A Cost-Effective Framework for Cloud-Based Data Privacy Preservation

Akshara R. Kulkarni, Dr. Shrikant G. Patil, and Rohan R. Deshmukh

Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology, Nagpur, India - 440010.

ABSTRACT

In recent the most of the companies are using the cloud to store their huge database. Cloud provides the large space for storage. Cloud is nothing but the pay-as-you-go is used as the economical aspect of privacy-preserving. Privacy for this data is provided by encryption of the data. But there are chances of attack so the violence of the data is possible. For data protection the anonymization and then encryption of data is held on. But whenever the user tries to re-access the data, it should be decrypted. At every time of accessing the encryption and the decryption of the data should be done, which increases the cost of the privacy-preservation also it is the time consuming process as the large number of keys used for encryption and decryption. To reduce this cost of privacy-preservation the privacy leakage constraint is used in which the problem is divided in subproblems and then finding the solution. Then the data is divided into the intermediate datasets. The threshold value is used to privacy-preservation which gives the low cost privacy-preservation. Here the privacy-preserving cost reducing heuristic algorithm is useful for the privacy leakage.

Keywords: *data storage privacy, privacy leakage, intermediate datasets.*

I. INTRODUCTION

Cloud computing referred as application delivered as services over internet and hardware-software systems. In recent, the number of IT business offers the IT services and large database is to be saved of the business. The companies' infrastructures invest a lot on core business. If the cloud is used for the database storage then business investment can be concentrated on only the core business. So in recent era cloud is most popular for database as the advantages of cloud like security and privacy.

In some cases data users often reanalyze results, and conducts new analysis of the data sets. But there are more chances of attack that causes the risk of violence. Generally the data on cloud is accessed and processed by multiple users not only the own holders. So there is need of data sets and privacy preservation. Privacy preservation with low cost to data owners. Encryption is one of the effective approaches preserving the privacy of different datasets in cloud. As the most of the applications are run on the unencrypted datasets only, so the encrypted datasets are challenging to process. There are some algorithms which are used to perform computations on the encrypted datasets. But these algorithms are theoretical, so are more expensive to apply due to efficiency. In some cases the partial information is required to be given to data users in applications like data mining and analytics. In these cases data is anonymized instead of instead of encrypting for data utility and the preservation of privacy for single data sets privacy preservation generalization technique is used but the multiple datasets problem is challenging. Thus, in this case all the datasets anonymized first and then encrypted before storing in the cloud. As the volume of intermediate dataset is huge, encryption will lead to high overhead and low efficiency when frequently accessed or processed. Thus first encrypt part of dataset rather than all which reduces the cost. First of all we must know which intermediate datasets need to be encrypted, which not, in order to satisfy privacy requirements by data holders. From the generation of relationships of intermediate datasets to analyze privacy preservation of datasets, a tree structure is modeled. As the quantifying joint privacy leakage of multiple datasets efficiently is challenging task. So we exploit the upper constraint to confine privacy of intermediate datasets. Based on their constraints we considered the problem of saving privacy preservation cost as constraint based problem. Then this problem is divided into series of sub-problems by decomposing the constraints.

II. RELATED WORK

The feature of cloud computing that is pay-as-you-go is used as economical aspect of privacy preserving. Once the data is identified to be encrypted, the keys must be to encrypt. But there may be problem of a lot number of keys for encryption. So we can encrypt all such data with single key and share key to all users, but this arises a problem that could obtain the key by posing as legitimate user. So the confidentiality of the date is compromised. Thus instead of this we can use different keys to encrypt different data gives robustness, but it raises the key management problem. Our goal is to conclude right granularity automatically for robustness and management complexity. So we partition data into subsets, where each data subset is accessed by the same group of users. We then encrypt each data subset using a different key, and distribute keys to groups of users that should have access.

Royet al. investigated the data privacy problem caused by Map Reduce and presented a system named Airavat[1] which gives mandatory access control for different privacy. Puttaswamy et al told a set of tools called Silverline[2] which identifies all encryptable data, then encrypts it to provide privacy. Zhang et al. invented a system named Sedic[3] which partitions MapReduce computing jobs in security labels of data they work on and then assigns the processes without sensitive data on a public cloud. The sensitivity of data is needed to be labeled to make the above approaches available. Ciriani et al.[4] suggested an approach that combines data fragmentation and encryption to achieve privacy for distributed data storage that encrypting only required part of datasets. We follow this, but integrate data encryption and anonymization together to fulfill cost-effective privacy preservation. The importance of retaining intermediate datasets in cloud has been recognized, but the research on privacy issues taken by such datasets. Davidson et al. concluded the privacy issues in work flow provenance, and proposed to achieve module privacy preserving and high utility of provenance information via hiding a subset of intermediate data. This idea is similar to ours, but our focus is on data privacy preservation from an economical cost perspective while theirs concentrates on functionality privacy of work flow modules instead of data privacy. Our focus also different from theirs in multiple aspects such as privacy quantification, data hiding techniques and cost models. But our approach can be used for selection of hidden data items in their research if economical cost is considered. Privacy principles for multiple datasets are also suggested, but they aim at particular scenarios such as continuous data publishing or sequential data releasing. The research in feat information theory to quantify the privacy by utilizing the maximum entropy principle. The privacy quantification is based on the work. Many anonymization techniques like generalization have been given to preserve privacy, but these methods singly fail to solve the problem of privacy preservation for multiple datasets. Our approach incorporates anonymization with encryption to achieve privacy preservation of multiple datasets.

III. APPROACHES

Our approach works by automatically identifying subsets of an application's data that are not directly used in computation, and exposing them to the cloud only in encrypted form.

- We present a technique to separate encrypted data into parts that are accessed by different sets of users. Sophisticated key assignment limits the damage possible from a given key compromise, and strikes a good trade off between robust and key management complexity.

- We present a technique that enables clients to store and use their keys safely while preventing cloud-based service from theft of the keys. Our solution works today on original web browsers.

IV. PRIVACY PRESERVING COST PROBLEM

Privacy-preserving cost of intermediate datasets from frequent en/decryption with charged cloud services. Cloud service venders have set up various pricing models to support the pay-as-you-go model, e.g., Amazon Web Services cost evaluation model. Practically, computation power, data storage and other cloud services are required for encryption or decryption. To avoid cost evaluation details and focus on the discussion of our core ideas, we combine the prices of various services required by en/decryption into one. This combined price is denoted as PPRR. PPRR indicates the overhead of en/decryption on per GB data per execution. Datasets in DD can be divided into two sets. One is for encrypted datasets, denoted as. The other is for unencrypted datasets, denoted as Then, the equations $DD_{enc} \cup DD_{une} = DD$ and $DD_{enc} \cap DD_{une} = \varphi$ hold. We define the pair, DD_{une} as a global privacy-preserving solution. The privacy-preserving cost incurred by a solution, DD_{une} is denoted as Cpp (DD_{enc} , DD_{une}). With the notations framed above, the cost

Cpp (,DD_{une}) in a given period [TT0 , T], can be deduced by the following formula:

Cpp(DD_{enc} ,$DD_{unerlangle}$)= ∫(∑dd_{ii} DD_{enc} SS_{ii} . PPRR.ff_{ii}.tt).dd_{tt}.(1)

tt=TT0

The privacy-preserving cost rate for Cpp(, DD_{une}), denoted as CRPpp, is defined as follows:

CRpp ∑dd_{ii} ϵ DD_{enc} SS_{ii} . PPRR

In the real world, SS_{ii} and ff_{ii} possibly vary over time, but we assume herein that they are static so that we can concisely present the core ideas of our approach. With this assumption, CRpp determines Cpp (langle , $DD_{une\ rangle}$) in a given period. That's why, we blur their meanings. The problem of how to make privacy-preserving cost as low as possible given a SIT can be modeled as an optimization problem on:

Minimize CRpp =∑dd_{ii} DD_{enc} SS_{ii} . PPRR. ff_{ii}, ϵ DD.

Meanwhile, the privacy leakage caused by unencrypted datasets in DD_{une} must be under a given threshold.

Privacy Leakage Upper Bound Constraint Based Approach For Privacy Preserving

We are proposing an upper bound constraint based approach which selects the appropriate subset of intermediate datasets which need to be encrypted. Specifying the relevant notations and elaborate the useful properties on SIT. The privacy leakage upper bound constraint is decomposed layer by layer. A constrained optimization problem with the PLC is then transformed into a recursive form and, a heuristic algorithm is designed for our approach. We extend our approach to a SIG.

V. PRIVACY-PRESERVING COST REDUCING HEURISTIC ALGORITHM

To justify the intermediate datasets to be encrypted with low cost privacy-preservation under privacy leakage constraint.

Here the input for the privacy-preservation cost reducing heuristic algorithm is the sensitive intermediate datasets tree (SIT) having root with attributes like size, frequencies, privacy leakage and the threshold e.

From this input we are going to calculate the privacy-preserving cost. The algorithm is as given below:

Step 1: Firstly search all the nodes with the root on SIT having the heuristic values, solution, the current cost and privacy leakage.

Step 2: Now check whether the node is empty. If yes then go to step 3 elsewhere perform the below steps:

2.1: check the sensitive columns to be encrypted with respect to threshold value. Encryption can be done by two ways:

a) if threshold e £ dataset values of the sensitive columns then encryption of the data is done.

b) if threshold e > dataset values of the sensitive columns then encryption of the data is not done on the data.

Step 3: So obtain the privacy-preserving cost by the current cost of the solution and also the other solutions.

VI. CONCLUSION

In this paper, we propose an approach to identify which intermediate datasets to be encrypted while others do not, that satisfy the privacy requirements of data holders. A tree structure is modeled from generation relationships of intermediate datasets to analyze privacy propagation of datasets. Based on such a constraint, as a

constrained optimization problem we model the problem of saving privacy-preserving cost. This problem is then divided into a series of sub-problems by decomposing privacy leakage constraints. So, we design a practical heuristic algorithm accordingly to identify the datasets that Experimental need to results on be encrypted. real-world datasets demonstrate that privacy preserving cost of intermediate datasets can be reduced with our approach over existing ones where all datasets are encrypted.

REFERENCES

1. *I. Roy, S.T.V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and Privacy for Mapreduce,"Proc. Seventh USENIX Conf. Networked Systems Design and Implementation (NSDI'10), p. 20, 2010.*
2. *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.*
3. *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.*
4. *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.*
5. *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.*
6. *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.*