www.ijiarec.com

ISSN: 2348-2079

MAR-2015

International Journal of Intellectual Advancements and Research in Engineering Computations

ROUGH OUTLIER AGENT BASED EFFICIENT QUERY SERVICES ON CLOUDS

[1]E. Saranya, [2]C. Gayathri.

ABSTRACT

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. 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. 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. 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. The proposed system 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 the approach on efficiency and security.

INTRODUCTION

Cloud computing is a recent trend in IT that moves computing and data away from desktop and portable PCs into large data centers. It refers to applications delivered as services over the Internet as well as to the actual cloud infrastructure — namely, the hardware and systems software in data centers that provide these services. The key driving forces behind cloud computing are the ubiquity of broadband and wireless networking, falling storage costs, and progressive improvements in Internet computing software. Cloud-service clients will be able to add more capacity at peak demand, reduce costs, experiment with new services, and remove unneeded capacity, whereas service providers will increase utilization via multiplexing, and allow for larger investments in software and hardware.

The main technical underpinnings of cloud computing infrastructures and services include virtualization, service-oriented software, grid computing technologies, management of large facilities, and power efficiency. Consumers purchase such services in the form of infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), or soft-ware-as-a-service (SaaS) and sell value-added services to users. Within the cloud, the laws of probability give service providers great leverage through statistical multiplexing of varying workloads and easier management — a single software installation can cover many users' needs.

We can distinguish two different architectural models for clouds: the first one is designed to scale out by providing additional computing instances on demand. Clouds can use

Author for Correspondence:

[1]Final year ME, Mahendra Institute of Technology, Mahendrapuri , Namakkal, Tamilnadu, India.

[2]Assistant Professor/CSE,Mahendra Institute of Technology, Mahendrapuri , Namakkal, Tamilnadu, India.

these instances to supply services in the form of SaaS and PaaS. The second architectural model is designed to provide data and compute-intensive applications via scaling capacity. In most cases, clouds provide on-demand computing instances or capacities with a "pay-as-you-go" economic model. The cloud infrastructure can support any computing model compatible with loosely coupled CPU clusters. Organizations can provide hardware for clouds internally, or a third party can provide it externally. A cloud might be restricted to a single organization or group, available to the general public over the Internet, or shared by multiple groups or organizations.

A cloud comprises processing, network, and storage elements, and cloud architecture consists of three abstract layers. Infrastructure is the lowest layer and is a means of delivering basic storage and compute capabilities as standardized services over the network. Servers, storage systems, switches, routers, and other systems handle specific types of workloads, from batch processing to server or storage augmentation during peak loads. The middle platform layer provides higher abstractions and services to develop, test, deploy, host, and maintain applications in the same integrated development environment. The application layer is the highest layer and features a complete application offered as a service.

In 1961, John McCarthy envisioned that "computation may someday be organized as a public utility." We can view the cloud computing paradigm as a big step toward this dream. To realize it fully, however, we must address several significant problems and unexploited opportunities concerning the deployment, efficient operation, and use of cloud computing infrastructures.

Cloud computing service's emergence suggests fundamental changes in software and hardware architecture. Computer architectures should shift the focus of Moore's law from increasing clock speed per chip to increasing the number of processor cores and threads per chip. Industry and academia must design novel systems and services that would exploit a high degree of parallelism. Software architectures for massively parallel, data-intensive computing, will grow in popularity. In terms of storage technologies, we'll likely shift from hard disk drives (HDDs) to solid-state drives (SDDs), such as flash memories, or, given that completely replacing hard disks is prohibitively expensive, hybrid hard disks — that is, hard disks augmented with flash memories, which provide reliable and high-performance data storage. The biggest barriers to adopting SSDs in data centers have been price, capacity, and, to some extent, the lack of sophisticated query-processing techniques. However, this is about to change as SSDs' I/O operations per second (IOPS) benefits become too impressive to ignore, their capacity increases at a fast pace, and we devise new algorithms and data structures tailored to them.

RELATED WORK

PROTECTING OUTSOURCED DATA

Order preserving encryption. Order preserving encryption 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. [10] 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 [8]. Wong et al. [12] 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) 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. [4] 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. [9] use private information retrieval methods enhance location privacy. Their approach does not consider protecting the confidentiality of data. Space Twist [13] 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 [1] considers both data confidentiality and query privacy, the detail of which has been discussed in our experiments.

OTHER RELATED WORK

Another line of research facilitates authorized users to access only the authorized portion of data, for example, a certain range, with a public key scheme. However, the underlying encryption schemes do not produce index able encrypted data. The setting of multidimensional range query in [11] is different from ours. Their approach requires that the data owner provides the indices and keys for the server, and authorized users use the data in the server. While in the cloud database scenario, the cloud server takes more responsibilities of indexing and query processing. Secure keyword search on encrypted documents [3], [5] scans each encrypted document in the database and finds the documents containing the keyword, which is more like point search in database. The research on privacy preserving data mining has discussed multiplicative perturbation methods [7], which are similar to the RASP encryption, but with more emphasis on preserving the utility for data mining.

QUERY SERVICES IN THE CLOUD

Hosting data-intensive query services in the cloud is increasingly popular because of the unique advantages in 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. 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]. 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.

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.

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. They do not satisfactorily address all of these aspects. 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. 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. We have summarized the weaknesses of the existing approaches.

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 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 polyhedral 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.

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.

PROBLEM STATEMENT

Object ranking is a popular retrieval task in various applications. In relational databases, rank tuples using an aggregate score function on the attribute values. For example, a real estate agency maintains a database that contains information of flats available for rent. A potential customer wishes to view the top-10 flats with the largest sizes and lowest prices. Here, the score of each flat is expressed by the sum of two qualities: size and price, after normalization to the domain.

In spatial databases, ranking is often associated to K-nearest neighbor (NN) retrieval. Given a query location, user interested in retrieving the set of nearest objects to it that qualifies a condition. Assuming that the set of interesting objects is indexed by an R-tree, Apply distance bounds and traverse the index in a branch-and-bound fashion to obtain the answer. There is a growing awareness of the importance of ranking in recommender systems. Furthermore, while, as mentioned earlier, there has been significant amount of work done on improving individual diversity, the issue of aggregate diversity in recommender systems has been largely untouched. When a data is based on ranking analysis is no longer has a business needed to access data centres his privileges to access data centres should be immediately revoked. All physical and electronic access to data centres by employees should be logged and audited routinely. The following problems are identified from the existing system

- It is becoming increasingly harder to find relevant content. The problem is not only widespread but also alarming.
- It does not serve the user as a prediction tool on cloud based pertubation data management.
- The data retrieval based on user requirement is not done.

- Unsecured data processing.
- Difficult to maintain large amount of data.
- It does not answer the business question.

ROUGH OUTLIER AGENT BASED QUERY SERVICES ON CLOUDS

Spatial ranking, which orders the objects according to the distance from a reference point, and Non-spatial ranking, which orders the objects by an aggregate function on the non-spatial values. Greedy k-NN spatial preference query integrates these two types of ranking in an intuitive way. Greedy k-NN spatial preference query integrates these two types of ranking in an intuitive way. As indicated by examples, the new query has a wide range of applications in service recommendation and decision support systems. To the knowledge, there is no existing efficient solution for processing the greedy K-NN spatial preference query.

A brute-force approach for evaluating it is to compute the scores of all objects in D and select the greedy K-NN ones. The method, however, is expected to be very expensive for large input datasets. In real world settings, recommender systems generally perform the following two tasks in order to provide recommendations to each user. First, the ratings of unrated items are estimated based on the available information using some recommendation algorithm. And second, the system finds items that maximize the user's utility based on the predicted ratings, and recommends to the user.

Ranking approaches proposed in the project are designed to improve the recommendation diversity in the second task of finding the best items for each user. The new Greedy K-NN algorithms that can improve the predictive accuracy of recommendations. The quality of recommendations can be evaluated along a number of dimensions, and relying on the accuracy of recommendations alone may not be enough to find the most relevant items for each user.

In particular, the importance of diverse recommendations has been previously emphasized in several studies. These studies argue that one of the goals of recommender systems is to provide a user with highly idiosyncratic or personalized items, and more diverse recommendations result in more opportunities for users to get recommended such items. The following modules are used in the system.

- Passive data disclosure data mining.
- Query Processing modeling.
- Testing the effectiveness of Data Points
- Security and distributional Data Points

PASSIVE DATA DISCLOSURE DATA MINING

Passive data disclosure is our major concern. Data Pointers can access the data at any of compromised virtual machines in the cloud. They might be interested in recovering or estimating the original data records or distributional information, based on the perturbed data. While distinguishing data mining are meaningful to traditional encryption systems, they are lessusefull in our context. We identify that the main data mining are based on statistical estimation. Data tampering or dishonest cloud service providers is not addressed by our study, which can be covered by integrity preserving techniques.

DATA POINTS ER MODELING

Active Data Pointers will try to obtain as much knowledge as possible to help recover the original data. To better analyze the security of the E-RASP(EFFICIENT RASP) perturbation, we define the adversarial power according to the levels of prior knowledge on the data.

TESTING THE EFFECTIVENESS OF DATA POINTS

This measure has an extra benefit in evaluating a new type of estimation data mining. It can be used to assess how serious the Data Points can be based on Data Points simulation. First, we randomly sample the dataset to generate a subset. The simulated Data Points will generate an estimation on the subset, which is used to calculate the estimation error E. Then, the LOC measure can be calculated. Repeating this procedure multiple times on different random sample sets, we can get a robust estimation on the effectiveness of the new Data Points .

SECURITY AND DISTRIBUTIONAL QUERY PROCESSING

Since the query transformation is deterministic, the same query is always mapped to the same perturbed query. The Query Processing can keep track of the frequencies of the perturbed queries. With the Data mining Process knowledge about the query distribution and counting a sufficiently large number of perturbed queries, the attacker can possibly build a mapping between the original queries and the perturbed queries. Thus, the privacy of some queries could be breached under the Data mining Process adversarial power

CONCLUSION

Here 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 using indices for efficient range query processing.

With the topology-preserving features, we are able to develop efficient range query services to achieve sublinear time complexity of processing queries. 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. It also conduct several sets of experiments to show the efficiency of query processing and the low cost of in-house processing. The studies will continue 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]. M.F. Mokbel, C. yin Chow, and W.G. Aref, “The New Casper: Query Processing for Location Services without Compromising Privacy,” Proc. 32nd Int’l Conf. Very Large Databases Conf. (VLDB), pp. 763-774, 2006.

[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]. C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, “Secure Ranked Keyword Search over Encrypted Cloud Data,” Proc. IEEE Int’l Conf. Distributed Computing Systems (ICDCS), 2010.

[4]. P. Williams, R. Sion, and B. Carbunar, “Building Castles Out of Mud: Practical Access Pattern Privacy and Correctness on Untrusted Storage,” Proc. ACM Conf. Computer and Comm. Security, 2008.

[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]. S. Papadopoulos, S. Bakiras, and D. Papadias, “Nearest Neighbor Search with Strong Location Privacy,” Proc. Very Large Databases Conf. (VLDB), 2010.

[10]. 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.

[11]. E. Shi, J. Bethencourt, T.-H.H. Chan, D. Song, and A. Perrig, "Multi Dimensional Range Query over Encrypted Data," Proc. IEEE Symp. Security and Privacy, 2007.

[12]. W.K. K, D.W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure KNN Computation on Encrypted Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 139-152, 2009.

[13]. M.L. L, C.S. S, X. Huang, and H. Lu, "SpaceTwist: Managing the Trade-Offs among Location Privacy, Query Performance, and Query Accuracy in Mobile Services," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 366-375, 2008.