ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Investigating the Privacy Preserving and Efficient Scheduling of Sensitive Datasets in Cloud Storage

[I]Pradeeba. K, [II]Baskaran.G
[I,II]Srinivasan Engineering College, Perambalur
Email : [I]pradeeba06@gmail.com, [II]gbaskarancse@gmail.com

Abstract
Cloud c*omputing is an emerging field in the development of business and organizational environment. Nowadays most of the organization is willing to store their large amount of data into cloud in a secure manner and get the application from it. In earlier day's encryption and decryption technique is used to preserve the intermediate data set in cloud storage. To encrypt entire data sets, this method shows increased time and computation cost, which is a drawback of this method. However, preserving the privacy of sensitive intermediate data sets becomes a challenging problem because adversaries may hack sensitive information by analysing multiple intermediate data sets. In this paper, Upper bound constraints based approach is proposed to identify sensitive intermediate data sets and non sensitive intermediate data sets and apply suppression technique on sensitive data sets only in order to reduce time and cost. Suppression of data is done in two ways that is full suppression and semi suppression. Value Generalization Hierarchy protocol is used to achieve the more security. Here number of user can access the data with privacy and to avoid privacy leakage.*

Keywords
Privacy Preserving, Privacy Leakage Constraint (PLC), Value Generalization Hierarchy (VGH), Sensitive intermediate data set, Upper Bound Constraint, Suppression, and Data sharing.

I. Introduction

Nowadays cloud computing provide the many more applications and also improve the utility on storage level, security level and computation level. It is a new business model, which enable on-demand self service. Cloud computing is delivery services rather than a product over internet such as software, resources etc. Number of participants involve this business model and can access benefit from it without investment of infrastructure [1]. Therefore, many huge organizations are going into cloud and build their applications and perform much computation. However, larger numbers of customers are still using cloud, to take advantage of Security and privacy concerns [4, 5]. Why more users like cloud storage means because the cloud storage does not affected by any natural disaster like earth quake, flood and electrical fire, So enormous amount users using cloud storage .

Usually, intermediate data sets in cloud are accessed and processed by multiple parties, but rarely controlled by original data set holders so the original data set to violated. This enables an adversary may collect intermediate data set together and got privacy sensitive information from them. It leads economic loss to data holder. Existing approaches for privacy preserving of intermediate data set in cloud storage mainly using the Encryption

And decryption technique, One hand, all data sets to be encrypted for preserve the security; it is effective approach, widely adopted in many scenarios [8, 9, 10]. However, existing applications, that are performs any computational process only on unencrypted data sets, is quite a challenging task in cloud. On the other hand, selected information to be encrypted and rest of the information do not encrypted. In this approach reduce the cost rather encrypting entire data set. Modern techniques for privacy preserving like generalization [12] can resist more security attacks on one data set, while preserving privacy for collective data set is still challenging problem in cloud computing [13].

However to encrypting all intermediate data set, when they are frequently accessed or processed then it will lead to high redundancy and inefficient. In existing approaches not support storage level security in cloud. An organization store their data into cloud using encryption/decryption techniques for privacy preserving, these techniques applicable only on cloud service provider itself not in cloud storage server. All users can store their original data set into cloud storage server through cloud service provider. In this paper, propose a novel approach to

Provide privacy preserving for multiple intermediate data set with minimum cost and fast computation. To preserve the privacy of intermediate data sets in cloud storage, Upper bound constraints based approach is used to identify which intermediate data set is more sensitive data set and which is none sensitive intermediate data set, then that most sensitive intermediate data set only to be encoded and the none sensitivity data set doesn't encoded. Encoding is done through the suppression technique; suppression technique on sensitive intermediate data set is done in two forms that are semi-suppression and full-suppression in order to reduce the privacy preserving cost and time. Hence semi-suppression means not fully encoded of individual sensitive data set, full-suppression means fully encoded for individual sensitive intermediate data set. Also propose Value Generalization Hierarchy (VGH) protocol in order to reduce overall time and cost.

II. Related Work

Here, briefly examine on privacy preserving or protection in cloud, privacy preserving for intermediate data set and Privacy-Preserving Data Publishing (PPDP).

The privacy requirement in cloud storage done through the intermediate data sets, any organizations are securely store their data into cloud and get computation services is based on the pay per usages [1]. Thus cloud user can stores valuable intermediate data set in order to avoid frequent re-computation on the same data set which is more expensive. The processing of large amount of intermediate data set is known data intensive applications. Example of data intensive application like medical report management, medical analysis. This application improves the efficiency of medical management. [6, 7].

Yuan et al [6] proposed Benchmarking technique, demonstrates the cost effectiveness over the representative storage strategies

and using Cost Transitive Tournament Shortest Path(CTT-SP) based algorithm that can find the minimum cost storage of the intermediate datasets in scientific cloud workflow system, Data stored in fixed location and data transfer is not efficient, not focus on security. Here original data is converted into intermediate data dependency graph (IDG) to reduce the cost of their storage in scientific cloud workflow system. Lin and Tzeng [8] investigate the data confidentiality over multiple storage servers, and proposed a threshold proxy re-encryption scheme incorporates it with a decentralized erasure code such that a secure distributed storage. This scheme supports encoding process on encrypted messages as well as forwarding technique on encoded and encrypted messages. Main drawback of these operations needs more time and cost. The privacy preserving for huge data sets in cloud using encryption and decryption techniques, that are differ from various functionality, this approaches are already exists in cloud environment.

Ciriani et al [19] described privacy of data collection that combines data fragmentation and encryption, initially data holder fragment the original data that is larger data divided into smaller data after, apply the encryption method only when explicitly demanded by the privacy requirements, which provides safeguards to companies and also control privacy breaches. The resulting fragments may be stored at the same or at different servers. Cao et al [9] proposed Multi-keyword Ranked Search over Encrypted cloud data (MRSE), to improve variety of privacy requirements, using co-ordinate matching principle i.e., as many matches as possible, to capture the inner product similarity between search query and data documents. According to their similarity, cloud server can provide ranked results to the users, then user can access data document with limitation and MRSE concern the security but not much stronger.

In this paper [10] to defined APKS over encrypted data based on Hierarchical Predicate Encryption (HPE) technique, supporting efficient multi-dimensional range queries, which is similar to MRSE method. The PPDP [13] research on privacy preserving issues and proceeds with various privacy methods and preserving models. Privacy-preserving data publishing (PPDP) provides and to develop the methods and tools for sharing useful information while preserving data privacy. There are two types of data publishers available one is un-trusted model and other one is trusted model, this data publishing privacy models are already exists in data mining environment.

Zhang et al [14] introduced an Upper-Bound Control Approach for privacy protection of intermediate data set storage in cloud computing, to identify which stored data set is need to be encrypted and which not. Privacy requirements of original data sets are decomposed and satisfied layer by layer, a tree model is leveraged to analyse privacy disclosure of datasets and this approach practice heuristic implementation for effectively shorten the privacy preserving cost of re-computing data sets. Our approach also reduces cost of storage and computation cost but differs from data hiding techniques; selectively attribute or intermediate data sets to be encoded in order to avoid privacy protection charges while encrypting entire data sets. Many techniques have been proposed to preserve secrecy but these methods failure to solve the problem of privacy preserving for multiple data sets. Our methodology to achieve privacy preserving of multiple data sets to incorporates anonymization with encoding technique.

III. System Architecture

The proposed system offered an approach is effective way to control privacy leakage of sensitive intermediate data set, while accessing and computation perform on them. The figure 3.1 shows overall system architecture. Initially data owner can generate original data object, then collects the intermediate data sets from original document which also made by data holder itself, because original data objects contain more and more noisy data, irrelevant and redundant data. Apply Upper-Bound Constraints Based Approach used to determine which intermediate data set is most sensitive and which data set is non sensitive.

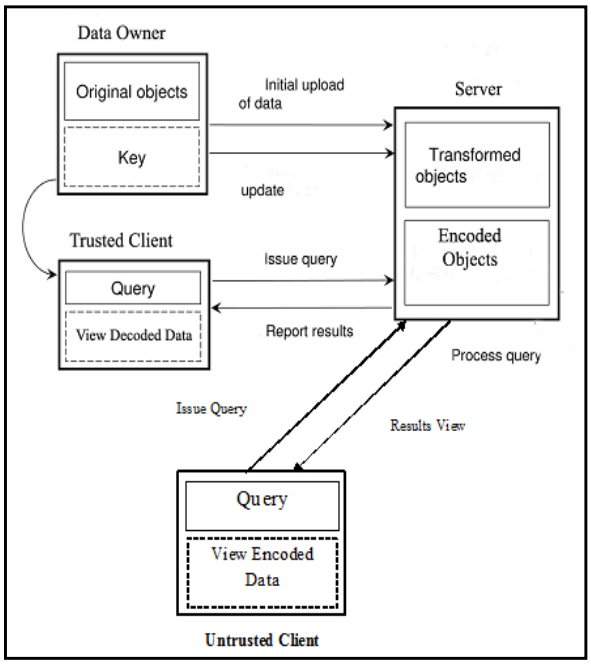

Fig. 1 : System Architecture

Here Upper-bound in the sense of some measurable threshold value, above the threshold value is denoted as sensitive and below the threshold value is denoted as none sensitive intermediate data sets, then apply suppression technique, is encoding or hiding on sensitive intermediate data sets. Encoding done through n-records with key value (K=2). Encoded the sensitive data sets then data owner initially upload to cloud storage server. Any update to the encoded data set is done by data owner alone, and then data owner granted a key value to trusted client for purpose of accessing or processing of original data set. Trusted client can access the stored data in cloud server by way of query, client issue query to server then server report resultant data set to corresponded client. Un-trusted client can access the data sets through query but view only encoded data object. Finally avoid privacy leakage of original data set. Hence to achieve privacy preserving for multiple data set with low cost and fast computation.

A. System Process

Initially data holder to make an authentication process, data owner will enter the user name and password in the login form, if valid means SQL server database connection will be established then the owner to generate authentic data sets. The major grants of our approach are threefold. First, formally determine the possibility of protect privacy leakage requirements without encrypting all intermediate data sets when encryption is incorporated with anonymization to preserve privacy of multiple data sets. Second, design a practical heuristic algorithm to identify which data sets need to be encoded for privacy preserving while the rest of them do not. Third, experiment results display that our approach powerful

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

to reduce privacy-preserving cost over existing approaches, which is quite benefits for the cloud users who utilize cloud services in a pay per usage fashion. These processes have done by four modules such as privacy data storage, privacy preserving, and privacy intermediate dataset, privacy upper bound.

B. Motivating Example

A motivating scenario is where an online health service provider, e.g., Microsoft Health Vault [27], store the valuable information or patient record maintenance into cloud for the purpose of economical benefits, and also help effortless analysis. Original intermediate data sets are encoded for privacy protection. Data users like governments or research centre's access or process part of original data sets after anonymization. Number of Intermediate data sets are generated during data access or process are retained for data reuse and cost saving. Two intermediate data sets are generated independently and anonymized to satisfy 2 diversity, i.e., minimum two individuals own the same quasi-identifier and each quasi-identifier corresponds to at least two sensitive values [23]. Knowing that a lady aged 25 living in 21,400 (corresponding quasi identifier is (214*; female; young) is in both data sets, an adversary can deduce that this personality suffers from HIV with high confidence collected together. Hiding the data set is a promising way to avoid such a privacy breach. Assume two intermediate data sets are same size; the frequency of accessing for first intermediate data set is 10 and second intermediate data set is 100, then hide on second intermediate data set only in order to reduce the privacy protection cost. In real-world applications, huge numbers of intermediate data sets are involved. Hence, it is challenging problem to identify which data sets should be encoded to ensure that privacy leakage requirements are satisfied while keeping the hiding with lowest expenses as far as possible.

C. Problem Analysis

1. Sensitive Intermediate Data Set Management

Data Provenance is employed to manage intermediate data sets in our research. Provenance defined as the source or basis, history of derivation of some objects and data, which can be evaluated as the information upon how data were generated. Recreate of data provenance can help to regenerate a data set from its nearest existing predecessor data sets rather than from scratch. Assume here in that the information recorded in data source to build up the generation relationships of data sets, define several basic notations below. Let d be a privacy-sensitive original data set. We use D = {d1, d2… dn} to denote a group of intermediate data sets created from d where n is the number of intermediate data sets. Directed Acyclic Graph (DAG) is a topological structure that exploited to capture generation relationships among these data sets.

DAG Representation of generation relationships among intermediate data sets D from d is defined as a Sensitive Intermediate data set Graph, that denoted as SIG. Formally SIG = {v ,e}, where v= {d} and e is set of directed edges. Sensitive intermediate data sets to convert as tree structure with the help of Sensitive Intermediate data set Graph (SIG) is defined as Sensitive Intermediate data set Tree, denoted as SIT, where the root of tree is d.

2. Cost Problem for Privacy Protection

A cloud service provides various pricing models to support the pay-as-you-go model, e.g., Amazon Web Services pricing model [29]. Privacy-preserving cost of intermediate data sets reduce from frequent en/decryption with charged cloud services. Encryption and decryption needs more computation power, data storage, and other cloud services. To avoid pricing details and focus on, combine the prices of various services required by en/decryption into one. This is denoted as PR, it describes the overhead of en/decryption on one GB data per execution.

D. Proposed Framework

The technique as well the new protocol included for privacy protection here is the Value Generalization Hierarchy Protocol which has the functionality of assignment of the common values for the unknown and original data values for general identification. This is later succeeded by the application of full suppression on the more important data sets which enhances the complete encoding of the entire data sets given.

Investigates privacy aware and efficient scheduling of intermediate data sets for minimum cost and fast computation. Suppression of data is done to reduce the overall computation time and cost, where VGH Protocol is also proposed to achieve it. In my base paper does not consider the full suppression, secure the more important data set though semi suppression only. The full suppression to achieve the high privacy/security of original data sets. And the original data set only viewed by owner. Here number user can access the data with security and to avoid privacy leakage. Minimizing the privacy-protection cost for intermediate data sets that needs to be encoded while using an upper bound constraint-based approach to select the necessary subset of intermediate data sets. The privacy concerns caused by retaining intermediate data sets in cloud are important

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 re-computation to obtain these data sets. Such scenarios are quite common because data users often re-analyse results, conduct new analysis on intermediate data sets, or share some intermediate results with others for collaboration. 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. However, processing on encrypted data sets efficiently is quite a challenging task, because most existing applications only run on unencrypted data sets. 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.

Data sets are divided into two sets. One is sensitive intermediate data set and another is non-sensitive intermediate data set. Sensitive data set is denoted as SD then non sensitive data set is denoted as NSD. The equations,

sd U NSD =D and SD ∩ NSD =Φ hold.

The pair of (SD, NSD) as a global privacy preserving of cloud data. Suppression technique done only on sensitive data sets in two ways such as semi suppression and full suppression, while full suppression apply on most important sensitive intermediate data set that is individual data set value fully encoded then semi suppression apply on selective sensitive data sets that is half of the data set value will be encoded. Also propose Value Generalization Hierarchy (VGH) protocol to reduce cost of data

storage and computation. VGH defined as assign the common value for different attribute value, this method used to achieve reduce the time delay, for example, pg courses like M.E., M.tech, mca, mba. These values in cloud data set simply denote pg instead of original courses.

1. Data Storage Privacy Module

The data owner and providers are set to create the original datasets. The privacy concerns caused by retaining intermediate datasets in cloud are important but they are paid little attention as far as they are concerned in the recent times. There exists a motivating scenario illustrated with an on-line health service provider which is incorporated by many existing organizations e.g., Microsoft Health Vault has moved data storage into cloud for economical and management benefits. The original datasets are encrypted for confidentiality purpose. Data users like governments or research centre's access and process a part of original datasets after anonymizing the initial data. The intermediate datasets generated during data access or process are retained for data reuse and cost saving. Here the proposal of an approach that combines encryption and data fragmentation achieves privacy protection for distributed data storage with encrypting only part of datasets.

2. Privacy Preserving Module

The privacy-preserving techniques like generalization can withstand most privacy attacks on one single dataset, whereas preserving privacy for multiple datasets is still a challenging problem till date. Here there arises a necessity to distinguish between the sensitive and non-sensitive datasets. As a result, for preserving privacy of multiple datasets, it is promising to anonymize all datasets first and then encrypt them before storing or sharing them in cloud medium of storage. Privacy-preserving cost of intermediate datasets stems from frequent encryption with charged cloud services.

3. Intermediate Dataset Module

An intermediate dataset is assumed to have been anonymized to satisfy certain privacy requirements which are given a very deep importance in cloud and various platforms. However, putting multiple datasets together may still invoke a high risk of revealing privacy-sensitive information. The suppression and non semi-suppression techniques are applied along with VGH protocol to assign a common value for different attributes. This results in violation of the privacy requirements of the available datasets. Data provenance is employed to manage intermediate datasets in research Areas. Here provenance is commonly defined as the origin, reference, source or history of derivation of some objects and data, which is reckoned as the information upon how data was generated and about the inception and evolution of overall process. The recurrence of data provenance can help to regenerate a dataset from its nearest existing predecessor datasets rather than from scratch.

4. Privacy Upper Bound Module

Privacy quantification of a single data-set is stated as a continuation of the intermediate dataset processing. In order to point out the challenge of privacy quantification of multiple datasets and then deriving a privacy leakage upper-bound constraint correspondingly, an upper-bound constraint based approach is proposed to select the necessary subset of intermediate datasets that needs to be encrypted for minimizing privacy-preserving cost. Aforesaid privacy leakage upper-bound constraint is decomposed layer by layer which makes the end of the stated module completion.

Upper Bound Constraint of Joint Privacy Leakage

The sum of privacy leakage of un-encoded data sets or witness data set can be deemed as an upper bound constraint. Based on replacing the PLC with such an upper bound constraint; propose an approach to address the optimization problem.

Definition 3 (Privacy leakage constraint): Let be the privacy leakage threshold allowed by a data holder, then a privacy requirement can be represented as PLm (NSD <ε) , NSD Є D. This privacy requirement is defined as a Privacy Leakage Constraint, denoted as PLC.

An heuristic Algorithm for cost based privacy Sustention

Based on the heuristics of the assumed statistics, we design a privacy-preserving cost reduction algorithm, which is denoted as H_PPCR for example. The basic idea is that the algorithm iteratively selects a state node with the highest heuristic value and then extends its child state nodes until it reaches a goal state node in prior the beginning of the action. As usual the privacy-preserving solution and corresponding cost are derived from the goal state in order to achieve the desired solution.

Algorithm 1 specifies the details of the proposed heuristic algorithm for better understanding. A priority queue is exploited to keep state nodes initially. Only the qualified state nodes that are added to the priority queue so that the corresponding partial global solutions are feasible in reality. To avoid the size of the priority queue increase gradually, the algorithm only retains the state nodes with top K highest heuristic values. The child nodes are determined to add search nodes in layer L_{ip1} into the priority queue, and the algorithm generates a local encryption solution from CDE_i at first. Later, the algorithm probably suffers from poor efficiency because it has to check all combinations of data sets in CDE_i. In order to avoid this situation, the privacy algorithm ascendingly sorts the data sets in CDEi according to the value $C_k=PL_s/d_k$ where d_k Є CDE_i and $C_k = S_k . PR . f_k$. If $|CDE_i|$ is larger than a threshold M, then the first M data sets in the sorted CDE_i will be examined while the remaining are set to be encoded. Intuitively, sensitive intermediate data sets with higher privacy-protection cost and lower privacy leakage are expected to remain unencrypted. As a result, the value $C_k=PL_s (d_k)$ can help to guide the algorithm to find these data sets with a higher possibility.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

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 ε.
Output	A vector of local solutions $\langle \pi_1, \ldots, \pi_{ll} \rangle$ 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* ε_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* Λ_i.	
2.5 *Select a solution from* Λ_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* $\pi = (ED_\pi, UD_\pi)$.	
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 $\langle \pi_1, \ldots, \pi_{ll} \rangle$.

Algorithm 1. Heuristic algorithm

Therefore, the algorithm is guided to approach the goal state in the state space as close as possible to overcome all drawbacks. Above all, in the light of heuristic information, privacy algorithm can achieve a near-optimal solution practically. The existing SORT and SELECT are two simple external functions as their names signify applies to the algorithm to proceed to further steps in future.

V. Conclusion

In this paper, focus is mainly contributed towards an approach that identifies the portions where the most sensitive intermediate data sets are present in cloud. That data sets needs to be encoded, in order to reduce the privacy preserving cost and mainly in the overall time reduction. For achieving it, a tree structure is modelled from the generation relationships of intermediate data sets to analyse privacy requirements among data sets stored in cloud. The privacy-preserving cost problem of modelling is changed as a constrained optimization problem which is addressed by the privacy leakage constraints. To address the issues a practical heuristic algorithm has been designed accordingly. Experimental results on real-world data sets and larger extensive data sets are done for demonstrating the cost of preserving privacy in cloud. This approach is an advantage over existing ones by having the encryption and decryption of all data sets in cloud transaction. Finally, in order to achieve the desired goal a high security provisioning is done with the help of full suppression, semi suppression and Value Generalization Hierarchy Protocol. This protocol is used to assign the common attribute for different attributes. Experimental results prove the efficient working of the real time implementation of data sets in cloud environment.

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] *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.*

[4] *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*

[5] *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.*

[6] *I. Roy, S.T.V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and Privacy for Map reduce," Proc. Seventh USENIX Conf. Networked Systems Design and Implementation (NSDI '10), p. 20, 2010.*

[7] *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.*

[8] *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.*

[9] *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*

[10] *X. Zhang, C.Liu, Surya Nepal, Suraj Pandey, and J. 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 system, vol.24, no.6, june 2013*

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