Privacy Preservation of Cloud Storage Data using Classification

Gurudayal Singh Bhandari, Asst. Prof. Abhishek Singh Chauhan

Abstract— Privacy Preserving Data Mining (PPDM) addresses the problem of developing accurate models about aggregated data without access to precise information in individual data record. Under the single trust level assumption, a data owner generates only one perturbed copy of its data with a fixed amount of uncertainty. This assumption is limited in various applications where a data owner trusts the data miners at different levels. Multilevel Trust (MLT) poses new challenges for perturbation-based PPDM. In contrast to the single-level trust scenario where only one perturbed copy is released, now multiple differently perturbed copies of the same data are available to data miners at different trusted levels. The more trusted a data miner is the less perturbed copy it can access; it may also have access to the perturbed copies available at lower trust levels. Moreover, a data miner could access multiple perturbed copies through various other means, e.g., accidental leakage or colluding with others. Here a new and efficient technique for privacy preservation is implemented for the cloud so that various security issues in the cloud is detected and prevented.

Index Terms— Cloud Computing, Privacy Preservation, PPDM, Multi Level Trust.

1 INTRODUCTION

Cloud computing acquires improvement of hardware virtualization to steadily and enthusiastically distribute physical resources such as computational power, storage, and networks to the users. Clouds resources are distributed to the end-users through Web services. These uncomplicated model consequences in the following good-looking features:

- *Elasticity:* Since physical resources are dynamically allocated to the consumers according to their needs, cloud services can scale on-demand.
- *Cost Effectiveness:* Resource sharing improves utilization of physical resources and thus reduces the associated cost.
- *Pay-as-you-go Pricing Model:* Cloud services have consumption-based metering and billing; this property makes them more affordable for small businesses and startups.
- *Global-scale Accessibility and Usability:* Cloud consumers have access to a virtually unlimited physical resource pool through Web.
- *Easy Maintenance:* All non-functional requirements of IT, such as maintenance of hardware and software, are addressed by cloud providers, therefore consumers can concentrate on their functional business requirements.

In addition, the large amount of cloud data and owner's confined computing competence additional makes the job of data correctness auditing in a cloud environment exclusive and even difficult for entity cloud customers. Consequently, facilitating public audit ability [1], [2], [3] for cloud storage is of critical importance so that owners can alternative to a concentrated third party auditor (TPA) to audit cloud storage services and sustain strong storage accuracy assurance, while saving their own valuable computing resources.

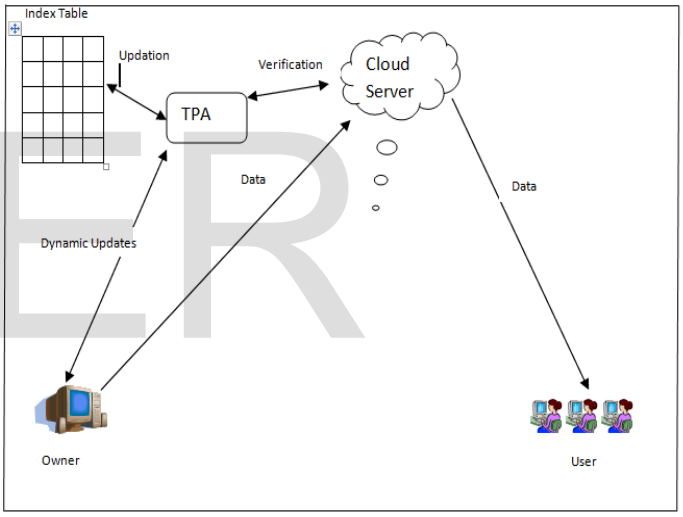

Figure.1: Dynamic TPA System

Secure Third Party Auditing framework and key components of the cloud computing environment are shown in Figure. The TPA Auditing Manager facilitates collaboration among different service providers by composing new desirable services. Each TPA Auditing Manager has components that are responsible for establishment and maintenance of trust between the local provider domains and between the providers and the users, provisioning desirable services and generating global policies.

PRIVACY PRESERVING AND ITS TYPES

Privacy is becoming an important issue in various data-mining applications such as health care, security, financial, and other types of sensitive data. It has become important in counter terrorism and homeland defense-related applications. Today the data storage requirements are becoming large, as the large number of data is evolving day by day. As the Large database is required to store the large amount of data, so the privacy of

the data is also an important factor. This has caused concerns that personal data may be used for a variety of intrusive or malicious purposes. Privacy preserving data mining [4], [5], [6] helps to achieve the aim of data mining by preserving the privacy of sensitive data here data mining goals without scarifying the privacy of the individuals and without learning underlying data values. Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. The previous privacy preserving solutions were limited to only single level trust, which was not sufficient to preserve the privacy of information. So by expanding the scope from single level trust, here in the proposed system, multilevel trust solution for privacy preservation is applied in which data owner generates the different perturbed copies of same data for data miners of different trust levels. In privacy-preserving data mining (PPDM), data mining algorithms are analyzed for the side-effects they incur in data privacy, and the main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process [7].

2 LITERATURE SURVEY

Yuan, Jiawei, and Shucheng Yu Proposed Privacy Preserving Back-Propagation Neural Network Learning Made Practical with Cloud Computing. The main idea of this scheme can be summarized as follows: each participant first encrypts her/his private data with the system public key and then uploads the cipher texts to the cloud; cloud servers then execute most of the operations pertaining to the learning process over the cipher texts and return the encrypted results to the participants; the participants jointly decrypt the results with which they update their respective weights for the BPN network. For the duration of this process, cloud servers discover no privacy data of a participant even if they collude with all the rest participants. During off-loading the computation tasks to the resource-abundant cloud, this scheme makes the computation and communication complexity on each participant independent to the number of participants and is thus extremely scalable. For privacy preservation they crumble most of the sub-algorithms of BPN network into simple operations such as multiplication, addition, and scalar product. To sustain these functions over cipher texts, they adopt the BGN (Boneh, Goh and Nissim) 'doubly homomorphic' encryption algorithm [9] and tailor it to split the decryption capability among multiple participants for collusion-resistance decryption [8].

Zhu, Yan et al [10] suggested efficient provable data possession for hybrid clouds. They focused on the construction of PDP scheme for hybrid clouds, supporting privacy protection and dynamic scalability. They first provide an effective construction of Cooperative Provable Data Possession (CPDP) using Homomorphic Verifiable Responses (HVR) and Hash Index Hierarchy (HIH). This construction uses homomorphic property, such that the responses of the client's challenge computed from multiple CSPs can be combined into a single response as the final result of hybrid clouds. By using this mechanism, the clients can be convinced of data possession without knowing what machines or in which geographical locations their files reside. More importantly, a new hash index hierarchy is proposed for the clients to seamlessly store and manage the resources in hybrid clouds. Their experimental results also validate the effectiveness of their construction.

Cong Wang et al [11] propose a privacy-preserving public auditing system for data storage security in cloud computing. They utilize the homomorphic linear authenticator and random masking to guarantee that the TPA would not learn any knowledge about the data content stored on the cloud server during the efficient auditing process that not only removes the burden of cloud user from the tedious and possibly expensive auditing task, although also assuages the users' fear of their outsourced data escape. Taking into consideration TPA may concurrently handle multiple audit sessions from different users for their outsourced data files, they further extend our privacy-preserving public auditing protocol into a multiuser situation, where the TPA can execute numerous auditing tasks in a batch manner for better efficiency [11].

To accomplish privacy-preserving public auditing, they suggest to uniquely integrating the homomorphic linear authenticator with random masking method. In this protocol, the linear combination of sampled blocks in the server's reaction is masked with randomness generated by the server. With random masking, the TPA no longer has all the essential information to construct up a accurate group of linear equations and for that reason cannot derive the user's data content, no issue how many linear combinations of the identical set of file blocks can be composed. Alternatively, the rightness corroboration of the block-authenticator pairs can still be accepted in a new way even with the occurrence of the randomness. Their design makes employ of a public key-based HLA, to provide the auditing protocol with public auditability [11].

Li, Yaping et al [12] offered Enabling Multi-level Trust in Privacy Preserving Data Mining. In particular, we focus on the additive perturbation approach where random Gaussian noise is added to the original data with arbitrary distribution, and make available a systematic solution. During a one-to-one mapping, this solution allows a data owner to generate distinctly perturbed copies of its data according to different trust levels. They expand the scope of perturbation based PPDM to multi-level trust, by relaxing the implicit assumption of single-level trust in existing work. MLT-PPDM introduces another dimension of flexibility which allows data owners to generate differently perturbed copies of its data for different trust levels. They identify a key challenge in enabling MLTPPDM services. In MLT-PPDM, data miners may have access to multiple disconcerted copies. By combining numerous perturbed copies, data miners may be capable to achieve assortment attacks to reconstruct the original data more accurately than what is allowed by the data owner [12].

In 2008 by Stephen S. Yau, et. al [13] gives a concept about warehouse for integrating data from various data sharing services without central authorities is existing with our warehouse, data sharing services can update and control the access and limit the usage of their shared data, as a substitute of submitting data to establishment, and our repository will support data sharing and

addition. The main differences between their storehouse and existing central authorities are: 1) repository collects data from data sharing services based on users' integration requirements rather than all the data from the data sharing services as existing central establishment. 2) While existing central establishment have full control of the collected data, the capability of warehouse is controlled to computing the integration results required by users and cannot get other information about the data or use it for other work. 3) The data composed by warehouse cannot be used to generate other results except that of the specified data addition request, and, hence, the cooperation of warehouse can only reveal the results of the specified data integration demand, while the compromise of central establishment will reveal all data and presented a privacy preserving repository to integrate data from various data distribution services. In contrast to existing data allocation techniques, warehouse only collects the least amount of information [13].

The proposed method [14] is competent to protect user's privacy in opposition to each one single authority. This paper presents an anonymous attribute-based privilege control scheme Anony Control to address only the data privacy problem in cloud storage, but also the user uniqueness privacy concerns in existing access control methods. The proposed scheme understands adjacent to authority give and take, and compromising of up to (N −2) multiple authorities does not bring the entire system downward. By using multiple authorities in cloud computing system, their method attains anonymous cloud data access and fine grained privilege control. They also make available specified analysis on security and performance to demonstrate likelihood of our method demonstrates that Anony Control is both secure and efficient for cloud computing situation. They first implement the genuine toolkit of multi-authority based encryption method. As well, their method stand for the give and take attack towards attributes authorities, which is not faced in many existing efforts [14].

3 Proposed Methodology

Privacy Preservation is a concept of keeping the data private for the external users. The concept of cloud computing means the sharing of resources or data between various users. But when any users shared data between users privacy is maintained between users so that the external users of the cloud can't access the data.

1. A broker needs to compute the data which contains the management of resources on the basis of certain criteria.
2. Now Broker contains a dataset with huge values. Now the broker needs a quick decision when ever any user asks a question.
3. The broker if pass the whole data to any user of the cloud for computation privacy discloses.
4. Hence broker divides the whole data set into two parts and distributes it to two users.
5. Now since the users don't have complete data hence decision is not taken by the user and the data is still private for the users.
6. The users can compute the data and send the information to the broker, where broker on the basis of data will creates a decision tree and is only known to the broker but not the users.
7. Hence a privacy is maintained by the broker for the users of the cloud.

ALGORITHM

1) Create a cloud simulation environment.
2) User Ui of the cloud sends the data to the Datacenter DCi.
3) The original data is partition in N parts and submit it N brokers BRi.
4) Each of the brokers BRi computes Information Gain using the algorithm below and sends the Gain to the Global Broker.
5) The Global broker on the basis of Information Gain generates a classification tree.

INPUT LAYER

- Party individually calculates Expected Information of every attribute.
- Party individually calculates Entropy of every attribute.
- Party individually calculates Information Gain of each attribute.

Assume there are two classes, P and N

Let the set of examples S contain p elements of class P and n elements of class N

The amount of information, needed to decide if an arbitrary example in S belongs to P or N is defined as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Assume that using attribute A set S will be partitioned into sets $\{S_1, S_2, \ldots, S_v\}$

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i)$$

- If S_i contains p_i examples of P and n_i examples of N, the entropy, or the expected information needed to classify objects in all subtrees S_i is,

The encoding information that would be gained by branching on A

$GAIN(A) = I(p, n) - E(A)$

OUTPUT LAYER

- All party sends Information Gain of each attribute to the UTP
- UTP compute the sum of Information Gain of all parties of all attributes (TotalInformationGain ( )).
- UTP find out the attribute with the largest Information Gain by using MaxInformationGain( )
- Create the root with largest Information Gain attribute and edges with their values, then send this attribute to all parties at Input Layer for further development of decision tree.

- Recursively do when no attribute is left.
- Assumptions
- The following assumptions have been set
- UTP computes the final result from the intermediate results provided by all parties at every stage of decision tree.
- UTP computes attribute with highest information gain and send to all party at every stage of decision tree.
- UTP has the ability to announce the final result of the computation publicly.
- Each party is not communicating their input data to other party.

The communication networks used by the input parties to communicate with the UTP are secure.

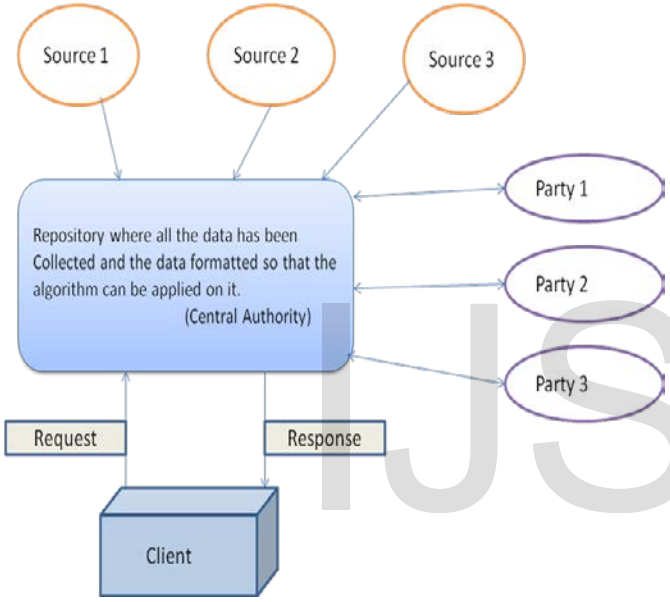

Figure 2. Proposed Methodologies

4 Result Analysis

The table shown below is the analysis and comparison of existing and proposed work on Encryption cost for Iris and Diabetes dataset.

	Encryption Cost	
Dataset	Existing work	Proposed Work
Iris	5	2
Diabetes	43	36

Table 1. Comparison of Encryption Cost

The table shown below is the analysis and comparison of existing and proposed work on learning time for Iris and Diabetes dataset.

	Learning Time (min)	
Dataset	Existing work	Proposed Work
Iris	10	6
Diabetes	22	13

Table 2. Comparison of learning time

The table shown below is the analysis and comparison of existing and proposed work on Communication cost for Iris and Diabetes dataset.

	Communication Cost	
Dataset	Existing work	Proposed Work
Iris	3.56	2.48
Diabetes	6.73	4.41

Table 3. Comparison of Communication Cost

The table shown below is the analysis and comparison of existing and proposed work on Error rate for Iris and Diabetes dataset.

	Error Rate	
Dataset	Existing work	Proposed Work
Iris	0.45	0.28
Diabetes	0.26	0.08

Table 4. Comparison of Error Rate

The figure shown below is the generation of decision tree using the prposed methodology. The decision tree is used here for the privacy preservation where the decision is taken only on the basis of decision tree and not o the actual dataset.

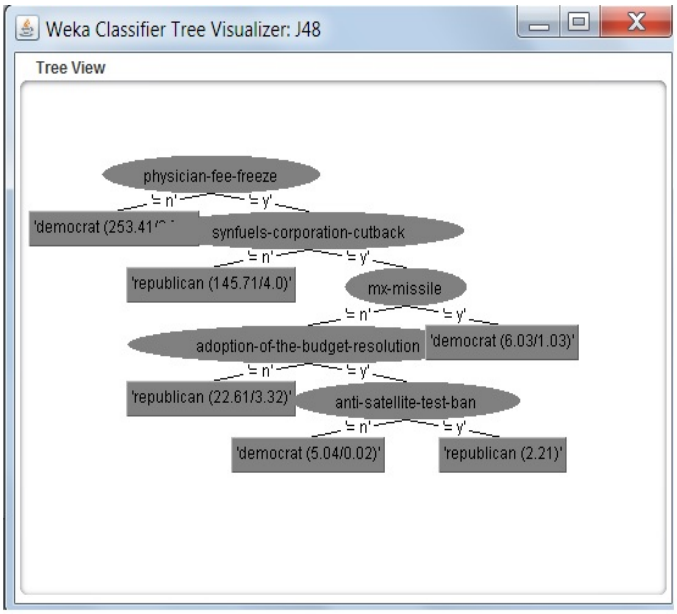

Figure 3. Generation of Decision Tree

The figure shown below is the analysis and comparison of existing and proposed work on Learning for Iris and Diabetes dataset.

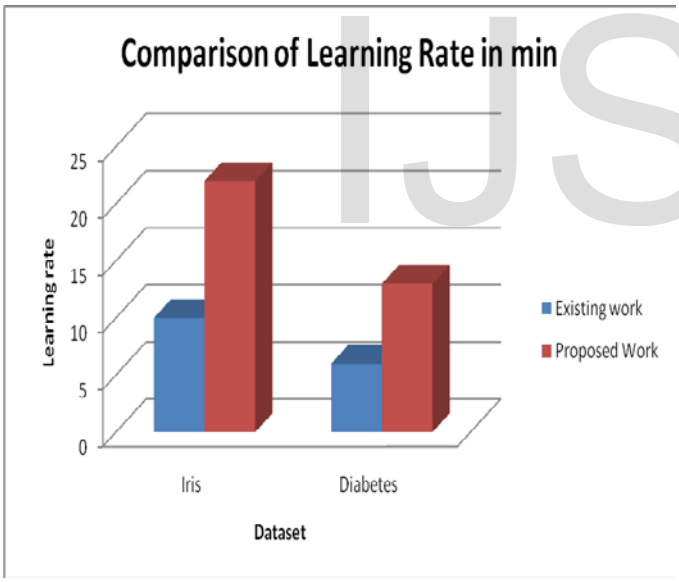

Figure 4. Analysis of Learning Rate

The figure shown below is the analysis and comparison of existing and proposed work on Encryption cost for Iris and Diabetes dataset.

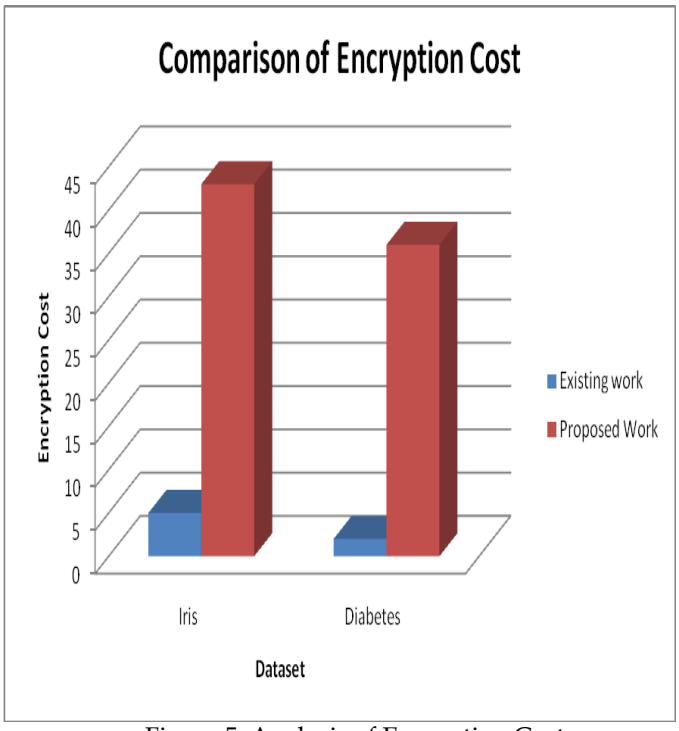

Figure 5. Analysis of Encryption Cost

The figure shown below is the analysis and comparison of existing and proposed work on Communication cost for Iris and Diabetes dataset.

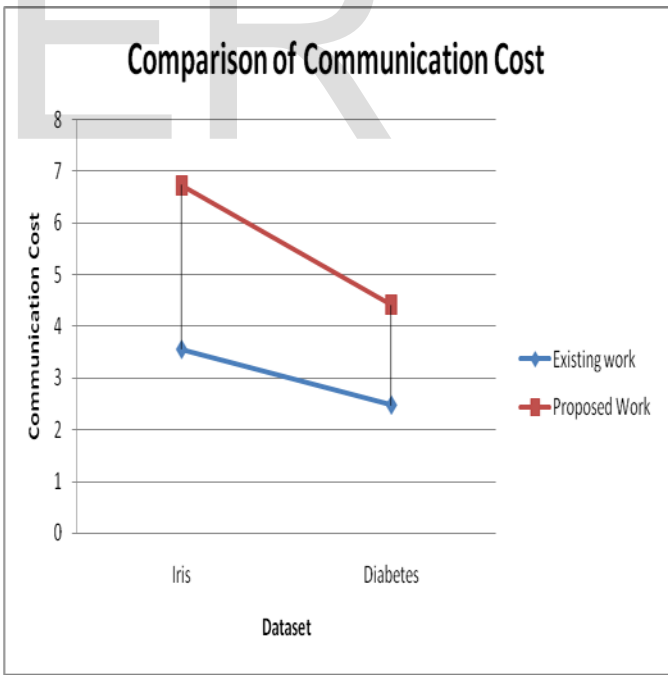

Figure 6. Analysis of Communication Cost

5 Conclusion & Future Work

The proposed methodology implemented here is efficient in terms of providing privacy to the user of the cloud when access and datacenter in the cloud. The technique implemented here using classification tree provides bandwidth utilization,

reduces computational cost and reduces overhead in cloud. Classification is based on Predicate over Encrypted data. Develop the prototype, conduct experiments and evaluate the approach. Authenticate without disclosing identifying information. Ability to securely use a service while on an untrusted host (VM on the cloud). Minimal disclosure and minimized risk of disclosure during communication between user and service provider (Man in the Middle, Side Channel and Correlation Attacks) Independence of Trusted Third Party.

REFERENCES

[1] Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, "Enabling Public Auditability and Data Dynamics for Storage Security in Cloud Computing," IEEE Trans. Parallel and Distributed Systems, vol. 22, no. 5, pp. 847-859, May 2011.

[2] Cong Wang, Sherman S.M. Chow, Qian Wang, Kui Ren, and Wenjing Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage," IEEE Transactions on Computers (TC), 2011 (A preliminary version of this paper appeared at the 29th IEEE Conference on Computer Communications (INFOCOM'10)).

[3] Cong Wang, Qian Wang, Kui Ren, and Wenjing Lou, "Towards Secure and Dependable Storage Services in Cloud Computing," To appear, IEEE Transactions on Service Computing (TSC). (A preliminary version of this paper appeared at the 17th IEEE International Workshop on Quality of Service (IWQoS'09)).

[4] D. Agrawal and C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Principles of DatabaseSystems (PODS '01), pp. 247-255, May 2001.

[5] R. Agrawal and R. Shrikant, "Privacy Preserving Data Mining,", Proc. ACM SIGMOD Int'l Conf. Management of Data 2000.

[6] Y. Lindell and Benny Pinkas, "Privacy Preserving Data Mining", Proc. Int'l Cryptology Conf. (CRYPTO), 2000.

[7] Verykios V.S., Bertino E., Fovino I.N., Provenza L.P., Saygin, Y. & Theodoridis Y.(2004a). State-of-the-art in privacy preserving data mining, SIGMOD Record, Vol. 33, No. 1, pp.50-57.

[8] Yuan, Jiawei, and Shucheng Yu. "Privacy Preserving Back-Propagation Neural Network Learning Made Practical with Cloud Computing", IEEE Transactions on Parallel and Distributed Systems, Vol. 25, Issue 1, pp. 212 – 221, 2014.

[9] D. Boneh, E.-J. Goh, and K. Nissim "Evaluating 2-dnf formulas on ciphertexts". In Proceedings of the Second international conference on Theory of Cryptography, TCC'05, pp. 325–341, Berlin, Heidelberg, 2005.

[10] Zhu, Yan, Huaixi Wang, Zexing Hu, Gail-Joon Ahn, Hongxin Hu, and Stephen S. Yau "Efficient provable data possession for hybrid clouds." In Proceedings of the 17th ACM conference on Computer and communications security, pp. 756-758. ACM, 2010.

[11] Cong Wang, Sherman S.M. Chow, Qian Wang, Kui Ren, and Wenjing Lou "Privacy-Preserving Public Auditing for Secure Cloud Storage", IEEE Transactions On Computers, Vol. 62, No. 2, pp. 362 – 375, February 2013.

[12] Li, Yaping, Minghua Chen, Qiwei Li, and Wei Zhang "Enabling multilevel trust in privacy preserving data mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, no. 9 pp. 1598-1612, 2012.

[13] Stephen S. Yau, Fellow And Yin Yin "A Privacy Preserving Repository For Data Integration Across Data Sharing Services", IEEE Transactions On Services Computing, Vol. 1, No. 3, July-September 2008 .

[14] Du, Wenliang, and Zhijun Zhan. "Building decision tree classifier on private data", In Proceedings of the IEEE international conference on Privacy, security and data mining, Vol. 14, pp. 1-8. Australian Computer Society, Inc., 2002.