Rank constraint based approach for cost effective data Security

[1]K. Arun Kumar, [2]M Vinay Kumar Reddy,[3]Madupoju Subodh,
[4]Mamillapalli Meghana,[5]Mavin Dilip, [6]Mohammad Waseem Afroz,[7]N Ranjith Raju
[1]Asst. Professor, Dept. of CSE, [2,3,4,5,6,7] B. Tech., (CSE)
Malla Reddy Engineering College (Autonomous), Secunderabad, Telangana

ABSTRACT: The original data and intermediate data are protected with the support of encryption and anonymization techniques. Intermediate data sets in cloud are accessed and processed by multiple parties, but rarely controlled by original data set holders. Encrypting all intermediate data sets will lead to high overhead and low efficiency. Single intermediate data privacy model is used to protect intermediate data under only one node. Multiple intermediate data sets is protected by using joint privacy leakage model. An upper bound privacy leakage constraint-based approach is used to identify which intermediate data sets need to be encrypted. Sensitivity relationship between multiple data set is represented under Sensitive Intermediate data set Graph (SIG). Privacy-Preserving Cost Reducing Heuristic algorithm is used to control privacy leakage in multiple data sets. Multiple intermediate data set privacy models is integrated with data scheduling mechanism. Privacy preservation is ensured with dynamic data size and access frequency values. Storage space and computational requirements are optimally utilized in the privacy preservation process. Data distribution complexity is handled in the scheduling process. Cloud computing provides massive computation power and storage capacity. Cloud enables users to deploy computation and data-intensive applications without infrastructure investment. Intermediate data are generated under the cloud applications and stored to save the cost of recomposing. Adversaries may recover privacy-sensitive information by analyzing multiple intermediate data sets. Encrypted data storage mechanism is used to secure cloud data values. Encrypting all intermediate data sets are neither efficient nor cost-effective one. In data intensive applications encrypt/decrypt operation requires high time and cost. Partial encryption mechanism is used to provide privacy on data with minimum resource levels.

Key words- Data Storage Privacy, Privacy Preserving, Intermediate Dataset, Privacy Upper Bound, Economics of scale

1. INTRODUCTION

A powerful underlying and enabling concept is computing through service-oriented architectures (SOA) – delivery of an integrated and orchestrated suite of functions to an end-user through

composition of both loosely and tightly coupled functions, or services – often network based. Related concepts are component-based system engineering, orchestration of different services through workflows, and virtualization. The key to a SOA framework that supports workflows is componentization of its services, an ability to support a range of couplings among workflow building blocks, fault tolerance in its data- and process-aware service-based delivery, and an ability to audit processes, data and results, i.e., collect and use provenance information. Component-based approach is characterized by reusability, substitutability, extensibility and scalability, customizability and compos ability. There are other characteristics that also are very important. Those include reliability and availability of the components and services, the cost of the services, security, total cost of ownership, economy of scale, and so on. In the context of cloud computing we distinguish many categories of components: from differentiated and undifferentiated hardware, to general purpose and specialized software and applications, to real and virtual "images", to environments, to no-root differentiated resources, to workflow-based environments and collections of services, and so on.

They are discussed later in the paper.

An integrated view of service-based activities is provided by the concept of a workflow. IT assisted workflow represents a series of structured activities and computations that arise in information-assisted problem solving. Workflows have been drawing enormous attention in the database and information systems research and development communities. Similarly, the scientific community has developed a number of problem solving environments, most of them as integrated solutions. Scientific workflows merge advances in these two areas to automate support for sophisticated scientific problem solving.

A workflow can be represented by a directed graph of data flows that connect loosely and tightly coupled processing components. One such graph is. It illustrates a Kepler-based implementation of a part of a fusion simulation workflow. In the context of "cloud computing", the key questions should be whether the underlying infrastructure is supportive of the work flow oriented view of the world. This includes on demand and advance-reservation based access to individual and aggregated computational and other resources, autonomics, ability to group resources from potentially different "clouds" to deliver workflow results,

appropriate level of security and privacy, etc.

2. RELATED WORK

We briefly review the research on privacy protection in cloud, intermediate data set privacy preserving and Privacy-Preserving Data Publishing (PPDP). Currently, encryption is exploited by most existing research to ensure the data privacy in cloud [8]. Although encryption works well for data privacy in these approaches, it is necessary to encrypt and decrypt data sets frequently in many applications. Encryption is usually integrated with other methods to achieve cost reduction, high data usability and privacy protection. Roy et al. [11] investigated the data privacy problem caused by MapReduce and presented a system named Airavat which incorporates mandatory access control with differential privacy. Puttaswamy et al. [10] described a set of tools called Silverline that identifies all functionally encryptable data and then encrypts them to protect privacy. Zhang et al. [7] proposed a system named Sedic which partitions MapReduce 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 sensitivity of data is required to be labeled in advance to make the above approaches available. Ciriani et al. proposed an approach that combines encryption and data fragmentation to achieve privacy protection for distributed data storage with encrypting only part of data sets. We follow this line, but integrate data anonymization and encryption together to fulfill cost-effective privacy preserving. The importance of retaining intermediate data sets in cloud has been widely recognized [6], but the research on privacy issues incurred by such data sets just commences. Davidson et al. [2] studied the privacy issues in workflow provenance, and proposed to achieve module privacy preserving and high utility of provenance information via carefully hiding a subset of intermediate data. This general idea is similar to ours, yet our 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. Our research also differs from theirs in several aspects such as data hiding techniques, privacy quantification and cost models. But, our approach can be complementarily used for selection of hidden data items in their research if economical cost is considered. The PPDP research community has investigated extensively on privacy-preserving issues and

made fruitful progress with a variety of privacy models and preserving methods. Privacy principles such as k-anonymity and l-diversity are put forth to model and quantify privacy, yet most of them are only applied to one single data set. Privacy principles for multiple data sets are also proposed, but they aim at specific scenarios such as continuous data publishing or sequential data releasing. Many anonymization techniques like generalization have been proposed to preserve privacy, but these methods alone fail to solve the problem of preserving privacy for multiple data sets. Our approach integrates anonymization with encryption to achieve privacy preserving of multiple data sets. Moreover, we consider the economical aspect of privacy preserving, adhering to the pay-as-you-go feature of cloud computing.

3. PRIVACY PRESERVING OF INTERMEDIATE DATA SETS IN CLOUD

Technically, cloud computing is regarded as an ingenious combination of a series of technologies, establishing a novel business model by offering IT services and using economies of scale. Participants in the business chain of cloud computing can benefit from this novel model. Cloud customers can save huge capital investment of IT infrastructure, and concentrate on their own core business [3]. Therefore, many companies or organizations have been migrating or building their business into cloud. However, numerous potential customers are still hesitant to take advantage of cloud due to security and privacy concerns [5].

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. 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 recomputation to obtain these data sets. 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 at risk 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 privacysensitive 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]. 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 performing computation on encrypted data sets, applying current algorithms are rather expensive due to their inefficiency. 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 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, 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 constrained optimization problem [4]. This problem is then divided into a series of subproblems 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 realworld and extensive data sets demonstrate that privacy-preserving 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. This paper is a significantly improved version of [12]. We mathematically prove that our approach can ensure privacy-preserving requirements. Further, the heuristic algorithm is redesigned by considering more factors. We extend experiments over real data sets. Our approach is also extended to a graph structure.

4. PRIVACY REPRESENTATION AND PRIVACY LEAKAGE UPPER BOUND CONSTRAINT

Single Intermediate Data Set Privacy Representation

The privacy-sensitive information is essentially regarded as the association between sensitive data and individuals [13]. We denote an original sensitive data set as do; an anonymized intermediate data set as doi, the set of sensitive data as SD; and the set of quasi-identifiers as QI. Quasi identifiers, which represent the groups of anonymized data, can lead to privacy breach if they are too specific that only a small group of people are linked to them. Let S denote a random variable ranging in SD, and Q be a random variable ranging within QI. Suppose s SD and q QI. The joint possibility of an association hs, qi, denoted as p(S = s, Q = q) (abbr. p(s, q)), is the information that adversaries intend to recover [13]. When an adversary has observed d* and a quasi-identifier q, the conditional possibility p(S = s|Q = q) representing intrinsic privacy sensitive information of an individual can be

inferred. If p(S = s|Q = q) is deduced as a high value or even 1.0, the privacy of the individual with q will be awfully breached.

We employ the approach proposed in [1] to compute the probability distribution P*(S,Q) of hs, qi in do after observing d*. More details can be found in Appendix A.1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2012.238 (All appendices are included in the supplemental file). Then, the privacy quantification of data set d* can be fulfilled. Formally, PLs(d*) is defined as

PLs(d*) H(S,Q) – H*(S,Q): (1)

H(S,Q) is the entropy of random variable ⟨S,Q⟩ before d* is observed, while H*(S,Q) is that after observation. P(Q, S) is estimated as a uniform distribution according to the maximum entropy principle [1]. Based on this, H(S,Q) can be computed by H(S,Q) = log(|QI| . |SD|). H*(S,Q) is calculated from distribution

Joint Privacy Leakage of Multiple Intermediate Data Sets

The value of the joint privacy leakage incurred by multiple data sets in D = {d1, d2, ... dn}, n ε N, is defined by

PLm(D) H(S,Q) - HD(S,Q) (3)

H(S,Q) and HD(S,Q) are the entropy of ⟨S,Q⟩ before and after data sets in D are observed, respectively. H(S,Q) = log(|QI| . |SD|). HD(S,Q) can be calculated once P(S,Q) is estimated after data sets in D are observed. Given the relationship between " and PLm(Dune) in PLC, " ranges in the interval [max1≤i≤n{PLs(di)}; log(|QI| . |SA|)].

Zhu et al. [9] proposed an approach to indirectly estimate P(S,Q) for multiple data sets with the maximum entropy principle. But this approach becomes inefficient when many data sets are involved because the number of variables and constraints possibly increase sharply when the number of data sets grows. According to the experiments in [9], it takes more than 200 minutes to quantify the privacy of two data sets with 6,000 records. Further, since Dune is uncertain before a solution is found, we need to try different Dune, where Dune ε 2D. So, the inefficiency will become unacceptable in many applications where a large number of intermediate data sets are involved. Fortunately, the PLC can be achieved without exactly acquiring PLm(Dune) because our goal is to control the privacy disclosure caused by multiple data sets. A promising approach is to substitute the PLC with its sufficient conditions. Specifically, our approach is to replace the exact value of

PLm(Dune) with one of its upper bounds which can be calculated efficiently.

Upper Bound Constraint of Joint Privacy Leakage

We attempt to derive an upper bound of PLm(Dune) that can be easily computed. Intuitively, if an upper bound B(PLm(Dune)) is found, a stronger privacy leakage constraint B(PLm(Dune))≤ε can be a sufficient condition of the PLC. Accordingly, PLm(Dune) will never exceed the threshold " if B(PLm(Dune))≤ ε holds. Let du and dv be two data sets whose privacy leakage are PLs(du) and PLs(dv), respectively. The joint privacy leakage caused by them together is PLm({du, dv}). As information gain is never negative, PLm({du, dv}) is not less than neither PLs(du) nor PLs(dv). Further, PLm({du, dv}) will not exceed the sum of PLs(du) and PLs(dv), i.e., PLm({du, dv}) ≤ PLm({du, dv}), where the equality holds if and only if the information provided by du and dv does not overlap. This property of joint privacy leakage can be extended to multiple unencrypted data sets in Dune: PLm(Dune) ≤ (4) Hence, the sum of privacy leakage of unencrypted data sets can be deemed as an upper bound of PLm(Dune). Based on replacing the PLC with such an upper bound constraint, we propose an approach to address the optimization problem.

5. ISSUES ON PRIVACY LEAKAGE UPPER BOUND CONSTRAINT-BASED APPROACH

The original data and intermediate data are protected with the support of encryption and anonymization techniques. Intermediate data sets in cloud are accessed and processed by multiple parties, but rarely controlled by original data set holders. Encrypting all intermediate data sets will lead to high overhead and low efficiency. Single intermediate data privacy model is used to protect intermediate data under only one node. Multiple intermediate data sets is protected by using joint privacy leakage model. An upper bound privacy leakage constraint-based approach is used to identify which intermediate data sets need to be encrypted. Sensitivity relationship between multiple data set is represented under Sensitive Intermediate data set Graph (SIG). Privacy-Preserving Cost Reducing Heuristic algorithm is used to control privacy leakage in multiple data sets. The following drawbacks are identified in the existing system.

Static privacy preserving model

Privacy preserving data scheduling is not focused

Storage and computational aspects are not considered

Load balancing is not considered

6. PRIVACY PRESERVED DATA SCHEDULING SCHEME

Multiple intermediate data set privacy model is integrated with data scheduling mechanism. Privacy preservation is ensured with dynamic data size and access frequency values. Storage space and computational requirements are optimally utilized in the privacy preservation process. Data distribution complexity is handled in the scheduling process. Cloud data sharing system provides security for original and intermediate data values. Data sensitivity is considered in the intermediate data security process. Resource requirement levels are monitored and controlled by the security operations. The system is divided into five major modules. They are data center, data provider, intermediate data privacy, security analysis and data scheduling.

The data center maintains the encrypted data values for the providers. Shared data uploading process are managed by the data provider module. Intermediate data privacy module is designed to protect intermediate results. Security analysis module is designed to estimate the resource and access levels. Original data and intermediate data distribution is planned under the data scheduling module.

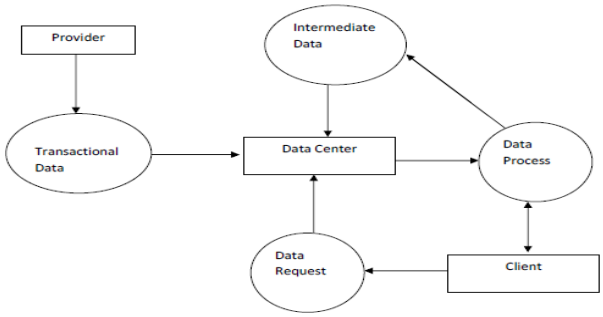

Fig1: Privacy Preserved Data Scheduling Scheme

Data Center

Database transactions are shared in the data centers. Data center maintains the shared data values in encrypted form. Homomorphic encryption scheme is used for encryption process. Key values are also provided by the data center.

Data Provider

Data provider uploads the database tables to the data center. Database schema is also shared by the provider. Encryption process is performed under the data provider environment. Access control tasks are managed by the providers.

Intermediate Data Privacy

Intermediate data values are generated by processing the original data values. Intermediate data values are stored under the data center or provider environment.

Encryption process is carried out on the intermediate data values. Sensitivity information is used for the intermediate data security process.

Security Analysis

Joint privacy leakage model is used for the security process. Storage requirements are analyzed in the intermediate data analysis. Computational resource requirements are also analyzed in the security analysis. Intermediate data encryption decisions are made with reference to the storage and computational resource requirements.

Data Scheduling

Data scheduling is used to plan the data distribution process. Computational tasks are combined in the scheduling process. Scheduling is applied to select suitable provider for data delivery process. Request levels are considered in the data scheduling process.

7. EXPERIMENTAL RESULT

Fig 2: Main Page

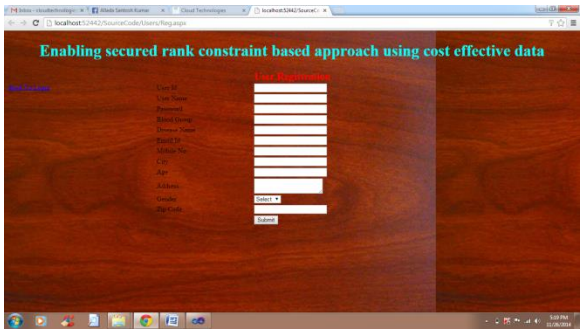

Fig 3:- New User Page

Fig 4: Result

8. CONCLUSION

Shared data values are maintained under third party cloud data centers. Data values are processed and stored in different cloud nodes. Privacy leakage upper bound constraint model is used to protect the intermediate data values. Dynamic privacy management and scheduling mechanism are integrated to improve the data sharing with security. Privacy preserving cost is reduced by the joint verification mechanism. Resource consumption is controlled by the support of sensitive data information graph. Data delivery overhead is reduced by the load balancing based scheduling mechanism. Dynamic privacy preservation model is supported by the system.

REFERENCES

[1] 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.

[2] 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.

[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] Xuyun Zhang, Chang Liu, Surya Nepal, Suraj Pandey, and Jinjun Chen, "A Privacy Leakage Upper Bound Constraint-Based Approach for Cost- Effective Privacy Preserving of Intermediate Data Sets in Cloud", IEEE Transactions On Parallel And Distributed Systems, Vol. 24, No. 6, June 2013.

[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] 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.

[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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.