A Novel Approach for Privacy Preservation of Intermediate Datasets in Cloud

Ms. Chanamthabam Monica Roy
2nd year Mtech,Dept of ISE
The Oxford College of Engineering
Bangalore,Karnataka,India

Mrs. Sowmya R
Assistant Professor,Dept. of ISE
The Oxford College of Engineering
Bangalore,Karnataka,India

***Abstract*—Massive computation power and storage capacity of cloud computing systems allow users to deploy computation and data intensive applications without infrastructure investment, where large application datasets can be stored in the cloud. Along the processing of data-intensive applications, a large number of intermediate datasets are generated which are often stored so as to save the cost of recomputing them. The impact of privacy requirements in the development of modern applications is increasing very quickly. For preserving privacy of datasets in cloud, encrypting all datasets are widely adopted in existing approaches. But the problem arises in preserving privacy of the large volume of intermediate datasets generated as this will neither be time efficient nor cost-effective. Hence, in this paper, we present a combination of anonymisation and encryption to enable securing of intermediate datasets. A heuristic algorithm is also used to identify which intermediate datasets are to be encrypted rather than encrypting all the intermediate datasets so as to enable cost effective security.**

Keywords— datasets, intermediate datasets, anonymisation, encryption,heuristic.

I. INTRODUCTION

There are various advantages of Cloud computing which includes storage, on-demand self-service, global network access because of which Cloud computing has a major role in the history of Information Technology [1]. Customers can save a huge amount of capital investment by using cloud IT infrastructure. The new trend in the IT business is to employ cloud infrastructure so as to save capital investment. When the original datas which are uploaded in the cloud are processed by various data users, intermediate datasets are generated. These intermediate datasets are retained so as to restrict the cost of recomputing them [8]. The privacy concerns caused by retaining intermediate data sets in cloud are important but they are paid little attention. The intermediate datasets are processed by many users but the privacy of these datasets are not taken care of by even the original dataset holder. This leads to privacy leakage of sensitive information which can lead to economic loss or social reputation damage to data owners [2]. Various approaches had been introduced for preserving the privacy of datasets which are stored in cloud, which includes anonymisation and encryption of all datasets. Anonymisation works well with single datasets and on the other hand, encrypting all the datasets generated is not an efficient approach as most applications run on unencrypted datasets. However, many progress has been made to improve the limitation of encryption by introducing the concept of homomorphism encryption[3], which supports computation without decrypting the input. But homomorphism encryption is theoretically very good but not so noble for practical use. In this paper, we combine anonymisation and encryption for preserving privacy of intermediate datasets. A heuristic algorithm is also used to identify which intermediate datasets are to be encrypted and which does not need to be encrypted so as to enable cost effective security.

II. PROBLEM STATEMENT

The privacy of datas which are uploaded in cloud needs to be preserved. As the volume of intermediate datasets generated by processing original datasets are huge, encrypting all the intermediate datasets to preserve the privacy does not become cost-effective.

III. EXISTING APPROACHES

Existing technical approaches for preservation of privacy of data sets stored in cloud mainly includes encryption and anonymisation [7]. Using Encryption to preserve privacy of datasets is widely adopted. But encrypting all datasets is not so efficient as most of the existing applications works only on unencrypted datas[5]. But recently progress has been made in homomorphism encryption which theoretically allows performing computation on encrypted datasets[4]. However, using the current homomorphic encryption algorithm is not very efficient as in most cloud applications like data mining and analytics, partial information of data sets is required to be exposed to data users. In cases like this, anonymisations of datasets are used [6]. Using anonymisation alone can preserve privacy of single data sets but it is still a challenging problem to preserve privacy for multiple datasets.

IV. PROPOSED WORK

In this section, we present a novel approach wherein we combine anonymisation and encryption to preserve privacy of multiple intermediate datasets[9]. At first, the intermediate datasets are anonymized and then encryption is done. Usually the volume of intermediate datasets generated in cloud are high, so it is a challenging task to identify which intermediate datasets should be encrypted and which are not needed to be encrypted. So, a privacy-preserving heuristic algorithm is proposed and applied to the anonymized intermediate

datasets so as to identify which part of intermediate data sets are to be encrypted and which are not so that the privacy-preserving cost is reduced.

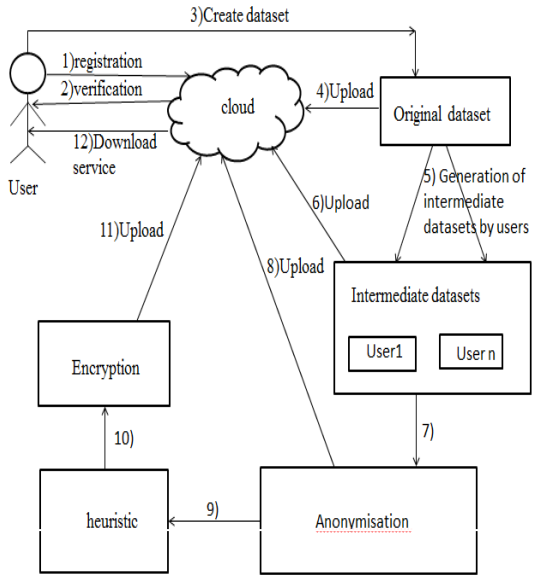

Fig .1. Proposed Architecture

In the proposed scheme, the user registers itself with any cloud service provider so as to access the cloud. The user then creates the original dataset and uploads it to the cloud. Intermediate datasets are generated when the original datasets are processed or accessed by other users. For Example, for an online Health Service Provider which has stored its datas on cloud , users can be some government hospitals or some research centres. These intermediate datasets are anonymized and then a heuristic algorithm is proposed which identifies which intermediate datasets are to be encrypted. Encryption is done on the intermediate datasets which are identified by the heuristic algorithm [10]. The cost of preserving privacy is reduced using the proposed approach as compared to the cost of encrypting all the intermediate datasets as done in the existing approaches.

V. RELATED WORK

In this section we briefly mention the previous work carried out by various researchers on privacy preserving of datas stored in cloud. In existing research, encryption is widely adopted to ensure privacy of data stored in cloud. But encrypting all datasets is not so efficient as most of the existing applications works only on unencrypted datas and also using encryption requires frequent encryption and decryption of data sets. So in order to improve the situation, Encryption is usually integrated with other methods to achieve cost reduction high data usability and privacy protection. Puttaswamy et al. [11] described a set of tools called Silverline that identifies all functionally encryptable data and then encrypts them to protect privacy. Zhang et al.[17] 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. Ciriani et al. [9] proposed an approach that combines encryption and data fragmentation to achieve privacy protection for distributed data storage with encrypting only part of data sets. Keeping in view the works carried out by the various researchers, in this paper we combine data anonymisation and encryption together to fulfill cost-effective privacy preserving. Using anonymisation alone can preserve privacy of single data sets but it is still a challenging problem to preserve privacy for multiple datasets. So, integrating data anonymisation and encryption together we overcome the limitations of using only anonymisation and encryption. Our approach preserves the privacy of datas stored in the cloud better than other previously existing approaches.

VI. CONCLUSION

The use of cloud computing is an important development in the world. Many IT business has started using cloud architecture because of its pay-as-you-go concept. Security is one most important factor that everyone thinks before uploading datas in the cloud. In this paper, we have proposed a combination of anonymisation and encryption to enable securing of intermediate datasets. A privacy preserving heuristic algorithm is also proposed to identify which intermediate datasets are to be encrypted rather than encrypting all the intermediate datasets so as to enable cost effective security.

ACKNOWLEDGEMENT

I take this opportunity to express my profound gratitude and deep regards to my guide Mrs.Sowmya R, Assistant Professor, The Oxford College of Engineering, Bangalore, for her exemplary guidance, and constant encouragement throughout.

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 CloudComputing,” Comm. ACM, vol. 53,no. 4, pp. 50-58, 2010.

[2] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, “Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data,” Proc.IEEE INFOCOM ’11, pp. 829-837, 2011.

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

[4] H. Takabi, J.B.D. Joshi, and G. Ahn, “Security and Privacy Challenges in Cloud Computing Environments,” IEEE Security & Privacy, vol. 8, no. 6, pp. 24-31, Nov./Dec. 2010.

[5] D. Yuan, Y. Yang, X. Liu, and J. Chen, “On-Demand Minimum Cost Benchmarking for Intermediate Data Set Storage in Scientific Cloud Workflow Systems,” J. Parallel Distributed Computing,vol. 71, no. 2, pp. 316-332, 2011.

[6] B. C. M. Fung, K. Wang, and P. S. Yu. Anonymizing classification data for privacy preservation. IEEE Transactions on Knowledge and Data Engineering, 19(5):711–725, 2007.

[7] M. Li, S. Yu, N. Cao, and W. Lou, “Authorized Private Keyword Search over Encrypted Data in Cloud Computing,” Proc. 31st Int’lConf. Distributed Computing Systems (ICDCS ’11), pp. 383-392, 2011

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

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

[10] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in Proc. of EUROCRYPT, 2004.

[11] 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