ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Building Private and Humanizing Performance of Query Services in the Cloud with RASP Data Perturbation

Guthula Savitri
M.Tech Student,
Department CSE,
D.N.R. College of Engineering & Technology.

M.Ch.Sekhar Varma
Assistant Professor,
Department CSE,
D.N.R. College of Engineering & Technology.

DDD.Suri Babu
HOD & Associate Professor,
Department CSE,
D.N.R. College of Engineering & Technology.

Abstract:

With the development of services computing and cloud computing, it has become possible to outsource large databases todatabase service providers and let the providers maintain the range-query service. However, some data might be sensitive that the data owner does not want to move to the cloud unless the data confidentiality and query privacy are guaranteed. We propose the Random Space Encryption (RASP) approach that allows efficient range search with stronger attack resilience than existing efficiency-focused approaches. The random space perturbation (RASP) data perturbation method to provide secure and efficient range query and kNN query services for protected data in the cloud. The RASP data perturbation method combines order preserving encryption, dimensionality expansion, random noise injection, and random projection, to provide strong resilience to attacks on the perturbed data and queries. It also preserves multi-dimensional ranges, which allows existing indexing techniques to be applied to speedup range query processing. The kNN-R algorithm is designed to work with the RASP range query algorithm to process the kNN queries.

Keywords:

query services in the cloud, privacy, range query, kNN query.

I.INTRODUCTION:

With the wide deployment of public cloud computing infrastructures, using clouds to host data query services has become an appealing solution for the advantages on scalability and cost-saving. With the cloud infrastructures, the service owners can conveniently scale up or down the service and only pay for the hours of using the servers. While new approaches are needed to preserve data confidentiality and query privacy, the efficiency of query services and the benefits of using the clouds should also be preserved. It will not be meaningful to provide slow query services as a result of security and privacy assurance. It is also not practical for the data owner to use a significant amount of in-house resources, because the purpose of using cloud resources is to reduce the need of maintaining scalable in-house infrastructures. Therefore, there is an intricate relationship among the data confidentiality, query privacy, the quality of service, and the economics of using the cloud.[1]Here we summarize these requirements for constructing a practical query service in the cloud as the CPEL criteria: data confidentiality, query privacy, efficient query processing, and low in-house processing cost. Satisfying these requirements will dramatically increase the complexity of constructing query services in the cloud. Some related approaches have been developed to address some aspects of the problem.

However, they do not satisfactorily address all of these aspects. For example, the cryptoindex and order preserving encryption (OPE) are vulnerable to the attacks. The enhanced cryptoindex approach puts heavy burden on the in-house infrastructure to improve the security and privacy. The New Casper approach uses cloaking boxes to protect data objects and queries, which affects the efficiency of query processing and the inhouse workload. We propose the random space perturbation (RASP) approach to constructing practical range query and k-nearest- neighbor (kNN) query services in the cloud. The proposed approach will address all the four aspects of the CPEL criteria and aim to achieve a good balance on them. The RASP kNN query service (kNN-R) uses the RASP range query service to process kNN queries.[1]

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The RASP perturbation is a unique combination of OPE, dimensionality expansion, random noise injection, and random projection, which provides strong confidentiality guarantee. We have carefully evaluated our approach with synthetic and real data sets. The results show its unique advantages on all aspects of the CPEL criteria.The RASP method and its combination provide confidentiality of data and this approach is mainly used toprotect 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. It will retrieve the records from the database where it can denotes some value between upper and lower boundary. The kNN query denotes k-Nearest Neighbor query. K denotes positive integer and this query are used to find the value of nearest neighbor to k. The RASP perturbation embeds the multidimensional data into a secret higher dimensional space, en- hanced with random noise addition to protect the confidentiality of data.

II.RELATED WORK:

We review the some most related methods like 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. OPE represents 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. It allows database indexes to be built over an encryption table. The drawback of this process is the encryption key is too large and implementation makes the time and space overhead. 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, Xi < ai might be replaced with

$$X'i \in [ID1, ID2, ID3]$$

If the attacker manages to know the mapping between the input original query and the output bucket-based query, the range that a bucket ID represents could be estimated. The width of the bucket determines how precise the estimation could be done. A bucket-diffusion scheme was proposed to address this problem, which, however, has to sacrifice 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 large burden on the network and the client system. Furthermore, due to the randomized bucket IDs, the index built on bucket IDs is notso efficient for processing range queries as the index on OPE encrypted data is.Distance-recoverable encryption : DRE is the most intuitive method for preserving the nearest neighbor relationship. Because of the exactly preserved distances, many attacks can be applied. Here, dot products are used instead of distances to find kNN, which is more resilient to distance targeted attacks. One drawback is the search algorithm is limited to linear scan and no indexing method can be applied.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. In this the searching process will done by ranking process. The drawback of this concept is because of ranking process in house processing time will be maximized. The research on privacy preserving data mining has multiplicative perturbation methods, which are similar to the RASP encryption, but with more emphasis on preserving the utility for data mining.

III. METHODOLOGY:

A.Query Services In The Cloud:

Query is mainly used to search. Queries are constructed by using structured query language. It is mainly used to retrieving the needed information from the database. Query services are the method for services that are exposed through an implementation of service provider. Here by using RASP, range query and kNN query in cloud provide secure, fast storing and retrieving process of encryption and decryption of a data from database.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 Xi, Xj, and Xk 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 Xi ϵ [ai,bi] and Xj ϵ (aj,bj) and Xk=ak

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

which calculates the number of records in the range defined by conditions on Xi, Xj, and Xk. 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 Xiє[ai,bi] can be described with two half space conditions Xi ≤ bi and −Xi ≤−ai. Without loss of generality, we will discuss how to process half-space conditions like Xi ≤ bi in this paper. A slight modification will extend the discussed algorithms to handle other conditions like Xi< bi and Xi = bi.

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.

B.System Architecture:

We assume that a cloud computing infrastructure, such as Amazon EC2, is used to host the query services and large data sets. 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.Each record x in the outsourced database contains two parts: the RASP-processed attributes D'=F(D,K) and the encrypted original records, Z = E(D,K'), where K and K' are keys for perturbation and encryption, respectively. The RASP-perturbed data D' are for indexing and query processing. Fig. 1 shows the system architecture for both RASP-based range query service and kNN service.

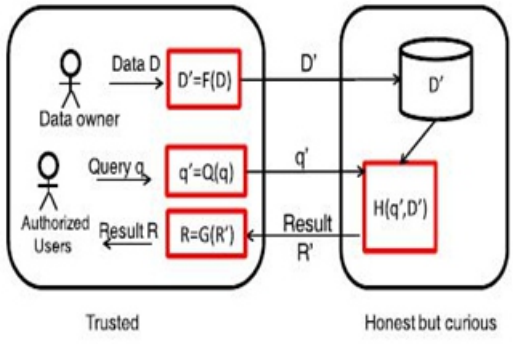

Fig. 1.The system architecture for RASP-based query services.

There are two clearly separated groups: 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 queriesto learn statistics or find some records. The untrusted parties include the curious cloud provider who hosts the query services and the protected database. The RASP-perturbed data will be used to build indices to support query processing.There are a number of basic procedures in this framework: 1) F(D) is the RASP perturbation that transforms the original data D to the perturbed data D'; 2) Q(q) transforms the original query q to the protected form q' that can be processed on the perturbed data; and 3) H(q',D') is the query processing algorithm that returns the result R'. When the statistics such as SUM or AVG of a specific dimension are needed, RASP can work with partial homomorphic encryption such as Paillier encryption [24] to compute these statistics on the encrypted data, which are then recovered with the procedure G'(R').

C.Threat Model:

The cloud server is considered as "honest-but-curious" in our model, which is consistent with related works on cloud security. Specifically, the cloud server acts in an "honest" fashion and correctly follows the designated protocol specification. However, it is "curious" to infer and analyze data (including index) in its storage and message flows received during the protocol so as to learn additional information.

Assumptions:

Our security analysis is built on the important features of the architecture. Under this setting, we believe the following assumptions are appropriate:

•Only the authorized users can query the proprietary database. Authorized users are not malicious and will not intentionally breach the confidentiality. We consider insider attacks are orthogonal to our research; thus, we can exclude the situation that the authorized users collude with the untrusted cloud providers to leak additional information.

•The client-side system and the communication channels are properly secured and no protected data records and queries can be leaked.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

•Adversaries can see the perturbed database, the transformed queries, the whole query processing procedure, the access patterns, and understand the same query returns the same set of results, but nothing else.

•Adversaries can possibly have the global information of the database, such as the applications of the database, the attribute domains, and possibly the attribute distributions, via other published sources (e.g., the distribution of sales, or patient diseases, in public reports).

Protected assets: Data confidentiality and query privacy should be protected in the RASP approach. While the integrity of query services is also an important issue, it is orthogonal to our study. Existing integrity checking and preventing techniques [33], [29], [18] can be integrated into our framework. Thus, the integrity problem will be excluded from the paper, and we can assume the curious cloud provider is interested in the data and queries, but it will honestly follow the protocol to provide the infrastructure service. Attacker modeling. The goal of attack is to recover (or estimate) the original data from the perturbed data, or identify the exact queries (i.e., location queries) to breach users' privacy. According to the level of prior knowledge the attacker may have, we categorize the attacks into two categories:

•Level 1: The attacker knows only the perturbed data and transformed queries, without any other prior knowledge. This corresponds to the cipertext-only attack in the cryptographic setting.

•Level 2: The attacker also knows the original data distributions, including individual attribute distributions and the joint distribution (e.g., the covariance matrix) between attributes. In practice, for some applications, whose statistics are interesting to the public domain, the dimensional distributions might have been published via other sources.

D.RASP: Random Space Perturbation:

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.RASP has several important features. First, RASP does not preserve the order of dimensional values because of the matrix multiplication component, which distinguishes itself from order preserving encryption schemes, and thus does not suffer from the distribution-based attack. Second, RASP does not preserve the distances between records, which prevents the perturbed data from distance based attacks. Because none of the transformations in the RASP: Eope, G, and F preserves distances, apparently the RASP perturbation will not preserve distances. Third, the original range queries can be transformed to the RASP perturbed data space, which is the basis of our query processing strategy. A range query describes a hypercubic area (with possibly open bounds) in the multidimensional space.

E.kNN Query Processing with RASP:

RASP denotes Random Space Perturbation. RASP is one type of multiplicative perturbation, with a novel combination of OPE, dimension expansion, random noise injection, and random projection.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.The RASP perturbation does not preserve distances (and distance orders), kNN query cannot be directly processed with the RASP perturbed data. In this section, we design a kNN query processing algorithm based on range queries (the kNN-R algorithm).

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

As a result, the use of index in range query processing also enables fast processing of kNN queries.The original distance-based kNN query processing finds the nearest k points in the spherical range that is centered at the query point. The basic idea of our algorithm is to use square ranges, instead of spherical ranges, to find the approximate kNN results, so that the RASP range query service can be used. There are a number of key problems to make this work securely and efficiently. 1) How to efficiently find the minimum square range that surely contains the k results, without many interactions between the cloud and the client? 2) Will this solution preserve data confidentiality and query privacy? 3) Will the proxy server's workload increase? to what extent ?The algorithm is based on square ranges to approximately find the kNN candidates for a query point, which are defined as follows.

IV Back Ground:

Definition 1 :"A square range is a hypercube that is centeredat the query point and with equal-length edges."Fig. 2 illustrates the range-query-based kNN processing with 2D data. The Inner Range is the square range that contains at least k points, and the Outer Range encloses the spherical range that encloses the inner range. The outer range surely contains the kNN results (see Proposition 2) but it may also contain irrelevant points that need to be filtered out.

Proposition 1:"The kNN-R algorithm returns results with100 percent recall."

Proof: The sphere in Fig. 2 between the outer range and the inner range covers all points with distances less than the radius r. Because the inner range contains at least k points, there are at least k nearest neighbors to the query points with distances less than the radius r. Therefore, the k nearest neighbors must be in the outer range.

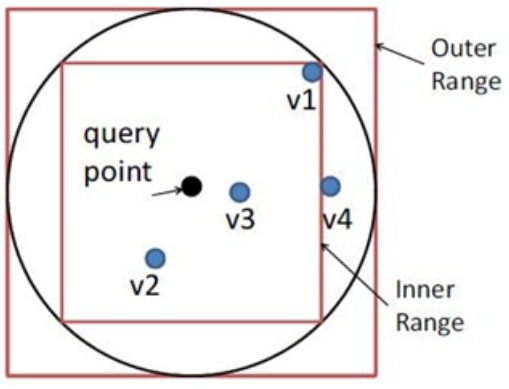

Fig. 2. Illustration for kNN-R Algorithm when k = 3

The kNN-R algorithm consists of two rounds of interactions between the client and the server. Fig. 3 demonstrates the procedure. 1) The client will send the initial upper bound range, which contains more than k points, and the initial lower bound range, which contains less than k points, to the server. The server finds the inner range and returns to the client. 2) The client calculates the outer range based on the inner range and sends it back to the server. The server finds the records in the outer range and sends them to the client. 3) The client decrypts the records and find the top k candidates as the final result.

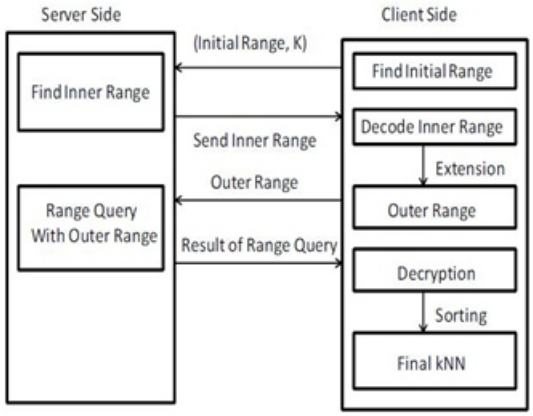

Fig. 3.Procedure of the KNN-R algorithm.

If the points are approximately uniformly distributed, we can estimate the precision of the returned result. With the uniform assumption, the number of points in an area is proportional to the size of the area. If the inner range contains m points, m > = k, the outer range contains q points, and the dimensionality is d, we can derive q = 2d=2m.

V EXPERIMENTS:

Here four sets of experimental resultsto investigate the following questions, correspondingly.1. How expensive is the RASP perturbation?2. How re0silient the OPE enhanced RASP is to the ICAbasedattack?3. How efficient is the two-stage range query processing?4. How efficient is the kNN-R query processing andwhat are the advantages?

Data Sets:

Three data sets are used in experiments. 1) A syntheticdata set that draws samples from uniform distribution inthe range [0, 1]. 2) The Adult data set from UCI machine-learning database.5 We assign numeric values to thecategorical values using a simple one-to-one mappingscheme, as described in Section 3.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

3) The 2D NorthEastlocation data from rtreeportal.org.6.2 Cost of RASP PerturbationIn this experiment, we study the costs of the componentsin the RASP perturbation. The major costs can be dividedinto two parts: the OPE and the rest part of RASP. Weimplement a simple OPE scheme [1] by mapping originalcolumn distributions to normal distributions. The OPEalgorithm partitions the target distribution into buckets.Then, the sorted original values are proportionally partitionedaccording to the target bucket distribution to createthe buckets for the original distribution. With the alignedoriginal and target buckets, an original value can bemapped to the target bucket and appropriately scaled.Therefore, the encryption cost mainly comes from thebucket search procedure (proportional to logD, where D isthe number of buckets).

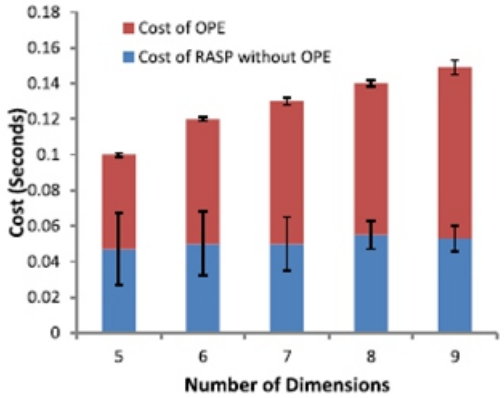

Fig. 3.The cost distribution of the full RASP scheme. Data: Adult (20K records, 5-9 dimensions).

IV. CONCLUSION:

We propose to study an outsourced service based on the CPEL criteria: data Confidentiality, query Privacy, Efficient query processing, and Low inhouse workload. With the CPEL criteria in mind, we develop the kNN-R approach for secure outsourced kNN query service. The kNN-R approach takes advantage of fast and secure RASP range query processing to implement kNN query processing. It can find high precision kNN results and also minimize the interactions between the cloud server and the inhouse client. High precision kNN results and minimized interactions result in low inhouse workload. We have conducted a thorough security analysis on data confidentiality and query privacy. Compared to the related approaches, the kNN-R approach achieves a better balance over the CPEL criteria.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 processin time of the query is minimized.

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. andEfficient 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.