IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 3, May 2014.

www.ijiset.com

ISSN 2348 - 7968

THRESHOLD BASED FILTERING APPROACH FOR COST EFFECTIVE OVER ENCRYPTED CLOUD DATA

T.Praveena[1], G. Raja[2]

[1]*Student ,Department of Computer Science and Engineering, Dhanalakshmi Srinivasan Engineering College, Perambalur.*

[2]*Asst Professor,Department of Computer Science and Engineering, Dhanalakshmi Srinivasan Engineering College, Perambalur.*

***Abstract*— Cloud computing is an appealing paradigm that provide users to store large volume of data and to use application over cloud without any infrastructure investment .When processing such application will generate large volume of intermediate data set which is often stored in the cloud to save the cost of recomputation. For preserving those generated intermediate dataset encrypting all the dataset is handled but this is cost consuming process, and consume lot of storage space and maintenance cost. Among those intermediate dataset some may be used in future and some may not but encrypting all kind of intermediate dataset is widely adopted in existing approach. Here in this approach Threshold based Filtering approach is Adopted to classify the dataset that are to be encrypted. A value $\in$ will be fixed for each intermediate dataset based on the privacy information present in the dataset. The intermediate dataset whose $\in$ value is higher than the threshold will be encrypted and remaining dataset is anonymized. For encryption two round searchable encryption(TRSE) is used to ease the searching and accessing the encrypted dataset. The encrypted dataset may be anonymized in future if the value is decreased. Evaluation results demonstrate that the privacy preserving cost of intermediate data sets can be significantly reduced over existing ones and also it guarantees high security and search efficiency.**

***Keywords*-cloud computing, data privacy, intermediate dataset, privacy upper bound.**

1.Introduction

Cloud computing provides massive computation power and storage capacity which enable users to deploy computation and data-intensive applications without infrastructure investment. computing services such as SaaS, PaaS and IaaS over the internet that are supervised by arbitrator at outback locations. Many applications such as emails, file storage, business data, etc. are outsourced to cloud server. Only authorized user can access the data from the cloud server. Outsourcing unencrypted data to cloud by the owner is not much secure because server may leak information to cyberpunks. In spite of encrypting, retrieval of data becomes an intriguing task when searching has to be made on vast data. The best way is to use keyword based search on encrypted data for data concealing.

All the datasets generated by the scientific applications can be stored, if the users are willing to pay for the required resources. Hence, for scientific applications in the cloud, whether to store the generated datasets or not is not an easy decision anymore. 1) Storing different datasets will lead to a very different cost. Due to the pay-as-you-go model, either storing or generating a dataset carries certain cost. The datasets vary in size, and have different generation costs and usage frequencies. On one hand, it is most likely not cost effective to store all these datasets in the cloud. On the other hand, if we delete them all, regeneration of frequently used datasets would normally impose a high computation cost. Hence we need a strategy to balance the generation cost and

ISSN 2348 - 7968

the storage cost of the application datasets in order to reduce the total application cost.

Existing technical approaches for preserving the privacy of data sets stored in cloud mainly include encryption and anonymization. On one hand, encrypting all data sets, a straightforward and effective approach, is widely adopted in current research. However, processing on encrypted data sets efficiently is quite a challenging task, because most existing applications only run on unencrypted data sets. Although recent progress has been made in homomorphic encryption which theoretically allows performing computation on encrypted data sets, applying current algorithms are rather expensive due to their inefficiency. On the other hand, partial information of data sets, e.g., aggregate information, is required to expose to data users in most cloud applications like data mining and analytics. In such cases, data sets are anonymized rather than encrypted to ensure both data utility and privacy preserving. Current privacy-preserving techniques like generalization can withstand most privacy attacks on one single data set, while preserving privacy for multiple data sets is still a challenging problem. Thus, for preserving privacy of multiple data sets, it is promising to anonymize all data sets first and then encrypt them before storing or sharing them in cloud. Usually, the volume of intermediate data sets is huge. Hence, we argue that encrypting all intermediate data sets will lead to high overhead and low efficiency when they are frequently accessed or processed. As such, we propose to encrypt part of intermediate data sets rather than all for reducing privacy-preserving cost.

2. Related Work

We briefly review the research on privacy protection in cloud, intermediate data set privacy preserving and Privacy-Preserving Data Publishing (PPDP). Currently, encryption is exploited by most existing research to ensure the data privacy in cloud. Although encryption works well for data privacy in these approaches, it is necessary to encrypt and decrypt data sets frequently in many applications. Encryption is usually integrated with other methods to achieve cost reduction, high data usability and privacy protection. **Roy et al**.[**3**]investigated the data privacy problem caused by MapReduce and presented a system named Airavat which incorporates mandatory access control with differential privacy. **Puttaswamy et al[4]**.described a set of tools called Silverline that identifies all functionally encryptable data and then encrypts them to protect privacy. **Zhang et al[6].**proposed a system named Sedic which partitions MapReduce computing jobs in terms of the security labels of data they work on and then assigns the computation without sensitive data to a public cloud. The sensitivity of data is required to be labeled in advance to make the above approaches available. **Ciriani et al[8]**. proposed an approach that combines encryption and data fragmentation to achieve privacy protection for distributed data storage with encrypting only part of data sets. We follow this line, but integrate data anonymization and encryption together to fulfill cost-effective privacy preserving. The importance of retaining intermediate data sets in cloud has been widely recognized ,but the research on privacy issues incurred by such data sets just commences. **Davidson et al[9].** studied the privacy issues in workflow provenance, and proposed to achieve module privacy preserving and high utility of provenance information via carefully hiding a subset of intermediate data.

3 Motivating Example and Problem Analysis

Section 3.1 shows a motivating example to drive our research.the problem of reducing the privacy-preserving cost incurred by the storage of intermediate data sets is analyzed in Section 3.2.

3.1 Motivating Example

A motivatingscenario is illustrated in figure, where on online all data are moved into cloud for economical benefits. Original data sets are encrypted for confidentiality. Intermediated data sets generated during data access or process are retained for datareuse and cost saving. Thus through the generated dataset an adversary may know the information present in the set.so encrypting the intermediate data set is essential for privacy but at the same time the cost that require for privacy should not be high.

ISSN 2348 - 7968

In most real-world applications, a large number of intermediate data sets are involved. Hence, it is challenging to identify which data sets should be encrypted to ensure that privacy leakage requirements are satisfied while keeping the hiding expenses as low as possible.

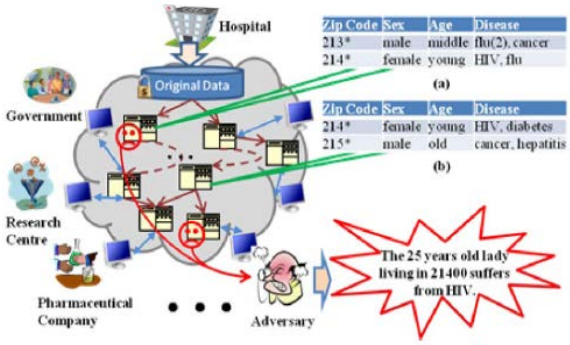

Fig 1. A Scenario showing privacy threats due to intermediate data sets

3.2 Problem Analysis

Encrypting all the data set is time consuming process and unnecessary. Storage will be consumed more when storing large amount of encrypted data. And finally the most important issue is economical cost.

4. The proposed system

Hence, it can be argued that encrypting all intermediate datasets will lead to high overhead and low efficiency when they are frequently accessed or processed. As such, I propose to encrypt part of intermediate datasets rather than all for reducing privacy-preserving cost. In this paper, a novel approach is introduced to identify which intermediate datasets need to be encrypted while others do not, in order to satisfy privacy requirements given by data holders. A tree structure is modeled from generation relationships of intermediate datasets to analyze privacy propagation of datasets.

As quantifying joint privacy leakage of multiple data sets efficiently is challenging, it exploit an upper bound constraint to confine privacy disclosure. Based on such a constraint, the problem is modelled for saving privacy-preserving cost as a constrained optimization problem. This problem is then divided into a series of sub problems by decomposing privacy leakage constraints. Finally, a threshold based filtering algorithm is adopted accordingly to identify the data sets that need to be encrypted. Experimental results on real-world and extensive data sets demonstrate that privacy-preserving cost of intermediate data sets can be significantly reduced with our approach over existing ones where all data sets are encrypted.The major contributions of the research are threefold. First, It is formally demonstrate the possibility of ensuring privacy leakage requirements without encrypting all intermediate data sets when encryption is incorporated with anonymization to preserve privacy. Second, a practical heuristic algorithm to identify which data sets need to be encrypted for preserving privacy while the rest of them do not. Third, experiment results demonstrate that this approach can significantly reduce privacy-preserving cost over existing approaches.

4.1 Architecture

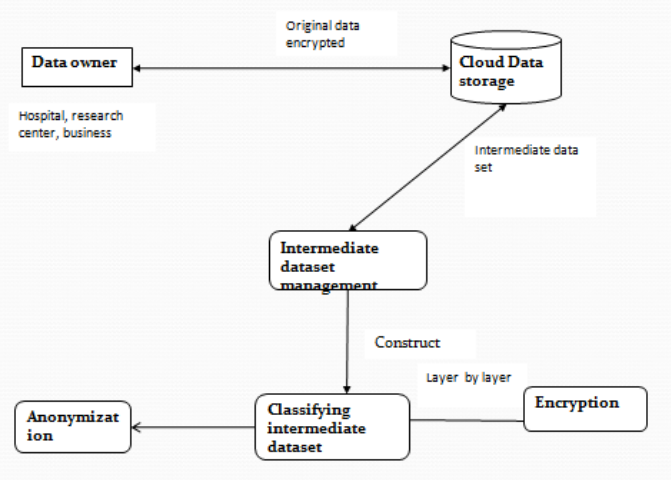

Fig.2 System Architecture

5.conclusion

In this work it deals with a new approach that identifies which part of intermediate data sets needs to be encrypted while the rest does not, in order to save the privacy preserving cost. For that a tree structure has been modeled from the generation relationships of intermediate datasets to analyze privacy

ISSN 2348 - 7968

propagation among data sets. Next a practical heuristic algorithm is designed to classify the intermediate dataset which will reduce the cost of maintaining the stored intermediate dataset.

5 References

[1] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," Comm. ACM, vol. 53,no. 4, pp. 50-58, 2010.

[2] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic,"Cloud Computing and Emerging It Platforms: Vision, Hype, and Reality for Delivering Computing as the Fifth Utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, 2009.

[3] I. Roy, S.T.V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel,"Airavat: Security and Privacy for Mapreduce," Proc. SeventhUSENIX Conf. Networked Systems Design and Implementation (NSDI'10), p. 20, 2010.

[4] K.P.N. Puttaswamy, C. Kruegel, and B.Y. Zhao, "Silverline:Toward Data Confidentiality in Storage-Intensive Cloud Applications,"Proc. Second ACM Symp. Cloud Computing (SoCC '11), 2011.

[5] D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues," Future Generation Computer Systems, vol. 28, no. 3, pp. 583-592, 2011.

[6] K. Zhang, X. Zhou, Y. hen, X. Wang, and Y. Ruan, "Sedic:Privacy-Aware Data Intensive Computing on Hybrid Clouds,"Proc. 18th ACM Conf. Computer and Comm. Security (CCS '11),pp. 515-526, 2011.

[7] V. Ciriani, S.D.C.D. Vimercati, S. Foresti, S. Jajodia, S. Paraboschi,and P. Samarati, "Combining Fragmentation and Encryption toProtect Privacy in Data Storage," ACM Trans. Information andSystem Security, vol. 13, no. 3, pp. 1-33, 2010.

[8] S.B. Davidson, S. Khanna, T. Milo, D. Panigrahi, and S. Roy,"Provenance Views for Module Privacy," Proc. 30th ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems(PODS '11), pp. 175-186, 2011.

[9] S.B. Davidson, S. Khanna, S. Roy, J. Stoyanovich, V. Tannen, and Y. Chen, "On Provenance and Privacy," Proc. 14th Int'l Conf. Database Theory, pp. 3-10, 2011.

[10] S.B. Davidson, S. Khanna, V. Tannen, S. Roy, Y. Chen, T. Milo, and J. Stoyanovich, "Enabling Privacy in rovenance-Aware Workflow Systems," Proc. Fifth Biennial Conf. Innovative Data Systems Research (CIDR '11), pp. 215-218, 2011.

[11] C. Gentry, "Fully Homomorphic Encryption Using Ideal Lattices,"Proc. 41st Ann. ACM Symp. Theory of Computing (STOC '09),pp. 169-178, 2009.

[12] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

[13] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey ," ACM Computing Survey, vol. 42, no. 4, pp. 1-53, 2010.

[14] X. Zhang, C. Liu, J. Chen, and W. Dou, "An Upper-Bound Control Approach for Cost-Effective Privacy Protection of Intermediate Data Set Storage in Cloud," Proc. Ninth IEEE Int'l Conf. Dependable,Autonomic and Secure Computing (DASC '11), pp. 518-525, 2011.

[15] W. Du, Z. Teng, and Z. Zhu, "Privacy-Maxent:integrating Background Knowledge in Privacy Quantification," Proc.ACMSIGMOD Int'l Conf. Management of Data SIGMOD '08), pp. 459-472, 2008.