as Jaccard function. It utilized to search similarities between data sets. The second method utilized one way Anova model and Barlett test was utilized to search the dependence of conditional variance between datasets. The third method utilized Euclidean and Cosine which calculate the dissimilarity between datasets. However, this method still needs an improvement in terms of energy consumption and network lifetime. A data masking scheme [2] was proposed for sensitive big data based on format preserving encryption. This scheme masked the sensitive information and ensured the masked data contains valid information. The format preserving algorithms applied to this scheme have passed security verification that is the encrypted data will not contained sensitive information which the user wants to mask. All format preserving algorithms which the system will use were added to the algorithm library. Based on the key length parameter in the library, key generation algorithm created a fixed length key for each algorithm. For each input field, the string was segmented into some substrings and algorithm was chosen for each substring. After the phase of initialization, it was created a mapping table according to the chosen algorithm rule. Before format preserving encryption, the substring was mapped to integer field. The encryption algorithm taken the substring after mapped as the input, and outputs the intermediate result. Then with inverse mapping, the result in specified format was obtained. However, the encryption speed is not speed enough. A conceptual data masking framework called as IMETU [1] was proposed for built in framework. This framework was comprised of five modules are Identify, Map, Execute, Test and Utilize. In the identify phase, sensitive data attributes were identified which needs masking. In the mapping phase, map the selected data attributes with the best fit masking algorithm. In the execute phase, run and apply the masking algorithm in efficient way. In the test phase, the masking results were tested if they were applied successfully. In the utilize phase, re-identification methods were used for querying and getting and original data. The new practical masking approach offered an alternative solution to preserve an individual re-identification over conventional masking or standard encryption methodologies. This new masking technique was required to be available in the early integration service stage as well as in the analytical layer to re-identify the sensitive data in the simple way. However, the computational complexity of this framework is high. An efficient and fine-grained access control mechanism [11] was proposed with privacy preserving policy. In this mechanism, whole attribute was hidden in the access policies. In this mechanism, data decryption was decrypted a novel attribute bloom filter was designed to compute whether an attribute is in the access policy and locate the exact position is in the access policy if it is in the security policy. It was located the precise row numbers of attributes in the access matrix. This scheme was selectively secure against chosen plaintext attacks.

III. PROPOSED METHODOLOGY

In this section, the proposed data aggregation method and data masking based on watermarking technique are described in detail. The access control is provided by based on Event-Role-Attribute based fine grained access control mechanism. While transmitting the data, the data are encrypted for security purpose. The encrypted data are called as ciphertext and the duplicate ciphertext improves the computational complexity and occupies high storage space in big data. This problem is sorted out by using AMH-IBCPRE with VHCGS framework. This framework de-duplicates the ciphertext by using two protocols. Then the data are aggregated by a secure data aggregation model where Fibonacci search to identify the invalid message. In order to provide more security, a watermarking construction is proposed which protects the data itself at the end points.

A. AMH-IBCPRE

Event-Role-Attribute based fine grained access control mechanism is used to provide flexible boundary for the accessible set of records in big data based on event, role an attribute. Based on the fine grained description for the role and events and environmental and temporal state of an event a flexible boundary is provided to the big data. Then the data are encrypted by AMH-IBCPRE technique. AMH-IBCPRE combines the advantages of proxy re-encryption with anonymous technique where a ciphertext can be conditionally and securely shared multiple times without leaking both the identity information of ciphertext senders or recipients and knowledge of underlying message.

B. AMH-IBCPRE-VHCGS

The encrypted data also contains some duplicate data which leads to high storage capacity and high computational complexity problem. These problems in big data are resolved by using AMH-IBCPRE-VHCGS method. 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.

C. ECPR-DAM

Retrieval Number: F1045376S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

After the encryption of data by using AMH-IBCPRE technique, the invalid data are determined by using Fibonacci search and then aggregate the data. In order to improve the security of data in cloud, a watermarking technique is applied in the data streams of big data. This process is named as Enhanced Conditional Proxy Re-encryption with Data Aggregation and Masking with Data Aggregation and Masking (ECPR-DAM). The Fibonacci search based data aggregation and the watermarking technique is briefly explained in the following subsection.

D. ECPRVS-DAM

After the encryption of data, the duplicated data are removed by using AMH-IBCPRE-VHCGS method. It resolved both high storage capacity and high computational complexity problem. Then the data are aggregated by using Fibonacci search and data security is improved by using watermarking technique. This process is named as Enhanced Conditional Proxy Re-encryption with Verifiable hash convergent group Signcryption, Data Aggregation and Masking (ECPRVS-DAM). The data aggregation and watermarking technique is briefly explained in the following sub section.

1) Data Aggregation

In one period, it is assumed that a node N receives $x_1$data $(c_n, P_{td}, T_n, \sigma_n)$ for $n = 1, 2, \dots x_1$. where c_n denotes the ciphertext of data n, P_{td}denotes the pseudonym of terminal device, T_n denotes the timestamp and σ_n denotes the signature. A node N does the following process to aggregate the data:

1. A node N verifies

$$e\left(\sum_{n=1}^{x_1}\sigma_n, P\right) = e\left(\sum_{n=1}^{x_1} H(c_n, I_n, P_{td}, T_n), Z\right)$$

where, e denotes the bilinear pairing, P denotes a generator of G_1, H denotes the cryptographic hash function, I_n denotes node's identifier and Z denotes terminal node public key or private key. If N does not hold, N looks for any invalid data by using Fibonacci search [8] and informs terminal node to retransmit until they are valid otherwise N goes on. The Fibonacci search is a univocal function has only one peak in a given interval. The one peak refers either maximum or minimum. Thus a function of one variable is said to be unimodal if, given that two values of the variable are on the same side of the optimum, then the one nearer the optimum gives the better functional value, i.e., the smaller value in the case on a minimization problem. Thus a unimodal function can be a non differentiable or even a discontinuous function. Based on the peak value the invalid data is determined.

2. Compute the aggregated ciphertext $c = (\overline{c_1}, \overline{c_2})$:

$$\overline{c_1} = \prod_{n=1}^{x_1} c_{i1}, \overline{c_2} = \prod_{n=1}^{x_1} c_{i2}$$

3. Create the signature

$$\sigma_n = yH(c, I_n, N, T)$$

where T denotes the time stamp.

4. Upload the data $(c, P_{td}, I_n, T, \sigma_n)$ in big data.

Thus the data are aggregated and uploaded in the big data by using the proposed data aggregation model.

2) Data masking

After the data aggregation process, a watermarking technique is introduced to provide security for data streams in big data. A local scaling factor called as watermark amplitude is calculated for every data point in the data stream to make the watermark invisible from noise floor. Then the scaled version of the shifted watermark pattern is added to the data stream. The following 2D notations, the watermark embedding procedure for data stream x, which is given as follows:

$$D_x^{wm}(m,n) = raster(d_x, l) + \Lambda_x(m,n) \times W_x^{hv}(m,n)$$

where, $W_x^{hv}(m,n) = rot2(W_x(m,n), h, v), W_x = \{b_{jk}\}_{r_x}$ is the 2D watermark pattern generated by Distinct Sum Array (DSA) construction for a unique r_x value, $\Lambda_x$denotes the random variable matrix, rot2() represents the 2D spatial shift by (h,v) and raster() operation takes a data stream d_x, window size l as inputs and make a square array of that data stream of size $l \times l$ and $D_x^{wm}(m,n)$ is watermarked data value with the 2D watermark pattern.The DSA construction generates a square array accumulating different cyclic shifts of a 'seed' sequence at multiple rows that are separated by a distance k. The requirement for the seed sequence is having a pseudorandom (PN) sequence of a prime length p with good auto-correlation properties. By appending rows of cyclic shifts of the seed sequence, matrices with good 2D auto- and cross-correlation are synthesized that offers additional diversity and security over binary 1D sequences and can be folded and unfolded, rendering them robust to cryptographic attacks. The DSA construction contains an adjustable parameter$r \in \{1,2, \dots p-1\}$, which is seed sequence independent. Therefore, for a seed sequence of length p, it is possible to construct p-1 square arrays. More specifically, the 2D square array b_{jk} is generated from the seed sequence a_k of prime length p as follows:

$$b_{jk} = a_k + \tau_j mod p$$

$$\tau_j = \frac{r_j(j-1)}{2}$$

where, τ_j is defined as the shift of the seed sequence relative to the top row from row j. Each node performs the watermarking embedding procedure based on the above equation. Then the watermarked data are sent to the aggregation model. Through the aggregation model, the watermarked data streams are added together to form a joint data stream.

IV. RESULT AND DISCUSSION

Retrieval Number: F1045376S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

In this section, the effectiveness of the proposed AMH-IBCPRE with Data Aggregation and Masking (AMH-IBCPRE-DAM) named as Enhanced Conditional Proxy Re-encryption with Data Aggregation and Masking (ECPR-DAM) and AMH-IBCPRE-VHCGS-DAM in terms of storage cost, retrieval time and search time. For the experimental purpose healthcare data is used which contain health event of patients.

A. 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. Table 1, shows the comparison of storage cost for AMH-IBCPRE, AMH-IBCPRE-VHCGS, Enhanced Conditional Proxy Re-encryption with Data Aggregation and Masking with Data Aggregation and Masking (ECPR-DAM) and Enhanced Conditional Proxy Re-encryption with Verifiable hash convergent group Signcryption, Data Aggregation and Masking (ECPRVS-DAM) under different number of uploads. From the Table 1, it is proved that the proposed ECPRVS-DAM requires less storage space than the other methods.

TABLE I. Comparison of Storage Cost

Methods	No. of uploads (k)						
	1000	2000	3000	4000	5000	6000	7000
AMH-IBCPRE	0.2	0.27	0.36	0.4	0.45	0.5	0.56
AMH-IBCPRE-VHCGS	0.14	0.2	0.27	0.34	0.39	0.42	0.47
ECPR-DAM	0.1	0.17	0.23	0.3	0.34	0.36	0.41
ECPRVS-DAM	0.07	0.14	0.2	0.25	0.29	0.32	0.36

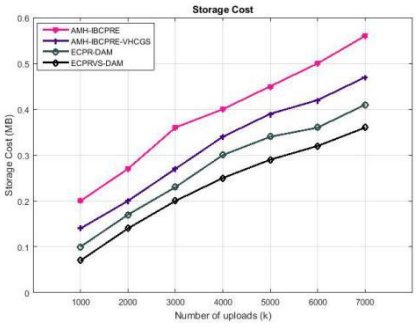

Fig. 1. Comparison of Storage Cost

Fig. 1 shows the comparison of storage cost between existing AMH-IBCPRE, AMH-IBCPRE-VHCGS and proposed ECPR-DAM and ECPRVS-DAM 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 ECPRVS-DAM method has less storage cost than the other methods.

B. 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. Table 2, shows the comparison of retrieval time for AMH-IBCPRE, AMH-IBCPRE-VHCGS, ECPR-DAM and ECPRVS-DAM. From the Table 2, it is proved that the proposed ECPRVS-DAM requires less retrieval time than the other methods.

TABLE II. Comparison of Retrieval Time

Methods	Retrieval Time
AMH-IBCPRE	0.033
AMH-IBCPRE-VHCGS	0.014
ECPR-DAM	0.009
ECPRVS-DAM	0.004

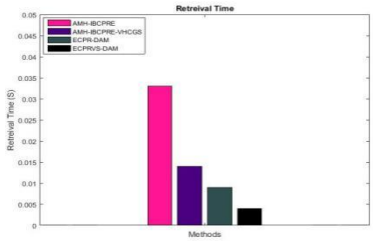

Fig. 2. Comparison of Retrieval Time

Fig. 2 shows the comparison of retrieval time between existing AMH-IBCPRE, AMH-IBCPRE-VHCGS and proposed ECPR-DAM and ECPRVS-DAM methods. X axis represents methods and Y axis represents the retrieval time in seconds. From the Fig. 2, it is proved that the proposed ECPRVS-DAM method has less retrieval time than the other methods.

C. 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. Table 3, shows the comparison of search time for AMH-IBCPRE, AMH-IBCPRE-VHCGS, ECPR-DAM and ECPRVS-DAM. From the Table 3, it is proved that the proposed ECPRVS-DAM requires less search time than the other methods.

TABLE III. Comparison of Search time

Methods	Search Time
AMH-IBCPRE	0.03
AMH-IBCPRE-VHCGS	0.01
ECPR-DAM	0.008
ECPRVS-DAM	0.005

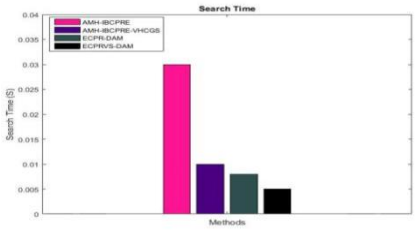

Fig. 3. Comparison of Search Time

Fig. 3 shows the comparison of search time between existing AMH-IBCPRE, AMH-IBCPRE-VHCGS and proposed ECPR-DAM and ECPRVS-DAM methods.

X axis represents methods and Y axis represents the search time in seconds. From the Figure 3, it is proved that the proposed ECPRVS-DAM method has less search time than the other methods.

V. CONCLUSION

In this paper data aggregation and data masking technique is proposed in big data environment which handles the problem of privacy preserving and security risks. Initially, the authentication is provided through Event-Role-Attribute based fine grained access control mechanism. Then the data are encrypted through AMH-IBCPRE technique and de-duplicate the ciphertext by using VHCGS. The ciphertext are aggregated through a proposed data aggregation technique where Fibonacci search is used to remove the invalid data in the dataset. It improves the computational efficiency and reduces the retrieval time and search time. Finally, while storing the data a data masking technique called as watermarking construction is applied to improve the privacy and security of big data. It provides end to end security. The experiment is conducted on healthcare data to prove the effectiveness of the proposed method. It shows that the proposed method has better performance in terms of storage cost, search time and retrieval time than the other methods.

REFERENCES

1. Ali, O., and Ouda, A., “A content-based data masking technique for a built-in framework in Business Intelligence platform”, Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference on IEEE, pp. 1-4, 2017.
2. Cui, B., Zhang, B., and Wang, K., “A Data Masking Scheme for Sensitive Big Data Based on Format-Preserving Encryption”, Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on IEEE, Vol. 1, pp. 518-524, 2017.
3. Guan, Z., and Si, G., “Achieving privacy-preserving big data aggregation with fault tolerance in smart grid”, Digital Communications and Networks, vol. 3, no. 4, pp. 242-249, 2017.
4. Harb, H., Makhoul, A., Tawbi, S., and Couturier, R., “Comparison of Different Data Aggregation Techniques in Distributed Sensor Networks”, IEEE Access, vol. 5, pp. 4250-4263, 2017.
5. Liang, K., Susilo, W., and Liu, J. K., “Privacy-preserving ciphertext multi-sharing control for big data storage”, IEEE transactions on information forensics and security, vol. 10, no. 8, pp. 1578-1589, 2015.
6. Meneka, M.,Meenakshisundaram,K., “An Enhancement Role and Attribute Based Access Control Mechanism in Big Data”, International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 5, 2018.
7. Schmitt, C., Shoffner, M., Owen, P., Wang, X., and Lamm, B., “Security and privacy in the era of big data. White Paper”, ARENCI/National Consortium for Data Science. ARENCI White Paper Series, 2013.
8. Subasi, M., Yildirim, N., and Yildiz, B., “An improvement on Fibonacci search method in optimization theory”, Applied mathematics and computation, vol. 147, no. 3, pp. 893-901, 2004.
9. Wang, H., Wang, Z., and Domingo-Ferrer, J., “Anonymous and secure aggregation scheme in fog-based public cloud computing”, Future Generation Computer Systems, vol. 78, pp. 712-719, 2018.
10. Wu, D., Yang, B., and Wang, R., “Scalable privacy-preserving big data aggregation mechanism”, Digital Communications and Networks, vol. 2, no. 3, pp. 122-129, 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 Journal, vol. 4, no. 2, pp. 563-571, 2017.
12. Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97.1 (2017): 1267-1289.
13. Rajesh, M., and J. M. Gnanasekar. "Sector Routing Protocol (SRP) in Ad-hoc Networks." Control Network and Complex Systems 5.7 (2015): 1-4.
14. Rajesh, M. "A Review on Excellence Analysis of Relationship Spur Advance in Wireless Ad Hoc Networks." International Journal of Pure and Applied Mathematics 118.9 (2018): 407-412.
15. Rajesh, M., et al. "SENSITIVE DATA SECURITY IN CLOUD COMPUTING AID OF DIFFERENT ENCRYPTION TECHNIQUES." Journal of Advanced Research in Dynamical and Control Systems 18.
16. Rajesh, M. "A signature based information security system for vitality proficient information accumulation in wireless sensor systems." International Journal of Pure and Applied Mathematics 118.9 (2018): 367-387.
17. Rajesh, M., K. Balasubramaniaswamy, and S. Aravindh. "MEBCK from Web using NLP Techniques." Computer Engineering and Intelligent Systems 6.8: 24-26.

Retrieval Number: F1045376S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication