ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Preserving Data Confidentiality and Query Privacy Using KNN-R Approach

Shruthi.K, Aruna Reddy.H, Dr K.N.Narasimha Murthy

PG Student, Dept of CSE, Vemana Institute of Technology, Bangalore, India

Assistant Professor, Dept of CSE, Vemana Institute of Technology, Bangalore, India

Professor & HOD, Vemana Institute of Technology, Bangalore, India

ABSTRACT: Cloud computing is one of the famous and well known technique that processes the data query efficiently. Since it is maintaining huge amount of resources, its privacy and security is an issue. Cloud service providers are not trust worthy, so data is to be secured. Whenever the data is sent to the cloud, it is encrypted because to protect the sensitive data such that query privacy and data confidentiality is assured. Cloud computing reduces the in-house resources .This doesn't mean processing of the query should be slow. To ensure query privacy and data confidentiality RASP approach is designed. The RASP Perturbation technique combines Order preserving Encryption, Dimensionality Expansion, random noise injection, random projection to provide strong safety to the perturbed data and query. RASP makes use of the KNN algorithm to process the query efficiently. KNN approach use the minimum square range to process the query. It transfers data to the multidimensional space where it uses indexing approach to process the minimum square range points.

KEYWORDS: Indexing, ranging, Data confidentiality, Privacy, Ranging.

I.INTRODUCTION

Cloud computing refers to the service which is accessed over the internet. It is based on pay-as-you-go manner. The goal of cloud computing is to provide high performance computing or super computing power with the cloud computing technologies, large pool of resources can be connected using public or private network. Cloud is maintaining huge amount of resources hence security and privacy are two main concepts which is to be preserved [15]. There are three different types of cloud computing.

Infrastructure as a service where hardware is accessed over the internet such as server or storage. Software as a service where complete application is running on other's computer can be accessed such as web based email and Google document is well known exam which offer many online application.Platform as a service means that the application can be developed using web based tools so they run on system software and hardware. Force.com and the Google map application are examples.

Parallel computing of query service in the cloud is very popular because of the advantages of scalability and cost saving. Using the cloud infrastructure, the cloud service providers/ owners can conveniently scale –up and down and pay for what they use. Cloud service providers are not trust worthy and hence the data confidentiality and query privacy should be preserved.

The new approach should be proposed to preserve the privacy of data resources. To enjoy the benefits of the cloud computing, it is not meaningful to provide slow query services, because of security and privacy issues. The main purpose of cloud is to reduce the significant amount of cloud resources [4]. There exists some co-ordination between the data privacy, query privacy, use of the cloud. This is referred as DQEL criteria: Data privacy, query privacy , efficient query processing and low processing cost[9].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

II.RELATED WORK

Cloud computing is one of the most important and unique technique because of the scalability and cost saving. The data owner is not trust worthy and hence preserving the privacy of the sensitive data is major problem. The data is not sent to the cloud unless data confidentiality and query privacy are assured [12]. It is not meaningful to provide slow query processing because of these issues one should resolve the in-house resources.

Drawback:-

Service provider make a copy of data base or may corrupt the user query, as a result efficient query processing has to present.

BACKGROUND

In this section, the definition and properties of the RASP(Random Space Perturbation) is introduced. In Random Space Perturbation, the set of data is securely transformed, so that the order is preserved but the distribution and domain are changed [3]. So that the attacker cannot effectively recover the original data and the derived properties are preserved.

RASP is the multidimensional and uses the techniques such as geometric perturbation, random noise injection.

2.1 Properties of RASP

RASP has many important features

It is convexity preserving. It transforms the range into another polyhedron.

It doesn't preserve the order of dimensional values and the proof is straightforward [5].

It doesn't consider the length between two records.

The original query can be transformed in to the RASP perturbed data space. A hyper-cube is transformed into polyhedron using RASP perturbation.

2.2. MEANING OF DATA PERTURBATION.

Data perturbation is a popular technique, in preserving the privacy of data processing[11]. A major challenge in data perturbation is to balance the privacy protection and data utility, which are pair of conflicting factors.

There are two types of data perturbation, namely probability distribution approach and value distribution approach [7].

2.3. RASP USING AND KNN-ALGORITHM

RASP does not pertain the distance between the records, KNN query cannot be directly processed with the RASP perturbed data.

KNN algorithm is based on the range queries and uses the index in range query processing and uses the index in range query processing and hence fast processing of range queries takes place.

2.4 Processing **of KNN algorithm**

The main goal of KNN algorithm is to find the KNN nearest point in the spherical range that centred at the query point. It uses the square range instead of spherical range. However it has to overcome the few issues such as whether the data privacy and query privacy are present, whether these are a increment in the service workload [13]. How to find the minimum square range that exactly contains the KNN-nearest points.

The KNN algorithm consists of three rounds of interaction between service and the client.

First, the client will sent, the upper-bound range which contains more than K points and the lower- bound range which contains less than K-points to the server. The server finds the inner range and than returns to the client.

Second, the client finds the outer range depending upon the inner range and returns to the server. The server finds the outer range and then returns to the client.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Third, the client decrypts the records and finds the first K-point as the resulting server side.

2.5 Procedure of KNN-R

The client find the initial range and sends to the server, the server finds the inner range and returns to the client. Client upon receiving the inner range it decodes and request for the outer range. The server process the request and determines the range query with outer range and sends to the client. Client decrypts and sorts the requested data according to the nearest neighbour.

III.SYSTEM ARCHITECTURE

3.1 Proposed system.

The proposed system uses the KNN-algorithm for processing the range queries in the perturbation space. It helps in parallel processing of the data

.

Advantages:-

It satisfies all the aspects of DQEL criteria. Data privacy, query privacy, efficient query processing and low in-house processing cost.

The utility of the processing range queries will be preserved.

It uses the concept of indexing to support and find the minimum square range.

It processes the range queries in the efficient way such that it provide fast processing of range queries without any problem.

The users upon obtaining the result, decrypts and uses for its purpose.

The purpose of the architecture is to extend the data base server to the public cloud and the private cloud.

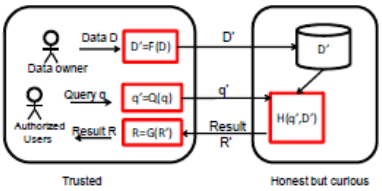
System architecture is having two groups:-

The trusted group and the honest group. The trusted group include the data owner and the authorized users. Data owner who have the ability to store the data into the cloud and authorized user who can query the data. The honest group include the cloud provider who host the data base and response the query services[8].

D' is the data stored in the cloud.

H(q',D') is the encrypted data over the service.

G(R') it is the encrypted result provided by the cloud server to the authorized user.



3.2 Threat model

Security analysis is considered as one of the important features, hence some assumption are made.

The data base is accessed only by the authorised user.

The communication between the client and server is properly protected and hence there will be no leakage of the data or the query[2].

The opponent can see the query processing, perturbed data base, the access pattern but not more than this.

The opponent can have the complete knowledge regarding the database, such as the attributes, application etc.

3.3 Security enhancement on transformation of query

The attacker will always target for transformed, hence it is necessary to describe the secure methods to preserve the queries. Whenever the admin login to process the necessary requirement ,admin will access the content and logout by changing the password. Similarly when the user login ,it checks for the registered user in the data base

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

then only it gives the permission to access the files or else it asks the user to register first in the database. Upon accessing the required files, the user logout by changing the password .in this manner the queries and data will be preserved.

IV. IMPLEMENTATION

4.1 (K-Nearest Algorithm) Algorithm For KNN

STEP-1 : Give the query keyboard
STEP-2 : Select all data related to that quary word
STEP-3 : If no result found
stop
STEP-4 : else
Retrieve all the post belongs to the keyword and in given range
STEP-5 : Take K value and range
If (k>no. of posts)
Repeat step 4 with increasing range
STEP-6 : calculate distance between all posts and user's (co-ordinates)
STEP-7 : Sort according to the distance (ascending order)
STEP-8: else
Take first k result
Send result
STEP-9: Exit

4.2 Algorithm for Distance Calculation

STEP 1: Compute the Euclidean distance for one dimension. The distance between two points in one dimension is simply the absolute value of the difference between their coordinates. Mathematically, this is shown as [p1-q1] where p1 is the first coordinate of the first point and q1 is the coordinate of the second point. We use the absolute value of this difference since distance is normally considered to have only a non – negative value.
STEP-2: Take two points p and s in two dimension Euclidean space. We will describe P with the coordinates (p1-p2) and Q with the coordinates (q1-q2). Now construct a line segment with the endpoint of P and Q .This line segment will form the hypotenuse of a right triangle. Extending the result obtained in step-1, we note that the length of the legs of this triangle are given by (p1-q1) and [p2-q2] .The distance between the two points will then be given as the length of the hypotenuse
STEP-3: use the Pythagorean theorem to determine the of the hypotenuse in step-2. this theorem states that c^2=a^2+b^2 where c is the length of the right triangle's hypotenuse and a,b are the length of the other two legs. This gives us c=(a^2+b^2)^(1/2)=((p1-q2)^2+ (p2-q2)^2)^(1/2). The distance between 2 points P =(p1,p2) and Q =(q1,q2) in two dimension space is therefore ((p1-q1)^2+(p2-q2)^2)^(1/2).
STEP-4: Extend the result of step 3 to there dimension space. The distance between points P= (p1,p2,p3) and Q =(q1,q2,q3) can then be given as ((p1-q1)^2+(p2-q2)^2+(p3-q3)^2)^(1/2)
STEP-5: Generalize the solution in step 4 for the distance between two points P= 9p1,p2……pn) and Q=9q1,q2……qn) in n dimensions. This general solution can be given as ((p1-q1)^2+(p2-q2)^2+…..+(pn-qn)^2(1/2)

4.3 Range logic

Let k =10
Let s_word= "ATM"
Let Range =500mts, dis=0, F=1,T=1
Let cur_loc is current location
If F=0 then
Filter all the postings with keyboards_word from location cur_location to distance position
End if

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

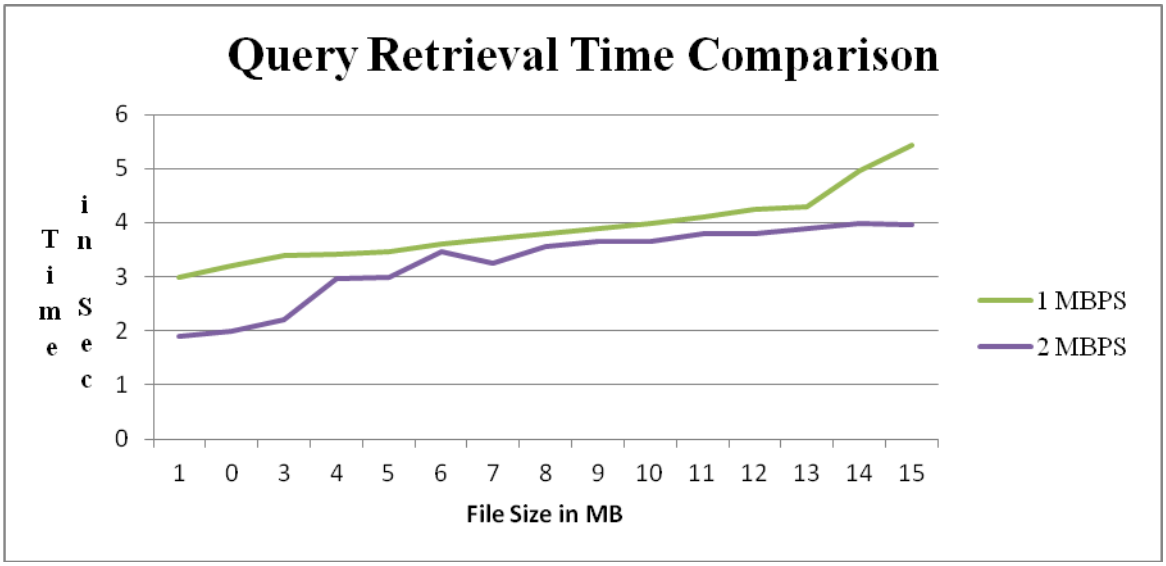
(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

If F = 1 then
Filter all the postingswith keyboards_word and location is cur_location
F=0
End if
Let c is no of search result
If c> k then
Display all the result and stop
Else
T=T+1
Dis = dis+ range
If T<=10 go to step x Stop
End if

V . RESULT AND ANALYSIS

All the secure approaches cannot use the indices to process the range queries, which may result in poor performance. Without the aid of indices, processing of KNN query will have to scan the entire database, leaving many optimization impossible to implement.



Performance of query processing

The figure represents time in seconds and file size. Here the database of 100 thousands of data points and 100 randomly selected queries is considered. The processing of the query is very fast and security is preserved very good.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

No of queries	1 MBPS	2MBPS
1	1	0.9
2	1.2	1
3	1.3	1.1
4	1.42	1.15
5	1.49	1.17
6	1.6	1.2

TABLE 1

Table 1 Compares the processing of range queries with different file size. As we see number of seconds to access the file size of 5MBPS is very less almost it takes just 1.49 seconds.

VI CONCLUSION AND FUTURE ENHANCEMENT

The RASP perturbation approach helps in processing of the queries very efficiently. It satisfies all the requirements of DQEL criteria it combines the features of order preserving encryption, random noise injection, random projection. The main benefit of using cloud computing is to reduce the amount of in-house workload. It uses KNN approach to process the query by deciding the range in the perturbed space and by allotment of indexing to the queries, so that the query can be processed very fastly.

Future Enhancement

(1) Further improve the performance of query processing for both range queries and KNN queries.
(2) Formally analyze the leaked query and access patterns and the possible effect on both data and query confidentiality.

REFERENCES

[1] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in Proceedings of ACM SIGMOD Conference, 2004.
[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. K. and Andy 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 Berkerley, 2009.
[3] J. Bau and J. C. Mitchell, "Security modeling and analysis," IEEE Security and Privacy, vol. 9, no. 3, pp. 18–25, 2011.
[4] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.
[5] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacypreserving multi-keyword ranked search over encrypted cloud data," in INFOCOMM, 2011.
[6] K. Chen, R. Kavuluru, and S. Guo, "Rasp: Efficient multidimensional range query on attack-resilient encrypteddatabases," in ACM Conference on Data and Application Security and Privacy, 2011, pp. 249–260.
[7] K. Chen and L. Liu, "Geometric data perturbation for outsourced data mining," Knowledge and Information Systems, 2011.
[8] K. Chen, L. Liu, and G. Sun, "Towards attack-resilient geometric data perturbation," in SIAM Data Mining Conference, 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," in Proceedings of the 13th ACM conference on Computer and communications security. New York, NY, USA: ACM, 2006, pp. 79–88.
[11] N. R. Draper and H. Smith, Applied Regression Analysis. Wiley, 1998.
[12] H. Hacigumus, B. Iyer, C. Li, and S. Mehrotra, "Executing sql over encrypted data in the database-service-provider model," in Proceedings of ACM SIGMOD Conference, 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," in Proceedings of Very Large Databases Conference (VLDB), 2004.
[15] H. Hu, J. Xu, C. Ren, and B. Choi, "Processing private queries over untrusted data cloud through privacy homomorphism," Proceedings of IEEE International Conference on Data Engineering (ICDE), pp. 601–612, 2011.
[16] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in Proceedings of ACM SIGMOD Conference, 2005.