International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

A Cost Effective Approach for Privacy Preservation in Cloud

Feby Cherian
M.E Computer and Communication
PSNA College of Engineering and Technology
Dindigul, India

Mrs. R. Sivakami,M.E,(Ph.D)
Associate Professor, IT Department
PSNA College of Engineering and Technology
Dindigul, India

***Abstract*—Cloud computing helps users to store enormous amount of data and permits them to process or retrieve the data whenever required without the aid of any infrastructure investment. But in the cloud, the confidentiality of the data is a major concern. When processing the data, it is to be classified into multiple data sets. So in order to uphold the privacy, the common method used is to encrypt all the data sets and store in the cloud. But this method will increase the cost and also will consume more time which would be a big task. In this paper, we recommend an approach which will calculate the sensitive information of all intermediate data sets and stored in a queue in the order of the importance of sensitive information. Based on a specific threshold value, user can select how much is the degree of encryption, and which will encrypt the most sensitive data sets and keep insensitive data sets without encryption. This will reduce the cost and also save the time without losing the privacy of data.**

Keywords—Cloud computing; confidentiality: data privacy; threshold; degree of encryption; sensitive data

I. INTRODUCTION

Cloud computing denotes to delivery of computer services which lets the users to use the software and hardware which are managed in remote locations by the Cloud Service Providers. This helps the cloud customers to save their money by investing in IT servers and thereby ponder in their own primary businesses.

Security and privacy are the major challenges that we are facing in cloud computing but they are paid little attention. The price for storing the data in the cloud is directly proportional to the volume of data. The cloud users may store intermediate data sets for computation purposes so that they can reduce expenses by computing the same data sets again and again[1]. The storage of these data sets increases the risk of privacy being disrupted. Multiple vendors are able to access the same data sets. This increases the chances for adversary to get the sensitive information by retrieving the multiple intermediate data sets. It might lead to financial losses or it might also affect the social status of the data owner.

Prevailing technical approaches for preserving the privacy of data sets stored in cloud mainly comprise of encryption and anonymization. On one point, encrypting all data sets, an up-front and effective approach, is commonly adopted in current exploration. Yet, processing on encrypted data sets competently is quite a challenging job, because most existing approaches only run on unencrypted data sets. Still, conserving the privacy of intermediate datasets becomes a challenging problem because adversaries may recover privacy-sensitive information by analyzing multiple intermediate datasets.

Encrypting all datasets in cloud is generally adopted in existing approaches to address this challenge [2]. But in our approach we are arguing that encrypting all intermediate datasets are neither efficient nor cost-effective because it consumes more time and not cost effective for data-intensive applications to en/decrypt datasets frequently while performing any operation on them.

We have categorized our research into three. First, we demonstrate how we can ensure privacy leakage requirements without encrypting all intermediate data sets. Second speaks about the algorithm which finds the data sets need to be encrypted for preserving privacy. Third one is that experiment results demonstrate how privacy-preserving costs get reduced by our approach over existing approaches.

II. RELATED WORK

Nowadays, popular scientific workflows are frequently deployed in data privacy preservation and privacy protection in cloud computing environments. These privacy and security related issues have been extensively studied in the research area and they have made prolific progresses with a variety of privacy models and privacy preserving methods. But most of the security algorithms are lacking scalability over the data. In the year of 2007, the idea of cloud computing was proposed [3] and it is to be considered as the next generation of IT platforms that can deliver computing as a kind of utility.

Encryption is usually combined with other methods to achieve cost reduction, high data usefulness and privacy protection. Roy et al. [4] put forward the data privacy problem that was reasoned by Map Reduce and presented a system called Airavat which incorporates mandatory access control with differential privacy. Puttaswamy et al. [5] proposed a set of tools called Silverline that encrypt all functionally data and then encrypts them to protect privacy. Zhang et al. [6] proposed a work called Sedic which partitions Map Reduce computing jobs in terms of the security labels of data they work on and then assigns the computation without including sensitive data to a public cloud.

The sensitivity of data in cloud is required to be labeled in advance to make the above techniques. Ciriani et al. [7] put forward a method that encryption and data fragmentation together combines to achieve the privacy protection for distributed data storage with encrypting only selected data sets. By taking this theory, we are integrating data anonymization and encryption together to fulfill cost-effective privacy preserving.

The significance of storing data sets in cloud has been widely recognized. Davidson et al. [8] proposed the privacy based issues in the workflow provenance and proposed to achieve module privacy preserving and high utility of provenance information via warily hiding a subset of intermediate data. This idea is similar to ours, yet our research mainly focus on data privacy preserving from an economical cost perspective while theirs concentrates mainly on functionality privacy of workflow modules rather than data privacy. Our technique also differs in several aspects such as cost models, privacy preservation and data hiding techniques. But our approach can be used for selection of hidden data items in their research if economical cost is considered.

The research community PPDP has broadly investigated on privacy-preserving issues and made fruitful progress with a diversity of privacy models and preserving methods. Many anonymization techniques like generalization [10] have been used to preserve privacy of data, but these techniques work alone will fail to meet the problem of preserving privacy for multiple data sets. Our concept combines anonymization with encryption to achieve privacy preserving of multiple data sets.

III. PROBLEM ANALYSIS

The privacy disquiets caused by retaining intermediate datasets in cloud are important but they are paid little attention. For an example let's consider the scenario, Microsoft Health Vault, has moved data storage into cloud for economic benefits. Then the original data sets are encrypted for maintaining confidentiality. Data users like government or pharmaceutical company access or process the part of original data sets after anonymization. Intermediate data sets that are generated during data processing is stored in cloud database for data reuse and cost saving.

Two independently created intermediate data sets in Fig.1 are anonymized to satisfy two diversity, that means at least two individuals own the same quasi-identifier and every quasi-identifier corresponds to at least two sensitive. A lady who has 25 years age living in 21400 (corresponding quasi-identifier is (214 * female, young) is in both data sets, an adversary analyzing the data can conclude that this individual suffers from HIV with high confidence if (a) and (b) are collected together. Hiding either (a) or (b) by encryption is a promising way to prevent such a Privacy leakage. Assume (a) and (b) are of the same size, the frequency of retrieving (a) is 1000 and that of (b) is 10000. We hide (a) to preserve privacy because this can be done in less cost than hiding (b). In all real-world approaches, a larger number of intermediate data sets are involved. Hence, it is certainly challenging to identify which data sets should be encrypted to ensure that privacy leakage requirements are satisfied and also by keeping the hiding expenses as low as possible

We suggested an approach that combines encryption and data fragmentation to attain privacy protection for distributed data storage with encrypting only part of datasets.

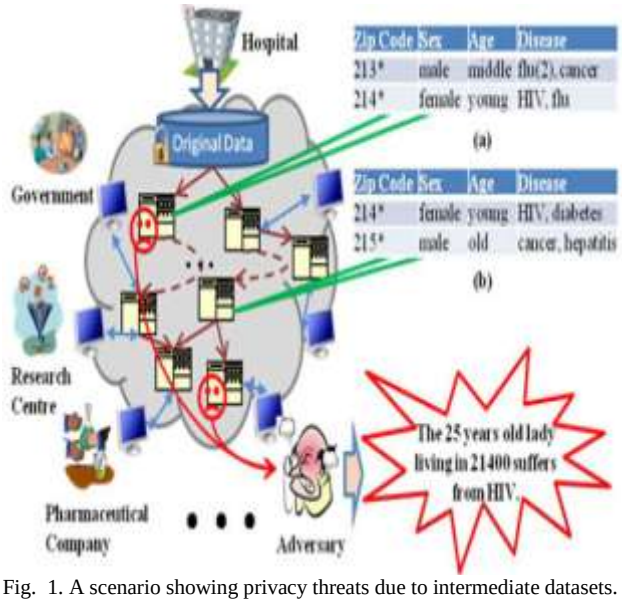

Fig. 1. A scenario showing privacy threats due to intermediate datasets.

The privacy concerns affected by retaining intermediate data sets in cloud are important but they are paid little attention. Original data sets are encrypted for attaining confidentiality and authentication. Data users like governments or other research centers access or process part of original data sets after anonymization. The Intermediate data sets generated during the data access or process are retained for data reuse and cost saving. We suggested an approach that combines encryption and data fragmentation to achieve privacy protection for distributed data storage with encrypting only part of datasets.

A. *Privacy Preserving Method*

The privacy-preserving methods like generalization can with-stand most privacy attacks on one single data set, though preserving privacy for multiple data sets is still a challenging task. The best thing to preserve privacy of multiple data sets is to anonymize all datasets first and then encrypt them before storing or sharing them in cloud. The fig.3 shows the steps that involved in the privacy preservation of intermediate data sets. The cost for Privacy-preserving of intermediate data sets stems from frequent en/decryption with charged cloud services.

B. *Intermediate Dataset Handling*

An intermediate data set is expected to have been anonymized to satisfy certain privacy requirements. But, putting multiple data sets together may still invoke a high task of revealing privacy-sensitive information, leads to the violation of privacy requirements. Data provenance is hired to manage intermediate datasets in our study.

Provenance [1] is defined as the origin, source or history of source of specific objects and data, which can be counted as the information upon how data was produced. Re-reducibility of data provenance can help to revive a data set from its nearest existing predecessor data sets rather than from scratch. Directed Acyclic Graph (DAG) is used to capture the topological structure of generation relationships among the data sets. Sensitive Intermediate data set Graph, represented as

SIG is defined as DAG representing the generation relationships of intermediate data sets DS from ds0.

Sensitive Intermediate data set Tree (SIT) is nothing but SIG in a tree structure (fig. 2). The root of the tree is ds0. An SIG or SIT not only represents the generation relationships of an original data set and its intermediate data sets, but also captures how the privacy-sensitive information moves among such data sets. Commonly, the privacy-sensitive information in ds0 is spread into its offspring data sets. Hence, an SIG or SIT can be used to analyze privacy disclosure of multiple data sets.

C. Privacy Upper Bound Constraint

Privacy quantification of a single data set is specified in the privacy upper bound constraint. The challenge of privacy quantification of multiple data sets and then derives a privacy leakage upper-bound constraint consistently.

An upper-bound constraint based method to select the necessary subset of intermediate data sets that needs to be encrypted for reducing the cost of privacy preserving. The fig.2 represents the privacy leakage upper-bound constraint is disintegrated layer by layer. Based on these properties of SIT tree generating method the initial constraint of the dataset depends upon the threshold value. Here the data is categorized into two as necessary to encrypt and unnecessary to encrypt as left and right side of the nodes respectively. Each layer maintains various threshold values in the tree.

D. Minimum Privacy-Preserving Cost

Usually, more than one viable global encryption solution occurs under the PLC constraints, as there are many other solutions in each layer. Each intermediate dataset having different frequencies of usage and size which leading to different overall cost with different solutions. The values are to be calculated for the privacy preserving cost. And the constraint for the resulted ideals should be lesser than the threshold. Finally, we are identifying the minimum privacy preserving cost.

E. Heuristic Cost

The goal state in our algorithm is to find a near-optimal solution in a limited search space. Based on this heuristic, we design a heuristic privacy preserving cost reduction algorithm. The proposed algorithm can attain a near-optimal solution practically.

SORT and SELECT are the simple external functions as their names signify. Lastly we can able to identify sensitive data set. Heuristic value is attained by heuristic function [11]. Heuristic function is used to calculate the heuristic value of state node (SNi) of a particular layer. The Heuristic value of one layer can be calculated by the following formula:

$$F(SNi)=Ccur/(€-€i+1+(€i+1.Cdes.BFAVG)/PrLeAVG \quad (1)$$

Where Ccur signifies the privacy preserving cost Cdes represents total cost of the data sets and BFAVG signifies average brand factor. This algorithm works by selecting a state node having the highest heuristic value and goes to its child state nodes until it ranges the goal. In this algorithm a priority queue is used to possess the nodes which add the qualified

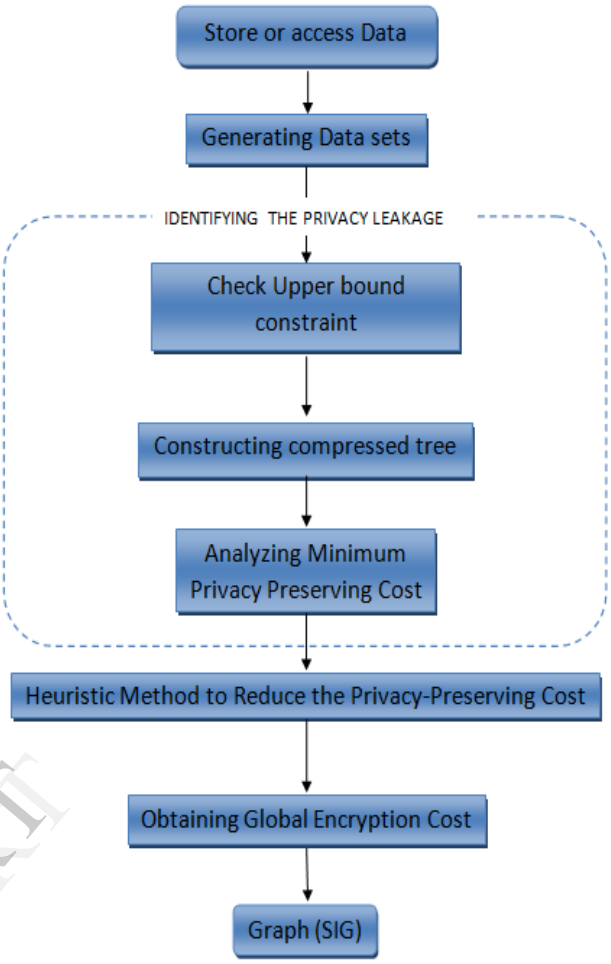

Fig. 2. The flow chart showing the construction of SIG graph.

state nodes. While adding the child nodes into its priority nodes the algorithm generates a local encryption solution. Based on the cost and the privacy leakage algorithm it sorts the data sets. Thus the data sets with lower privacy leakage are estimated to remain not encrypted.

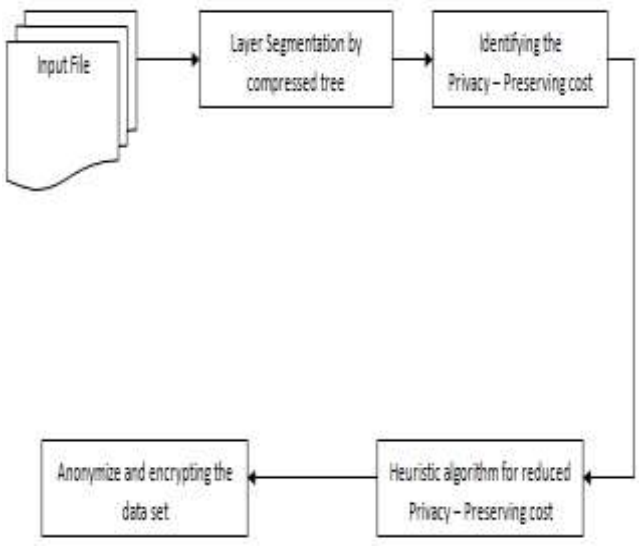

Fig. 3. Block Diagram

IV. ANONYMIZATION AND ENCRYPTION

After identifying the data need to be encrypted we will split the intermediate data set. The intermediate dataset which contains valuable information will be encrypted and anonymized.When an adversary login to view the intermediate data cannot able to identify the original data (fig.3). Generate the intermediate data set with the encrypted and anonymized values.

V. CONCLUSION AND FUTURE WORK

In this paper, we put forward a distinct approach to identify which intermediate data sets need to be encrypted while others do not, in order to accomplish privacy requirements given by data holders. A tree structure is modeled from generation relationships of intermediate data sets to assess privacy propagation of data sets. Centered on such a limit, we design the problem of saving privacy-preserving cost as a constrained optimization problem. This problem is then allocated into a series of sub-problems by crumbling privacy leakage constraints. At the end, we suggest a practical heuristic algorithm accordingly to identify the data sets that want to be encrypted. The experimental results on real-world and extensive data sets exhibit that privacy preserving cost of intermediate data sets can be considerably reduced with our method over existing ones where all data sets are encrypted.

REFERENCES

[1] 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.

[2] Liu C, Zhang X, Yang C, Chen J. Ccbke—session key negotiation for fast and secure scheduling of scientific applications in cloud computing. Future Generation Computer Systems 2013; 29(5):1300–1308.

[3] A. Weiss, Computing in the cloud, ACM Networker 11 (2007) 18–25.

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

[5] Puttaswamy KPN, Kruegel C, Zhao BY. Silverline: Toward Data Confidentiality in Storage-Intensive Cloud Applications. Proceedings of the 2nd ACM Symposium on Cloud Computing (SoCC'11), 2011; Article 10.

[6] K. Zhang, X. Zhou, Y. Chen, 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 to Protect Privacy in Data Storage," ACM Trans. Information and System Security, vol. 13, no. 3, pp. 1-33, 2010.

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

[9] Cao N, Wang C, Li M, Ren K, Lou W. Privacy-preserving Multi-Keyword Ranked Search Over Encrypted Cloud Data. Proceedings of the 31st Annual IEEE International Conference on Computer Communications (INFOCOM'11), 2011; 829–837.

[10] 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.

[11] 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.