Building Confidential and Efficient Query Services in the Cloud

MS. DIPALI S. SHINTRE[1] & DR. S. M. JAGADE[2]

[1]Department of M.E.(C.S.E.), [2]Principal T.P.C.T.'s College Of Engineering, Osmanabad, India.

Abstract : **Cloud computing infrastructures are popularly used by peoples now a day. By using cloud users can save their cost for query services. But some of the data owners are hesitate to put their data's in cloud because, sometimes the data may be hack when they use in cloud unless the confidentiality of data and secure query processing will be provided by the cloud provider. In cloud if the user can get secured query service then the efficiency of query processing will be increased and the workload of the query processing will also be saved. To provide the confidentiality and efficient query service here we proposed RASP method. RASP denotes RAndom Space Perturbation. It also combines order preserving encryption, random projection and random noise injection. In order to process the range query to kNN query here we used kNN-R algorithm. We also analyze the RASP method will secure the multidimensional range and it will increase the working process of query.**

Keywords : cloud computing, query services, confidentialityy, kNN query

I. INTRODUCTION

Cloud computing is the internet based storage method. It is mainly used for storing the files and applications in it infrastructures. Peoples uses the cloud because of its attractive features like secure service, infinite of storage, it will satisfy the user experience, low cost and multiple user can access the files and applications. In cloud, the query service process are frequently used because, the user can save their cost. The owners in the cloud will pay the amount only for their using time of server. This is an important feature because, the working time of query service in cloud is very high and it is more expensive. New process are need for the cloud to protect the data and query privacy, so by that new process the query service can be protected.

We recently proposed the RAndom Space Perturbation (RASP) method [5] for the protection of tabular data, which is secure under the assumption of limited adversarial knowledge - only the perturbed data and the data distributions are known by adversaries.

This assumption is appropriate in the context of cloud computing. The RASP perturbation is a unique combination of Order Preserving Encryption (OPE) [1], dimensionality expansion, noise injection, and random projection, which provides sufficient protection for the privacy of query services in the cloud. It has a number of unique features, such as preserving the topology of range query, non-deterministic results for duplicate records, and resilience to distributional attacks.

We develop the secure half-space query transformation method that casts any enclosed range in the original space to an irregularly shaped range in the perturbed space. Therefore, we are able to use a two-stage range query processing method: an existing multidimensional index, such as R*- Tree in the perturbed space is used to find out the records in the bounding box of the irregularly shaped range, which is then filtered with the transformed query condition. This processing strategy is fast and secure under the security assumption.

To allow the readers to fully appreciate the intuition and the ideas behind the RASP based perturbation and query processing, we propose this RASP Query Services (RASP-QS) demonstration system. This system consists of the following major components: (1) the user interface for perturbation parameter generation that allows users to observe the details of RASP perturbation, (2) the visualization of the two stage range query processing procedure to understand the transformed

Manuscript received Dec, 2015.

Ms. Dipali S. Shintre*, Master of Computer Science & Engineering, T.P.C.T.'s College of Engineering Osmanabad, India.*

Dr. S. M. Jagade*, Ph.D(E & TC), Principal, T.P.C.T.'s College of Engineering Osmanabad, India.*

query ranges and the query results, (3) the visualization of the progressive steps in the kNN query processing that is based on RASP range query processing, and (4) the performance comparison on index-aided processing on non data, linear-scan query processing on encrypted data [2], and the RASP query processing.

II. RELATED WORK

Order preserving encryption : Order preserving encryption technique, encryption is a well-known technology for protecting sensitive data Integration of encryption method with database system cause performance reduction example if a column of contain sensitive information is encrypted, and is used in query predicate with a comparison operator an entire table scan would be needed to evaluate the query. Reason is that current encryption techniques do not preserve order and there data base indices such as B-tree cannot be used. Order preserving encryption allows comparison operation to be applied directly on encrypted data without decrypting the operands MAX, MIN, and COUNT queries can be directly processed over encrypted data. “Groupby” and “order by” operations can also be applied. SUM, AVG and “group by” values need to be decrypted. A value in the column can be modified or a new value can be inserted in a column without requiring changes in the encryption of other values. OPES can easily be integrated with the existing database system as it has been designed to work with the existing indexing structure such as B-trees.

Cryptoindex : Cryptoindex is also based on column-wise bucketization. It assigns a random ID to each bucket; the values in the bucket are replaced with the bucket ID to generate the auxiliary data for indexing. To utilize the index for query processing, a normal range query condition has to be transformed to a set-based query on the bucket IDs. Crypto index method is vulnerable to attacks but the working system of the crypto index has many difficult processes to provide the secured encryption and security and also the New Casper approach is used to protect data and query but the efficiency of the query process will be affect.

For example, $X_i < a_i$ might be replaced with $X'i \in [ID1, ID2, ID3]$.

OPE and crypto-index assumes the attacker knows only the ciphertext. However, If the attacker has some prior knowledge, such as the attribute domains (maximum and minimum values), the attribute distributions, and even a few pairs of plaintext and ciphertext, these encryption methods will be vulnerable to attacks.

Private Information Retrieval : There is a significant risk to the privacy of the user, since a curious database operator can follow the user's queries and infer what the user is after. One thing a user can do to preserve his privacy is to ask for a copy of the whole database (B. Chor *et al.*)[4]. Same database is replicated at several sites, viewing the database as binary string $x=x_1,x_2\ldots\ldots\ldots x_n$ of length n. identical copiesof strings are stored by $k\geq 2$ servers.The user has some index i, and he is interested in obtaining the value of bit x_i. To achieve this goal, the user queries each of the server and gets replies from which the direct bit x_i can be computed. Drawback is cost increases because of creating more than one copy of database.

III. PROBLEM FORMULATION

A. Range Query Processing with RASP

Based on the RASP encryption we proposed, we target to provide two kinds of query services: range query and k nearest neighbors.

Range query is an important type of query for many data analytic tasks from simple aggregation to more sophisticated machine learning tasks. Let T be a table and X_i, X_j, and X_k be the real valued attributes in T, and a and b be some constants. Take the counting query for example. A typical range query looks like

select count (*) from T
where $X_i \in [a_i,b_i]$ and $X_j \in (a_j,b_j)$ and $X_k=a_k$

which calculates the number of records in the range defined by conditions on X_i, X_j, and X_k. Range queries may be applied to arbitrary number of attributes and conditions on these attributes combined with conditional operators “and”/“or.” We call each part of the query condition that involves only one attribute as a simple condition. A simple condition like $X_i \in [a_i,b_i]$ can be described with two half space conditions $X_i \leq b_i$ and $-X_i \leq -a_i$. Without loss of generality, we will discuss how to process half-space conditions like $X_i \leq b_i$ in this paper. A slight modification will extend the discussed algorithms to handle other conditions like $X_i < b_i$ and $X_i = b_i$.

kNN query is to find the closest k records to the query point, where the euclidean distance is often used to measure the proximity. It is frequently used in locationbased services for searching the objects close to a query point, and also in machine learning algorithms such as hierarchical clustering and kNN classifier. A kNN query consists of the query point and the number of nearest neighbors, k.

ISSN: 2278 – 7798

B. System Framework

The purpose of this architecture is to extend the proprietary database servers to the public cloud, or use a hybrid private public cloud to achieve scalability and reduce costs while maintaining confidentiality.

The trusted parties and the untrusted parties. The trusted parties include the data/service owner, the in-house proxy server, and the authorized users who can only submit queries. The data owner exports the perturbed data to the cloud. Meanwhile, the authorized users can submit range queries or kNN queries to learn statistics or find some records. The untrusted parties include the curious cloud provider who hosts the query services and the protected database.

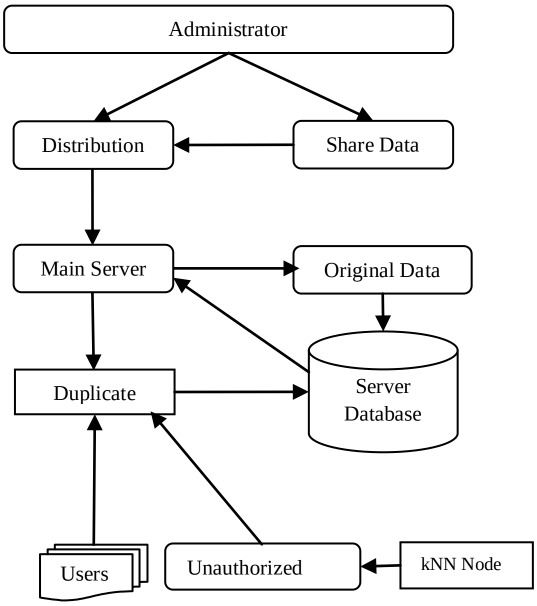

Fig. 1 RASP System Architecture

The RASP-perturbed data will be used to build indices to support query processing.

C. Security Analysis

The security analysis in the architecture shows the following

- Users have been authorized by using the key value provided by the owner. So an authorized user is not being a malicious and only those users can send the queries for retrieving the data.
- The communication process between the user, owner and cloud and client system are well secured, the data and queries cannot be leaked from the cloud.
- RASP method is used to protect the query privacy and confidentiality of the data.

Attacker Process: The main process of attacker is to hack the data from the database and they will try to find the perturbed data and they will try to find the queries.

D. Modules

Three modules are used. They are RASP, range query and kNN query.

RASP :

RASP denotes Random Space Perturbation. It also combines OPE, random projection and random noise injection. Here OPE denotes Order Preserving Encryption is used for data that allows any comparison and that comparison will be applied for the encrypted data; this will be done without decryption. Random projection is mainly used to process the high dimensional data into low dimensional data representations. It contains features like good scaling potential and good performances.

Random noise injection is mainly used to adding noise to the input to get proper output when we compare it to the estimated power. The RASP method and its combination provide confidentiality of data and this approach is mainly used to protect the multidimensional range of queries in secure manner and also with indexing and efficient query processing will be done. RASP has some important features. In RASP the use of matrix multiplication does not protect the dimensional values so no need to suffer from the distribution based attack.

RASP prevents the data that are perturbed from distance based attacks; it does not protect the distances that are occurred between the records. And also it won’t protect more difficult structures it may be a matrix and other components. The range queries can be send to the RASP perturbed data and this range query describes open bounds in the multidimensional space.

In random space perturbation, the word perturbation is used to do collapsing this process will happen according to the key value that is given by the owner. In this module the data owner have to register as owner and have to give owner name and key value. And then the user have register and get the key value and data owner name from the owner to do access in the cloud. Here user can submit their query as range query or kNN query and get their answer. We analyze and show the result with encrypted and also in decrypted format of the data for the query construct by the user.

Range Query :

Range query is an important type of query for many data analytic tasks from simple aggregation to more sophisticated machine learning tasks.

The range search is mainly used to return the values that are present between the two specified values given in the query. For example database name is AAAworkers2012 then

```
Go
          SELECT product id
          FROM AAAworkers2012.production
          WHERE price BETWEEN 40 and 60
```

The above example will show an another example of range query search it will provide the entries of what are product id that are present in production database with price above 40 and within 60. So by using range query user can easily retrieve the data's from records and this query process will be done in secure manner and the speed of the query process will also increased.

kNN Query :

kNN query represents k-Nearest Neighbor query. This query is mainly used to retrieve the nearest neighbor values of k. here k used to denote positive integer value. kNN algorithm is mainly used for classification and regression. In this it uses kNN-R algorithm to process the range query to kNN query. This algorithm consists of two methods. That is used to make interaction between the client and the server. The client will send the query to the server with initial upper bound and lower bound. This upper bound range has to be more than the k points and the lower bound range have to be less than the k points.

IV. EXPERIMENTAL RESULTS

In this experiment, we study the costs of the components in the RASP perturbation. The major costs can be divided into two parts: the OPE and the rest part of RASP. We implement a simple OPE scheme [1] by mapping original column distributions to normal distributions. The OPE algorithm partitions the target distribution into buckets.

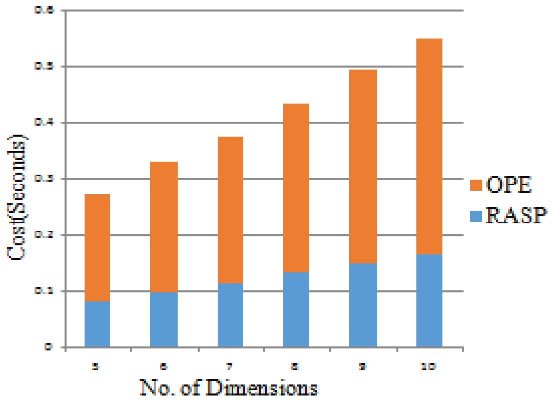

Fig. 2 The cost distribution of the full RASP scheme.

Then, the sorted original values are proportionally partitioned according to the target bucket distribution to create the buckets for the original distribution. With the aligned original and target buckets, an original value can be mapped to the target bucket and appropriately scaled. Therefore, the encryption cost mainly comes from the bucket search procedure (proportional to logD, where D is the number of buckets).

Fig. 2 shows the cost distributions for 10K records at different number of dimensions. The dimensionality has slight effects on the cost of RASP perturbation.

V. CONCLUSION

We proposed RASP method with range query and kNN query. This method mainly used to perturb the data given by the owner and saved in cloud storage it also combines random injection, order preserving encryption and random noise projection and also it has contains CPEL criteria in it. By using the range query and kNN query user can retrieve their data's in secured manner and the processing time of the query is minimized. And also we continue our studies to improve the effect of query.

VI. ACKNOWLEDGMENT

I would like to express my deep gratitude to my research, Dr. S. M. Jagade has a research guide, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to extend my thanks to the technicians of the laboratory of the Computer department for their help in offering me the resources in running the program.

Finally, I wish to thank my friends and parents for their support and encouragement throughout my project tenure.

REFERENCES

[1] Xu, H., Guo, S., and Chen, K. "Building confidential and efficient query services in the cloud with RASP data perturbation", IEEE Transactions on Knowledge and Data Engineering 26, 2 (2014).

[2] K. Chen, R. Kavuluru, and S. Guo, "RASP: Efficient Multidimensional Range Query on Attack-Resilient Encrypted Databases," Proc. ACM Conf. Data and Application Security and Privacy, pp. 249-260, 2011.

[3] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order Preserving Encryption for Numeric Data,"

Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2004.

[4] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.K. Andy Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," technical report, Univ. of Berkerley, 2009.

[5] J. Bau and J.C. Mitchell, "Security Modeling and Analysis," IEEE Security and Privacy, vol. 9, no. 3, pp. 18-25, May/June 2011.

[6] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOMM, 2011.

[7] K. Chen and L. Liu, "Geometric Data Perturbation for Outsourced Data Mining," Knowledge and Information Systems, vol. 29, pp. 657- 695, 2011.

[8] K. Chen, L. Liu, and G. Sun, "Towards Attack-Resilient Geometric Data Perturbation," Proc. SIAM Int'l Conf. Data Mining, 2007.

[9] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private Information Retrieval," ACM Computer Survey, vol. 45, no. 6, pp. 965-981, 1998.

[10] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions," Proc. 13th ACM Conf. Computer and Comm. Security, pp. 79-88, 2006.

[11] R. Marimont and M. Shapiro, "Nearest Neighbour Searches and the Curse of Dimensionality," J. Inst. of Math. and Its Applications, vol. 24, pp. 59-70, 1979.

[12] H. Hacigumus, B. Iyer, C. Li, and S. Mehrotra, "Executing SQL over Encrypted Data in the Database-Service-Provider Model," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2002.

[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer-Verlag, 2001.

[14] B. Hore, S. Mehrotra, and G. Tsudik, "A Privacy-Preserving Index for Range Queries," Proc. Very Large Databases Conf. (VLDB), 2004.

[15] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2005.

[16] A. Hyvarinen, J. Karhunen, and E. Oja, Independent Component Analysis. Wiley, 2001.

[17] I.T. Jolliffe, Principal Component Analysis. Springer, 1986.

[18] F. Li, M. Hadjieleftheriou, G. Kollios, and L. Reyzin, "Dynamic Authenticated Index Structures for Outsourced Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.

[19] K. Liu, C. Giannella, and H. Kargupta, "An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining," Proc. 10th European Conf. Principle and Practice of Knowledge Discovery in Databases (PKDD), 2006.

[20] M.L. Liu, G. Ghinita, C.S. Jensen, and P. Kalnis, "Enabling Search Services on Outsourced Private Spatial Data," The Int'l J. Very Large Data Base, vol. 19, no. 3, pp. 363-384, 2010.

[21] Y. Manolopoulos, A. Nanopoulos, A. Papadopoulos, and Y. Theodoridis, R-Trees: Theory and Applications. Springer-Verlag, 2005.

[22] R. Marimont and M. Shapiro, "Nearest Neighbour Searches and the Curse of Dimensionality," J. Inst. of Math. and Its Applications,vol. 24, pp. 59-70, 1979.

AUTHOR PROFILE

Ms. Dipali Sambhaji Shintre, is a M.E student of Computer Science & Engineering from T.P.C.T.'s College of Engineering, Osmanabad India. She graduated in Computer Science and Engineering from BAMU University, Aurangabad. Her current research work focuses on "Building Confidential and Efficient Query Services in the Cloud".

Dr. S. M. Jagade received Ph.D. in (Electronics & Telecommunication) from SGGS IE & T Research center Nanded, ME (Ec) specialization in Computer Science from SGGS, College of Engineering & Technology, Nanded. completed BE (Electronics and Telecommunication) Degree from Govt. Engineering College, Aurangabad. He is currently working as a Principal in T.P.C.T.'s College of Engineering, Osmanabad, India.