INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 3)

Available online at: www.ijariit.com

Construction of private methodical query services in the cloud with RASP data commotion

Shubhashree Sahoo
shubhashree1451990@gmail.com
Teegala Krishna Reddy Engineering College, Hyderabad, Telangana

Gogu Swathi
goguswathi@gmail.com
Teegala Krishna Reddy Engineering College, Hyderabad, Telangana

ABSTRACT

As Digital technology is fast evolving and becoming an essential tool for businesses, the concept of Cloud is evolved. The Phenomenon of the cloud is described in terms of private and public. The proposed approach is based on the public cloud domain, which consists, numerous nodes with distributed computing resources in many different geographic locations. This approach leads the public cloud domain into several cloud partitions. The approach of distributed computing in the cloud simplifies the load balancing and allows database indexes to build over an encryption table. Many times, data into the cloud is stored by maintaining confidentiality, query privacy, efficient query processing at low cost (CPEL Criteria). However, the data owners always desire to submit their quires after realizing the privacy assurance of the cloud. In this aspect, researchers have introduced few techniques such as RASP (Random Space Perturbation), k-NN (k-Nearest Neighbor) Algorithm etc. The main problem across RASP technique is, generating the encryption key which is too large and its implementation makes the time and space overhead. The existing RASP data perturbation technique along with k-NN algorithm is exploited to furnish privacy to the cloud. Wherein, issues such as categorical data and leaked query in the model are identified and addressed, by holding no change in designing the k-NN-R algorithm.

Keywords: *Algorithm, k Nearest Neighbor distance, Privacy, Confidentiality, and Range Query.*

1. INTRODUCTION

With the wide deployment of public cloud computing infrastructures, using clouds to host data query services has become an attractive solution for the advantages on scalability and cost-saving. [XX] The service owners can conveniently scale up or down the service and only pay for the useable hours with the help of cloud infrastructures. [1]

The data confidentiality and query privacy could be preserved by using the efficiency of query services and the benefits of clouds. The aim of using cloud resource is to maintain the CPEL criteria i.e. data confidentiality, query privacy and efficient query processing at low cost. The complexity of constructing query services in the cloud dramatically increases to fulfill the above requirements. However, they do not address all of these aspects satisfactorily. For example, the crypto-index and order-preserving encryption (OPE) are vulnerable to the attacks. The enhanced crypto-index approach puts heavy burden on the in-house infrastructure to improve the security and privacy. Hence the approach of random space perturbation (RASP) was proposed to construct practical range query and k-nearest- neighbor (k-NN) query services in the cloud. Also, the approach addresses all the four aspects of the CPEL criteria and aim to achieve a commendable balance on them. The k-NN-R query service uses the RASP range query to process k-NN queries. [4]

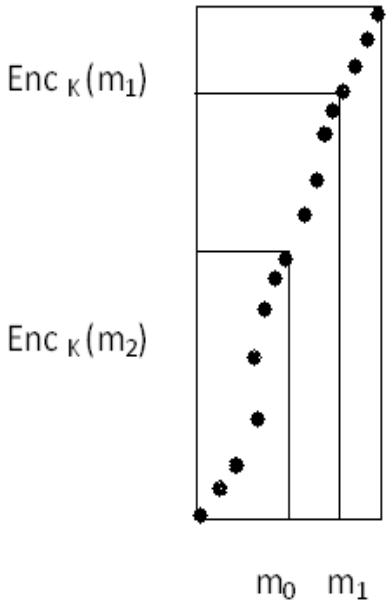
The RASP perturbation technique is a unique combination of OPE (Order Preserving Encryption), dimensionality expansion, random noise injection, and random projection, wherein the aim is to provide strong confidentiality guarantee. Authors have carefully evaluated the current approach with synthetic and real data sets whose results maintain the CPEL criteria. The RASP method provides confidentiality of data and this approach is mainly used to protect the multidimensional range of queries in secure manner, with indexing and efficient query processing. The range query is used in database for retrieving the stored data's from the database where it can denote some value between upper and lower boundary. The k-NN query denotes k-Nearest Neighbor query where K is the positive integer and this query are used to find the value of nearest neighbor to k. [5] The RASP perturbation embeds the multidimensional data into a secret higher dimensional space, enhanced with random noise addition to protect the confidentiality of data.

2. RELATED WORK

A public cloud refers to a computing service model used for the provisioning of storage and computational services to the general public over the Internet. Hamming code distance is very much effective across categorical data within RASP process. [7], because Hamming Code distance is mainly used to detect and correct bit(short for binary digit, which is a smallest unit of data in a computer ex. Binary digit 0 or 1)errors that can occur when computer data is moved or stored. In another way, it measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other. The performance of query processing is examined by enhancing the capacity of threat model. However, the effect of data and query confidentiality were analysed using leaked query. Authors addressed few related methods such as OPE, Crypto-index, DRE, and PIR.

Order Preserving Encryption:

The order-preserving encryption (OPE) preserves the dimensional value order after encryption. Thus, it can be used in most database operations, such as indexing and range query. The drawback of this process is too large encryption key and its implementation leads to time and space overhead. [8] A symmetric encryption scheme is order preserving, only when the encryption is deterministic and strictly increasing. For $K \leftarrow Key\ Generetion$;



Where, $\mathbf{m_0}$, $\mathbf{m_1}$ is the Plane text and

Enc $_{K\,(m1)}$, **Enc** $_{K\,(m2)}$ is the cipher text.

Crypto-index: Crypto-index is based on column-wise bucket-load. 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 [9]. For Example let database DB with a set of tables T having a number of records R consisting of several sets of attributes A. Then it can be classified as **A={a_1,a_2,…, a_n}** be the set of attributes and clustered into a set of Buckets **B= {B_1, B2,…,B_n}**. Each bucket can be replaced with some ID values like (change the equation with **B'$_n$** Ɛ [ID_1, ID_2.…IDn**]**. [10].

If the attacker manages to know the mapping between input original query and output bucket-based query, the range that a bucket ID represents could be estimated. A bucket-diffusion scheme was proposed to address the problem that sacrifices the precision of query results. Another drawback of this method is that the client, not the server, has to filter out the query result. Low precision results raise the large burden on the network and the client system. Furthermore, due to the randomized bucket IDs, the index built on bucket IDs is not so efficient for processing range queries as the index on OPE encrypted data is Distance-recoverable encryption.

Distance-recoverable encryption (DRE):

This approach DRE [10] is the most intuitive method for preserving the nearest neighbor relationship as the exactly preserved distances, used in the approach causes the scope for many attacks. Here, dot products are used instead of distances to find k-NN, which is more resilient to distance targeted attacks. However, limiting the search algorithm to linear scan and also the non existence of indexing method are *left un-addressed.*

Private information retrieval (PIR): PIR tries to fully preserve the privacy of access pattern, while the data may not be encrypted. PIR schemes are normally very costly. This privacy preserving multi keyword search is based on the plain text search. This can be carried out by using the ranking process. The drawback of this concept is ranking process that maximizes the in-house processing time. The research on privacy-preserving data mining has multiplicative perturbation methods, which are similar to the RASP encryption. [10]

3. QUERY SERVICE

Queries are formulated by using structured query language. It is mainly used to fetch the essential information from the database. Query services are the methods which are exposed through an implementation of the service provider. **RASP**, **range query,** and the **k-NN** query are used in Cloud to provide fast storing, retrieving process of encryption and decryption of a data from the database.

3.1 SYSTEM ARCHITECTURE

The system architecture of RASP method is shown in figure1, given below. Basically, the Cloud computing infrastructures store large datasets and query services. The architecture represents two main parts such as data owner and accredited user, where the data owner can store the data in the cloud.

Data d = n(d, k)

Where, **d** represents data, an **n**-normal form of data and k-key value given by the data owner.

The data will be saved into the cloud as encrypted form i.e. $\mathbf{D^1}$ **= e(D, k),** where **e, D,** and $\mathbf{D^1}$ denotes encryption, decrypted data and encrypted data respectively. When user requests for retrieving the data then decryption H is performed with encrypted data $\mathbf{D^1}$ and key value k.

D = H(D^1, k);

All metadata stored in the cloud database is in perturbative form. Hence, any application of the legitimate client can transparently issue SQL operations (e.g., **SELECT, INSERT, UPDATE** and **DELETE**) to the cloud database in perturbative form. Data which is transferred between the user application and the index engines are in plain format. Information is always in encrypted form before sending to the cloud database. When an application issues a new SQL operation, the perturbated database interface contacts the perturbated engine to retrieves the metadata. After obtaining the metadata, the encryption engine is able to execute the SQL operation on encrypted data, and then decrypt the results. The results are returned to the user application through the encrypted database interface. Also, in order to improve its performance, the plain metadata are cached locally by the client as volatile information.

Subsequently, the accredited user who is the trusted party stores their data in the cloud.

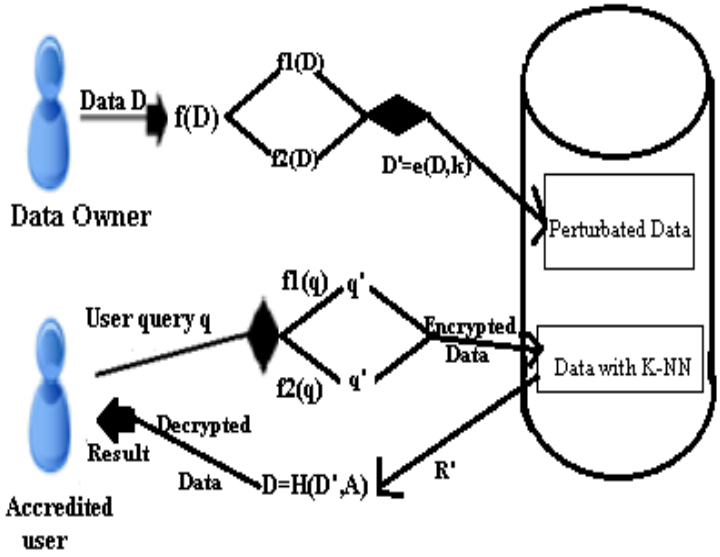


Fig. 1 System Architecture using RASP method

The owners can store their data in the cloud, where the data will be encrypted in the cloud and stored in the cloud database. The data owner could provide the key value, which will encrypt the data by using random space perturbation method. The user will send a query to retrieve the data from the cloud. The user can also send range query and k-NN query to get the data. The cloud providers have to host the user query services and protect the data stored in the cloud database.

The basic procedure involves following steps: (1) The owner sends the data to store in the cloud. The data could be encrypted using random space perturbation (RASP) method and stored in cloud database based on datatype (VARCHAR OR INT).
(2) The user will send the range query or k-NN query to retrieve the data. The query could be encrypted and send to cloud storage.
(3) The cloud storage will send the data for that query after processing the query inside cloud storage. Subsequently, the data could be decrypted and finally send to the user. [11], [6].

3.2 SECURITY ANALYSIS

The security analysis in the architecture shows the following;

i) Users are authorized by using the key value provided by the owner. An authorized user is not being malicious can send the queries for retrieving the data.
ii) The communication process between the user, owner, cloud and client system are well secured. The data and queries cannot be leaked from the cloud.
iii) RASP method is used to protect the query privacy and confidentiality of the data.
Attacker Process: The main process of the attacker is to hack the data from the database. They will try to find the perturbed data and queries. [13]

4. MODULES

Three modules are used. They are RASP, range query and k-NN query.

4.1 RASP

RASP denotes Random Space Perturbation. [1, 2] It also combines OPE, random projection and random noise injection. Here OPE denotes Order Preserving Encryption is used for data comparison. The comparison will be applied to the encrypted data which could be carried out without decryption. Random projection is a suitable approximation technique for distance-based method. The random projection matrices are controlled. Therefore, it is used to preserve the pair wise distance between any two samples of the dataset. Random noise injection is mainly used for 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 a secure manner. Additionally, indexing and efficient query processing will be carried out. In RASP the use of matrix multiplication does not protect the dimensional values. Therefore, the owners no need to suffer from the distribution based attack.

4.2 RANGE QUERY

A range query is the query used to retrieve the data from the database. It will retrieve the data value between the upper bound and lower bound. The range query is not usual because user doesn't know in advance about the result for the query and how much entries come as result for the query. [6] For example Employee_table is the real table which contains all the record
about employee where I need to retrieve the data by using range query which is given below.

	empid	fname	lname	sex	ssn	salary	deptno
1	501	JOHN	DOE	M	500000001	30000	4001
2	502	JOHN	SMITH	M	500000002	40000	4001
3	503	SEAN	LEE	M	500000003	30000	4001
4	504	EVAN	SEAN	M	500000004	50000	4002
5	505	REBECCA	SEAN	F	500000005	30000	4002
6	506	TIM	DUNCAN	M	500000006	30000	4002
7	507	ROBERT	DUVAL	M	500000007	30000	4002
8	508	CLINT	JOHNSON	M	500000008	30000	4002
9	509	SARRAH	MCMILLAN	F	500000009	60000	4003
10	510	DAVID	LIMB	M	500000010	30000	4003
11	511	DAVID	BOWE	M	500000011	30000	4003
12	512	SMITH	CLARK	M	500000012	50000	4003
13	513	TED	KENNEDY	M		30000	4003
14	514	RONALD	REAGAN	M	500000014	30000	4003
15	515	FRANKL...	ROSSEV...	M	500000015	30000	4003
16	516	George	BUSH	M	500000016	30000	4004
17	517	SAM	MALONE	M		30000	4004
18	518	NANCY	REAGAN	F	500000018	30000	4004
19	519	HILLARY	CLINTON	F	500000019	30000	4004
20	520	MARRY	GEORGIA	F	500000020	30000	4004

SELECT TOP 3 * FROM table_name where salary>30000

	empid	fname	lname	sex	ssn	salary	deptno
1	502	JOHN	SMITH	M	500000002	40000	4001
2	504	EVAN	SEAN	M	500000004	50000	4002
3	509	SARRAH	MCMILLAN	F	500000009	60000	4003

The above example shows the sample query for range query. The above example query is to retrieve the entries from the table. It will retrieve the employee whose salary is above 30000 in the top 3 list from the record of Employee_table. The range search is mainly used to return the values that are present between two specified values given in the query. For example database name is

SELECT empid (column_name) FROM Empdb1 (table_name) WHERE salary (column_name) BETWEEN 35000 and 50000(value1 and value2);

	empid
1	502
2	504
3	512

The above example will show another example of range query search. It will provide the entries of ids that are present in employee database with a salary above 35000 and within 50000. Therefore, the user can easily retrieve the data are from records by using range query. This query process will be carried out in a secure manner and the speed of the query process also increases.

4.3 k-NN QUERY

k-NN query represents the k-Nearest Neighbor query. This query is mainly used to retrieve the nearest neighbor values of k, where k denotes positive integer value. the k-NN algorithm is mainly used for classification and regression. It uses k-NN-R algorithm to process the range query to k-NN query. This algorithm consists of two methods and used to make the interaction between the client and server. The client will send the query to the server with initial upper bound and lower bound. [13]

This upper bound range has to be more than the k points and the lower bound range has to be less than the k points. The above process is used to give the inner range of database by the server. The client will calculate outer range with the inner range and send this outer range to the server. The server will search and find the records in the outer range from the database and send it to the client. Subsequently, the client will decrypt the record and find the top k files to provide the final result. This algorithm is used to find the compact inner square range for providing high precision values. It has two difficult processes such as to find the number of points that are present in the square range and updating of the boundary (i.e.) upper bound and lower bound. This is because the range queries are well secured by using random space perturbation. The security of k-NN query and range query is equal. [7]

5. STUDY OF IMPLEMENTATION

Problem 1: Using Manhattan (Hamming code) distance to employ across categorical data within RASP process**.**
Solution: n Categorical Data, a set of data is sorted or divided into different categories, according to the attributes of the data.
Example: Number of users {set} є Number of data

When there is a mixture of numerical and categorical variables present in the data set, Hamming Code Distance is used for standardization.

Manhattan\Hamming Code Distance

$$D_H = \sum_{i=1}^{K} |X_i - Y_i| \qquad x = y \Rightarrow D = 0 \qquad x \neq y \Rightarrow D = 1$$

If the data have a much higher influence on the distance calculated. For example, if one variable is based on annual income in dollars and other is based on age in years then income will have a much higher influence on the distance calculated, then distance can be calculated by using the Equation,

$$X_Z = \frac{X - Min}{Max - Min}$$

We describe the details of the adaptive layer mechanism by referring to an example.

Problem 2

The performance of query processing will be changed by enhancing the capacity of a threat model and increasing the query confidentiality by using leaked query.

Solution:

Let us consider a table T with columns id of type int, the name of type string and a tenant client preparing to issue the following statement to the encrypted cloud database: SELECT * FROM T WHERE id < 10. The client perturbation engine analyzes the SQL statement and identifies that the operation id < 10 has to be perturbated on the encrypted

database. Then, the client reads the metadata and checks whether there is INT attribute associated with the column id because this is the only INT data type.

For INT data type

```
Input: Select Data D;
        Output: data D1;
        Let String S=null;
        S=D;
    foreach ( char c in S)
        c → byte b;
        temp = b;
       temp = temp - rand ();
   b=temp + b;
        endfor;
    D1=rand(c) + b;
  return D1;
```

For varchar data type

```
Input: Select Data D;
      Output: data D1;
      Let b=null;
    temp = Timestamp (Date and Time);
    b=temp;
D1=rand (char) + b;
return D1;
```

Threat model: It is a process for optimizing network/application harmed the internet security by identifying objective and vulnerability. A threat is a potential or actual undesirable event that may be malicious such as DOS attack or incidental failure of a storage device. The threat model is a panel activity for identifying and assessing application threats and vulnerabilities.

Leaked Query: It reveals that the application is vulnerable. Finally, the accidental leaking of sensitive information through data queries occurs.

6. EXPERIMENTAL SETUP

In the current research, work authors explore various perturbation techniques by using Manhattan\Hamming code distance, which is advantageous for k-NN. In details, it provides extensive coverage of perturbation techniques over existing technology. Nevertheless, authors do highlight some general trends with regard to search effectiveness for the categorical and leaked data.

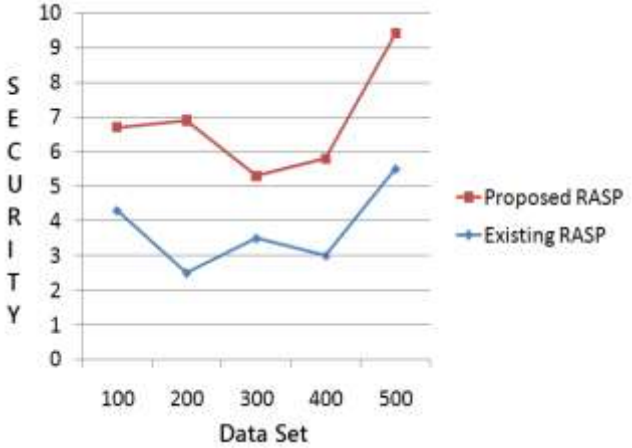


Fig: 2 Performance Graph

Fig. 2 shows security vs data set graph with two perturbation techniques. The graph is comparable to the number of datasets by an existing and proposed RASP. This metric has been used in a number of previous evaluations. The figure reveals two interesting trends. First, the security levels of existing **RASP** and proposed **RASP** with two level of security.

7. CONCLUSIONS

A RASP method is proposed with range query and k-NN query. This method mainly used to perturb the data given by the owner and saves in cloud storage. It also combines random injection, order-preserving encryption and random noise projection and contains CPEL criteria. The user can retrieve their data's in a secured manner using the range query and k-NN query. The processing time of the query is minimized. Our studies will also continue to improve the effect of a query.

8. REFERENCES

[1] 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 5th utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599–616, 2009.

[2]. H.-L. Truong and S. Dustdar, "Composable cost estimation and monitoring for computational applications in cloud computing environments," Procedia Computer Science, vol. 1, no. 1, pp. 2175 – 2184, 2010, iCCS 2010.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. K. andAndy Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A Berkeley view of cloud computing," Technical Report, University of Berkeley, 2009.

[4]https://www.google.co.in/?gfe_rd=cr&ei=Qf0wVdnhG9iDuASKsIDwBw&gws_rd=ssl#q=abstract+about+the+building+of+confidential+data+stored+in+clouds+with+RASP+data+perturbation

[5]https://www.google.co.in/?gfe_rd=cr&ei=WQMxVd-LFNOGoAOzp4HgAQ&gws_rd=ssl#q=cpel+criteria+followed+by+cloud

[6]http://www.ijreat.org/Papers%202014/Issue7/IJREATV2I1021.pdf

[7]https://www.youtube.com/results?search_query=explanation+about+the+working+principle+of+k-NN+k-NN

[8]https://www.google.co.in/?gfe_rd=cr&ei=oQ0xVeb8G4aBoAPR_YDYAQ&gws_rd=ssl#q=explanation+about+Manhattan+(Hamming+code)+distance

[9] A. Boldyreva, N. Chenette, and A. O'Neill, "Order-preserving encryption revisited: Improved security analysis and alternative solutions," in Proc. Advances in Cryptology – CRYPTO 2011. Springer, Aug. 2011.

[10] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in Proc. Advances in Cryptology – EUROCRYPT99. Springer, May 1999.

[11] http://arxiv.org/pdf/1212.0610.pdf

[12]https://www.google.co.in/?gfe_rd=cr&ei=ahYxVfHoFojFoAPF4YDIDQ&gws_rd=ssl#q=meaning+of+cpel+criteria

[13]http://www.researchgate.net/publication/233824247_Building_Confidential_and_Efficient_Query_Services_in_the_Cloud_withRASP_Data_Perturbation.

