INTERNATIONAL JOURNAL OF RESEARCH
In Advanced Engineering Technologies

Volume 6, Issue 1 MAY 2017

Cost-Effective Privacy Preserving Of Intermediate Data Sets in Cloud Environments through Upper Bound Privacy Leakage Constraint

[1]KURAKULA RAMAKOTESWARA RAO,[2]MANIKANTA REDDY.T

[1]M. Tech Student, Department of CS,Nalanda Institute Of Engineering &Technology, Village Kantepudi, Mandal Sattenapalli, District Guntur, Andhra Pradesh, India

[2]Assistant Professor, Department of CSE,Nalanda Institute Of Engineering &Technology, Village Kantepudi, Mandal Sattenapalli, District Guntur, Andhra Pradesh, India

Abstract— Cloud computing is an evolving paradigm with tremendous momentum, but its unique aspects exacerbating security and privacy challenges. Cloud computing provides massive computation power and storage capacity which enable users to deploy computation and data-intensive applications without infrastructure investment. Along the processing of such applications, a large volume of intermediate data sets will be generated, and often stored to save the cost of recomputing them. However, preserving the privacy of intermediate data sets becomes a challenging problem because adversaries may recover privacy-sensitive information by analyzing multiple intermediate data sets. Encrypting ALL data sets in cloud is widely adopted in existing approaches to address this challenge. But we argue that encrypting all intermediate data sets are neither efficient nor cost-effective because it is very time consuming and costly for data-intensive applications to end decrypt data sets frequently while performing any operation on them. In this paper, we propose a novel upper bound privacy leakage constraint-based approach to identify which intermediate data sets need to be encrypted and which do not, so that privacy-preserving cost can be saved while the privacy requirements of data holders can still be satisfied. Evaluation results demonstrate that the privacy-preserving cost of intermediate data sets can be significantly reduced with our approach over existing once where all data sets are encrypted.

Index Terms --Cloud Computing, Data Sets, Privacy Preserving, Data Privacy Management, Privacy Upper Bound;

I. INTRODUCTION

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams Figure1: Structure of cloud computing Cloud computing entrusts remote services with a user's data, software and computation. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services. These services typically provide access to advanced software applications and high-end networks of server computers The goal of cloud computing is to apply traditional supercomputing, or high-performance computing power, normally used by military and research facilities, to perform tens of trillions of computations per second, in consumer-oriented applications such as financial portfolios, to deliver personalized information, to provide data storage or to

power large, immersive computer games The cloud computing uses networks of large groups of servers typically running low-cost consumer PC technology with specialized connections to spread data processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together. Often, virtualization techniques are used to maximize the power of cloud computing.

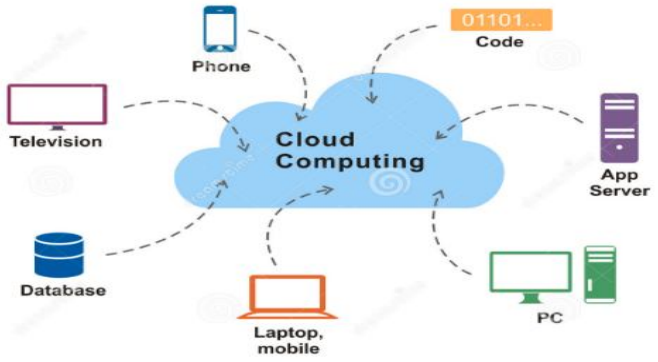


Figure1: Cloud Computing Architecture

Cloud users can store their valuable intermediate data sets selectively when processing original data sets in a data intensive application in order to curtail the overall expenses by avoiding frequent re-computation to obtain these data sets. Data users often reanalyze results, conduct new analysis, or share some intermediate results with others for collaboration. The secure encryption of privacy preserving of dynamic data sets are used to identify which intermediate data sets need to be encrypted and which do not, so that privacypreserving cost can be saved. The technical approaches for preserving the privacy of data sets stored in cloud mainly include encryption and anonymization. On one hand, encrypting all data sets, an effective approach, is widely adopted in current research. However, processing on encrypted data sets efficiently is a challenging task, because most of the applications run on unencrypted data sets. Although homomorphism encryption which theoretically allows performing computation on encrypted data sets, applying algorithms are rather expensive due to their inefficiency. On the other hand, partial information of data sets, example 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, encrypting all intermediate data sets will lead to high overhead and low efficiency when they are frequently accessed or processed. To address this issue, the system proposes to encrypt a part of intermediate data sets rather than all for reducing privacy preserving cost.

II. RELATED WORK

H. Lin et al. proposed a general encryption schemes to protect data confidentiality, but also limit the functionality of the storage system because a few operations are supported over encrypted data. Constructing a secure storage system that supports multiple functions is challenging when the storage system is distributed and has no central authority. This system proposed a threshold proxy re-encryption scheme and integrates it with a decentralized erasure code such that a secure distributed storage system is formulated. The distributed storage system not only supports secure and robust data storage and retrieval, but also lets a user forward his data in the storage servers to another user without retrieving the data back. The main technical contribution is that the proxy re-encryption scheme supports encoding operations over encrypted messages

as well as forwarding operations over encoded and encrypted messages. T. Praveena, G. Raja, in this they derived that, the approach is a Threshold Filtering adopted to classify the dataset that are to be encrypted. A value for each intermediate dataset will be fixed and based on the privacy information present in the dataset .If the value of intermediate dataset is higher than the threshold value then it will be encrypted and remaining dataset anonymized. For encryption two round searchable encryption (TRSE) is used for easy searching and accessing the encrypted dataset. For 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. Ms. C. Celcia, Mrs. T. Kavitha, in this paper they projected that in existing data intensive application of cloud provide massive computation power and storage space. In this surroundings they are more number of user can accessed or processed original data sets frequently due to this intermediate data sets are generated from original one. For privacy preserving heuristic approach identifies which intermediate data set need to be encrypted while others not. In this approach encryption is incorporated with anonymization for cost effective preserving.

III. FRAME WORK

This system is designed to identify only the important and critical intermediate datasets that needs to be encrypted for security purposes hence reducing encryption/decryption cost and thus maintaining data privacy. It is based on identifying the least frequent table using least frequent pattern mining algorithm and thereby encrypting it by advanced encryption algorithm. From the least frequent table, the reference attribute between the data tables are found out and imposing a privacy leakage constraint to it in order to identify the sensitive information. For each constraint, the maximal possible value for any of these values is an upper bound and may recover privacy-sensitive partial column level encryption. Hence a column wise encryption to the unencrypted table of the intermediate datasets is proposed. Additional feature of encrypting on the basis of reference attribute between the data tables are achieved to reduce the cost complexity when accessing the data. An automatic scheduling strategy is involved to maintain a log report of the frequent and infrequent usage of intermediate dataset under time conditions as the data in cloud are dynamic in nature. Based on the frequency of accessing, the tables are scheduled according to it and segregated on the least and most frequent table. Therefore this process is repeated to handle the data in cloud in a dynamic manner. When scheduling is done to the datasets the tables are modified and updated to the current situation to handle the dynamic nature of cloud. The data owner can store valuable intermediate data sets selectively when processing original data sets in data intensive applications, in order to curtail the overall expenses by avoiding frequent re-computation to obtain these data sets. Such scenarios are quite commonbecause data users often reanalyze results, conduct new analysis on intermediate data sets, or share some intermediate results with others for collaboration. Usually, intermediate data sets in cloud are accessed and processed by multiple parties, but rarely controlled by original data set holders.

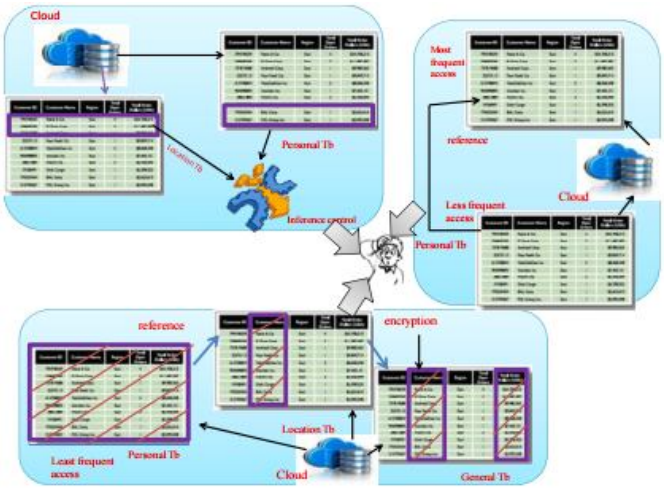
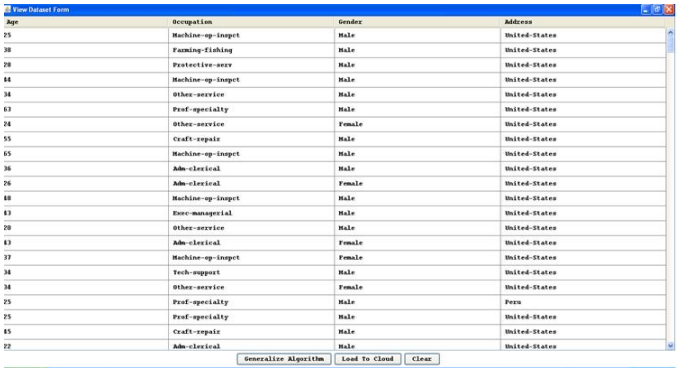


Figure 2: Architecture of Privacy Preserving Intermediate Datasets

In the proposed system an intermediate datasets are created for a Government application where all the people related information is present. When an original dataset is being processed, the intermediate datasets are created such as General Table, Industry Table, Location Table and Personal table. When these intermediate datasets are collected together by an adversary it can menace the privacy-sensitive information from them, bringing considerable economic loss. Therefore an inference analysis can be made from these datasets. In order to avoid an inference analysis from these intermediate datasets, the system uses a Least Frequent Pattern Matching algorithm to identify the least frequent tables. The reason for identifying the least frequent table is due to less encryption/decryption computational cost. As a result, the least frequent table will have least frequency of access to the intermediate datasets and therefore it incurs less computation cost rather than the most frequent table. Therefore the least frequent tables will be encrypted using Advanced Encryption Standard algorithm.

IV. EXPERIMENTAL RESULTS

In our experiments,admin login into the system after login into the system and upload the dataset into the system after successful uploading the dataset into the system apply the generalize algorithm it will anonymize the required data here we are anonymize age here after that data set will be upload to the cloud after that admin can access the data into the cloud to search the dataset candidate like United-Sates it will generate a intermediate dataset of United-Sates candidates after to apply the privacy preserving algorithm to preserve the privacy for intermediate dataset after that run the user application in that any number of users register into the system after successful registering into the system, authorized user login into the system after login into the system authorized user access the dataset to access the dataset user search the United-Sates keyword it will generate the United-Sates peoples data in that occupation details of United-States candidates are encrypted because the same data has been accessed by many numbers of candidates so it crossed certain heuristic value so it is encrypted to shown in below screens



Through our implementation we have implemented an efficient system for privacy preserving of intermediate datasets through Upper Bound Privacy Leakage Constraintin efficient manner with low cost when compare to existing techniques.

V.CONCLUSION

This paper has 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. The problem of saving privacy-preserving cost as a constrained optimization problem which is addressed by decomposing the privacy leakage constraints has been modeled. 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 this approach over existing ones where all data sets are encrypted.

REFERENCES

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

[2] M. Li, S. Yu, N. Cao, and W. Lou, “Authorized Private KeywordSearch over Encrypted Data in Cloud Computing,”Proc. 31st Int’lConf. Distributed Computing Systems (ICDCS ’11),pp. 383-392, 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] B.C.M. Fung, K. Wang, and P.S. Yu, “Anonym zing ClassificationData for Privacy Preservation,”IEEE Trans. Knowledge and Data Eng. vol. 19, no. 5, pp. 711-725, May 2007.

[5] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, “Privacy-PreservingData Publishing: A Survey of Recent Developments,”ACMComputing Survey,vol. 42, no. 4, pp. 1-53, 2010.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] P. Samarati, "Protecting Respondents' Identities in MicrodataRelease,"IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6,pp. 1010-1027, Nov. 2001.

[13] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond K-Anonymity," ACM Trans.Knowledge Discovery from Datival. 1, no. 1, article 3, 2007.

[14] G. Wang, Z. Zutao, D. Wenliang, and T. Zhouxuan, "InferenceAnalysis in Privacy-Preserving Data Re-Publishing,"Proc. EighthIEEE Int'l Conf. Data Mining (ICDM '08), pp. 1079-1084, 2008.

[15] E.T. Jaynes, "Information Theory and Statistical Mechanics,"Physical Rev.,vol. 106, no. 4, pp. 620-630, 1957.