Reduce cost and efficient access for cloud storage Using Intermediate Cloud Datasets

Dr.R.Mala[*], **M.Lakshmi**[**], **R.Rajakumar**[***]

[*] Guest Lecturer, Department of Computer Science
[**] M.phil Research Scholar, Department of Computer Science, Marudupandiyar college of arts and science, Thanjavur
[***] Assistant Professor, Department of Computer Science, Annai Group of Institutions, Kumbakonam.

Abstract- 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. 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 en/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 ones where all data sets are encrypted.

Index Terms- Cloud computing, data storage privacy, privacy preserving, intermediate data set, privacy upper bound

I. Introduction

Technically, cloud computing is regarded as an inge-nious combination of a series of technologies, establish-ing a novel business model by offering IT services and using economies of scale [1], [2]. 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 [4], [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 [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 collabora-tion. 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 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 anonymized 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 [13]. 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 challen-ging, 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. 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 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.

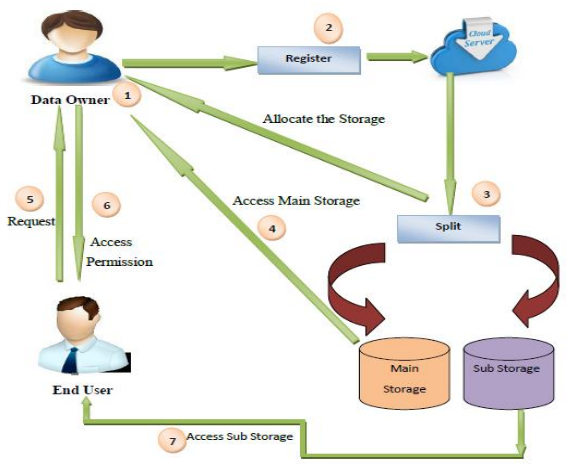
This paper is a significantly improved version of [14]. Based on [14], we mathematically prove that our approach can ensure privacy-preserving requirements. Further, the heuristic algorithm is redesigned by considering more factors. We extend experiments over real data sets. Our approach is also extended to a graph structure.

The remainder of this paper is organized as follows: The related work is reviewed in the next section. A motivating example and problem analysis are given in Section 3. In Section 4, we present the fundamental privacy representa-tion of data sets and derive privacy leakage upper bound constraints. Section 5 formulates our approach. In Section 6, we evaluate the proposed approach by conducting experi-ments on both real-world data sets and extensive data sets. Finally, we conclude this paper and discuss our future work in Section 7.

II. RELATED WORKS

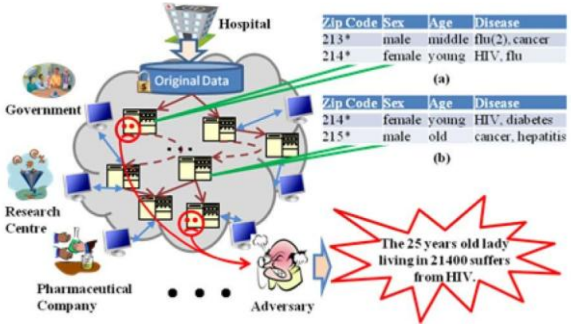
We briefly review the research on privacy protection in cloud, intermediate data set privacy preserving and Privacy-Preserving Data Publishing (PPDP).

Currently, encryption is exploited by most existing research to ensure the data privacy in cloud [8], [9], [10]. 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. [15] investigated the data privacy problem caused by MapRe-duce and presented a system named Airavat which incorporates mandatory access control with differential privacy. Puttaswamy et al. [16] described a set of tools called Silverline that identifies all functionally encryptable data and then encrypts them to protect privacy. Zhang et al. [17] 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. [18] 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], [7], but the research on privacy issues incurred by such data sets just com-mences. Davidson et al. [19], [20], [21] 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 exten-sively on privacy-preserving issues and made fruitful progress with a variety of privacy models and preserving methods [13]. Privacy principles such as k-anonymity [22] and l-diversity [23] 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 [13]. The research in [22], [21] exploits information theory to quantify the privacy via utilizing the maximum entropy principle [22]. The privacy quantification herein is based on the work in [22], [21]. Many anonymization techniques like generalization [12] have been proposed to preserve privacy, but these methods alone fail to solve th Problem of preserving privacy for multiple data .Our



Zip Code	Sex	Age	Disease
213*	male	middle	flu(2), cancer
214*	female	young	HIV, flu

(a)

Zip Code	Sex	Age	Disease
214*	female	young	HIV, diabetes
215*	male	old	cancer, hepatitis

(b)

Fig. 1. A scenario showing privacy threats due to intermediate data sets.consider the economical aspect of privacy preserving, adhering to the pay-as-you-go feature of cloud computing.

III. MOTIVATING EXAMPLE AND PROBLEM ANALYSIS

Section 3.1 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 3.2.

3.1 Motivating Example

A motivating scenario is illustrated in Fig. 1 where an online health service provider, e.g., Microsoft HealthVault [27], 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 [20]. 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.

3.2 Problem Analysis

3.2.1 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 [28]. Reproducibility of data provenance can help to regenerate a data set from its nearest existing predecessor data sets rather than from scratch [6], [20]. We assume herein that the information recorded in data provenance is leveraged to build up the generation relationships of data sets [6].

We define several basic notations below. Let d$_o$ be a privacy-sensitive original data set. We use D ¼ fd$_1$; d$_2$; . . . ; d$_n$g to denote a group of intermediate data sets generated from d$_o$ where n is the number of intermediate data sets. Note that the notion of intermediate data herein refers to both intermediate and resultant data [6]. Directed Acyclic Graph (DAG) is exploited to capture the topological structure of generation relationships among these data sets.

Definition 1 (Sensitive intermediate data set graph). A DAG representing the generation relationships of intermediate data sets D from d$_o$ is defined as a Sensitive Intermediate data set Graph, denoted as SIG. Formally, SIG ¼ hV ; Ei, where V ¼ fd$_o$g [D, E is a set of directed edges. A directed edge hd$_p$; d$_c$i in E means that part or all of d$_c$ is generated from d$_p$, where d$_p$; d$_c$ 2 fd$_o$g [D.

In particular, an SIG becomes a tree structure if each data set in D is generated from only one parent data set. Then, we have the following definition for this situation.

Definition 2 (Sensitive intermediate data set tree (SIT)).

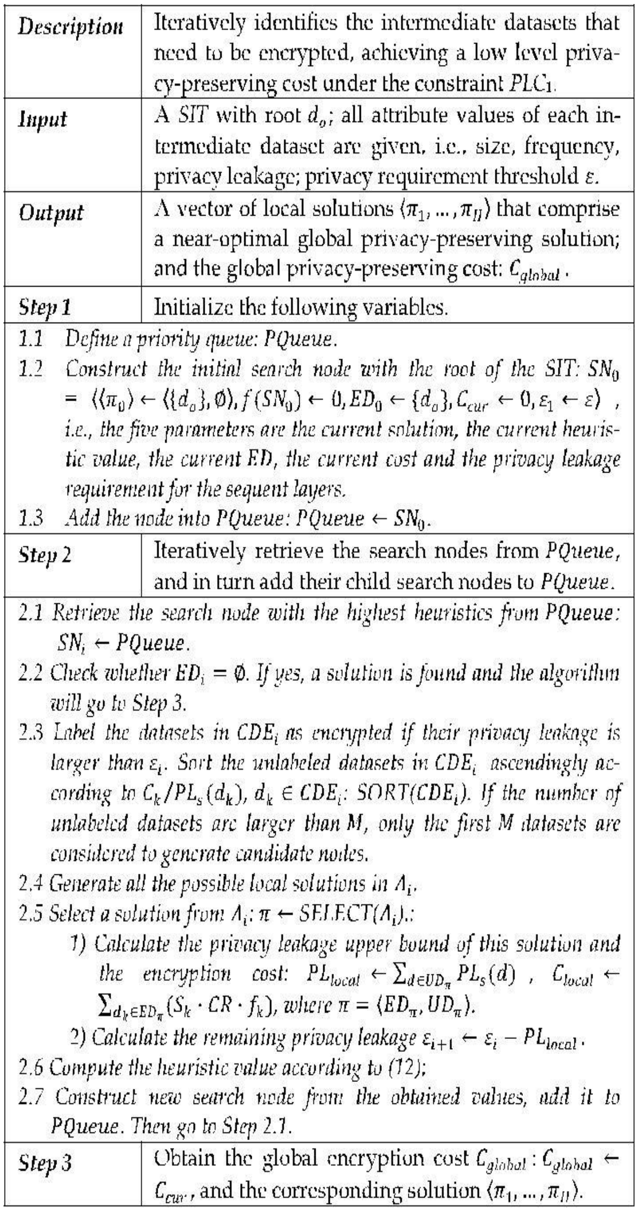
An SIG is defined as a Sensitive Intermediate data set Tree if it is a tree structure. The root of the tree is d$_o$.

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 d$_o$ 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 5.

3.2.2 Privacy-Preserving Cost Problem

Privacy-preserving cost of intermediate data sets stems from frequent en/decryption with charged cloud services. Cloud service venders have set up various pricing models to support the pay-as-you-go model, e.g., Amazon Web Services pricing model [21]. Practically, en/decryption needs computation power, data storage, and other cloud services. To avoid pricing details and focus on the discussion of our core ideas, we combine the prices of various services required by en/decryption into one. This combined price is denoted as PR. PR indicates the overhead of en/decryption on per GB data per execution.

Algorithm 1. Privacy-Preserving Cost Reducing Heuristic



```
Description  Iteratively identifies the intermediate datasets that need to be encrypted,
             achieving a low level privacy-preserving cost under the constraint $PLC_1$.
Input        A $SIT$ with root $d_o$; all attribute values of each intermediate dataset
             are given, i.e., size, frequency, privacy leakage; privacy requirement threshold $\varepsilon$.
Output       A vector of local solutions $\langle \pi_1, \dots, \pi_H \rangle$ that comprise a near-optimal
             global privacy-preserving solution; and the global privacy-preserving cost: $C_{global}$.
Step 1       Initialize the following variables.
  1.1  Define a priority queue: PQueue.
  1.2  Construct the initial search node with the root of the SIT: $SN_0 = \langle \langle \pi_0 \rangle \leftarrow \langle \{d_o\}, \emptyset \rangle, f(SN_0) \leftarrow 0, ED_0 \leftarrow \{d_o\}, C_{cur} \leftarrow 0, \varepsilon_1 \leftarrow \varepsilon \rangle$,
       i.e., the five parameters are the current solution, the current heuristic value, the current ED,
       the current cost and the privacy leakage requirement for the sequent layers.
  1.3  Add the node into PQueue: $PQueue \leftarrow SN_0$.
Step 2       Iteratively retrieve the search nodes from PQueue, and in turn add their child
             search nodes to PQueue.
  2.1 Retrieve the search node with the highest heuristics from PQueue: $SN_i \leftarrow PQueue$.
  2.2 Check whether $ED_i = \emptyset$. If yes, a solution is found and the algorithm will go to Step 3.
  2.3 Label the datasets in $CDE_i$ as encrypted if their privacy leakage is larger than $\varepsilon_i$.
      Sort the unlabeled datasets in $CDE_i$ ascendingly according to $C_k / PL_s(d_k)$, $d_k \in CDE_i$: $SORT(CDE_i)$.
      If the number of unlabeled datasets are larger than M, only the first M datasets are
      considered to generate candidate nodes.
  2.4 Generate all the possible local solutions in $\Lambda_i$.
  2.5 Select a solution from $\Lambda_i$: $\pi \leftarrow SELECT(\Lambda_i)$.:
        1) Calculate the privacy leakage upper bound of this solution and the encryption cost:
           $PL_{local} \leftarrow \sum_{d \in UD_\pi} PL_s(d)$ , $C_{local} \leftarrow \sum_{d_k \in ED_\pi} (S_k \cdot CR \cdot f_k)$, where $\pi = \langle ED_\pi, UD_\pi \rangle$.
        2) Calculate the remaining privacy leakage $\varepsilon_{i+1} \leftarrow \varepsilon_i - PL_{local}$.
  2.6 Compute the heuristic value according to (12);
  2.7 Construct new search node from the obtained values, add it to PQueue. Then go to Step 2.1.
Step 3       Obtain the global encryption cost $C_{global}$ : $C_{global} \leftarrow C_{cur}$, and the corresponding
             solution $\langle \pi_1, \dots, \pi_H \rangle$.
```

IV. CONCLUSIONS AND 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.

ACKNOWLEDGMENT

Part of this work was conducted under the Research centre of Marudupandiyar College of arts and science and bharathidasan university cloud laboratory, Tirchirappalli laboratory manners.

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,"

[9] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOM '11, pp. 829-837, 2011.

[10] 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.

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

[12] 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.

[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.

[14] 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.

[15] 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.

[16] K.P.N. Puttaswamy, C. Kruegel, and B.Y. Zhao, "Silverline: Toward Data Confidentiality in Storage-Intensive Cloud Applica-tions," Proc. Second ACM Symp. Cloud Computing (SoCC '11), 2011.

[17] 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), 515-526, 2011.

[18] 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.

[19] 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.

[20] 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.

[21] S.B. Davidson, S. Khanna, V. Tannen, S. Roy, Y. Chen, T. Milo, and J. Stoyanovich, "Enabling Privacy in Provenance-Aware Work-flow Systems," Proc. Fifth Biennial Conf. Innovative Data Systems Research (CIDR '11), pp. 215-218, 2011.

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

AUTHORS

First Author – Dr.R.Mala, received her Phd degree and also working in Guest Lecturer, in Dept of Computer Science and supporting guide of M.phil and Phd students in Maruduapandiyar arts and science ,Thanjavur., E-Mail: murugan.dcdrf@gmail.com

Second Author – M.Lakshmi, received her M.CA, degree in sastra university, kumbakonam and also doing M.Phil research degree in marudupandiyar college of arts and science,thanjavur
E-Mail:lakshmi.guru86@gmail.com

Third Author – R.Rajakumar received his M.sc., M.Phil. M.Tech Degree in computer science and computer science and engineering from various recognized universities. He is currently pursuing PHD degree in computer science at Bharathiar University., Email:raja2012mtech@gmail.com