Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Approach for Privacy Preserving Cost of Intermediate Datasets in Cloud

[1]Ms.C.Lakshmi (M.Tech),
PG Scholar
Department of Computer Science& Engineering,
G. Pulla Reddy Engineering College,
Kurnool, Andhra Pradesh, India

[2]Dr.N.Kasiviswanath
Professor& Head of Dept
Department of Computer Science& Engineering,
G. Pulla Reddy Engineering College,
Kurnool, Andhra Pradesh, India

Abstract--- ***In cloud computing at the time of processing a data intensive application a large volume of intermediate datasets are generated, instead of regenerating these intermediate datasets, we can store these datasets in the cloud for future purpose. If any adversary should analyze these intermediate datasets they can get the information. In existing technologies to provide privacy for these intermediate datasets, we are encrypting all these datasets. But it is very time consuming and cost-effective, to reduce this privacy preserving cost we are proposed a heuristic method. In this method by encrypting only part of intermediate datasets, we are reducing privacy preserving cost and also satisfying the privacy requirements of data holders.***

Keywords--- cloud computing, intermediate datasets, privacy, Anonymization, encryption

I. INTRODUCTION

Cloud computing [1] should provide massive computation power and storage space for the users. The users can use these resources in pay as you go manner [2], instead of buying the required hard-disk or processors for their business. Because of this the business persons can reduce their investment cost and concentrate on their business development. Due to this so many users are very interested to use this cloud computing technology. But some of the users are very hesitant to store their data into the cloud according to security.so to provide security [3] for the data we are encrypting the entire data and allowing only authenticated users. At the time of executing any data intensive applications some intermediate datasets [4] or resultant data sets are generated, these are stored in the cloud for future purpose, instead of re-computing each and every whenever they need.

If any adversary should access these datasets then there is a chance of analyzing the information, so we need to provide privacy for these datasets. For providing security in the existing technologies we are encrypting all the intermediate datasets. But the computations are performed only on the readable data, so to perform any operations each and every time we need to decrypt the data set, perform the computation and then encrypt and store the dataset. For this purpose we need some extra storage space and also it is time consuming. There is a technology homomorphic encryption [5] by using theoretically proved not implemented practically.

For some data mining or analysis areas there is a need of revealing some aggregate information to the public. Publishing some data by satisfying the privacy requirements of data holders can be done by Anonymization [6]. Anonymization is one of the privacy technique like encryption. For a single dataset there is a privacy, but multiple datasets are not secure.so, in our proposed system to provide privacy for multiple datasets we are using both Anonymization and encryption technologies.

In the proposed system constructing a Sensitive Intermediate Dataset Tree(SIT) based on generation relationship among the intermediate datasets and finding privacy leakage for each and every intermediate dataset and then by using heuristic method we can identify which intermediate dataset we need to encrypt and find the minimum privacy preserving cost. Based on this we can prove that comparing with existing technologies our proposed system should reduce this privacy preserving cost.

II. CORRELATED EFFORT

There is a need of retaining the intermediate datasets in the cloud for that we need to provide privacy, so for that use the data hiding technique, privacy quantification and cost models. In this section we are initially to protect privacy for the datasets using encryption. It provides privacy but each and every time we need to decrypt for many applications. But it is very cost effective so it should be failed. Based on privacy preserving data publishing (PPDP) [7] finding sensitive data and converted into another form by Anonymization techniques like generalization and suppression approaches, but considering multiple datasets there is no privacy it should be failed.

So, in our approach we are providing privacy and also reducing the cost of intermediate datasets by encrypting only part of intermediate datasets.

III. PRIVACY PRESERVING COST of INTERMEDIATE DATASETS

In this section we are finding the effective privacy preserving cost of intermediate datasets in the cloud by using the SIT, privacy representation and construction of compressed tree, minimum privacy preserving cost and heuristic method as follows.

A. Sensitive Intermediate Dataset Management

Data provenance is used to manage the intermediate datasets, here provenance should contain the history from which the intermediate dataset should be generated. Which helps to generate a dataset from its nearest dataset rather than from original dataset. By using this data provenance we can manage the intermediate datasets based on generation relationship and construct a SIT.

Let us consider an original dataset as d_o and the intermediate datasets as $\{d_1, d_2, d_3 \ldots d_n\}$, where n represents the number of intermediate datasets generated. Based on generation relationship we can construct a SIT i.e. if d_5 contains all or part of fields in d_3 then we can say that d_5 is generated from d_3, so form an edge from d_3 to d_5.

1) Property on an SIT: If a dataset d_r is unencrypted then there is no need to check the datasets in parent dataset i.e. $PD(d_r)$.the sub tree with root dr unencrypted is called as Unencrypted Sub Tree (UST).

B. Privacy Representation

To protect the intermediate datasets using the Anonymization technique, it should protect only single intermediate dataset but multiple intermediate datasets should reveal the privacy sensitive information.

1) Single Intermediate Dataset Privacy Representation: The privacy leakage of a single intermediate dataset is $PL_s(d)$, i.e. an adversary should get $PL_s(d)$ after analyzing a dataset. Association between quasi identifiers and sensitive data is nothing but privacy information. Let us consider d_0 as original dataset and d* as anonymized intermediate dataset, QI is a set of quasi identifiers and SD is sensitive data.

Let Q be a variable in QI and S be a variable in SD. Assume that $s \epsilon S$ and $q \epsilon Q$, then an adversary should analyze the information as P(s=S,q=Q).the privacy quantification for a single dataset can be calculated by using a maximum entropy principle as follows.

$PL_s(d^*) = H(S, Q) - H^*(S,Q)$

Where H(S, Q) is the entropy value before an adversary observing an intermediate dataset. H(S, Q) can be calculated by using the equation as $H(S,Q) = \log(|QI|.|SD|)$.$H^*(S,Q)$ is the entropy value after an adversary should analyzing the intermediate dataset P*(S,Q) is defined as

$H^*(S,Q) = -\sum p(s, q). \log(p(s, q))$

2) *Privacy Representation for Multiple Datasets:* The privacy leakage for multiple datasets in $D=\{d_1,d_2,d_3,\ldots,d_n\}$, $n \epsilon N$ is defined by $PL_m(D)=H(S,Q) - H_D(S,Q)$.

Where H(S,Q) is entropy value before an adversary should observe the dataset and $H_D(S,Q)$ is after observing for datasets in D. H(S,Q) is calculated by $H(S,Q)=\log(|QI|.|SD|)$ and $H_D(S,Q)$ is calculated based on P(S,Q). But the maximum entropy value for more variables and constraints is very difficult, so we are considering the privacy leakage constraint as condition for D^{une} datasets and calculate the $PL_m(D^{une})$.The constraint can be taken by considering a threshold value ε and the privacy requirement is $PL_m(D^{une}) \leq \varepsilon$.

C. Construction of Compressed Tree

A compressed tree can be constructed by decomposing the privacy leakage constraint for each and every layer as shown in figure1. Let us consider L_i represents layers, where $1 \leq i \leq H$, H denotes height of the tree.

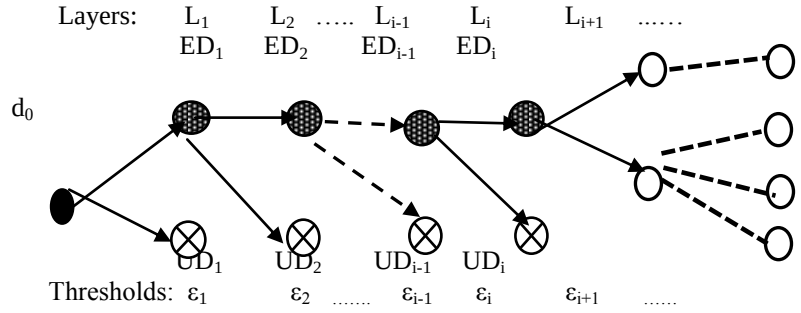

Fig. 1 Construction of compressed tree

Initially in an SIT at layer L_i each and every intermediate dataset privacy leakage value can be compared with the thresh hold value ε then the datasets less than threshold value can be un-encrypted and greater than threshold value are encrypted. The encrypted datasets are represented by ED_i and unencrypted datasets as UD_i and the constraint is PLm $(UD_i) \leq \varepsilon_i$. Each and every layer should consists a set of encrypted and unencrypted datasets as $\pi_i = \{ED_i, UD_i\}$. For the next lower layers the ε can be calculated by

$\varepsilon_i = \varepsilon_{i-1} - \sum PL_s(d)$, where $d \epsilon UD_{i-1}$

$\varepsilon_1 = \varepsilon$.

D. Minimum Privacy Preserving Cost

The cost per one GB of data storing in cloud can be considered as price PR. For every intermediate dataset there is a vector as $\{s_i, f_i, PL_i, flag\}$, i.e. size, frequency, privacy leakage value and a flag which indicates the dataset is encrypted or not.

The privacy preserving cost for an SIT with height H in a layer Li with a solution $\pi_i = \{ED_i, UD_i\}$ can be calculated for encrypted datasets only by

$C(\pi_i) = \sum S_k.PR.f_k$, where $d_k \in ED_i$, $1 \le i \le H$.

The minimum privacy preserving cost after the layer L_{i-1} threshold value ε_i can be calculated by using a recursive formula is as follows.

$CM_i(\varepsilon_i) = \min \{\sum_{dk \in EDi} Sk.PR.fk + CM_{i+1}(\varepsilon_i - \sum_{dk \in UDi} PL_s(d_k))\}$, $CM_{H+1}(\varepsilon_{H+1}) = 0$.

For large intermediate datasets to get the optimal solution we are using heuristic algorithm as described in the following section.

E. Heuristic Method

The heuristic algorithm can be used to find which intermediate dataset we need to encrypt and then for obtaining the optimal privacy preserving cost solution. Each and every layer having local solutions $\wedge^i = \{\pi_{ij}\}$ where i represents layer number and j indicates the number of local solutions in that particular layer L_i. The global encryption solutions $\pi^k = \{\pi_{1j1} \dots \pi_{HjH}\}$, where $\pi_{iji} \in$

$\wedge_i$, $1 \le i \le H$, $1 \le k \le \prod_{i=1}^{H} |\wedge i|$.

Heuristic value can be calculated for selecting a dataset with small cost and large privacy leakage value to encrypt. Heuristic function for a state node SN_i is $f(SN_i) = g(SN_i) + h(SN_i)$, where $g(SN_i)$ is the heuristic information from starting state to current state and $h(SN_i)$ is the estimated information from current state to the goal state. The $g(SN_i) = C_{cur} / (\varepsilon - \varepsilon_{i+1})$, where c_{cur} is the privacy preserving cost up to current state and ε is the threshold value and ε_{i+1} is the threshold value for the next layer. The $h(SN_i)$ is calculated as $h(SN_i) = (\varepsilon_{i+1}.C_{des}.BF_{AVG})/PL_{AVG}$, where C_{des} represents the total cost of the tree and BF_{AVG} is the branch factor of SIT, it can be computed by $BF_{AVG} = NE/NI$, where NE denotes the number of edges in the tree and NI represents the number of intermediate datasets in the tree. PL_{AVG} is the average of privacy leakage of all the intermediate datasets.

Based on the above description the heuristic value can be calculated for a search node is as follows.

$f(SN_i) = C_{cur} / (\varepsilon - \varepsilon_{i+1}) + (\varepsilon_{i+1}.C_{des}.BF_{AVG})/PL_{AVG}$.

The heuristic value can be used in the heuristic algorithm for selecting the dataset with highest heuristic value and then retrieving its child nodes finding which intermediate dataset we need to encrypt and adding the child nodes one by one to the priority queue up to the goal state. Finally the global privacy preserving solution and corresponding costs are derived.

1) Heuristic Algorithm for Reducing Privacy Preserving Cost:

Description: finding which intermediate dataset we need to encrypt and then getting minimum privacy preserving cost based on privacy leakage constraint.

Input: An SIT with root d_0, for every intermediate dataset frequency, size and privacy leakage value and threshold value as ε.

Output: optimal privacy preserving cost.

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: $SN_0 = ((\Pi_o) \leftarrow (\{d_o\}, \emptyset), f(SN_o) \leftarrow 0, ED_o \leftarrow \{d_o\}, c_{cur} \leftarrow 0, \varepsilon_1 \leftarrow \varepsilon)$.

1.3. Add the node into PQueue: PQueue $\leftarrow SN_o$.

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 algorithm will go to step 3.

2.3 label the data sets in CDE_i as encrypted if their privacy leakage is larger than ε_i

2.4 calculate the local privacy leakage and encryption cost and remaining privacy leakage cost.

2.5 calculate the heuristic value.

2.6 construct new search node from the calculated values and add it to priority queue. Then go to step 2.1

Step 3: obtain the global encryption cost.

IV. RESULTS

The comparison of privacy preserving cost for encrypting all the intermediate datasets in existing system and encrypting only part of intermediate datasets in our approach shows that we are reducing the privacy preserving cost by using our approach as shown in the figure2.

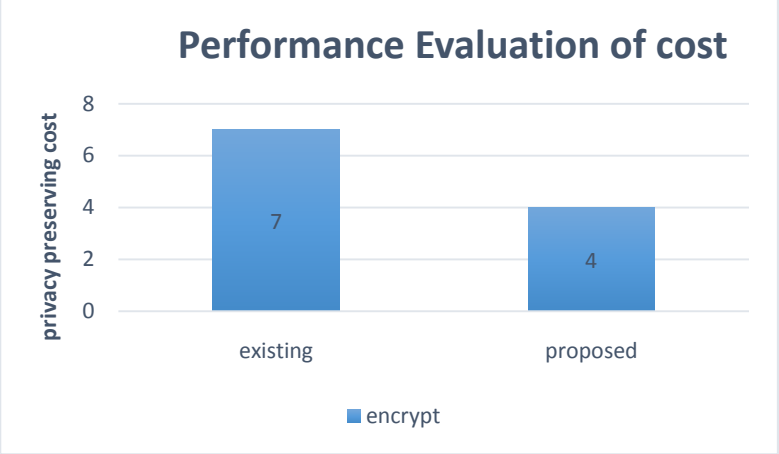

Fig. 2 Reducing the privacy preserving cost by our approach

In the figure2 the vertical axis represents the cost required to encrypt the datasets and the horizontal axis shows the two categories existing and proposed, the shaded bars in the graph shows the encryption cost required. By observing the existing and proposed encryption costs we can evaluate the performance of our approach.

By using our approach we can also prove that the time consuming is very less for encrypting only part of intermediate datasets compared with the existing approaches can be shown in the figure3.

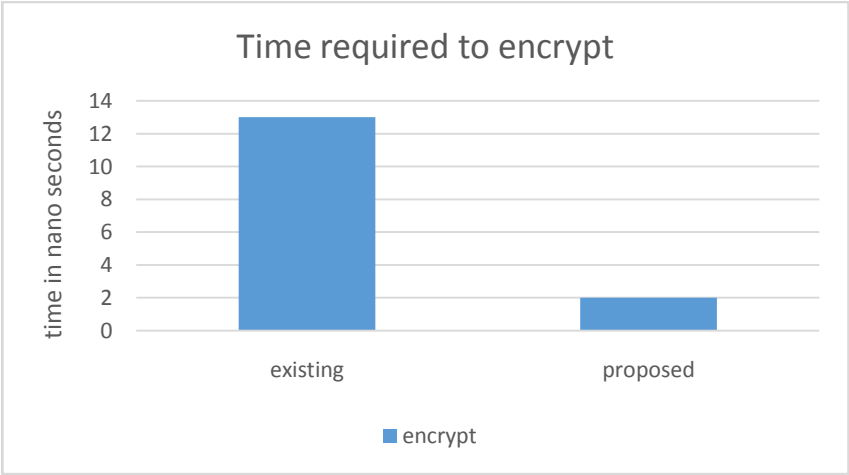

Fig. 3 Result of time comparison for existing and our approach

In the figure3 the vertical axis represents the time required to encrypt the dataset in nanoseconds and the horizontal axis has two categories existing and proposed and the shaded bars shows the encrypt time. We can easily analyze the time required to encrypt the datasets in the existing and our 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.

$$C_{SAV}=C_{ALL}-C_{HEU}$$

Here C_{SAV} is the privacy preserving cost saved, C_{ALL} is the privacy preserving cost for encrypting all the intermediate datasets and C_{HEU} is the privacy preserving cost for encrypting only part of intermediate datasets in the cloud. The resultant of our approach shows that the saving cost should be increases going on increasing the threshold value.

V. CONCLUSION

The privacy preserving cost of intermediate datasets in cloud can be reduced over existing approaches instead of encrypting all the intermediate datasets by encrypting only part of intermediate datasets in our approach by using SIT, compressed tree and heuristic algorithms.

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] L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?," IEEE Trans. Parallel and Distributed Systems, vol. 23, no.2, pp. 296-303, Feb.2012.

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

[4] D. Yuan, Y. Yang, X. Liu, and J. Chen, "On-Demand Minimum Cost Benchmarking for Intermediate Data Set Storage in Scientific Cloud Workflow Systems," J. Parallel Distributed Computing,vol. 71, no. 2, pp. 316-332, 2011.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] H. Lin and W. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding," IEEE Trans.Parallel and Distributed Systems, vol. 23, no. 6, pp. 995-1003, June2012.

[11] 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.

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

[13] 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.

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

[15] G. Wang, Z. Zutao, D. Wenliang, and T. Zhouxuan, "Inference Analysis in Privacy-Preserving Data Re-Publishing," Proc. Eighth IEEE Int'l Conf. Data Mining (ICDM '08), pp. 1079-1084, 2008.

[16] W. Du, Z. Teng, and Z. Zhu, "Privacy-Maxent: Integrating Background Knowledge in Privacy Quantification," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08), pp. 459-472, 2008.

