IJSEAT International Journal of Science Engineering and Advance Technology

A Novel Approach to Build Reliable and Efficient Query Services in Cloud with RASP Data Perturbation

Mohammad Junaid[1], V. Padmaja[2]
[1] M.Tech (CSE), Nimra College of Engineering and Technology, A.P., India.
[2]Asst. Professor, Dept. of Computer Science & Engineering, Nimra College of Engineering and Technology, A.P., India.

Abstract — Inspired and motivated by the widespread development and deployment of public cloud computing infrastructures, using clouds to host data query services has become an fascinating solution for the advantages on scalability and cost-saving. However, Data owners show reluctance to share the sensitive and confidential data unless the Data integrity is guaranteed. On the other hand, a secured query service should still provide efficient query processing and significantly reduce the in-house workload to fully realize the benefits of cloud computing. We propose 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 multidimensional 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. We have carefully analyzed the attacks on data and queries under a precisely defined threat model and realistic security assumptions. Extensive experiments have been conducted to show the advantages of this approach on efficiency and security

Keywords — **Query services in the cloud, privacy, range query,**
kNN query

I. Introduction

Cloud offers unique advantages in scalability and cost-saving. Hence, hosting data-intensive query services in the cloud become increasingly popular. Service owners can conveniently scale up or down the service and only pay for the hours of using the servers with the cloud infrastructures, the. This is an attractive feature because the workloads of query services are highly dynamic, and it will be expensive and inefficient to serve such dynamic workloads with in-house infrastructures [2]. However, because the service providers lose the control over the data in the cloud, data confidentiality and query privacy have become the major concerns. Adversaries, such as curious service providers, can possibly make a copy of the database or eavesdrop users' queries, which will be difficult to detect and prevent in the cloud infrastructures.

There must be a trade-off between security and performance. Performance should not be compromised on the cost of security. 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 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.

Requirements have been summarized 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 [12] and order preserving encryption (OPE) [1] 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 in-house workload.

II. Problem Statement

We propose the random space perturbation (RASP) approach to constructing practical range query and knearest-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 basic idea is to randomly transform the multidimensional data sets with a combination of order preserving encryption, dimensionality expansion, random noise injection, and random project, so that the utility for processing range queries is preserved. The RASP perturbation is designed in such a way that the queried ranges are securely transformed into polyhedra in the RASP-perturbed data space, which can be efficiently processed with the support of indexing structures in the perturbed space. The RASP kNN query service (kNN-R) uses the RASP range query service to process kNN queries. The key components in the RASP framework include:

1. the definition and properties of RASP perturbation;
2. the construction of the privacy-preserving range query services;
3. the construction of privacy-preserving kNN query services; and
4. an analysis of the attacks on the RASP-protected data and queries.

In summary, the proposed approach has a number of unique contributions:

- The RASP perturbation is a unique combination of OPE, dimensionality expansion, random noise injection, and random projection, which provides strong confidentiality guarantee.
- The RASP approach preserves the topology of multidimensional range in secure transformation, which allows indexing and efficiently query processing.
- The proposed service constructions are able to minimize the in-house processing workload because of the low perturbation cost and high precision query results. This is an important feature enabling practical cloud-based solutions.

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.

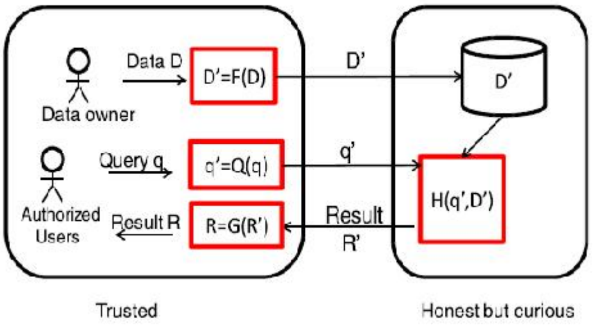

Fig. 1 The system architecture for RASP-based query services

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

RASP: RANDOM SPACE PERTURBATION

RASP is one type of multiplicative perturbation, with a novel combination of OPE, dimension expansion, random noise injection, and random projection. Let's consider the multidimensional data are numeric and in multidimensional vector space. The database has *k* searchable dimensions and *n* records, which makes a *d* x *n* matrix X. The searchable dimensions can be used in queries and thus should be indexed. Let *x* represent a *d*-dimensional record, *x* € IR^d. Note that in the *d*-dimensional vector space IR^d, the range query conditions are represented as half-space functions and a range query is translated to finding the point set in corresponding polyhedron area described by the half spaces [4]. The RASP perturbation involves three steps. Its security is based on the existence of random invertible real-value matrix generator and random real-value generator. For each *k*-dimensional input vector *x*,

Cost of RASP Perturbation

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. 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. Fig. 2 shows the cost distributions for 20K records at different number of dimensions. The dimensionality has slight effects on the cost of RASP perturbation. Overall, the cost of processing 20K records is only around 0.1 second.

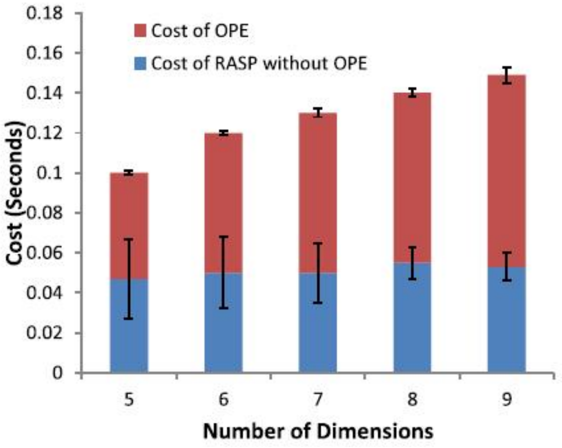

Fig. 2 The cost distribution of the full RASP scheme

Resilience to ICA Attack

We have discussed the methods for countering the ICA distributional attack on the perturbed data. In this set of experiments, we evaluate how resilient the RASP perturbation is to the distributional attack.

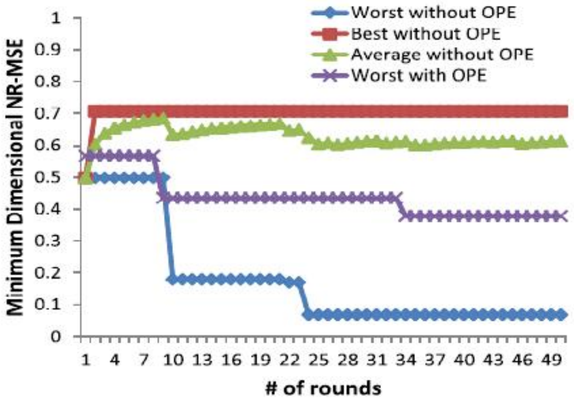

Fig. 3 Randomly generated matrix A and the progressive resilience to ICA attack.

Results. We simulate the ICA attack for randomly chosen matrices A. The data used in the experiment is the 10D Adult data with 10K records. Fig. 3 shows the progressive results in a number of randomly chosen matrices A. The xaxis represents the total number of rounds for randomly choosing the matrix A; the y-axis represents the minimum dimensional NR_MSE among all dimension. Without OPE, the label “Best-without-OPE” represents the most resilient A at the round i, “Worst-without-OPE” represents the A of the weakest resilience, and “Average-without-OPE” is the average quality of the generated A matrices for i rounds. We see that the best case is already close to the upper bound 0.7. With the OPE component, the worst case can also be significantly improved.

III. Related Work

Protecting Outsourced Data

Order preserving encryption. Order preserving encryption [3] preserves the dimensional value order after encryption. A well-known attack is based on attacker’s prior knowledge on the original distributions of the attributes. If the attacker knows the original distributions and manages to identify the mapping between the original attribute and its encrypted counterpart, a bucket-based distribution alignment can be performed to break the encryption for the attribute [6]. There are some applications of OPE in outsourced data processing. For example, Yiu et al. [8] use a hierarchical space division method to encode spatial data points, which preserves the order of dimensional values and thus is one kind of OPE.

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

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 [7]. Wong et al. suggest preserving dot products 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.

Preserving Query Privacy

Private information retrieval (PIR) [9] tries to fully preserve the privacy of access pattern, while the data may not be encrypted. PIR schemes are normally very costly. Focusing on the efficiency side of PIR, Williams et al. [11] use a pyramid hash index to implement efficient privacy preserving data-block operations based on the idea of Oblivious RAM. It is different from our setting of high throughput range query processing. Papadopoulos et al. [5] use private information retrieval methods [10] to enhance location privacy. However, their approach does not consider protecting the confidentiality of data. SpaceTwist proposes a method to query kNN by providing a fake user's location for preserving location privacy. But the method does not consider data confidentiality, as well. The Casper approach considers both data confidentiality and query privacy, the detail of which has been discussed in our experiments.

IV. Conclusion

We propose the RASP perturbation approach to hosting query services in the cloud, which satisfies the CPEL criteria: data confidentiality, query privacy, efficient query processing, and low in-house workload. The requirement on low in-house workload is a critical feature to fully realize the benefits of cloud computing, and efficient query processing is a key measure of the quality of query services.

RASP perturbation is a unique composition of OPE, dimensionality expansion, random noise injection, and random projection, which provides unique security features. It aims to preserve the topology of the queried range in the perturbed space, and allows to use indices for efficient range query processing. With the topology-preserving features, we are able to develop efficient range query services to achieve sub linear time complexity of processing queries. We then develop the kNN query service based on the range query service. The security of both the perturbed data and the protected queries is carefully analyzed under a precisely defined threat model. We also conduct several sets of experiments to show the efficiency of query processing and the low cost of in-house processing.

We will continue our studies on two aspects: 1) further improve the performance of query processing for both range queries and kNN queries; and 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," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2004.

[2] 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.

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

[4] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge Univ. Press, 2004.

[5] 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.

[6] 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.

[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 nformation 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 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," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2002.