ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Privacy Preserving Based on Encrypting Selective Data Sets in Cloud

Anlin Helbah. R[*1], Ahamed Ali. S[2]

[*1] PG Student, [2] Assistant Professor, Department of IT, Velammal Engineering College, Chennai-600066, Tamil Nadu, India

anlinhelbah@gmail.com

Abstract

Cloud computing provides massive computation power and storage capacity which enable users to deploy computation and data-intensive applications without infrastructure investment. The data are stored in the data sets in the cloud. However, preserving the privacy of intermediate data sets becomes a challenging problem. In existing approaches encrypting all data sets in cloud is widely a greater overhead. So by encrypting all intermediate data sets are neither efficient nor cost-effective. Because it is very time consuming and greater cost consuming. The time consuming and greater cost problem is resolved by encrypting only the selected intermediate data sets. For selecting which intermediate data sets need to be encrypted and which do not, a proposal is made that a novel upper bound privacy leakage constraint-based approach. By using this approach the privacy-preserving cost can be reduced and also time consuming for encryption also get reduced.

Keywords: Data Intensive, Data storage privacy, Intermediate dataset, Privacy preserving, Privacy upper bound.

Introduction

Cloud computing relies on sharing of resources to achieve coherence and economies of scale similar to a utility over a network. The basement of cloud computing is the broader concept of converged infrastructure and shared services. The cloud focuses on maximizing the effectiveness of shared resources. Cloud resources are not only shared for multiple users but are also dynamically re-allocated per demand. The privacy concerns[11] caused by retaining intermediate data sets in cloud are important but they are paid little attention. For preserving privacy[10] of multiple data sets, it is important to anonymize all data sets first and then encrypt them before storing or sharing them in cloud. Usually, the weightage of intermediate data sets[12] is huge. Cloud users can store valuable intermediate data sets selectively when processing original data sets in data-intensive applications such as medical diagnosis[15], in order to reduce the overall expenses by avoiding frequent re-computation to obtain these data sets. Such methods are quite common because data users often re-analyse results, conduct new analysis on intermediate data sets, and also share some intermediate results with others for collaboration. Without loss of majority, the notion of intermediate data set herein refers to intermediate and resultant data sets. The storage of intermediate data enlarges attack surfaces so that privacy requirements of data holders are at risk of being violated. The intermediate data sets in cloud are accessed and processed by multiple parties, but not frequently controlled by original data set holders. This enables a dispute to collect intermediate data sets together and menace privacy-sensitive information from them, bringing the economic loss or severe social reputation impairment to data owners. But, little important has been paid to such a cloud-specific privacy issue.

Related Work

In all existing works the data in cloud are stored after encryption only. While users accessing some data from cloud the server will again encrypt and send that information. Some sensitive information and related data are stored in intermediate dataset. Intermediate data sets in cloud are accessed and processed by multiple parties, but not frequently controlled by original data set holders. For accessing the data from cloud a private keyword[9] is needed in-order to give privacy to the data present in the cloud.

Encryption works well for data privacy[2] in those approaches, it is necessary to encrypt and decrypt data sets frequently in many applications. Encryption is usually coordinated with other methods to achieve cost reduction, high data accessibility and privacy protection. The data privacy problem caused by MapReduce and presented a system named Airavat which incorporates mandatory access control with differential privacy. A set of tools called Silverline that identifies all functionally encryptable data and then encrypts them to protect privacy. A system named Sedic[8] which partitions Map Reduce 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. The quality of data is required to be noted in advance to make the above approaches available.

An approach that combines encryption and data fragmentation[13] to achieve privacy protection for distributed data storage with encrypting only part of data sets, but integrate data anonymization and encryption together to fulfil cost-effective privacy preserving[4]. The importance of maintaining intermediate data sets in cloud has been widely identified, but the research on privacy issues[5] incurred by such data sets just commences. Privacy principles such as k-anonymity and l-diversity[1] are put forth to model and express privacy, yet most of them are only applied to one single data set. The research in exploits information theory to express the privacy via utilizing the maximum entropy principle .Many anonymization techniques like generalization have been proposed to preserve privacy[2], but these techniques alone fail to solve the problem of preserving privacy for multiple data sets.

Some of the disadvantages in existing works are: The sensitivity of data is required to be labelled in advance. Encrypting all datasets is cloud be a high overhead. The computing cost is very expensive and very low efficiency.

Selecting Intermediate Dataset

Cloud computing provides massive computation power and storage capacity which enable users to deploy computation and data-intensive applications without infrastructure investment. The data are stored in the data sets in the cloud. The time consuming and greater cost problem is resolved by encrypting only the selected intermediate data sets. For selecting which intermediate data sets need to be encrypted and which do not, a proposal is made that a novel upper bound privacy leakage constraint-based approach[15]. By using this approach the privacy-preserving cost can be reduced and also time consuming for encryption[3] also get reduced.

Upper Bound Constraints

An upper bound constraint-based approach to select the necessary subset of intermediate data sets that needs to be encrypted for minimizing privacy-preserving cost. The security leakage upper bound constraint is decomposed layer by layer. An overly controlled optimization problem with the PLC is then transformed into a recursive form a heuristic algorithm is designed. The privacy issues in workflow history, and proposed to achieve module privacy preserving[6] and high utility of provenance information[10] via carefully hiding a subset of intermediate data. This general idea is similar to research mainly focuses on data privacy preserving from an economical cost perspective while theirs concentrates majorly on functionality privacy of workflow modules rather than data privacy. The model also differs from theirs in several aspects such as data hiding methods[14], privacy quantification and cost models. But, this model can be complementarily used for selection of hidden data items in their research if economical cost is considered. The research community has investigated extensively on privacy-preserving issues and made fruitful progress with a variety of privacy models and preserving methods.

To avoid payment details and focus on the discussion of the core ideas, here the combination of the prices of various services required by en/decryption into one is done.

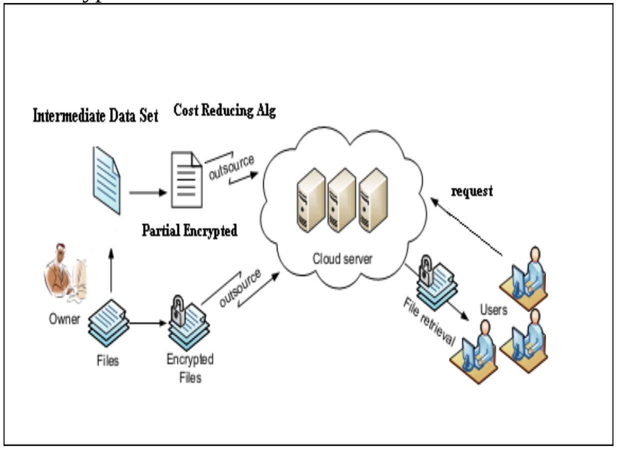

Fig.1: A process undergoes between user and owner

Create Intermediate Dataset

A quasi-identifier is a set of attributes that, in merging, can be linked with external information to re-identify (or reduce uncertainty about) all or some of the respondents to whom information refers. An underlying assumption for this type of risk is that there is an intruder who has two pieces of information: (a) the actual data set that has been disclosed and (b) some background information about one or more people in this data set. The less important information is described by a set of variables. These set of variables are the quasi-identifiers. One of the means to anonymize a database is generalization; i.e., replacing the values that appear in the database with subsets of values, so that each entry $R_i(j)$, $1 \leq i \leq n$, $1 \leq j \leq r$, which is an element of A_j, is overfilled by a subset of A_j that includes that element.

Sensitive Intermediate Dataset

A Directed Acyclic Graph (DAG) representing the generation relationships of intermediate data sets D from d_o is defined as a Sensitive Intermediate data set graph. It can be denoted as SIG. Formally SIG=(V,E) where V={d_o}U D,E a set of directed edges is. A directed edge ($d_{p,}d_c$) in E means that part or all of d_c is generated from d_p, where d_p,d_c €{d_o}U D Privacy leakage of a data set d is denoted as PL_s (d), meaning the privacy-sensitive information obtained by an adversary after d is observed.

RSA Encryption Algorithm

RSA is a widely used asymmetric encryption algorithm that, if properly implemented, so far cannot be cracked in acceptable time. RSA algorithm uses two separate keys. One of these two keys is used to encode (encrypt) the message on the sender side, another is used to decode (decrypt) the message on the receiver side. One of these two keys is usually kept secret; restricting access to it (private key) and another is public and can be known to everyone. RSA can be used in our proposed system, Sender (Hospital) uses widely known public key to encrypt the message. Only receiving person (research Centre) who also knows the private key can decrypt it.

Cost Reducing Heuristic Algorithm

Design a heuristic algorithm to reduce privacy-preserving cost. Each intermediate data set has various size and frequency of usage, guiding to different overall cost with different solutions. Therefore, it is important to find a feasible solution with the minimum privacy-preserving cost under privacy leakage constraints. 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. But the accurate solution can be exactly minimal in the sense of the PLC_1constraints.

Optimized Balanced Scheduling

This scheduling is used for the best mapping solution to meet the system load balance to the greatest extent or to make the cost of load balancing the lowest. The best scheduling solution for the current scheduling through genetic algorithm. First need to compute the cost through the ratio of the current scheduling solution to the best scheduling solution, and then decide the scheduling strategy according to the cost. So that it has the least influence on the load of the system after scheduling and it has the lowest cost to reach load balancing. In this way, the best strategy is formed.

Experiment Evaluation

The experimental result on real-world data sets is depicted in Fig.2 from which we can see that how total cost get reduced after applying heuristic algorithm.

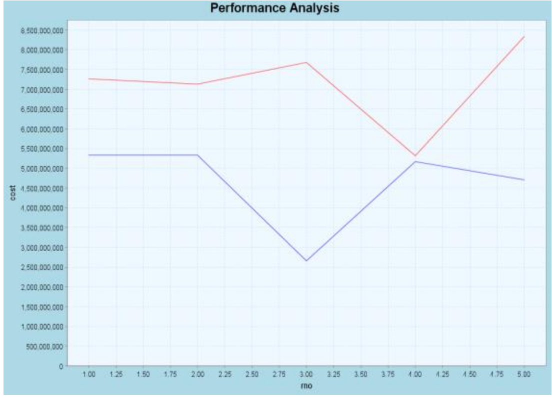

Fig.2: Experimental result about real-world.

The heuristic algorithm is applied only for the sensitive information which got encrypted using RSA algorithm. So that by applying heuristic algorithm the cost get reduced.

Conclusion

In this project, an approach is used that identifies which part of intermediate data sets needs to be encrypted and which does not, in order to save the privacy preserving cost. A tree structure is modelled 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. A heuristic algorithm has designed for reducing the privacy preserving cost. In accordance with various data and computation intensive applications on cloud, management of intermediate data set is becoming an important research area. Privacy for intermediate data sets is one of important yet challenging research issues, and needs intensive investigation. Optimized balanced scheduling strategies are developed toward overall highly efficient privacy aware data set scheduling.

References

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

ISSN: 2277-9655
Impact Factor: 1.852

[2] 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.

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

[4] 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

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

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

[7] 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.

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

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

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

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

[13] 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.

[14] W. Du, Z. Teng, and Z. Zhu, "Privacy-Maxent: Integrating Background Knowledge in Privacy Quantification," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08), pp. 459-472, 2008.

[15] 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.