International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

Searching over Secured Cloud by Preserving Privacy:SSPP

Chippy Mariam Thomas
Department of Computer Science and Engineering
Jawaharlal College of Engineering and Technology
Palakkad, Kerala, India

Sreejith R
Department of Computer Science and Engineering
Jawaharlal College of Engineering and Technology
Palakkad, Kerala, India

Ambika Devi Amma T
Department of Computer Science and Engineering
Jawaharlal College of Engineering and Technology
Palakkad, Kerala, India

***Abstract* — Cloud Computing has a great role in providing many flexible services to the users. The complexity in the data management is solved by outsourcing the relevant and sensitive data to the cloud. As privacy protection is a main concern in cloud storage, it can be maintained by keeping them encrypted. The sensitive documents are kept encrypted by using the symmetric key encryption. Earlier the works on searchable encryption was focusing on the single-keyword and the Boolean search. Later the multiple keywords are allowed to include in the search query. The search result obtained on considering each query which is included in the request. The relevance of the requested query keyword on the documents is evaluated by the similarity measure, co-ordinate matching. The synonyms of the requested queries are considered to increase the efficiency of the retrieval process. The co-occurrence probability of the keyword is taken in account for the sematic relationship. On analyzing the real world experiments the computational cost for the retrieval process over the secured data is reduced.**

Keywords — Cloud; keyword search; ranking; searchable encryption; sematic relation

I. INTRODUCTION

Cloud computing denotes a model on which a computing infrastructure is viewed as a cloud, from which business and individuals access applications from anywhere in the world on demand. The main principle behind this model is offering computing, storage and software as service. Cloud is a parallel and distributed computing systems consisting of a collection of inter- connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service level agreements established between service provider and consumer [1]. Cloud computing has gained great attention from both industries and academics. With goal of providing users more flexible services in a transparent manner, all services are allocated in a cloud that actually is a collection of devices and resources connected through internet. One of the core services provides by cloud computing is the data storage. This poses new challenges in creating secure and reliable data storage and access facilities over remote service providers in the cloud. This is a long dreamed vision of computing as a utility, where cloud customers can remotely store their data into the cloud so as to enjoy the on-demand high-quality applications and services from a shared pool of configurable computing resources. The benefits brought by this new computing model include but are not limited to: relief of the burden for storage management, universal data access with independent geographical locations, and avoidance of capital expenditure on hard-ware, software, and personnel maintenances. The security of data storage is one of the necessary tasks to be addressed before the blueprint for cloud computing is accepted.

However cloud computing provides efficient, flexible and cost effective services it is not 100% secure. The major concerns in cloud computing are privacy and security. There are various security issues in cloud computing. Of many security issues [2][3], data security seems to be the major obstacle towards the adaption of cloud computing. Data security in cloud is must, in order to ensure that the data has not been accessed by any unauthorized person. To address the concerns in the adoption of public cloud storage, a virtual private storage service based on recently developed cryptographic techniques is used. Such a service should provide confidentiality, integrity, availability, reliability, efficient retrieval and data sharing. Cryptographic mechanisms applied to data offer the best solution for data protection. This means the customer must encrypt the data locally prior to uploading to cloud. Predicate encryption is a new paradigm, generalizing Identity-Based Encryption. In this paradigm secret keys correspond to predicates and the encrypted text to attributes. Predicate encryption (PE) is used for searching over encrypted data, evaluating certain encrypted attributes. Such encryption schemes work only for a few classes of predicates. PE has two important properties: does not reveal any information about the cipher text and no information about the query predicate. The homomorphic encryption permits processing of encrypted data on a remote

storage without decrypting it. This method offers many advantages especially for the cloud paradigm because the users can benefit of the advantages offered by the Cloud and they also protect the data confidentiality and privacy [4].

Searching encrypted data over cloud storage is a challenging task. An optimal security is achieved by the oblivious RAM model [5] which does not leak any information to the server. But this model is impractical for real world scenarios due to the excessive computational costs. Searching with sequential scan [6] scheme works by computing the bitwise exclusive or (XOR) of the clear-text with a sequence of pseudorandom bits which have a special structure. This structure allow to search on the data without revealing anything else about the clear text. A sequential scan without an index may not be efficient enough when the data size is large. Regular private-key encryption prevents one from searching over encrypted data; clients also lose the ability to selectively retrieve segments of their data. To address this, several techniques have been introduced for provisioning symmetric encryption with search capabilities. Private-key searchable encryption [7], the users himself encrypts the data and the additional data structures and stored on the server so that only someone with the private key can access it. The searching on public-key-encrypted data [8], users who encrypt the data and send it to the server can be different from the owner of the decryption key. Searchable encryption allows data owner to outsource his data in an encrypted manner while maintaining the selectively search capability over the encrypted data. Traditional searchable encryption schemes allow a user to securely search over encrypted data through keywords without first decrypting it, these techniques support only conventional Boolean keyword search [9], without capturing any relevance of the files in the search result. The Bloom filter [10] is used as a per document index to track words in each document. The advantage of using an index is that it may be faster than the sequential scan when the documents are large. The disadvantage of using an index is that storing and updating the index can be of substantial overhead. So the approach of using an index is more suitable for mostly-read-only data. Using edit distance, the definition of fuzzy keyword [11] search can be formulated. The execution of fuzzy keyword search returns a set of file IDs whose corresponding data files possibly contain the searched word w. The string matching algorithms was applied to the context of searchable encryption by computing the trapdoors on a character base within an alphabet. Ranked search [12] greatly enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria like keyword frequency. Both system security and usability are achieved by bringing together the advance of both crypto and IR community. The statistical measure approach from IR and text mining to embed weight information (i.e., relevance score) of each file during the establishment of searchable index before outsourcing the encrypted file collection is used here. All provided solution to the secure ranked search over encrypted data problem consist the queries with single keyword. To design an efficient encrypted data search mechanism that supports multi-keyword without privacy breaches remained as challenge and Multi-keyword Ranked Searchable Encryption [13] scheme was introduced. This scheme had chosen the coordinate matching measure for the similarity matching.

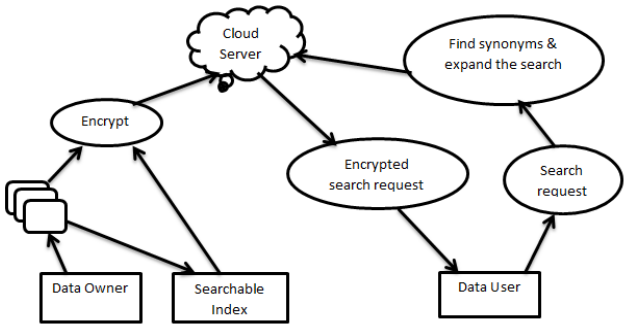

Fig. 1. Architecture of the Proposed System

To enable ranked search for effective utilization of outsourced cloud data, a system model is designed which simultaneously achieve security and performance guarantee. The data owner has a collection of n files and the owner encrypts the searchable index with homomorphic encryption. When the cloud server receives a query consisting of multi-keywords, it computes the scores from the encrypted index stored on cloud and then returns the encrypted scores of files to the data user. Data user can search for relevant files using multi keyword query. When the cloud server receives query consisting of multi-keyword, it computes the scores from the encrypted index stored on cloud, and then returns the encrypted scores of files to the data user. Next, the data user decrypts the received file. The cloud with data storage service system model involves three different entities: data owner, data user and cloud server. Data owner uploads a collection of n text files $F = \{F_1, F_2, F_3, \cdots, Fn\}$ in encrypted form C, together with the encrypted metadata set, to the cloud server. Note that, a corresponding file metadata is constructed for each file. Each file in the collection is encrypted with the symmetric encryption algorithm RC4. Before outsourcing the encrypted data, the owner first build an encrypted searchable index I from F and both the index and file is outsourced to the cloud server. Data user provides a search trapdoor Tw for keyword w to the cloud server. The authorization between the data owner and users is appropriately done. Upon receiving T from a data user, the cloud server is responsible to search the index I and return the corresponding set of encrypted documents. To improve the document retrieval accuracy, the search result should be ranked by the cloud server. The retrieval efficiency is increased by considering the synonyms of the requested keyword. Moreover, to reduce the communication cost, the data user may send an optional number k along with the trapdoor T so that the cloud server only sends back top-k documents that are most relevant to the search query. With the help of access control mechanism the decryption of the received file is done at the user side.

II. SECURING CLOUD DATA

The popularity of cloud computing is increasing very fast and that leads to centralize the sensitive information such as e-mails, personal health records, company finance data, and government documents into the cloud. The motivation to outsource is to reduce their complexity in the data management which provides great flexibility and economic strength. The fact that data owners and cloud server are no longer in the same trusted domain may put the outsourced unencrypted data at risk: the cloud server may leak data

information to unauthorized entities or even be hacked. The security of data storage is one of the necessary tasks to be addressed before the blueprint for cloud computing is accepted. It follows that sensitive data have to be encrypted prior to outsourcing for data privacy and combating unsolicited accesses. The data owner has got a collection of data documents. These have to be outsourced to the cloud server. In order to provide privacy for the data, those sensitive data have to be encrypted before outsourcing. The data owner employs a symmetric key encryption to encrypt the documents before uploading.

The encryption algorithm used is RC4 encryption algorithm [14]. It is a variable key size stream cipher with byte-oriented operations. The algorithm is based on the use of a random permutation. It is remarkably simple and quite easy to explain and also incorporate. A variable length from 1 to 255 is used to initialize 256 bytes state vector, S. At all times S contains a permutation of 8 bit number from 0 to 255. A byte k is generated from S by selecting one of the entries in a symmetric fashion. As each value of k is generated, the entries in S are permuted again. Once the S vector is initialized, the input key is no longer used. Key stream generation involves swapping of the elements in S starting with S[0] and going through to S[255] according to a scheme dictated by the current configuration of S. For the encryption the document is XORed with the value k. The same algorithm is used for the decryption; the document selected by the user from the retrieved files is decrypted by XORing it with the value k. The retrieval of files that matches with the query is obtained with help of an index based searching over the encrypted data.

III. SEARCHING OVER ENCRYPTED CLOUD DATA

The proposed system deals with retrieval of the documents on regarding the keywords requested by the data user. To improve the efficiency in the searching process over the outsourced data an index has to be created for those documents before encrypting it. The function of an index [15] is to provide data owner with an effective and systematic means for locating the document units that are relevant to the requests. The index is a logical view where documents in a collection are represented through a set of index terms or keywords, i.e., any word that appears in the document text. Here for indexing the term selection method is used. The index terms which represent the topic or the feature of the document are extracted. The textual pre-processing is done on the document and the index terms are yielded. Lexical analysis tokenizes the parsed document into words. Next the words mainly the prepositions and articles that got high frequency are removed. The stemming and lemmatization is done for the words left after the stop word removal. A weightage is given for the index terms constructed in account with the significance in the document. Inverted index is used to store the mapping from keywords to the files along with the weightage given each.

In order to maintain the data privacy the index table is also encrypted. This is to avoid the chance of deducing a relation with the index and the outsourced document by the cloud server. Homomorphic encryption is considered here for the protection of the searchable index. Homomorphic encryption [4][16] is expected to play an important part in cloud computing, allowing data owners to store encrypted data in a public cloud and take advantage of the cloud provider's analytic services. Homomorphic encryptions allow complex mathematical operations to be performed on encrypted data without compromising the encryption. To provide easiness in the searching partial homomorphic encryption method is used. Here the somewhat homomorphic encryption supports a limited number of operations i.e. any amount of addition but only one multiplication. SHE is faster and compact than the FHE cryptosystems. The permission for using multiple keywords to search the documents is the highlighting portion of the proposed system.

IV. MULTI-KEYWORD RANKED SEARCH

The core part of the proposed systems lies here. The authorized user requests for a file which is stored in the cloud by giving the keyword to the cloud server. Multiple keywords are allowed here in the request. As the user to keep his search from being exposed the requests are made hidden. The same cryptographic method used for protecting the index is used for the query keywords. The trapdoor corresponding to the keywords is passed on to the cloud server. The cloud sever should not be able to find the relationship between the trapdoors. A ranking function is used to calculate the scores of the matching files to the given search request. To improve the efficiency of the retrieval the synonyms of the requested queries are also considered. The combination of two search semantics, TFxIDF rule [9] and co-ordinate matching with inner product computation is the ranking function used in the proposed system. Term Frequency (TF) is the number of times a keyword present in the files which is used to measure importance of keyword in the file as well as the Inverse Document Frequency (IDF) is found by dividing number of collection by the number of files where terms appearing. Co-ordinate matching [17] is an intermediate similarity measure which quantifies the relevance of the document to the query by considering number of query appearing in the document.

The similarity score between the index term and the trapdoor keyword is as follows:

$$I*T = r(Score(F,Q)+€)+t \quad (1)$$

Here *I* is the index term, *T* the trapdoor function. *Score(F,Q)* is the relevance score of the query *Q* in the file document *F*. This gives the better result of the data users request with most related documents for the query passed. To reduce the computational cost the optional number k is given along with requested query.

V. CONCLUSION

In this paper, the problem of authorized keyword searches over encrypted data in cloud computing is considered. The multiple data owners encrypt their records along with a keyword index and allow searching by multiple users. The records are encrypted using RC4 and whereas the index is by homomorphic encryption. The proposed scheme returns the files including the terms semantically related to the query keyword. The co-occurrence probability of terms is used to get the semantic relationship of keywords in the dataset. The principle coordinate matching which effectively capture the similarity between query keywords and outsourced documents is included in the proposed system. An efficient

and most accurate retrieval of the documents from the cloud is given by the system. This system has got low overhead on the computation works.

REFERENCES

[1] L.M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A Break in the Clouds: Towards a Cloud Definition," ACM SIGCOMM CCR, vol. 39, no. 1, pp. 50-55, 2009

[2] Neha Rawat, Ratnesh Srivastava, Binay Kumar Pandey, Poonam Rawat, Shikha Singh and Awantika Sharma, "Data Security Issues in Cloud Computing", Open Journal Of Mobile Computing And Cloud Computing,Volume 1, Number 1, August 2014

[3] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," Proc. 14th Int'l Conf. Financial Cryptography and Data Security, Jan. 2010

[4] Laurenţiu Burduşel, "New Cryptographic Challenges In Cloud Computing Era" Proceedings Of The Romanian Academy, Series A, Volume 14, Number 1/2013, pp. 72–77

[5] O. Goldreich and R. Ostrovsky, "Software protection and simulation on oblivious RAMs". Journal of the ACM, 43(3):431–473, 1996

[6] D. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," Proc. IEEE Symp. Security and Privacy, 2000

[7] D. Boneh, G. di Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search". In Advances in Cryptology – EUROCRYPT '04, volume 3027 of Lecture Notes in Computer Science, pages 506–522. Springer, 2004

[8] R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," in Proc. of ACM CCS'06, 2006

[9] E.-J. Goh, "Secure Indexes," Cryptology ePrint Archive, http:// eprint.iacr.org/2003/216. 2003

[10] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy Keyword Search Over Encrypted Data in Cloud Computing," Proc. IEEE INFOCOM, Mar. 2010

[11] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," Proc. IEEE 30th Int'l Conf. Distributed Computing Systems (ICDCS '10), 2010

[12] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-PreservingMulti-Keyword
Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOM, pp. 829-837, April, 2014

[13] William Stallings, "Cryptography and Network Security: Principles and Practical" Prentice Hall, 2011

[14] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, "An Introduction to Information Retrieval" Online Edition 2009, Cambridge University Press

[15] Feng Zhao, Chao Li, Chun Feng Liu, "A cloud computing security solution based on fully homomorphic encryption" IEEE 16th International Conference for Advanced Communication Technology, February 2014

[16] J. Zobel and A. Moffat, "Exploring the similarity space," SIGIR Forum, vol. 32, no. 1, pp. 18–34, 1998

[17] I.H. Witten, A. Moffat, and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishing, May 1999