IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

Efficient Information Retrieval System for Cloud Environments

Jyothi T

Assistant Professor, Department of ISE, GSSSIETW, Mysore

Abstract: Cloud computing as an emerging technology trend is expected to reshape the advances in information technology. In a cost-efficient cloud environment, a user can tolerate a certain degree of delay while retrieving information from the cloud to reduce costs. The most important aspect in the cloud environment is maintaining privacy and efficiency. In this paper an efficient information retrieval query (EIRQ) scheme to reduce the querying overhead in the cloud is presented. In EIRQ, queries are classified into multiple ranks, where a higher ranked query can retrieve a higher percentage of matched files. A user can retrieve files on demand by choosing queries of different ranks. The rank shows the percentage of files that will be returned to the user.

Keywords: Cloud computing, cost efficiency, differential query services, privacy.

I. INTRODUCTION

Cloud computing is the delivery of computing resources over the Internet. It has been widely adopted in broad applications and is becoming more pervasive. The main reasons behind cloud computing sharp growth are increases in computing power and data storage, exponential growth of social network data, and modern data centres, some of which can suffer from high maintenance costs and low utilization. There are also challenges in the development of reliable and cost-effective cloud-based systems. Cloud computing presents a new way to supplement the current consumption and delivery model for IT services based on the Internet, by providing for dynamically scalable and often virtualized resources as a service over the Internet. Cloud computing is the use of computing resources (hardware and software) which are available in remote location and accessible over the network. Users are able to buy these computing resources as a utility, on demand. The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation.

Cloud computing as an emerging technology is expected to reshape information technology processes in the near future [1]. Due to the overwhelming merits of cloud computing, e.g., cost-effectiveness, flexibility and scalability, more and more organizations choose to outsource theirdata for sharing in the cloud. As a typical cloud application, an organization subscribes the cloud services and authorizes its staff to share files in the cloud. Each file is described by a set of keywords, and the staff, as authorized users, can retrieve files of their interests by querying the cloud with certain keywords. In such an environment, how to protect user privacyfrom the cloud, which is a third party outside the security boundary of the organization, becomes a key problem.

User privacy can be classified into search privacy and access privacy [2]. Search privacy means that the cloud knows nothing about what the user is searching for, and access privacy means that the cloud knows nothing about which files are returned to the user. When the files are stored in the clear forms, a proper solution to protectuser privacy is for the user to request allof the files from the cloud; this way, the cloud cannot know which files the user is really interested in. While this does provide the necessary privacy, the communication cost is high. Private searching was proposed by Ostrovsky et al. [3][4] which allows a user to retrieve files of interest from an untrusted server without leaking any information. However, the Ostrovsky scheme has a high computational cost, as it requires the cloud to process the query on everyfile in a collection. Otherwise, the cloud will assume that certain files, without processing, are of no interest to the user. It will quickly become a performance bottleneck when the cloud needs to processthousands of queries over a collection of hundreds of thousands of files.To make private searching applicable in a cloud environment, the previous work [7] designed a cooperateprivate searching protocol (COPS), where a proxy server, called the aggregation and distribution layer (ADL), is introduced between the users and the cloud. The ADL deployed inside an organization has two main functionalities: aggregating user queries and distributing search results. Under the ADL, the computation cost incurred on the cloud can be largely reduced, since the cloud only needs to execute a combined query once, no matter how many users are executing queries.

Furthermore, the communication cost incurred on the cloud will also be reduced, since files shared by the users need to be returned only once.Motivated by this goal, the new scheme, named Efficient Information retrieval for Ranked Query (EIRQ), in which each user can provide his own percentage along with the query to determine the percentage of matched files to be returned. The basic idea of EIRQ is to construct a privacy preserving mask matrixthat allows the

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

cloud to filter out a certain percentage of matched files before returning to the ADL. This is not a trivial work, since the cloud needs to correctly filter out files according to the rank of queries without knowing anything about user privacy.

II. RELATED WORK

A number of methods have been proposed in recent years to provide user privacy and also regarding private searching schemes.

Private searching on streaming data (2005).In this paper, R. Ostrovsky and W. Skeith [1] considered the problem of private searching on streaming data. He showed that in this model we can efficiently implement searching for documents under a secret criteria (such as presence or absence of a hidden combination of hidden keywords) under various cryptographic assumptions.The results can be viewed in a variety of ways: as a generalization of the notion of a PrivateInformation Retrieval as positive results on privacy-preserving datamining; and as a delegation of hidden program computation. Searchable symmetric encryption allows a party to outsource the storage of his data to another party in a private manner, while maintaining the ability to selectively search over it. This problem has been the focus of active research and several security definitions and constructions have been proposed. In this paper we begin by reviewing existing notions of security and propose new and stronger security definitions. R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky [2] presented two constructions that show secure under new definitions. Interestingly, in addition to satisfying stronger security guarantees, the new constructions are more efficient than all previous constructions. Further, prior work on SSE only considered the setting where only the owner of the data is capable of submitting search queries. They also consider the natural extension where an arbitrary group of parties other than the owner can submit search queries. The SSE is formally defined in this multi-user setting, and presents an efficient construction. Private searching on stream data Journal of Cryptology .Private searching on streaming data is a process to dispatch to a public server a program, which searches streaming sources of data without revealing searching criteria and then sends back a buffer containing the findings. From an Abelian group homomorphic encryption, the searching criteria can be constructed by only simple combinations of keywords, for example, disjunction of keywords.

The recent breakthrough in fully homomorphic encryption has allowed us to construct arbitrary searching criteria theoretically. Here consider a new private query, which searches for documents from streaming data on the basis of keyword frequency, such that the frequency of a keyword is required to be higher or lower than a given threshold. This form of query can help us in finding more relevant documents. Based on the state of the art fully homomorphic encryption techniques, we give disjunctive, conjunctive, and complement constructions for private threshold queries based on keyword frequency. Combining basic constructions, further presented a generic construction for arbitrary private threshold queries based on keyword frequency. The protocols are semantically secure as long as the underlying fully homomorphic encryption scheme is semantically secure.

Hierarchical attribute-based encryption and scalable user revocation for sharing data in cloud servers .Access control is one of the most important security mechanisms in cloud computing. Attributed based encryption provides an approach that allows data owners to integrate data access policies within the encrypted data. However, little work has been done to explore flexible authorization in specifying the data user's privileges and enforcing the data owner's policy in cloud based environments. In this paper, G. Wang, Q. Liu, J. Wu, and M. Guo [4] propose a hierarchical attribute based access control scheme by extending cipher text-policy attribute-based encryption (CP-ABE) with a hierarchical structure of multi authorities and exploiting attribute-based signature (ABS). The proposed scheme not only achieves scalability due to its hierarchical structure, but also inherits fine-grained access control with authentication in supporting write privilege on outsourced data in cloud computing. In addition, it showed decoupling the task of policy management from security enforcement by using the extensible access control mark up language (XACML) framework. Extensive analysis shows that this scheme is both efficient and scalable in dealing with access control for outsourced data in cloud computing. Efficient information retrieval for ranked queries in cost-effective cloud environments. Cloud computing as an emerging technology trend is expected to reshape the advances in information technology. In this paper, it addresses two fundamental issues in a cloud environment: privacy and efficiency. Here first review a private keyword-based file retrieval scheme proposed by Ostrovsky et. [5]. Then, based on an aggregation and distribution layer (ADL), presented a scheme, termed efficient information retrieval for ranked query (EIRQ), to further reduce querying costs incurred in the cloud. Queries are classifiedinto multiple ranks, where a higher ranked query can retrieve a higher percentage of matched files. Extensive evaluations have been conducted on an analytical model to examine the effectiveness of this scheme. New constructions and practical applications for private stream searching (2013).A system for private stream searching allows a client to retrieve documents matching some search criteria from a remote server while the server evaluating the request remains provably oblivious to the search criteria. In this extended abstract, we give a high level outline of a new scheme for this problem and an experimental analysis of its scalability. The new scheme is highly efficient in practice. We demonstrate the practical applicability of the scheme by considering its performance in the demanding scenario of providing a privacy preserving version of the Google News Alerts service.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

III. PROPOSED SYSTEM

Here the new proposed scheme called Efficient Information retrieval System is introduced. This new system uses the method of Flexible ranking mechanism which allows users to provide a rank and can personally decide how many matched files will cloud returns.
The basic idea is to construct a matrix that allows the cloud to filter out certain percentage of matched files. The new scheme reduces the querying overhead and also computational costs. The EIRQ system protects user privacy which allows each user to retrieve matched files on demand. This is not an easy work, because the cloud needs to correctly filter out files according to the rank of queries without knowing anything about user privacy. This has two extensions: the first extension shows the least amount of modifications from the Ostrovsky scheme, and the second extension provides privacy by leaking the least amount of information to the cloud

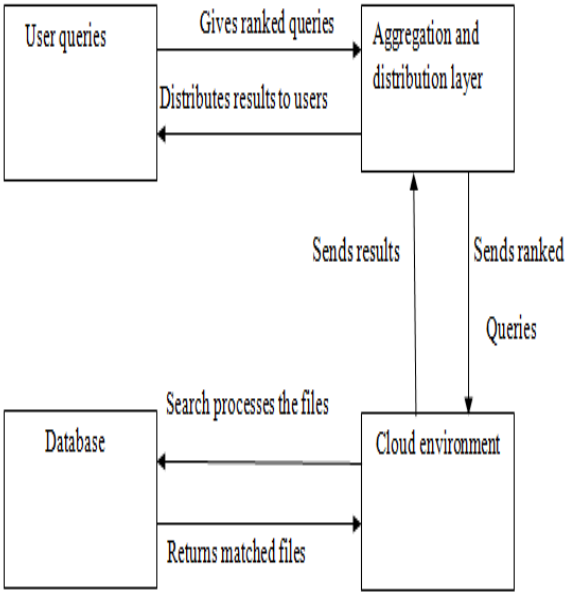


Figure 1: Architecture of the System

The proposed system has following four modules

Differential Query Services

The novel concept proposed here is a differential query service, to COPS, where the users are allowed to personally decide how many matched files will be returned. This is motivated by the fact that under certain cases, there are a lot of files matching a user's query, but the user is interested in only a certain percentage of matched files. To illustrate, let us assume that Alice wants to retrieve 2% of the files that contain keywords "A, B", and Bob wants to retrieve 20% of the files that contain keywords "A, C". The cloud holds 1,000 files, where {F1, . . . , F500} and {F501, . . . , F1000} are described by keywords "A, B" and "A, C", respectively. In the Ostrovsky scheme, the cloud will have to return 2,000 files. In the COPS scheme, the cloud will have to return 1,000 files. In our scheme, the cloud only needs to return 20 files. Therefore, by allowing the users to retrieve matched files on demand, the bandwidth consumed in the cloud can be largely reduced.

Efficient Information Retrieval for Ranked Query

The new scheme proposed is termed as Efficient Information retrieval for Ranked Query (EIRQ), in which each user can choose the rank of his query to determine the percentage of matched files to be returned. The basic idea of EIRQ is to construct a privacy preserving mask matrixthat allows the cloud to filter out a certain percentage of matched files before returning to the ADL. This is not a trivial work, since the cloud needs to correctly filter out files according to the rank of queries without knowing anything about user privacy. Focusing on different design goals, we provide two extensions: the first extension emphasizes simplicityby requiring the least amount of modifications from the Ostrovsky scheme, and the second extension emphasizes privacyby leaking the least amount of information to the cloud.

Aggregation and Distribution Layer

An ADL is deployed in an organization that authorizes its staff to share data in the cloud. The staff members, as the authorized users, send their queries to the ADL, which will aggregate user queries and send a combined query to the cloud. Then, the cloud processes the combined query on the file collection and returns a buffer that contains all of

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

matched files to the ADL, which will distribute the search results to each user. To aggregate sufficient queries, the organization may require the ADL to wait for a period of time before running our schemes, which may incur a certain querying delay.

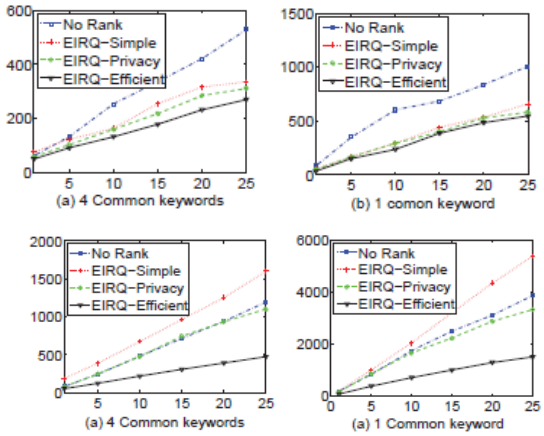
Ranked Queries

To further reduce the communication cost, a differential query service is provided by allowing each user to retrieve matched files on demand. Specifically, a user selects a particular rankfor his query to determine the percentage of matched files to be returned. This feature is useful when there are a lot of files that match a user's query, but the user only needs a small subset of them.

IV. EXPERIMENTAL RESULTS & EVALUATION

In this section, we will compare three EIRQ schemes, from the following aspects: file survival rate and computation/ communication cost incurred on the cloud. Then, based on the simulation results, we deploy our program in Amazon Elastic Compute Cloud (EC2) to test the transfer-in and transfer-out time incurred on the cloud when executing private searches. In the previous scheme there was a large querying overhead and consumes more bandwidth.

First, we test the transfer-in time in the real cloud, which is mainly incurred by receiving queries from the ADL. Then, we test the transfer-out time at the cloud, which is mainly incurred by returning files to the ADL. The results are shown below



Therefore, EIRQ-Efficient is most suitable to be deployed to a cloud environment. For example, the time to transfer a query from the ADL to the cloud consumes less than 100 seconds, and the time to transfer the buffer from the cloud to the ADL consumes less than 500 seconds, fewer than 4 common keywords.

V. CONCLUSION

In this paper, we proposed three EIRQ schemes based on an ADL to provide differential query services while protecting user privacy. By using our schemes, a user can retrieve different percentages of matched files byspecifying queries of different ranks. By further reducing the communication cost incurred on the cloud, the EIRQ schemes make the private searching technique more applicable to a cost-efficient cloud environment. However, in the EIRQ schemes, we simply determine the rank of each file by the highest rank of queries it matches. As a future enhancement a flexible ranking mechanism can be designed for the EIRQ schemes.

REFERENCES

[1] R. Ostrovsky and W. Skeith, "Private searching on streaming data," in Proc. of CRYPTOLOGY, 2005.
[2] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," in Proc. ofACM CCS, 2006.
[3] "Private searching on streaming data," Journal of Cryptology, 2007.
[4] G. Wang, Q. Liu, J. Wu, and M. Guo, "Hierarchical attribute-based encryption and scalable user revocation for sharing data in cloud servers," Computers & Security, 2011.
[5] Q. Liu, C. C. Tan, J. Wu, and G. Wang, "Efficient information retrieval for ranked queries in cost-effective cloud environments," in Proc. of IEEE INFOCOM, 2012.
[6] J. Bethencourt, D. Song, and B. Waters, "New constructions and practical applications for private stream searching," in Proc. ofIEEE S&P, 2013.