ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Intermediate Data Security and Privacy Preserving in Cloud

[I]P.Raghavan, [II]J.Princess Aanish
[I]P.G Scholar, Kalasalingam Institute of Technology, Krishnan Koil
[II]Assistant Professor, Kalasalingam Institute of Technology, Krishnan Koil

Abstract
Distributed computing shows another approach to supplement the present utilization and conveyance show for it administrations dependent upon the internet, by accommodating progressively versatile and regularly virtualized assets as an administration over the internet. Information taking care of might be outsourced by the immediate cloud service provider (csp) to different elements in the cloud and propositions substances can likewise appoint the assignments to others et cetera. The utilization of distributed computing has expanded quickly in numerous associations. Commonly little and medium organizations utilize distributed computing administrations for different views, incorporating since these administrations give quick access to their requisitions and diminish their framework costs. Cloud suppliers might as well address protection and security issues as a matter of high and earnest necessity. Protecting the security of moderate datasets turns into a testing issue since foes might recoup protection touchy data by dissecting numerous halfway datasets. Encoding all datasets in cloud is generally received in existing methodologies to address this test. Different in which moderate datasets need to be encoded and which don't, so protection safeguarding cost might be spared while the security prerequisites of information holders can in any case be fulfilled. Protection safeguarding cost lessens heuristic calculation utilized for security spillage demands and sensitive intermediate information set tree/graph (sit/sig) technique is utilized.

Keywords
Cloud computing, data storage privacy, privacy preserving, intermediate data set, privacy upper bound

I. Introduction

The word "cloud" in "cloud computing" is a metaphor for the internet, and cloud computing means using the internet to compute, or to use the internet to serve your computing needs. Furthermore, cloud computing is like a huge network of computers that serve as a single computer, and its size is growing and increasing each and every day.

There are basically three parts in cloud computing. The first is the foundation; which is called the infrastructure. The infrastructure is a number of computers connected to each other and called hosts. These hosts are built and held for cloud computing. Second part is the platform; the platform is a cloud server. Moreover, a cloud server is just like a dedicated server. You can use the cloud server to put applications on the internet or in another word the cloud, and those applications are the third part of cloud computing.

More than 60% of people use cloud computing. Furthermore, most of them are not aware of the term "cloud computing".

Webmail services, online storage and software programs are all means to use cloud computing, if they are located in the web .For example, Google documents or Adobe Photoshop Express are online applications and they use cloud computing to serve people. The privacy concerns caused by retaining intermediate data sets in cloud are important but they are paid little attention. Storage and computation services in cloud are equivalent from an economical perspective because they are charged in proportion to their usage [1]. Thus, cloud users can store valuable intermediate data sets selectively when processing original data sets in data-intensive applications like medical diagnosis, in order to curtail the overall expenses by avoiding frequent re-computation to obtain these data sets [6], [7].Such scenarios are quite common because data users often reanalyze results, conduct new analysis on intermediate data sets, or share some intermediate results with others for collaboration. Without loss of generality, the notion of intermediate data set herein refers to intermediate and resultant data sets [6]. However, the storage of intermediate data enlarges attack surfaces so that privacy requirements of data holders are atrisk of being violated. Usually, intermediate data sets in cloud are accessed and processed by multiple parties, but rarely controlled by original data set holders. This enables an adversary to collect intermediate data sets together and menace privacy-sensitive information from them, bringing considerable economic loss or severe social reputation impairment to data owners. But, little attention has been paid to such a cloud-specific privacy issue. 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 [8], [9], [10]. 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 per-forming computation on encrypted data sets, applying current algorithms are rather expensive due to their inefficiency [11]. 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 anonym zed rather than encrypted to ensure both data utility and privacy preserving. Current privacy-preserving techniques like generalization [12] 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 [6]. 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.

In this paper, we propose a novel approach to identify which intermediate data sets 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 data sets to analyze privacy propagation of data sets. As quantifying joint privacy leakage of multiple data sets efficiently is challenging, we exploit an upper bound constraint to

confine privacy disclosure. Based on such a constraint, we model the problem of saving privacy-preserving cost as a con-strained optimization problem. This problem is then divided into a series of sub problems by decomposing privacy leakage constraints. Finally, we design a practical heuristic algorithm accordingly to identify the data sets that need to be encrypted. Experimental results on real-world and extensive data sets demonstrate that privacy-preser-ving 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 our research are threefold. First, we 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, we design 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 our approach can significantly reduce privacy-preserving cost over existing approaches, which is quite beneficial for the cloud users who utilize cloud services in a pay-as-you-go fashion.

II. Motivating Example and Problem Analysis

Section 2.1 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 2.2

A. Motivating Example

A motivating scenario is illustrated in Fig. 1 where an online health service provider, e.g., Microsoft HealthVault , has moved data storage into cloud for economical benefits. Original data sets are encrypted for confidentiality. Data users like governments or research centres access or process part of original data sets after anonymization.

Intermediate data sets generated during data access or process are retained for data reuse and cost saving. Two independently generated intermediate data sets (Fig. 1a) and (Fig. 1b) in Fig. 1 are anonymized to satisfy 2-diversity, i.e., at least two individuals own the same quasi-identifier and each quasi-identifier corresponds to at least two sensitive values.

Knowing that a lady aged 25 living in 21,400 (corresponding quasi-identifier is h214_; female; youngi) is in both data sets, an adversary can infer that this individual suffers from HIV with high confidence if Fig. 1a and Fig. 1b are collected together. Hiding Fig. 1a or Fig. 1b by encryption is a promising way to prevent such a privacy breach.

Assume Fig. 1a and Fig. 1b are of the same size, the frequency of accessing Fig. 1a is 10 and that of Fig. 1b is 100. We hide Fig. 1a to preserve privacy because this can incur less expense than hiding Fig. 1b.

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.

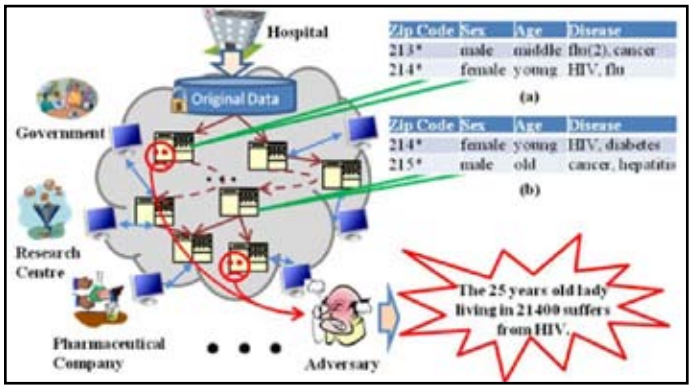


Fig. 1. A scenario showing privacy threats due to intermediate data sets.

B. Problem Analysis

1. Sensitive Intermediate Data Set Management

Similar to [6], data provenance is employed to manage intermediate data sets in our research. Provenance is commonly defined as the origin, source or history of derivation of some objects and data, which can be reckoned as the information upon how data were generated . Reproducibility of data provenance can help to regenerate a data set from its nearest existing predecessor data sets rather than from scratch [6]. We assume herein that the information recorded in data provenance is leveraged to build up the generation relationships of data sets [6]. (Sensitive intermediate data set tree (SIT)).An SIG is defined as a Sensitive Intermediate data set Tree if it is a tree structure. The root of the tree is do. An SIG or SIT not only represents the generation relationships of an original data set and its intermediate data sets, but also captures the propagation of privacy-sensitive information among such data sets. Generally, the privacy-sensitive information in do is scattered into its offspring data sets. Hence, an SIG or SIT can be employed to analyze privacy disclosure of multiple data sets. In this paper, we first present our approach on an SIT, and then extend it to an SIG with minor modifications in Section V.

An intermediate data set is assumed to have been anonymized to satisfy certain privacy requirements. How-ever, putting multiple data sets together may still invoke a high risk of revealing privacy-sensitive information, result-ing in violating the privacy requirements. Privacy leakage of a data set d is denoted as PLs(dÞ), meaning the privacy-sensitive information obtained by an adversary after d is observed.The value of PLs(dÞ) can be deduced directly from d, which is described in Section IV.A. Similarly, privacy leakage of multiple data sets in D is denoted as PLm(DÞ), meaning the privacy-sensitive information obtained by an adversary after all data sets in D are observed. It is challenging to acquire the exact value of PLm(DÞ) due to the inference channels among multiple data sets [24].

III. Minimum Privacy-Preserving Costs

Usually, more than one feasible global encryption solution exists under the PLC constraints, because there are many alternative solutions in each layer. Each intermediate dataset has various size and frequency of usage, leading to different overall cost with different solutions. The type value generated in the compressed tree helps to categorize the dataset. Thus privacy preserving cost is calculated only for the layer level less than the threshold value. As the field greater than the threshold values are omitted.

Here the values after some restrictions are allowed to identify the

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

privacy preserving cost value. Such categorization of the minimum value from the dataset under privacy leakage threshold value is dais to minimum privacy preserving cost.
These values will be identified with the help of size; price allocated for the transaction in GB or Mb, frequency of the dataset is taken. As these values iteratively find for all records under the field. Hence the dataset size remains same no further elimination is performed in this module. Finally we identify the minimum privacy preserving cost.
That the minimum solution mentioned herein is somewhat pseudo minimum because an upper bound of joint privacy leakage is just an approximation of its exact value. It is necessary to turn to heuristic algorithms for scenarios where a large number of intermediate datasets are involved, in order to obtain a near Optimal solution with higher efficiency than the optimal one.

A. Heuristic Cost

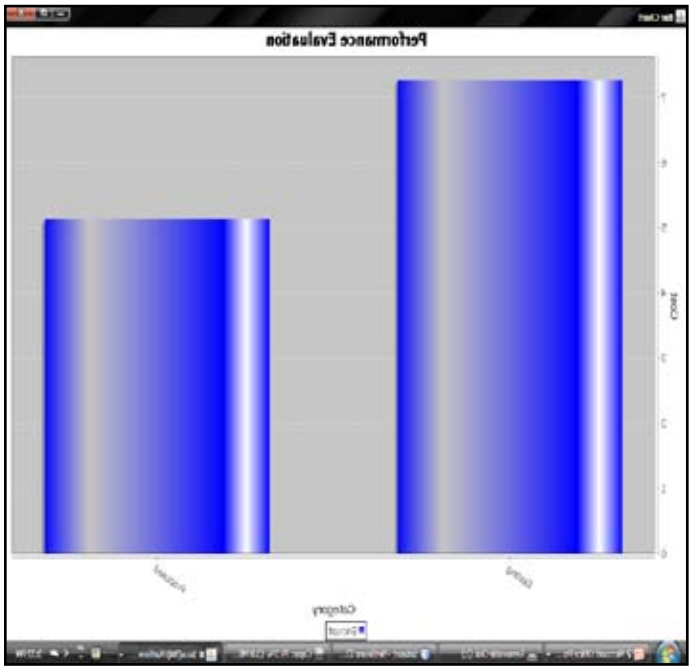
The state-search tree generated according to tan SIT is different from the SIT itself, but the height is the same. The goal state in our algorithm is to find a near-optimal solution in a limited search space. Based on this heuristic, we design a heuristic privacy preserving cost reduction algorithm. The basic idea is that the algorithm iteratively selects a state node with the highest heuristic value and then extends its child state nodes until it reaches a goal state node.The privacy-preserving solution and corresponding cost are derived from the goal state. The algorithm is guided to approach the goal state in the state space as close as possible. Above all, in the light of heuristic information, the proposed algorithm can achieve a near-optimal solution practically.SORT and SELECT are two simple external functions as their names signify.
Thus each value identified in the minimum privacy preserving cost is further undergoes the phase of heuristic algorithm to identify the optimized dataset need the privacy. Finally we can able to identify the solutions that needed to be encrypted including the sensitive dataset.

IV. Anonymization and Encryption

Here we under the process of converting the dataset that which is needed to be store at cloud storage space. As we able to analyze and find the strings to be encrypted, such process over analyze the integer's values is waste of time, because we can't categorize the integer values. To make such problem to be solved we use a technique name Anonymization. As both encryption and Anonymization for a dataset will surely reduce the privacy preserving cost as we proposed in earlier. And finally we transfer the dataset which is encrypted and anonym zed will be stored in a cloud space. Further we comparing the result set with the existing and proposed and also produce the graph for the heuristic privacy preserving cost value. When adversary user login to view the dataset uploaded by the data holder will produce the anonym zed and encrypted dataset instead of showing fully encrypted datasets.
We can see that both CALL and CHEU go up when the number of intermediate data sets is getting larger. That is, the larger the number of intermediate data sets is, the more privacy-preserving cost will be incurred. CALL Increases notably because it is proportional to the number of intermediate data sets. Given "d, CHEU also increases with the increase of the number of data sets because more data sets are required to be encrypted. Moreover, also shows that CHEU drops when the privacy leakage degree becomes larger while CALL keeps invariable.

V. Evaluation

A. Comparison



This tendency Complies with that shown in Most importantly, we can see from that the difference CSAV between CALL and CHEU becomes bigger and bigger when the number of intermediate data sets increases. That is, more expense can be reduced when the number of data sets becomes larger. This trend is the result of the dramatic rise in CALL and relatively slower increase in CHEU when the number of data sets is getting larger. In the context of Big Data, the number and sizes of data sets and their intermediate data sets are quite large in cloud. Thus, this trend means our approach can reduce the privacy preserving cost significantly in real-world scenarios.

VI. Future Work

In this paper, we have proposed an 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. A tree structure has been modeled from the generation relationships of intermediate data sets to analyze privacy propagation among data sets. We have modeled the problem of saving privacy-preserving cost as a constrained optimization problem which is addressed by decomposing the privacy leakage constraints.
A practical heuristic algorithm has been designed accordingly. Evaluation results on real-world data sets and larger extensive data sets have demonstrated the cost of preserving privacy in cloud can be reduced significantly with our approach over existing ones where all data sets are encrypted. In accordance with various data and computation intensive applications on cloud, intermediate data set management is becoming an important research area. Privacy preserving for intermediate data sets is one of important yet challenging research issues, and needs intensive investigation. With the contributions of this paper, we are planning to further investigate privacy-aware efficient scheduling of intermediate data sets in cloud by taking privacy preserving as a metric together with

other metrics such as storage and computation. Optimized balanced scheduling strategies are expected to be developed toward overall highly efficient privacy aware data set scheduling.

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] *L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 2, pp. 296-303, Feb. 2012.*

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

[6] *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.*

[7] *S.Y. Ko, I. Hoque, B. Cho, and I. Gupta, "Making Cloud Intermediate Data Fault-Tolerant," Proc. First ACM Symp. Cloud Computing (SoCC '10), pp. 181-192, 2010.*

[8] *H. Lin and W. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 6, pp. 995-1003, June 2012.*

Raghavan is currently pursuing M.E in Kalasalingam Institute of Technology, under Anna University Chennai, TamilNadu. His research interests include Cloud Computing, Data Security, and Cryptography.