International Journal of Computer Science
and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 3, May 2014, 171-178

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

PARTIAL ENCRYPTION AND PARTIAL INFERENCE CONTROL BASED DISCLOSURE IN EFFECTIVE COST CLOUD

K. AYSHWARYA

P.G Student, Saveetha Engineering College, Chennai, Tamil Nadu, India

ABSTRACT

Cloud computing is one of the most pre-dominant paradigm in recent trends for computing and storing purposes on data-intensive applications without infrastructure investment. It introduces an optimized approach towards management flexibility and economic savings for distributed applications. As an advantage in the computing world and storage resources offered by cloud service providers, the data owners must place their valuable information into the public cloud servers which are not within their trusted domains. Along the processing of such applications, a large volume of intermediate data sets that get generated are stored so that it need not be recomputed. Privacy-preserving database in cloud would allow a database owner to outsource its encrypted database to a cloud server. Due to this, data security and privacy of data is one of the major concerns in the cloud computing world. Encryption of data sets of all the content in cloud is widely adopted in existing approaches to address this challenge. But encrypting all intermediate data sets are neither efficient nor cost-effective because it is very time consuming and costly for data-intensive applications to en/decrypt data sets frequently while performing any operation on them. Even evaluated results demonstrate that the privacy-preserving cost of intermediate data sets is significantly high on existing ones where all data sets are encrypted. The proposed method has an optimized solution of restricting the data based on the request from the users. So that, predicting the data cannot be done at any case, this provides a highest level of security to the system.

KEYWORDS: Cloud Computing, Privacy Preserving, Intermediate Data Set, Inference Control, Partial Encryption

INTRODUCTION

Privacy is an increasingly important aspect of data publishing. Sensitive data, such as medical records in public databases, are recognized as a valuable source of information for the allocation of public funds, medical research and statistical trend analysis. However, if personal private information is leaked from the database, the service will be regarded as unacceptable by the original owners of the data. There are two approaches to avoiding leaks of private information from public databases: generalization methods and perturbation methods. Generalization methods modify the original data to avoid identification of the records. These methods generate a common value for some records and replace identifying information in the records with the common value. However, detailed information is lost during this process. On the other hand, perturbation methods add noise to data. While perturbed data usually retains detailed information, it also normally includes fake data.

An important issue is how to evaluate these methods with regard to privacy leakage. In particular, when performing such an evaluation, it is difficult to model the background knowledge of an adversary trying to obtain private information from a database. Even if some fields of records in a database have been anonymized in some manner, an adversary may still be able to identify a record through background knowledge. For example, even if ZIP codes are

generalized to include just the highest level of regional information in a medical database, this may still be enough to identify a record if there is only one case of a particular disease in that region and an adversary knows that a particular target has had that disease and lives in that region. Since the generalization and perturbation methods take such different approaches, it is very difficult to compare them. The proposed methodology is based on the notion of differential privacy but is applicable to k-anonymized data sets. The objective is to put forward the inference control strategy to avoid privacy leakage thereby performing double encryption in the datasets based on the severity. Further investigation of privacy aware efficient scheduling of intermediate data sets in cloud is considered to preserve the dataset in case of dynamic change.

PRELIMINARIES

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 re computing them. Preserving the privacy of intermediate data sets becomes a challenging problem because adversaries may recover privacy-sensitive information by analyzing multiple intermediate data sets. The proposed methodology has a novel usability matrix on dynamic datasets for cloud storage solution framework to identify which intermediate data sets need to be encrypted so that privacy-preserving cost can be saved while the privacy requirements of data holders can still be satisfied. The scope of the project is the option of privacy preserving in the intermediate datasets of cloud to reduce the accessibility cost of data.

Inference Control

An Inference Attack is a data mining technique performed by analyzing data in order to illegitimately gain knowledge about a subject or database. A subject's sensitive information can be considered as leaked if an adversary can infer its real value with a high confidence. This is an example of breached information security. An Inference attack occurs when a user is able to infer from trivial information more robust information about a database without directly accessing it. The object of Inference attacks is to piece together information at one security level to determine a fact that should be protected at a higher security level.

Inference control (disclosure control) aim at protecting data from indirect detection. It ensures queries of non-sensitive data when put together do not reveal sensitive information. The goal of inference control is to prevent the user from completing any inference channel.

RELATED WORKS

An enhanced scientific public cloud model (ESP) that encourages small or medium scale research organizations rent elastic resources from a public cloud provider. On a basis of the ESP model we design and implement the Dawning Cloud system that can consolidate heterogeneous scientific workloads on a Cloud site. An innovative emulation methodology and perform a comprehensive evaluation. We found that for two typical workloads. Two typical workloads: high throughput computing (HTC) and many task computing (MTC), Dawning Cloud saves the resource consumption maximally by 44.5% (HTC) and 72.6% (MTC) for service providers, and saves the total resource consumption maximally by 47.3% for a resource provider with respect to the previous two public Cloud solutions. To this end, we conclude that for typical workloads:

HTC and MTC, Dawning Cloud can enable scientific communities to benefit from the economies of scale of

public Clouds. A prominent shortcoming of the dedicated system model: for peak loads, a dedicated cluster system cannot provide enough resources, while lots of resources are idle for light loads.

A cloud computing model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three delivery models, and four deployment models. The drawbacks of this approach is that consumers might not completely trust measurements provided solely by a service provider, which might require agreed-upon third-party mediators to measure the SLA's critical service parameters and report violations Also that in clouds, service providers usually do not know their users in advance, so it is difficult to assign users directly to roles in access control policies.

The k-anonymization as an important privacy protection mechanism in data publishing. While there has been a great deal of work in recent years, almost all considered a single static release. Such mechanisms only protect the data up to the rst release or rst recipient. In practical applications, data is published continuously as new data arrive; the same data may be anonymized differently for a different purpose or a different recipient. In such scenarios, even when all releases are properly k-anonymized, the anonymity of an individual may be unintentionally compromised if recipient cross-examines all the releases received or colludes with other recipients. Preventing such attacks, called correspondence attacks, faces major challenges. In this paper, we systematically characterize the correspondence attacks and propose an efficient anonymization algorithm to thwart the attacks in the model of continuous data publishing.

This paper provides a systematic way to characterize the correspondence attacks and propose an efficient anonymization algorithm to thwart the attacks in the model of continuous data publishing. All the possible attacks like F-attack, C-attack and B-attack is discussed in this paper. The drawbacks of this approach is data with frequent changes like frequent updates / inserts were not considered in this paper and it adds a major drawback in this paper. Due to dynamic environment, the frequent data release makes this project void.

An emerging problem of continuous privacy preserving publishing of data streams which cannot be solved by any straightforward extensions of the existing privacy preserving publishing methods on static data. To tackle the problem, we develop a novel approach which considers both the distribution of the data entries to be published and the statistical distribution of the data stream.

An extensive performance study using both real data sets and synthetic data sets verifies the effectiveness and the efficiency. Distribution of the data entries to be published and the statistical distribution of the data stream is the core idea of this project and it's not handled ever before in this perception. A concrete model and an anonymization quality measure, and developed a group of randomized methods.

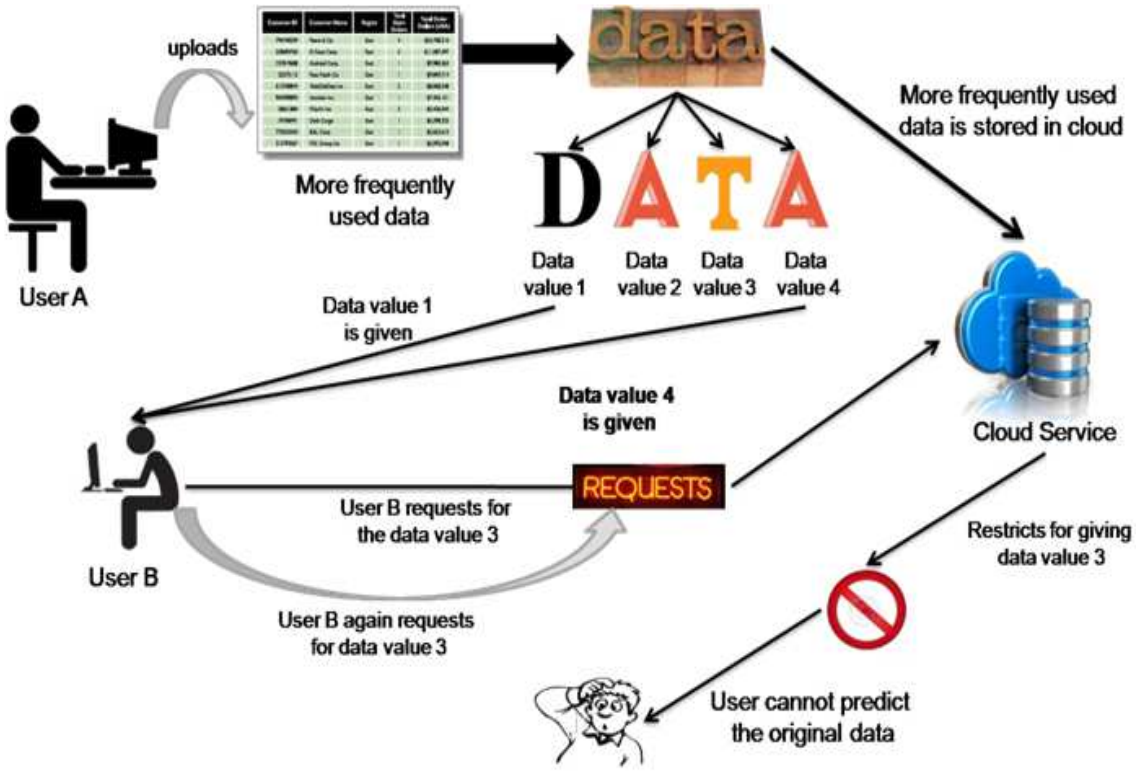

Figure 1: System Architecture for Partial Encryption

The over-provisioning for a service whose popularity does not meet their predictions, thus wasting costly resources, or under-provisioning for one that becomes wildly popular, thus missing potential customers and revenue. Moreover, companies with large batch-oriented tasks can get results as quickly as their programs can scale, since using 1000 servers for one hour costs no more than using one server for 1000 hours. This elasticity of resources, without paying a premium for large scale, is unprecedented in the history of IT. To reduce that confusion by clarifying terms, providing simple figures to quantify comparisons between of cloud and conventional computing, and identifying the top technical and non-technical obstacles and opportunities of cloud computing. The drawbacks of this approach is usage-based pricing is not renting. Renting a resource involves paying a negotiated cost to have the resource over some time period, whether or not you use the resource.

PARTIAL ENCRYPTION

We study the strategies for efficiently achieving data staging and data storage for privacy concern on a set of vantage to reduce the computational cost of encryption or decryption of data sets in a cloud system with a minimum outlay. Surplus data used for improvising the efficient optimal solutions is based on the dynamic upper bound privacy which is polynomial bounded by the number of service requests and the number of distinct data items in cloud. This is partial as most of the existing staging or privacy upper bound targets towards a class of services that access and process the decrypted data and thereby inherit the severity of data when access time sequence is more. Alternatively, a constraint optimization problem can be defined as a regular constraint to find a solution to the problem whose cost, evaluated as the sum of the cost functions, is minimized. Third parties who have privilege over intermediate datasets are created in order to reduce the frequent access of data from cloud directly that increases the cost. Hence the procedure of anonymization and homomorphic type of encryption are done in the system. In turn, avoids the possibility of inference channel analysis.

The major problem of the system is the relation between a particular sensitive data with the other data should be identified properly and it should be anonymized. Frequent access pattern on the data may get changed in timely manner.

The important and critical intermediate datasets that needs to be encrypted for security purposes, hence reducing encryption/decryption cost and thus maintaining data privacy. One way for evaluating this upper bound for a partial solution in our existing paradigm is to consider each constraint separately and mining the data in order to restrict access when the user claims to find the original information. For each constraint, the maximal possible value for any of these values is an upper bound may recover privacy-sensitive partial column level encryption. Hence a column wise encryption in the unencrypted data's of intermediate datasets is proposed. Additional a 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 a result, the algorithm requires an upper bound on the cost that can be obtained from extending a partial solution, and this upper bound should be significantly reduced with our approach over existing ones.

The advantages of the proposed system are that automatic scheduling process may enable the system synchronized as it is with the current situation and finding all possible data and encrypt based on the relationships will provide more value and wait age to the entire system.

RESULTS AND DISCUSSIONS

Information is captured to find out the dataset. The datasets have a collection of related information's with separate elements that can be used as an unit. Further an end user application is created in order to access the datasets accordingly. The term end user distinguishes the user for which the information is transferred from other users, who are transferring the information needed for the end user. The input data are listed in a table. These data in this table can be linked and referred to data in another table. All the required data can be obtained from these tables. These tables are the main basis of the datasets. Further, these datasets are prone to many degrading scenarios. The main scenario in this factor is the high severity. High severity data are prone to be critical and has to be counterfeited immediately to make sure the transfer process is done without any deteriorating activity. Datasets have to made free of those high severity data initially before carrying on the process.

Datasets are analyzed to find the data that has high severity. Data are prone to have high severity due to the fact that they are fed up to the table by random users. These random users has high probability to involve unauthorized users. When these data are added in to the table, predominantly it will start to deteriorate the entire operation. The data are mainly added to the table through the cloud. People who have direct access to that cloud can add any data in to the datasets. There might be users who are connected to the cloud but won't have authorization to the particular data set. When those users add up data to the table it will lead to high severity. Accessing high severity data will lead to adverse effects. These high severity data will not have any particular details about the sender of the data and the operation that it has to perform. It will be injected in to the dataset mainly to override the regular operation. This high severity data can be found anywhere in the dataset and can be injected in to a dataset at any time.

They have to be analyzed periodically to emphasize accuracy. In this proposed method the high severity data accessed by the requestor from the data owner through cloud are analyzed. The data accessed through the cloud are subjected to a particular pattern. These pattern will be familiar with the people who are linked to the respective dataset.

All the data that are being added up to a particular dataset will have the same pattern. So a small change in pattern and leave a hint that something abnormal has happened. So these patterns are analyzed periodically to ensure retrieval of the correct data and to find the data that has high severity. There are high chance of a leakage of tables due to high severity. The high severity data can be passed on to other tables if they are linked or referred.

Data present in table are profound to have links with other data in another table. Certain dataset are prone to be complex and would require branching of data to make it simple. So eventually, almost most of the data in a dataset will have links with the other table. These links are relationally shared. So if a data has high severity in one table and if it has relational link with another table's data then there's high chance leakage of severity to the another unaffected table's data. In this method, the data in a particular table are analyzed and the links that are referred by this table's data are identified. The identified links are further analyzed to check leakage of severity to the other table data. In similar way all the referred links are identified and analyzed to ensure that all the data having high severity are encapsulated.

The data that are used often and those that are common will make an ease of operation to increase the severity. Predictive data will lead to high severity. The data present in a dataset should be totally not understandable by the unauthorized user who access this dataset through the cloud. Making the data to be less predictable can emphasize hard chance for high severity. Complex data will be hard to identify and invoke. This can be ensured by adding unpredictable datasets. Unpredictable datasets will usually contain anonymous data. The unpredictable datasets have data that are more complex and with no particular details. This anonymous data in datasets will ensure that the real description behind all those data is not disclosed. If the real operation to be done is not disclosed then high severity data can be controlled. Identifying this unpredictable datasets is a tedious process and it is prone to ensure high security.

Datasets prone to leakage are identified using the respective scenarios. Datasets with leakage will lead to deterioration of the entire dataset along with the other datasets that are linked to it. Apart from leakage there's another factor as high severity. Both this leakage and high severity has to be encrypted to improve the security.

This encryption criteria will prevent the high severity data from getting spread to the other table that are linked. For the case of high encryption process more sophisticated double encryption is proposed. Double encryption algorithm works by encrypting the leaking and high severity dataset in two distinct encrypting methods. In order to obtain the dataset, both the encryption has to be decrypted. Data are transferred through cloud and so it is prone to have dynamic changes. Analyzing during the dynamic changes will require more sophisticated procedures. Dynamic changes have to be analyzed frequently to keep the data in exact order. On contrary, frequent analysis will cause more lag in the process so a particular time interval has to be proposed. Time stamps can be used to counterfeit. A particular interval will be selected and it will be recorded periodically in the form of time stamp. Dynamic changes can be analyzed during each timestamp and it can be checked against its severity.

CONCLUSIONS AND FUTURE WORK

The proposed method identifies which part of intermediate data sets needs to be encrypted while the rest does not, in order to save the privacy preserving. The problem of saving privacy-preserving cost as a constrained optimization problem is addressed by decomposing the privacy leakage constraints. Also 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 cost.

In future, privacy and cost optimization of datasets that are accessible through cloud by considering many other factors such time span of usage, availability of servers and so on can be executed effectively.

REFERENCES

1. X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen," A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud", IEEE Trans. Parallel and Distributed Systems, Vol. 24, no. 6, June 2013.
2. 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.
3. 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.
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. 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.
6. 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.
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. 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.
9. B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.. *(references)*