Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Encrypted Data Retrieval in Cloud Computing

K. Sudha, (M.Tech. MISTE), B. Anusuya, P. Nivedha, A. Kokila
CSE Department, Pondicherry University
Pondicherry, India

Abstract- With the advent of cloud computing, data owners are motivated to outsource their complex data management systems from local sites to the commercial public cloud for great flexibility and economic savings. Every owner data is encrypted before stored in the cloud server. Considering the large number of data users and documents in the cloud, it is necessary to allow multiple keywords in the search request and return documents in the order of their relevance to these keywords. We use ranked search scheme (co-ordinate matching algorithm) to find the similar data or information stored in the cloud and also it is used to retrieve the encrypted data. We use many technologies and binary data generation is the process which is used to create index value for the data and store it in the cloud. Data ciphering is the process of creating the cipher text for the owner data which is data owner wants to store in the cloud. Data user access control technique which provides access right to the data user before getting the full detail of the data user. Data user is enabled to create queries to retrieve data by the Data user query. The index value is attached to every data and it is stored in the cloud. If the user wants to retrieve the data from cloud then the given index value is to be matched with the index value in the service provider.

Keywords: cloud computing, multiple keywords, service provider, search request, ranked search

I. INTRODUCTION

Cloud computing is similar to the computer networks which includes the collection of more than computing resources commonly referred as a server and the computing resources are connected through a communication networks such as an internet, an intranet, local area network (LAN)or wide area network(WAN).Instead of using personal computer for every time to application, we can use the cloud to run the application of the user from anywhere at any time in the world and the processing power for the application is provided by the cloud server.

For example, European country user mainly get serves from cloud computer facility in European business hours with a specific application (e.g., email)and also it provides same services to the North American country users in their business hours with a different application(e.g., a web server).

Cloud computing increases the computer usage and reduce the power and also reduce the space needed to maintain the personal computer. Multiple users can save, retrieve and update their own information with single server is possible only because of cloud computing concept. It also allow multiple user to access different application.

The main aim of the cloud computer is too maximize the use of shared resources. In olden days, many organization moving from capex model to opex model to reduce the use of more dedicated PC's for every organization and it makes the organization to use the shared resources and we can pay only for our use. It reduces the maintaining and managing cost for PC's in the organization.

Searchable Symmetric encryption technique is used to encrypt the original data in plaintext into cipher text using key. It provides security to the owner data and then the data is stored on the third party. It also allows multiple users to search the same data at the same time in the cloud.

They provide more security to the outsourced data with the help of the encryption mechanism. These mechanism and controlled search technique does not allow the server to learn the information or get the details of the cloud data is not possible and they also support hidden queries to search secret word in the untrusted server. The algorithms are simple, fast(for a document of length n, the encryption and search algorithms only need O(n) stream cipher and block cipher operations), and reduces the space and communication overhead.

The advantage of the cloud is to allow user to store their data on the cloud and at the same time , it provide more integrity ,data privacy and security to the data on the cloud. The third party auditor is used to manage the outsourced data and also check the data integrity. The efficient Third party auditor must satisfies the main two requirement and they are:1)Third party auditor must audit cloud data efficiently and 2) Third party auditor should not make any weak point in the document stored in the cloud. And also it can perform multiple auditing tasks simultaneously.

In this, we increase the security and privacy of the cloud data and we effectively retrieve large amount of data on demand from cloud and provide the relevance result based on search query instead of getting unwanted result. These ranked search system enables data user to find the most relevant information quickly rather than perform some sorting through every match in the content collection. The ranked search can also reduce the network traffic by back only the most relevant data. For privacy ranking operation does not leak any information.

II. RELATED WORK

Traditional data utilization method is based on the plain text keyword search. On demand users and the large amount of data in cloud is the major problem. Boolean keyword search is also used to search the data in the cloud. But it is not efficient to provide the relevant data from the cloud. It does not provide high security to the data stored in the cloud or service provider. They use an 'inner-product similarity' to store and retrieve the data from cloud.

The major disadvantage in this existing system of the cloud computing is that the security that plays a vital threat. Downloading and decrypting all the data which is stored in the cloud is not possible. Encrypted data search with multi keyword is still a problem in cloud. Security is weak stage because the unauthorized person can able to access the data stored in the cloud or server. Security still lies as a disadvantage over the cloud as unauthorized person would get the access over the data on the cloud.

The distinction between the cloud and grid computing [1] deals with the study of the difference between cloud computing and grid computing with comparing all its features. It also discusses about the concepts of cloud computing. Its aim is to achieve a complete definition of what a cloud is and it pays attention to the distinction of cloud computing with grid computing. It provides scalability and virtualization but does not provide high level services and gives only limited support for availability.

A study about advantages and challenges in cloud adoption [2] tells about the difficulty and complexity in adopting a cloud. It deals with knowing what the problems in cloud are computing. It also discusses the problems in data management, system integration and management of costs. Cloud computing is used for two major reasons. One for its latest trend in Information Technology and the other for data owners could outsource their sensitive data. New classes of applications and parallel batch processing are the main advantages and it does not provide features of interoperability and privacy. Also it is not reliable.

Process of information retrieval [3] tells how to access the archived written information. It deals with string and then retrieving the written information that is being stored. The main idea of the concept is providing automatic access to large amounts of stored knowledge. For this a computer is needed for searching text. It is being proposed by using words as indexing units for documents and measuring word overlap for retrieval. It provides three models for this purpose and they are the Vector Space model, the Inference Network model and the Probabilistic models. The techniques being used are Term Weighting and Query modification also it includes a Cluster Hypothesis. NLP Natural Language Processing is a tool that is used to achieve information retrieval. The information retrieval can be made easier and thus faster information discovery can be achieved. The major disadvantage is that any unauthorized person could also retrieve the sensitive data.

Public key encryption is also done with the keyword search [4]. It studies the problem of searching on data that is encrypted using a public key system. If a mail is being sent from one user to another, the mail has to cross over the gateway. During the transfer of mail the gateway tests whether the email contains any keyword 'urgent' and if so, it routes email accordingly. It tells about the concept of public key encryption with keyword search. It is designed in such a way that the gateway should identify the keyword 'urgent' alone and act accordingly but should not learn anything else from the email. PEKS (Public Key Encryption with Keyword Search) implies the IBE (Identity Based Encryption) but the converse need not apply. It uses an algorithm called Searchable Encryption algorithm and a technique called cipher text security. The searchable encryption and the cipher text security are the advantages of this but it needs support for the gateway to identify the keyword alone without learning anything else from the email.

Privacy-Preserving Public Auditing for Secure Cloud Storage [5] explains the data outsourcing of users into cloud remotely. They need not worry about the problem of data integrity protection in cloud computing. Public auditing for cloud is enabled by a TPA (Third Party Auditor) to check the integrity of outsourced data. The TPA looks after by not bringing in new vulnerabilities towards user data privacy. The two schemes used here are the MAC based solutions scheme and he HLA based solution scheme. IT provides extensive security for the data being outsourced. It gives a very good and fast performance and the other advantages are its storage facility and communication tradeoff. But the downloading data is not possible and also batch auditing could not be done in this system.

Secure Key-Evolving Protocols for Discrete Logarithm Schemes [6] provides security and efficiency of key-evolving protocols. There are two protocols being used here, they are two protocols namely Z_p^* and Z_n^* protocols. The Z_p^* protocol is used for the forward secrecy and Z_n^* protocol is used for the backward secrecy. Both of these protocols are used for achieving the security goals. Z_p^* protocol is used for secret key holder and Z_n^* protocol is used for public key holder. These two protocols use three algorithms namely Public/Secret Key base generation algorithm, Public key evolving algorithm and secret key evolving algorithm. The features that are being satisfied by these two protocols are security, correctness and efficiency. But the disadvantage is that it cannot be applied to any algebraic structure or any other RSA related schemes and also it cannot be applied for any multi party computations.

Searchable Symmetric Encryption [7] SSE allows outsourcing data into cloud providing solutions for several security issues. They present two constructions that show the property of security. Private key storage outsourcing allows either limited resources or limited enterprise to be stored. Private key encryption prevents one from searching over the encrypted data, clients lose ability to retrieve segments of their data. Searchable encryption is provided for symmetric encryption and it is used to balance the need for both privacy and security. The models used in SSE are Private key searchable encryption model and the Public key searchable encryption model. Padding also be done using the concept of SSE. Efficient storage and access of sparse trees could be achieved using this SSE but it depends on size of data that is returned by server and multi user SSE is not possible.

Secure group communication [8] has received much attention with challenges revolving around secure and efficient group key management. Centralized methods are being used for key distribution. Group key agreement is blended by binary key trees with Diffie-Hellman key exchange. The resultant protocol is secure, simple and fault-tolerant. The secure group communication has made group communication and group key agreement possible but it brings complexity in tree management. It uses the TGDH protocols.

New techniques for remote searching on encrypted data using a distrusted server and provide proofs of security for the resulting crypto system. There are many advantages and they are provably secure, hidden search and query isolation. Simple and fast, very flexible. It is desirable to store data on data storage servers such as mail servers and file servers in encrypted form to reduce security and privacy risks. These techniques have a number of crucial advantages.

III. TECHNIQUES AND ALGORITHM

Some of the models, techniques and algorithms being used in the existing system are discussed and summarized as follows.

A. *Vector Space Model*: This model is used to represent the text by a vector of functions. The terms are the words and phrases. If words are considered as terms, every word becomes an independent dimension in a very high dimension vector space. If term represents a text, it gets a non- zero value in the text-vector along the dimension corresponding to the term. Text vectors are very space and no term is assigned a negative value.

B. *Probabilistic Model*: The principle of probabilistic model is that the documents in a collection should be ranked by decreased probability to query relevance. This principle is called as the probabilistic ranking principle. The ranking criteria is monotonic under log-odd transformations. Each probabilistic model that is proposed is based on a different probabilistic estimation technique.

C. *Inference Network Model*: A model that is used for a document to instantiate a term. The credit from multiple terms is accumulated given to compute the equivalent of a numeric score for data.

D. *Term Weighting* : Term weighting is a technique that rely upon the better estimation of various probabilities. The main three factors play in term weight formulation are

1. Term Frequency – Words that repeat multiple times in a document.
2. Document Frequency – Words that appear in many documents are considered common.
3. Document Length – When collection have documents of varying lengths, longer documents tend to score higher since they contain more words and more repetition.

E. *Query Modification*: Query modification is a technique which found a solution when it was hard for users to formulate effective search requests. It is done by adding synonyms of query words to query would improve search effectiveness. The synonyms relied on a thesaurus. So they developed a technique called query modification to automatically generate thesauri for query modification.

F. *Cluster Hypothesis*: Cluster hypothesis is a true fact which states that document that cluster together will have a similar relevance profile for a given query. So document clustering techniques were being used to improved search effectiveness.

G. *NLP-Natural Language Processing* : NLP is a tool that is used to enhance retrieval effectiveness but had only very limited success. The advantages are easier and faster information retrieval and information discovery.

H. *Searchable Encryption Algorithm*: An algorithm that consists of the polynomial time randomized algorithms. They are

1. KeyGen(s): s is a security parameter taken and used to generate a key pair either public or private.
2. PEKS(A_{pub} ,w): A_{pub} is a public key and w is a word which are used to produce a searchable encryption.
3. Trapdoor(A_{priv} ,w): A_{priv} is a private key and w is a word which are used to produce a trapdoor T_w.

I. *Cipher text Security*: It is a technique that is used to provide security for the encrypted data. A cipher text attacker could easily break semantic security by reordering the keywords and submitting the resulting cipher text for decryption. A standard technique is used to break this and this technique is called the cipher text security.

J. *MAC Based solution scheme* : It just uploads the data blocks with MACs to server and send the secret key to TPA. For this,

1. The number of times a particular data file can be audited is limited by the number of secret keys that must be fixed a prior.
2. The TPA also has to maintain and update state between audits.
3. It can only support static data and cannot deal with dynamic data.

K. *HLA Based Solution Scheme*: This is a scheme to find solutions for effectively supporting public auditing without having to retrieve the data blocks themselves. The difference is that HLAs can be aggregated and it becomes an advantage of this.

L. Z_p^* *Protocol*: It is a protocol based on difficulty of computing discrete logarithms in Z_p^*. It applies the technique of Feldman and also uses the secret key sharing scheme of Shamir. The secret key holder executes key base generation algorithm and publishes the public key base. He publishes a hash function that works as a simple randomizer and takes the index of a number. The public key holder retrieves and verifies it for assurance.

M. Z_n^* *Protocol*: It is a protocol based on difficulty of computing discrete logarithms in Z_n^* , where n is the product of several large primes. It uses the technique of Maurer-Yacobi which is suggested by Anderson to build a forward secret signature. Factoring n is hard and g is element of maximal order in Z_n^*.

Both the Z_p^* and Z_n^* protocols use three algorithms in each of them. They are

1. Public/Secret key base generation
2. Public key evolving algorithm
3. Secret key evolving algorithm

N. *Private Key Searchable Encryption*: A model called private key searchable encryption is used to search on a private key encrypted data. The user himself encrypts data, so as to organize in an arbitrary way.

O. *Public Key Searchable Encryption:* Public key searchable encryption is a model that allows user to encrypt data and send it to the server. The owner provides decryption key may be different.

P. *Padding:* Padding is defined as a search pattern, the total size of the encrypted document collection, and the number of documents it contains. To achieve this a certain amount of padding is done to the array and the tables that are necessary.

Q. *TGDH Protocols*: Each group member shares to group key, a function of current group member. This share is secret and is never revealed. When group grows, new members' share into group key but old members' share remain unchanged. When group shrinks, departing members, share are removed from new key.

R. *TDGH Events*:

1. Join protocol – a new member is added to group
2. Leave protocol – a member is removed from group
3. Merge protocol – a subgroup is added to group
4. Partition protocol – a subgroup is split from group
5. Key refresh protocol – group key is updated

Table1 Comparison of cloud techniques

PAPER	**ADVANTAGES**	**DISADVANTAGES**
A Break in the Clouds: Towards a Cloud Definition	Scalability Virtualization	No high level services Limited support for availability
Advantages and challenges of adopting cloud computing from an enterprise perspective	New classes of applications Parallel batch processing	Interoperability Privacy Reliability
Modern information retrieval: A Brief Overview	Easier information retrieval Faster information discovery	Unauthorized person retrieving data
Public Key Encryption with keyword Search	Searchable Encryption Cipher text security	Needs support for gateway to identify keyword alone
Privacy-Preserving Public Auditing for Secure Cloud Storage	Storage Communication tradeoff Security and performance	Downloading data is not possible Batch auditing
Secure Key-Evolving Protocols for Discrete Logarithm Schemes	Correctness Security Efficiency	Not applicable for algebraic structure and RSA schemes No multiparty computations
Searchable Symmetric Encryption: Improved Definitions and Efficient Construction	Efficient storage Access of sparse tables	Depends on size of data being returned Multi user SSE
Simple and Fault-Tolerant Key Agreement for Dynamic Collaborative Groups	Group communication Group key agreement	Complexity in tree management

IV. CONCLUSION

We overcome the disadvantages of the existing paper and also improve the security features of the cloud. We encrypt the original data before outsource it using advanced encryption technique and retrieve the original data by using co ordinate matching algorithm. This algorithm allows only user to retrieve the original data and it does not allow the unauthorized person to retrieve the data. It provides more security to the original data.

REFERENCES

[1] "A Break in the Clouds: Towards a Cloud Definition".Luis M. Vaquero, Luis Rodero-Merino, Juan Caceres, Maik Lindner.
[2] "Advantages and challenges of adopting cloud computing from an enterprise perspective" Maicela-Georgiana Avram (Olaru).
[3] "Modern information retrieval: A Brief Overview" Amit Singhal, Google, Inc.
[4] "Public Key Encryption with keyword Search" Dan Bonch, Giovanni Di Crescenzo, Rafail Ostrovsky, Giuseppe Persiano.
[5] "Privacy-Preserving Public Auditing for Secure Cloud Storage" Cong Wang, Member, IEEE, Sherman S.M. Chow, Qian Wang, Member, IEEE, Kui Ren, Senior Member, IEEE, and Wenjing Lou, Senior Member, IEEE.
[6] "Secure Key-Evolving Protocols for Discrete Logarithm Schemes" Chenf-Fen Lu and ShiuhPyng Winston Shieh
[7] "Searchable Symmetric Encryption: Improved Definitions and Efficient Construction" Reza Curtmola, Juan Garay, Seny Kamara, Rafail Ostrovsky.
[8] "Simple and Fault-Tolerant Key Agreement for Dynamic Collaborative Groups" Anonymous submission.
[9] "LT Codes-Based Secure and Reliable Cloud Storage Service," N. Cao, S. Yu, Z. Yang, W. Lou, and Y. Hou.
[10] "Cryptographic Cloud Storage," S.Karmara and K.Lauter.
[11] "Deterministic and Efficiently Searchable Encryption," M. Bellare, A. Boldyreva and A. ONeill.

