IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 5 Issue 6 June, 2017

# Cloud Based Secured Top Ranked Document Identification & Privacy-Preserving Keyword Search

**Kavya sri.M[1], Vidhya Prakash.R[2]**

[1]PG Scholar, Department of CSE, R.M.K. College Of Engineering, Anna University, Kaveraipettai, TamilNadu, India
Email-id:sri.anu94@gmail.com
[2]Assistant Professor, Department of CSE, R.M.K. Engineering College, Anna University, Kaveraipettai, TamilNadu, India
Email-id:rvp.cse@rmkec.ac.in

## ABSTRACT

*Cloud data owners store their data on cloud in an encrypted format to maintain privacy. Therefore various search techniques on encrypted data where found. Previously on the existing paper, the relationship between the documents are considered and the hierarchical clustering method is created in order of grouping the same category of documents together. This makes a huge difference in the search time. This system involves, when the data owner uploads the data into the cloud server it is divided in to four parts encrypts it and stores it in four difference cloud servers and the replica of each part is created and stored separately. The index of each file in created and each index consist if the main keywords and its corresponding threshold value which facilities for searching the document using the keyword which user enters and it also impacts the search time .The search time is variably decreased because of the usage of index. The top k value is used to display the top occurrences of the documents in the cloud server according the user's interest. The system makes use of TTP and MHT to validate and verify the data.*

**Keywords: Encrypts, index, replica and privacy**

## 1. INTRODUCTION

Theenterprises and users who own a large amount of data choose to store their data in the cloud to cut down the cost of storage facility and the data management. As the result the cloud is flooded with proliferating rating data and the data in the cloud is increased drastically. The cloud provides various security measures but the traditional method of decreasing information leakage and maintain data security is encrypting the data. There are various types of encryption and this encryption becomes the challenging task [1], [2].on the other and the search technique of cipher text also becomes challenging task. Various techniques are introduced [3], [4], [5]. These methods have proven security but there is a lot of time complexities. The relationship between the documents are taken into consideration and these relationships are taken as the properties of the documents [1].the category of each document is identified by its relationship with other documents and its properties. By this way the documents are grouped together in order to search and trace any particular document. The proposed system, which involves the data owner uploads or stores its data into the cloud server when the data is uploaded the index of each document in created and stored. This Index documents consists of the man

Keywords of that particular document and its corresponding threshold values using the stemming algorithm where else here threshold value represents as the number of occurrences of the particular keyword in that document. After that the document is split into four parts and encryption of AES algorithm takes place and stored in to four different cloud severs. On the other hand, once the document is split into four parts using the MHT it produces the hash value of the parts separately in data owner store that values in the TTP. The user



 

who wants to search a document from the cloud server should become authenticated before searching the document in thecloud. When the user enters the particular keyword it is also entered the topk value which retrieves the top occurrences of the keyword. When 5he user enters the top k value which represents as top number of results that particular key word may occur in any number of files .That particular keyword will connect to the cloud server and the cloud server will download the indexes f the files that matches the keyword with the index files. The index files which matches are retrieved separately and ranks itself according to the threshold value of the particular keyword which facilitates in displaying the topk results. When the user enters the particular file to be downloaded the request is sent to the data owner for the approval of the download .once the approval is done .A key is sent to the mail id of the user .and by entering that particular key the user can download the part car data which is uploaded by the data owner. While downloading it connects itself to the cloud server where the data are stored into four separate cloud server retrieves the data which is encrypted format and combines together and decrypts it again and retrieved to the user. And the verification of the data is done using the TTP. Previously the data owner while uploading the data it produces the hash value of the four different parts and it is saved into the TTP now for the verification purpose when the document is requested for download it goes to the cloud server and takes anyone part of the particular document and generate the hash value. If the hash value which is generated is equal to the initial hash value that is stored by the data owner in the TTP the data is not corrupted or lost. There is no data loss. If the hash values are not equal then there is data loss or data corruption.

## 2. SYSTEM MODELS

The system which includes data owner and user registration, cloud server, keyword ranked search, encryption and data splitup, replication, retrieving the data , trusted third party and MHT, and data verification.

### 2.1 Data Owner and User Registration

The data owner who is flooded with large amount of data chooses to store the data in the cloud server which helps in data management so the data owner enter into the cloud and configure cloud according to the need it can create any number of cloud and name it and save the documents into the separate clouds. In this system the privacy of the documents that is stored in the cloud is maintained. Without the approval of the admin which means cloud server no user is allowed to retrieve the document of the data owner .And it makes sure the data user is an authenticated one. Once after registration only it is allowed to search the documents in the cloud using the keyword.

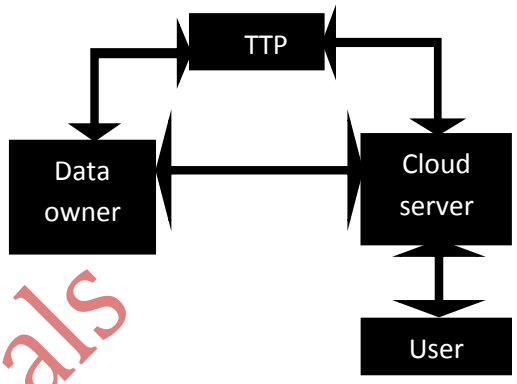


**Figure1.The system process**

### 2.2 Cloud Server

The cloud server is the server which works as an admin for both the data owner and the user. This is the place where the data owner stores and manages the large amount of day. Now a days in this modern world due to various technologies and social network the large amount of data is produced day by day.an d the data are considered to be entered into the bug data era and we have to design the storage are of such data which is called as cloud. Many enterprises and users who are proliferating data makes use of the cloud for their storage. This cloud server is the medium for uploading, storing and retrieving the data.

### 2.3 Keyword Ranked Search

Once the file is uploaded by the data owner the index of the file is created and stored into the drop box. The index file consists of the main keywords of the document and the corresponding threshold values. And this is done using the stemming algorithm or the word stemmer .When the document is uploaded the word stemmer removes all the conjunctions in the document and retrieves

only the main key word and it makes the count of the repetition of that particular word throughout the document. And that repetition count is saved as the threshold value of the keywords and this complete collection of main keyword and the threshold values are stored in the index file which is created for the document.Likewise when each document is uploaded by the data owner an Index file is created.

When the user enters the keyword to be searches the user will also enters the topk value which is labeled as search .This top k value is used to display the top results of the search result. When user enters the keyword it is not possible to match the keyword to every word of all the documents which is stored in the cloud. And that will result in time complexity. So when the user now enters the keyword that particular keyword may occur in any number of files in cloud server. The keyword will be sent to server and the server in turns connect with drop box and downloads the indexes of the files that is stored into the cloud. Since the index file consists of the keywords and threshold this the word is matched with the index files and retrieves the files where the keyword id exactly matched.

Now there will be a lots of files that will be matched now using the threshold value of the same word in every file it ranks itself from the higher number of occurrences to the lower number of occurrences. The ranking is done by the total number of search term frequency divided by the total number of the keyword in that index. For each search of the user it is not necessary to display all the files in cloud server that matches the entered term. It is enough to display the top k results according to the user's interest. This is referred as keyword ranked search.

## 2.4 Encryption and Data Splitup

The data which is uploaded by the data owner is split in to four parts and it is encrypted using the AES algorithm which uses the block size of 128 bits. The steps involved in the advanced encryption standard algorithm are the key expansion, initial round which consists of four steps subbytes, shiftrows, mix columns and addroundkey and the final round which consists of three steps they are subbytes, shiftrows and addroundkey. This AES algorithm will be the most secured form of encryption. Each bit in the file undergoes 64 rounds [6].Using this encryption techniques the data which are split into four parts are encrypted and stored into the separate cloud servers.

## 2.5 Replication

The data owner when uploading the file the file is split into four separated pats and stored into four separate cloud server in which the replica for parts of the cloud server and it is stored separately. When data is corrupted or lost the document that is stored in the replica server will be retrieved and displayed to the user. The replication is used to achieve data availability of the cloud.

## 2.6 Retrieving the Data

When the user enters the particular keyword and topk value the server results with the documents according to the value. From the displayed value user selects the document which is need to be downloaded. By clicking the download the file will not be directly downloaded the request will be passed on the admin for downloading the file. The admin should approve the file download request .once after accepting the file download request it checks whether the user is an authenticated person by the cloud server .thus it sends a key to the user's mail-id. When clicking on download the user will be asked to enter the key that is send to him. Once he entered the key.

If the key is a valid it will retrieve the divided four parts of the requested document and merged together and decryption takes place and then user requested documented will be downloaded .if the key is not valid the user cannot retrieve the requested file. This mechanism makes sure the privacy of the document is preserved.

This uses the HMAC SHA algorithm. This HMAC SHA uses a key $k_i$to transform an input array of bytes a[ ]. The key kiis the secret one and never be accessible to any hacker and the input is the challenge. So the key is challenge response authentication. This secret key is 20 bytes at least and the challenge is usually counter of 8 bytes whichleaves quite some time before value is exhausted [7].

This algorithm is generally used in the platform that required authentication. This is a hashing algorithm that converts a set of bytes to another set of bytes. And this algorithm is irreversible. It uses MD4 and MD5 hash algorithms. This algorithm takes 20bytes key and 8 bytes counter to create an

8 digits number. And this OTP can only be valid to a couple of minute.

### 2.7 Trusted Third Party and MHT

When the data owner uploads the file and when it is divided into four parts .The hash value is generated for the four parts of the file separately using merkle hash tree(MHT). And the generated hash value is stored in to the trusted third party at the initial stage of uploading the file. The input file is taken as the steam of binary tree and splits the files in to different parts of different size, for all the odd number of chunks choose the non-zero value to complete the pair. For rest of the chunk produce a hash digest value .Now arrange again according to the order of the file content [8].This is how the hash values of each parts of the file is generated and stored into TTP.

### 2.7 Data Verification

In this system the data can be validated or verified to check whether there is any loss of data. Sometimes data owner can upload more than one file at a same point or the files are stored into the cloud for a very long time and there are so many constraints are available and there so many reasons for data loss and data corruption .and this can be found using the TTP. When the user request a particular file, this TTP retrieves the any one part of the file parts of that document stored in the cloud server. Taken that one part as the input the TTP using the MHT generates a hash code for that particular part and it matches that code with the already generated code which is saved into TTP by data owner while uploading the file. If the hash code matches, then the data is not corrupted if not matched the particular file is corrupted and it refers the replica for that particular part retrieves it and displayed to the user. This is how data verification takes place.

## 3. PROBLEM STATEMENT

Previously, lot of search techniques over encrypted data are introduced and those techniques were time consuming. The search time of all those techniques are very time consuming and the search time is increasing exponentially. And again there introduces a method of hierarchical clustering of data where the relationship between the files are considered and are clustered together and search is done which reduces the search time to a level of linearity from the the exponential level. But still this is not enough because cloud is the platform for large amount of data which is expending exponentially day to day .And every such extension of the data sets it requires some technique where the search time is minimal so that the user may use the technique for the easy retrieval. In this system which uses the indexes to search the document. The index is created for each and every document in the cloud and it helps in matching the keywords. This decreases the search time variably to a great extent and the privacy preserving technique is also well considered .Thus the user retrieved file is privacy preserving.

## 4.THREAT MODEL

The adversaries' ability can be found in the following two threat models.

*Known cipher model*. This model states that the cloud server can get the collection of three. They are encrypted document collection, encrypted data index, and encrypted query keywords.

*Known background model*. Here he cloud server knows more information than the previous model. To launch statistical attack to identify specific keyword the document frequency and term frequency information are known by the cloud server[9], [10], which in turn relives the content of the plain document in these two threat models[1].

## 5. DESIGN GOALS

- Search efficiency. The time complexity of the search time will be varied according to the index files in the drop box.
- Retrieval accuracy. The relevance to the keyword involves with relevance precision
- Integrity of search results: The three aspects which is based on includes:

1) Correctness. If the documents are corrupted the returned results will be Sam as uploaded using replica.
2) Completeness. All the indexes all taken into consideration.
3) Freshness. The latest documents are only returned.

- Privacy factors. The requirements of privacy are as follows.

1) Data privacy. Data privacy represents protection and privacy of data. The user can

search the data only after documents and only after the approval of the cloud server the required document is allowed to download by the user.

2) Index privacy. Index privacy was maintained by the admin in the droop and it can only be accessed by the admin itself.
3) Rank privacy. Rank privacy is ordering of documents according to threshold value entered by the user as the top k value. Document privacy. All users cannot download the document of their requirement. They have to make a request to the cloud server to for downloading the required document. And once the request is received by the server it checks with the user authentication and approves the request. And a key is sent to the user. The user have to check his mail for the key. The key is received by the user only if he is an authenticated user. And only after enter that particular key the document can be downloaded.

- Security. The admin will maintain its own set of elements. And the user who searches the documents with keyword will also have its corresponding credentials and the document to be downloaded will be downloaded only after the approval.
- Authentication. The user who is searching the documents can only retrieve the documents only after the proper signing up in to the server.
- Confidentiality. The documents uploaded by the data owner can only be downloaded by the data user only after the approval.
- Authorization. The approval of the document is all needed the key which I s sent to the data user is to be entered into the server to retrieve the document.

## 5.SECURITY ANALYSES

### 5.1 Multi-User Request Attack

When the user wants to download a document it have to request the cloud server for the approval. When multiple user enters and get authenticated by the server. The request is increased and the cloud server only after approving all the request that is sent by the multiple users after that only the documents are allowed to download. The entry of multiple user can never change the system process. Even the users gave request or a single user gives more than one request at the same point of time the process remains the same.

### 5.2 Impersonation Attack

Only the registered users are allowed to entered the system and search for the required documents using the keyword that after a specific authentication process. The person who is not registered with the system will not be allowed for search.

## 6. IMPLEMENTATION

This implementation which consists of a laptop which includes the data owner side and clouds server which drop box is used and the data user is also enrolled. The laptop has a 2.5 GHz CPU processing speed and 4 GB ram. For demonstration purpose all the cloud, data owner and data user are enrolled in the same laptop and it is connected via a common Wi-Fi access portal .all the process in the cloud system cannot be proceeded offline. All the process are possible only when all the systems are connected to the internet only in the online mode.

## 7. EXPERIMENTAL RESULTS

The system is experimented and the implementation of the system process is given as follows in the snapshots. And this is developed in Net Beans platform. Because of the enhancement of trusted third party and the document verification in the system the user is benefited in a best way. The user is given with the option of validating and finding the loss or corruption of data in the application. Configuring cloud and connecting to drop box are shown in Fig.1.The Drop box connection and Uploading file are shown in Fig.2.The Index and User signup/login are shown in Fig.3.The User signup and Login are shown in Fig.4. Search box and Search result are shown in Fig.5.Download request and Admin login shown in Fig.6. User request and Approval are shown in Fig.7. Key entering and validating document are shown in Fig.8.



 


**Figure1.Configuring cloud and connecting to drop box**

**Figure2.Drop box connection and Uploading file**


**Figure3.Index and User signup/login**


**Figure4.User signup and Login**

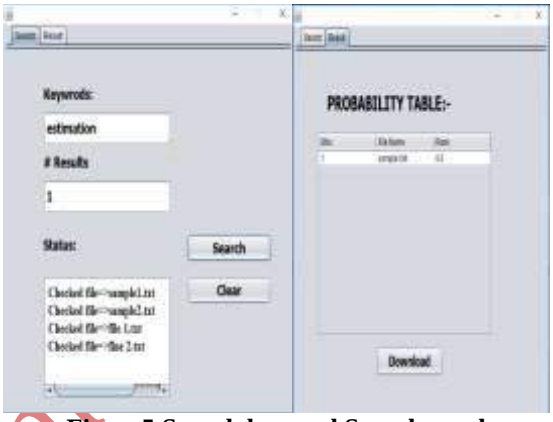


**Figure5.Search box and Search result**


**Figure6.Download request and Admin login**

**Figure7.User request and Approval**

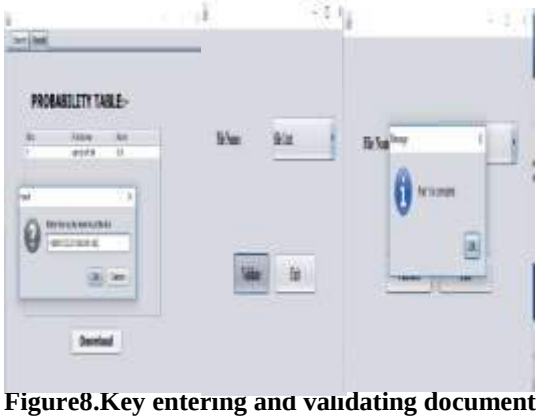

**Figure8.Key entering and validating document**

## 8. CONCLUSION

The investigation of cipher text search in the scenario of cloud storage. The problem of maintaining the semantic relationship between different plain documents over the related encrypted documents and give the design method to enhance the performance of the semantic search. The MRSE-HCI architecture to adapt to the requirements of data

explosion, online information retrieval and semantic search is also proposed. At the same time, a verifiable mechanism is also proposed to guarantee the correctness and completeness of search results. In addition, we analyze the search efficiency and security under two popular threat models. An experimental platform is built to evaluate the search efficiency, accuracy, and rank security. The experiment result proves that the proposed architecture not only properly solves the multi-keyword ranked search problem, but also brings an improvement in search efficiency, rank security, and the relevance between retrieved documents.

## 9. FUTURE ENHANCEMENTS

From the result, it is evaluated the system efficiency by extensive real experiments and show that the search time is relatively fast. It is also found that the reliability and accuracy of search result can be improved future by incorporating more techniques into the system and also the security of the cloud environment can also be improved.

## 10. REFERENCES

[1] Chi Chen Jiankun and Hu, "An Efficient Privacy-Preserving Ranked Keyword Search Method," Proc. IEEE Trans. Parallel And Distributed Systems, Vol. 27, No. 4, April 2016.

[2] D. X. D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. IEEE Symp. Security Priv., BERKELEY, CA, 2000, pp. 44–55.

[3] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search[C]," in Proc. Adv. Cryptol., Berlin, Heidelberg, 2004, pp. 506–522.

[4] D. Cash, S. Jarecki, C. Jutla, H. Krawczyk, M. Rosu, and M. Steiner, "Highly-scalable searchable symmetric encryption with support for Boolean queries," in Proc. Adv. Cryptol,. Berlin, Heidelberg, 2013, pp. 353–373.

[5] S. Kamara, C. Papamanthou, and T. Roeder, "Dynamic searchable symmetric encryption," in Proc. Conf. Comput. Commun. Secur., 2012, pp. 965–976.

[6] https://en.wikipedia.org/wiki/Advanced_Encryption_Standard.

[7] http://www.codeproject.com/Articles/592275/OTP-One-Time-Password-Demistified.

[8] http://www.swirl-project.org/content/merkle/

[9] C.Wang, N. Cao, K.Ren, and W. J. Lou, "Enabling secure and efficient ranked keyword search over outsourced cloud data," IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 8, pp. 1467–1479, Aug. 2012.

[10] A. Swaminathan, Y. Mao, G. M. Su, H. Gou, A. Varna, S. He, M. Wu, and D. Oard, "Confidentiality-preserving rank-ordered search," in Proc. ACM ACM Workshop Storage Security Survivability, Alexandria, VA, 2007, pp. 7–12.