ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

Cost-Effective Privacy Preserving Of Intermediate Data Sets in Cloud by Privacy Leakage Upper Bound Constraint-Based Approach

Abhishek R. Ladole[1], Prof. Amit M. Sahu[2]

M.E (CSE) Scholar, GHRCEAM, Amravati, India[1]

Assistant Professor, CSE Dept. GHRCEAM, Amravati, India[2]

Abstract: Cloud computing provides massive computation power and storage capacity which enable users to deploy computation and data intensive applications without infrastructure investment. Besides the processing of such applications, a large volume of intermediate datasets will be generated, and frequently stored to save the cost of re-computing them. Though, preserving the privacy of intermediate datasets becomes a challenging problem because adversaries may recover privacy-sensitive information by analysing multiple intermediate datasets. Encrypting all datasets in cloud is widely approved in existing approaches to tackle this challenge. But we dispute that encrypting all intermediate datasets are neither competent nor cost-effective because it is very time overwhelming and costly for data-intensive applications to en/decrypt datasets frequently while performing any operation on them. This paper, proposes a novel upper-bound privacy leakage constraint based approach to identify which intermediate datasets 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 exhibit that the privacy-preserving cost of intermediate datasets can be significantly reduced with our approach over existing ones where all datasets are encrypted.

Keywords: Cloud computing, data storage privacy, privacy preserving, intermediate data set, privacy upper bound

I. INTRODUCTION

Cloud computing derives as the collection of hardware, networks, storage, services, and interfaces that can be group together deliver aspects of computing as a service. This service mainly includes the delivery of software, infrastructure, and storage over the internet either as separate components or a complete platform based on user demand. It provide users to store large volume of data and to perform application over cloud without need of any also provides greater flexibility of storing and computation of data but, Such applications can be processed, huge volume processing data sets are to be generated. For Storing some valuable intermediate datasets has been considered in order to avoid the high recomposing them.

Cloud computing is a subscription-based service to obtain network storage space and computer resources. The Cloud helps to access the information from anywhere at any time but Internet connection is necessary to access the Cloud. E.g. Email client, if it is Yahoo!, Gmail, Hotmail, and so on, takes care of housing all of the hardware and software necessary to support client personal email account. Cloud also called as "pay-as-you-go" meaning that if technological needs change at any point to purchase more storage space from Cloud provider.

When processing on the original datasets in cloud applications, intermediate data sets are generated. Users in cloud environment have to pay-as-you-go manner for storage and computation. Therefore they may store valuable data sets in the cloud in order to reduce the cost of regenerating datasets again and again. Storing intermediate data in cloud elaborates the attack surfaces so that privacy requirement of data holder is at risk of being violated. Mostly multiple parties can access and process intermediate data, but it is rarely controlled by original data sets holder. In this scenario an advisory can collect intermediate datasets together to collect sensitive information from it. This will tends to severe social reputation impairment to data owners.

In existing approach mainly include encryption and anonymization techniques for the preserving privacy of whole intermediate data sets. A traditional counter measure is to hide ALL intermediate datasets by encryption. But these valuable datasets are usually shared by multiple users or accessed frequently, which requires repeat en/decryption.

To overcome the lacunas of the existing system encrypt part of intermediate data sets rather than all for reducing privacy preserving cost. 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 the data holders. For preserving privacy of multiple data sets, it is promising to anonymize all data sets first then encrypt them before storing them in cloud.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

II. RELATED WORK

X. Zhang, C. Liu, S. Pandey and J. Chen [1], 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. They employed data provenance to manage intermediate data sets.

T. Praveena, G. Raja [3], in this they derived that, the approach is a Threshold Filtering adopted to classify the dataset that are to be encrypted. A value for each intermediate dataset will be fixed and based on the privacy information present in the dataset .If the value of intermediate dataset is higher than the threshold value then it will be encrypted and remaining dataset anonymized. For encryption two round searchable encryption (TRSE) is used for easy searching and accessing the encrypted dataset. For 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.

Ms. C. Celcia, Mrs. T. Kavitha [5], in this paper they projected that in existing data intensive application of cloud provide massive computation power and storage space. In this surroundings they are more number of user can accessed or processed original data sets frequently due to this intermediate data sets are generated from original one. For privacy preserving heuristic approach identifies which intermediate data set need to be encrypted while others not. In this approach encryption is incorporated with anonymization for cost effective preserving.

K. Zhang, X. Zhou, Y. Chen, X. Wang, Y. Ruan [10], designed a new technique called Sedic. It is designed to protect data privacy during map-reduce operations. Sedic leverages the special feature of Map Reduce to automatically partition a computing job according to the sensitivity level of the data it works on. It provides a solution to the privacy threat that is to split a task , maintaining the computation on the private data within an organization's private cloud while moving the rest to the public commercial cloud.

B. Fung, K. Wang, L. Wang and P.C.K. Hung [12], in this paper demonstrate the problem of releasing person-specific data for cluster analysis while protecting privacy of the datasets. The proposed solution for this is to mask unnecessarily specific information into a less specific but semantically consistent version, so that person-specific identifying information is masked but essential cluster structure remains. The key idea in this is to encode the original cluster structure into the class label of data records and subsequently preserve the class labels for the corresponding classification problem

Marten van Dijk, Ari Juels [14], in this article author argue that cryptography alone cannot enforce the privacy required by common cloud computing services, even along with such powerful tools as fully homomorphic encryption. They proposed that users of cloud services will also need to rely on other forms of privacy enforcement, such as tamperproof hardware, distributed computing, and complex trust ecosystems.

III. EXISTING SYSTEM AND PROBLEM DEFINITION

A. Existing System

In past all users were directly interact with original database, so every time user interact with original database an intermediate data set is to be generated. The privacy concerns caused by retaining intermediate data sets in cloud are important but they are paid little attention. Storage and computation services in cloud are equivalent from an economical perspective because they are charged in proportion to their usage. Thus, cloud users can store valuable intermediate data sets selectively when processing original data sets in data intensive applications like medical diagnosis, in order to restrict the overall expenses by avoiding repeated recomputation to obtain these data sets.

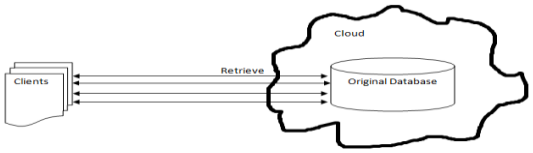

Fig.1 Existing Approach

B. Problem Definition

In current encrypting all the data sets is very costly and time consuming approach. Current privacy-preserving techniques like generalization can withstand most privacy attacks on one single data set. Preserving privacy for multiple data sets is still a challenging problem. And finally the most important issue is economical cost.

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, straightforward and effective approach , also processing on encrypted data sets efficiently is quite a challenging task, because most existing applications only run on unencrypted data sets.

IV. MOTIVATING EXAMPLE

A motivating scenario is illustrated in Fig. 4.3 where an online health service provider, e.g., Microsoft Health Vault, has moved data storage into cloud for economical benefits. Original data sets are encrypted for confidentiality. Data users like governments or research centers access or process part of original data sets after anonymization.

After frequent processing of original data sets different intermediate data sets are generated. Intermediate data sets generated due to data access or process by different users are retained for data reuse and also for cost saving. Due to retaining of intermediate data sets in cloud there is no need to access original data sets again and again.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

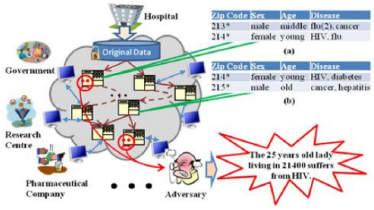

Fig. 2 Microsoft Health Vault

Two independently generated intermediate data sets (Figure 2 a) and (Figure 2 b) in Figure 2 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. 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 infer that this individual suffers from HIV with high confidence if Figure 2a and Figure 2b are collected together. Hiding Figure 2a or Figure 2b by encryption is a promising way to prevent such a privacy breach. Assume Figure 2a and Figure 2b are of the same size, the frequency of accessing Figure 2a is 10 and that of Figure 2b is 100. We hide Figure 2a to preserve privacy because this can incur less expense than hiding Figure 2b. In most real-world applications, a large number of intermediate data sets are involved. Hence, it is challenging to identify which data sets should be encrypted to ensure that privacy leakage requirements are satisfied while keeping the hiding expenses as low as possible.

V. PROPOSED SYSTEM

In this paper, to improve the performance of Privacy preserving of intermediate data sets, we propose, privacy leakage upper bound constrained based approach system that utilizes cost effective privacy preserving of intermediate data sets in cloud.

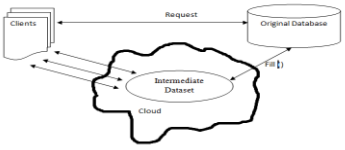

Fig. 3 Proposed Approach

The proposed system design requires a Cloud service for assistance; we opted for implementing compatibility with Cloud. In this we propose an 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 modelled from generation relationships of intermediate data sets to analyse privacy propagation of data sets.

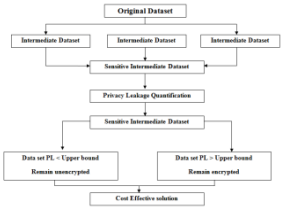

Fig. 4 Data Flow Diagram

VI. METHODOLOGY

Below figure 4 shows the various techniques implemented to achieve desired output. In this for the purpose of intermediate data sets management we have used techniques such quasi identifier, data provenience, encryption with anonymization.

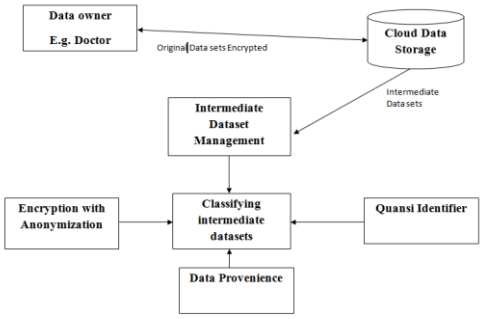

Fig. 5 Techniques Implemented

A. Data Provenience

It is 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. Take example given below. In the Sig given in below figure it is stated that D5 is generated from the Source D2, D3, and D4.

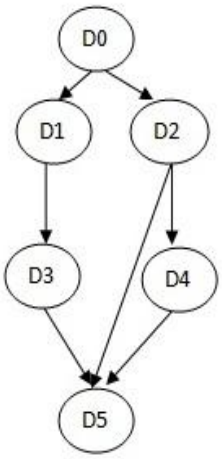

Fig. 6 Data Provenience

B. Quasi Identifier

Quasi Identifier (QI) is the pieces of information that are not themselves unique identifier, but are sufficiently well co-related with an entity that they can combine with other QI to create unique identifier. When we combine two or more QI then it will become personally indentifying information. For example neither gender, nor age and zip code uniquely identify an individual, but the combination of all three sufficient to identify 87% of individuals.

C. Encryption with Anonymization

In the existing technique to provide privacy all the intermediate data sets are encrypted first which is straightforward approach. However, processing on encrypting data set is an quite challenging task. Because every time for processing on encrypted data sets it need to be decrypted it first so it is somewhat expensive technique.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

VII. PRIVACY PRESERVING COST OF INTERMEDIATE DATA SETS

A. Privacy preserving cost Problem

As privacy preserving cost of intermediate data sets can be generated from frequent en/decryption along with charges provided by different cloud service providers. Cloud service seller such as Amazon Web Service model has their own pricing models according to services. Herein we combine the prices of various services required for en/decryption in to one. The privacy preserving cost can be calculated by below formula.

Privacy Preserving Cost= ∑ Si * PR *Fi

Where Si is the size of intermediate data set is which to be encrypted. Combine the prices of various services required by en/decryption into one. This combined price is denoted as PR. And Fi indicates the frequency of accessing or processing intermediate data set.

B. Sensitive Intermediate Tree (SIT).

We design here Sensitive Intermediate Tree (SIT). On the basis of privacy leakage through data set. While designing different layers of SIT encryption and Data provenience is employed to reduce the privacy preserving cost and to satisfy privacy requirement given by data holder.

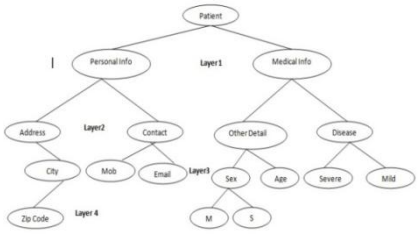

Fig. 7 SIT of Patient management System

An SIT is generated is 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. We have derived an SIT of patient information system layer-by-layer depending upon the privacy leakage through it as given above.

VIII. HEURISTIC ALGORITHM

Description	Iteratively identifies the intermediate datasets that need to be encrypted, achieving a low level privacy-preserving cost under the constraint PLC_0.
Input	A *SIT* with root d_0; 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 $\langle \pi_1, \ldots, \pi_H \rangle$ 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: $SN_0 = \langle \langle \pi_0 \rangle \leftarrow \langle \langle d_0 \rangle, \emptyset \rangle, f(SN_0) \leftarrow 0, ED_0 \leftarrow \{d_0\}, C_{cur} \leftarrow 0, \varepsilon_1 \leftarrow \varepsilon \rangle$, 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 $C_k/PL_{\varepsilon_i}(d_k)$, $d_k \in CDE_i$: $SORT(CDE_i)$. 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 A_i.*
2.5	*Select a solution from A_i: $\pi \leftarrow SELECT(A_i)$.:*
	1) Calculate the privacy leakage upper bound of this solution and the encryption cost: $PL_{local} \leftarrow \sum_{attr_{ik} \in \pi} PL_{\varepsilon}(d)$, $C_{local} \leftarrow \sum_{d_k \in ED_n} (S_k \cdot CR \cdot f_k)$, where $\pi = \langle ED_n, UD_n \rangle$.
	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 $\langle \pi_1, \ldots, \pi_H \rangle$.

Fig. 8 Heuristic Algorithm

IX. RESULT ANALYSIS

A. Time Comparison

Below graph indicate the comparison between time required to encrypt data set depending upon existing and proposed approach. From the below graph we can analyzed that time required to encrypt part of data set using proposed system significantly reduce over existing approach.

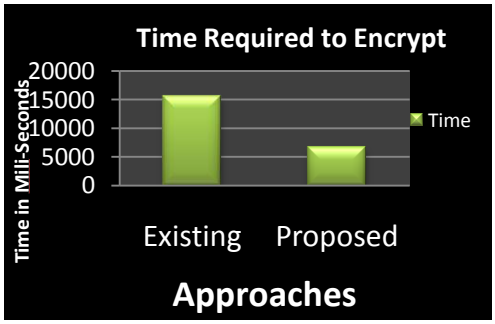

Fig. 9 Time Comparison for Existing and our Approach

B. Layer-by-Layer Decomposition of Time Required for Privacy Preserving

Below graph shows the time required to encrypt part of intermediate data set by using our approach from Layer 1 to Layer 4 and by existing approach in Layer 5.After analyzing graph we can easily say that, time required by our approach is less than existing one.

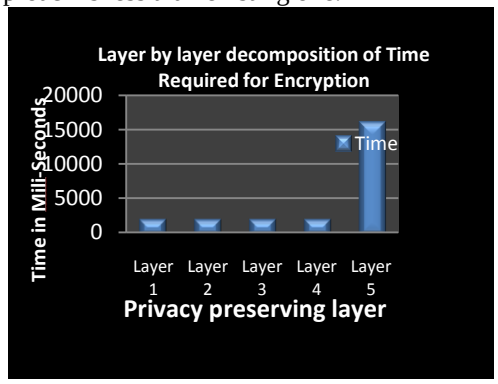

Fig. 10 Layer-by-Layer time required for preserving privacy

C. Size Comparison

Below graph indicate the comparison between size of data sets depending upon existing and proposed approach. From the below graph we can analysed that size of data set using proposed system is less than existing approach.

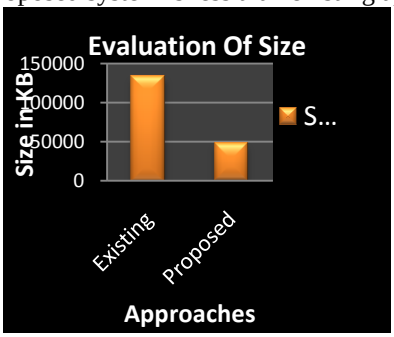

Fig. 11 Size Evaluations for Existing and Proposed Approach

DOI 10.17148/IJARCCE.2015.45112

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

D. Cost Comparison

In the below graph vertical axis represents the cost required to encrypt the datasets and the horizontal axis shows the two categories existing and proposed, the shaded bars shows the encryption cost required. After analysing the graph it shows that cost required by proposed approach is less than existing one.

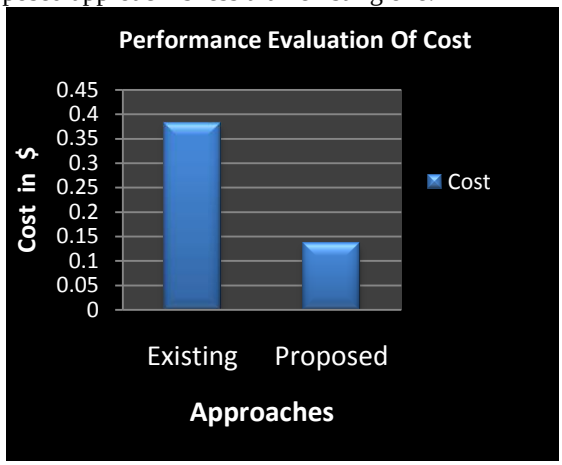

Fig. 12 Reducing the Privacy Preserving Cost by Proposed Approach

By comparing the cost for encrypting all the intermediate datasets and only part of intermediate datasets in the cloud we are saving the privacy preserving cost it can be shown in the following equation.

CSAV = CALL - CHEU

Here C**SAV** is the privacy preserving cost saved, CALL is the privacy preserving cost for encrypting all the intermediate datasets and CHEU is the privacy preserving cost for encrypting only part of intermediate datasets in the cloud.

X. CONCLUSION

In proposed system, overcome the limitation of Existing approach of encrypting all data sets instead of it we identifies which part of intermediate data sets need to be encrypted while rest of does not, in order to save the privacy preserving cost.

Evaluation results on real world data sets have demonstrated that the cost of privacy preserving in cloud can be reduced significantly with this approach over existing ones where all the data sets are encrypted. This is quite beneficial for the cloud users who utilize cloud services in a pay-as-you-go fashion.

REFERENCES

[1]. X. Zhang, C. Liu, 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, Vol 24, No. 6, June 2013.

[2]. A. Balachandra Rao and Sai Satyanarayana Reddy, "A Heuristic Privacy Leakage Control Approach For Profitable Privacy Protection Of Processing Datasets In Cloud ", weekly science aug 2014.

[3]. T. Praveena, G Raja, " Threshold Based Filtering approach For Cost Effective over Encrypted Cloud Data ", IJISET, Vol. 1 Issue 3, May 2014.

[4]. Ms. C. Lakshmi, Dr. N. Kasiviswanath, "An Approach for privacy Preserving Cost of Intermediate Datasets in Cloud ," IJARCSSE,Vol 4 issue 5, May 2014 ISSN 2277 128X.

[5]. Ms. C. Celcia, Mrs. T. Kavitha, "Privacy Preserving Heuristic Approach fpr Intermediate Data Sets in Cloud, " IJETT, Vol 9, March 2014 ISSN 2231-5381.

[6]. A.Thanapaul Pandi, M. Varghese, "Intermediate Data Scheduling in Cloud Environment with Efficient Privacy Preserving," IJERT, Vol 2 Issue 12, December 2013 ISSN 2278-0181.

[7]. S.Hemalatha, S. Alaudeen Basha, "Enabling for Cost-Effective Privacy Preservation of Intermediate Data Sets in Cloud," IJSRP, Vol 3 Issue 10, Octonber 2013 ISSN 2250-3153.

[8]. Lei Wang, Jianfeng Zhan, Weisong Shi, Yi Liang, "In Cloud, Can Scientific Communities benefit from the Economies of Scale?" Parallel and Distributed Systems, IEEE Transactions 2012.

[9]. K. P. N. Puttaswamy, C. Kruegel, B.Y. Zhao, "Silverline: Towards the data confidentiality in Storage-Intensive Cloud Applications," Proc. Second ACM Symp. Cloud Computing (SoCC"11), pp. 515-526, 2011.

[10]. K. Zhang, X. Zhou, Y. Chen, X. Wang, Y. Ruan, "Sedic: Privacy Aware Data intensive Computing on Hybrid Clouds," Proc. 18th ACM Conf. Computer and Comm. Security (CCS"11), pp 515-526, 2011.

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

[12]. B. Fung, K. Wang, L. Wang and P.C.K. Hung, "Privacy-Preserving Data Publishing for Cluster Analysis," Data Knowl. Eng., vol.68, no.6, pp. 552-575, 2009.

[13]. K. LeFevre, D.J. DeWitt and R. Ramakrishnan, "Workload- Aware Anonymization Techniques for Large-Scale Datasets," ACM Trans. Database Syst., vol. 33, no. 3, pp. 1-47, 2008.

[14]. Marten van Dijk, Ari Juels, "On the Impossibility of Cryptography Alone for Privacy-Preserving Cloud Computing ", Hot Sec. Tenth USENIX Conf. Hot topics in Security, Article No. 1-8, 2010.

[15]. Microsoft HealthVault, http:// www.microsoft.com/ health/ww/ products/Pages/ healthvault.aspx, July 2012.

[16]. Amazon Web Services, "Aws Service Pricing Overview," http://aws.amazon.com/pricing/, July 2012.

ACKNOWLEDGEMENT

I would like to thank **Prof. Amit M. Sahu**, my professor for his assistance and constant support. I greatly appreciate his valuable guidance. I like to thank all the website and paper which I have gone through and have refer to create my reaserch paper successful.

BIOGRAPHIES

Mr. Abhishek R. Ladole : Pursuing M.E (CSE) From G.H Raisoni College of Engineering And Management and completed his engineering From Rashtra Sant Tukdoji Maharaj Nagpur University, Amravati. His area of research is cloud security, and Network Security .

Prof. Amit M. Sahu: Received the B.E from SGBAU Amravati University and M.E. from SGBAU Amravati University. He is currently an Assistant Professor with the G.H Raisoni College of Engineering and Management, Amravati SGBAU Amravati University. His research interest includes, Network Security, Data mining and Cloud Computing .He has contributed to more than 10 research paper.