International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Cryptographic Data Retrieval from Cloud by Hashing and Indexing Technique

Arpita Porwal, Deepak Shukla

Department of Computer Science & Engineering, IES, IPS Academy, Indore, Madhya Pradesh, India

ABSTRACT

Since in this era, Cloud computing becoming more and more popular by accelerating huge quantity of data storage on server space i.e. CLOUD. The data owner uploads data on cloud space for increasing convenience and reduces the lot in organization of the data. The private data must be encrypting before outsource. The data which is being uploaded on cloud space is of supreme priority to enable the encryption of the data. Here the paper initiates the analysis about the similar approaches that are recently developed for cryptographic data retrieval from cloud. The system uses concept of TF-IDF model by which effective keywords are being extracted from text files that are repeatedly occurs, and their hashes get stored in Inverted Index. Thus the feature makes data accessible by searching through keywords, we propose model, by using SHA1 (Secure hash Algorithm) hash generation methodology, and DES (Data Encryption Standard) algorithm for the encryption of data. For the relevant retrieving of data from cloud, KNN (K-nearest neighbour) based searching is used. The proposed scheme achieves the identical security level by comparing the system with the existing ones and better performance. By doing this query complexity, functionality and the systems efficiency get improved.

Keywords: Cloud Computing, Storage Infrastructure, Cryptographic Cloud, Secure Search, Keyword Based Search.

I. INTRODUCTION

Cloud computing a place where vast number of systems is connected to a single pool so that dynamically scalable infrastructure could be made available for privately or publicly for data or file storage on the network. Cloud allows customers to store the data and information remotely to the on-demand cloud servers. Cloud computing uses hardware and software to provide a service in the network. In place of a local server or a private computer, the data users get benefit from the on-demand network anywhere for using dynamically network of hosted remote servers for managing, storing and processing data. By acquiring and reducing the economic upper part, you can easily access the common pool of resources for configurable computing.

Cloud computing fall in the 3 broad categories i.e. SAAS (Software as a Service) is a consumer model. It delivers services but doesn't manage and control. PAAS (Platform as a Service) can be used for controlling applications, configuration settings and for the deployment of the applications on cloud server. IAAS (Infrastructure as a Service) work as a host. As it controls operating system, storage and application deployment and select networking components [7].

A. Data Storage

Now a day, it is immense issue of storing data on remote system for long term accessing. The new opportunity for the long time data storage is introduced, known as cloud computing. Data Storage means it kept the records of data, files and information on cloud space for the effective and efficient storing of data and information. It helps in reducing the memory space of the remote system. The data and information will be preserved in text, audio, video, files and tabular form.

B. Cloud Storage

Cloud storage is a resolution for very long term data or information storage and the memory management. This structure can be used for storing the digitalized data in logical pool, and the environment is physically owned and maintained by the hosting companies. Cloud storage looks like saving data over off-site storage system and may be maintained by third party. In spite of using the hard-drive or the other local storage devices for storage, save the data and information on cloud. [8]

C. Cryptography

Cryptography is that study of concealing the data files and information for protected communication with the availability of unknown parties. Security is obtained by electronic messaging and data encoding to make non-readable for intruders and non-data house owners. [6][9]

D. Cryptographic Cloud

In cryptographic cloud the data must kept in the cryptographic or non-readable format which suggests that can't be used directly by other users. It designed for virtual secret storage devices for using knowledge and data for achieving security between the multiple access networks. [10]

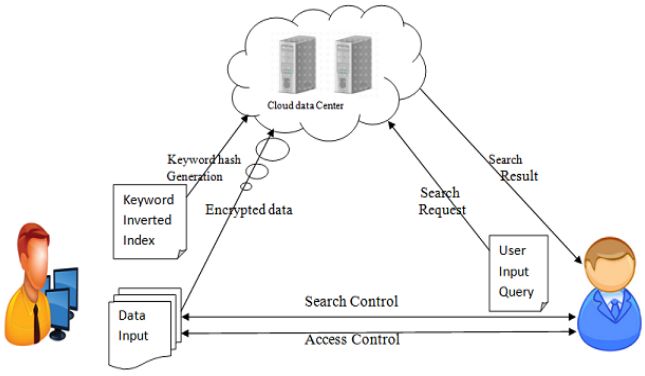

Figure1 Cryptographic Cloud

Advantages of Cryptographic Cloud:

- Resource sharing and data management can be provided.
- This server helps us for providing data integrity, security and confidentiality.
- It can ease the hardware, software cost and system maintenance overheads.
- For preserving data over cloud in confidential format.

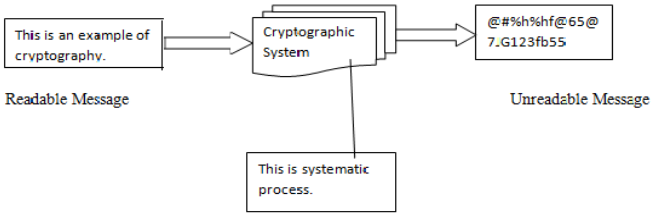

Figure2 Cryptographic System

II. RELEVANT WORK

This section includes the different approaches and techniques that are recently developed by supporting the proposed approach of cryptographic data searching.

Zhihua Xia et al., [1] A paper present a search plan using multi-keywords for accessing encrypted data from cloud server, that dynamically supports deletion and entry of documents. Not only has it supported accurate Multi-Keyword Rank Search (MKRS). The paper suggest the "greedy depth first search" algorithm for attaining better and efficient search results using special keywords for comparing balanced binary tree to linear search. The primary idea for Multi-Keyword search plan is base on a secure internal product computation and then provides major improvements in the MRSE plan to gain more confidentiality.

Zhangjie Fu et al., [2] in progress cloud computing state; keyword-based search information for confidential outsourced data becomes significant tool for data retrieval. Existing techniques focused on multi-keyword exact match or single keyword fuzzy based search. They tried to create multi-keyword fuzzy search plans. The necessity of multi-keyword fuzzy search for assembling hash function by using bloom-filtering and locality-sensitive was accounted by Wang. A new means of keyword was introduced on the idea of uni-gram, which might facilitate in the accuracy to seize other spelling mistakes. Stemming algorithm used basic keywords, which can consider keyword weight while selecting a suitable matching file set.

Sumathi Sivaraj et al., [3] Worked for existing, sensitive data on cloud computing and slowly centralized data over cloud. Before outsourcing private data it should be protected by encryption, the two disadvantages of this approach are not having pre-knowledge of cloud data, additionally to presenting the duplicate file with the identical keyword with clearing the large data to that.

Ning Cao et al., [4] Search encryption with single and Boolean keywords that rarely sorts search results. To preserve this confidential data on cloud by maintaining

confidentiality through Multi-Keyword Search Algorithms, they have selected from coordinate matches. Multi-Keyword Text Search is based on Equality-Based Ranking, which helps to match most of the matches possible to find relevant data.

Mayank Kudle et al., [5] for the recovery of data and documents it's necessary that it accumulates the searcher to the database holder, although there's sensitive data that ought to be store secret. The confidentiality should be maintained of document search with history. The author suggested ranking keyword which allows multiple keyword ranking that permits inquiries to be store confidential for maintaining user and its data privacy.

These give an overview of different approach and method that are frequently used for keyword based data retrieval.

E. *Text Mining & Its Techniques*

The method of obtaining high-quality information such as statistical process and preparation of patterns from text is known as text mining. This method allocates to highlight frequently utilize keywords in texts.

Information such as extraction, classification, clustering, visualization, and summarization can be used for analyses, understands and generates text through natural language resource processing.

1. *Information Extraction*

Information extraction for the computer is the primary step to identify unread text by identifying keywords and relevance within the text. This model is used for the matching i.e. predefined sequences in the text. It includes tokenization, identification of named institutions, punishment split, and part-of-speech assignment.

2. *Categorization*

The technique of supervised learning helps in classifies free text documents which are based on input-output for identifying the new documents. It will be used for categorise the document on the basis of their common property.

3. *Clustering*

The clustering is used for finding a group of documents with the similar content. A number of such documents form a cluster and the cluster is much more different than the quality of the clustering, they are considered better. Since for maintaining and managing thousands of organizational data clustering is used.

4. *Visualization*

The basis of the visual hierarchy is to keep the visual text for the improvement and simplification of the relevant information. The government used information visualization to recognize the terrorist network and for getting information about crimes through zooming and scaling in documents.

5. *Summarization*

Summary of documents is used to assemble the needs of user rather on set of documents. Summary and highlighting of main points are incorporated in following steps:

- To get structured representation of the original text by pre-processing to apply in subsequent step.
- Obtain the last summary from the summary structure.

F. *Text Feature Extraction Techniques*

1. *Bag-of-words (BOW)*

The algorithm is used for counting how frequently a word occurs in a text file or a document. Preparing for investment in a deep-learning network.[10]

2. *Term frequency/inverse document frequency (TF-IDF)*

The method frequently emphasizes words on a given file or a document, while deemphasising the repeated words in many documents at the similar time. [11] Then TF-IDF is calculated

$$\textbf{tf-idf=tf(T,D)*idf(T,D)}$$

3. *N-gram*

An n-gram is a contiguous sequence or co-occurring of words in given text or speech.[11]

If X=Num of words in a given sentence K, the number of n-grams for sentence K would be:[11]

$$\mathbf{Ngrams_k = X - (N - 1)}$$

G. *K-nearest-neighbor (KNN) algorithm*

The KNN algorithm is the simplest in the entire machine learning algorithms. Non-parametric techniques used for learning and classification based on example. [10]

The algorithm approximate the distance between fixed scenarios in the database and processes the query set. A distance function has been estimated between the distance using the D (X, Y), where x, y is the landscape developed through the facilities [12]

$$X = \{x_1, x_2, x_3, \dots\}$$
$$Y = \{y_1, y_2, y_3, \dots\}$$

The frequently used distance functions are absolute distance measuring using:

$$\mathbf{d_A(x,y) = \sum_{i=1}^{N} |x_i - y_i|}$$

And Euclidean distance measuring with:

$$d_A(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$$

III. REVIEW OF TECHNIQUES

H. Index and Inverted Index

An index is a listing of words that appears frequently in the document, whereas an inverted index is a listing of those words, in addition to documents or the files in which they appear. The inverted index contains all unique words in the database which appear in the many documents.

How to create Inverted Index?

Creating an Inverted Index to maintain any Search System you must follow several steps while parsing pages or documents.

Steps to build an inverted index:

Step 1Taking the document.
Step 2 Remove the stop words.
Step 3 Stem to root words.
Step 4 Record the document Ids into the list.
Step 5 Merge and store the terms.

I. Indexing and Hashing

Indexing is the simplest and easiest way to sort out many records on many areas, while hashing is used to index and retrieve keywords from the database for the finding of data from database by the hashed key for getting the original value.

1. Types of Hashing Techniques:

1. *MD-5:* MD-5 is Message Digest algorithm being invented by Ron Rivest and its digest length is 128 bits / 16bytes and the processing unit is 512 bits. It does 80 steps (20 rounds) for processing and the maximum message size is 264 - 1 bits. MD5 is cheap but it is less secure.
2. *SHA-1:* Secure hash algorithm being invented by National Institute of Standards and Technology (NIST) for the safety of digital signature algorithms (DSA), its digest length is 160 bit / 20 bytes and processing unit is 512 bits. It takes 64 steps (4 rounds of 16) and maximum size of the message depends on the infinity.

J. Comparison of Hashing Techniques:

1. SHA-1 is strong against brute force attack.
2. MD-5 is weak for cryptanalytic attacks, but SHA-1 is not all weak because its power is more complex to judge.
3. SHA-1 should be implemented gradually in comparison to MD-5 on same hardware.
4. Both algorithms are simple to design and implement.
5. SHA-1 extended change, an additional goal, and MD-4 through the addition of superior avalanche effects.

K. Comparison between Different Encryption Algorithms

Factors	AES (Advanced Encryption Standard)	DES (Data Encryption Standard)	3DES (3-Data Encryption Standard)
Key length	128,192 or 256 bits	56 bits	56 bits
Cipher type	Symmetric block cipher	Symmetric block cipher	Symmetric block cipher
Block size	128,192 or 256 bits	64 bits	64bits
Cryptanalysis Resistance	Strong against differential, truncated differential, linear, interpolation and square attacks.	Sensitive for insecure difference of differential and linear cryptanalysis;	Brute force attacker can analyse plain text using differential cryptanalysis.
Security	AES is more secure than its predecessors DES & 3DES.	Proven insufficient	One only weak which Exits in DES.

Table.1 [14]

IV. PROPOSE WORK

Input data Stream: The data model is displayed for secure privacy-protection data retrieval system with cloud data. Initially, the user selects text file from his system for uploading to the server. During this time,

the system uses the TF-IDF concept to achieve effective keywords from the first text files.

Keyword Extraction: After keyword extraction, the file gets encrypted using a hybrid cryptographic algorithm. Finally, the data uploaded over cloud server location. The extraction ability is implemented before collecting the actual storage on cloud system.

Keyword Hash generation: The algorithm for each keyword evaluates the hash values. Prepare a hash-code based indexing plan, through which keywords match the keyword exactly. To compare data with keywords, the data structure is used for effective management in the field of content and related keyword hash codes.

Inverted Index preparation: File and their keyword entry are prepared in these interfaces; by using the Inverted Index the user generates keywords from the file. Inverted index data has been uploaded to the trusted server. This process shows how the data is stored on the server.

KNN based Search: By using generated hash values K-NN based search on the stored inverted index for matching relevant keywords. These keywords are sophisticated and keywords are used with the same hash generation algorithm which protects data and information in intermediate attacks and searches the privacy of end user.

User Input Query: In sort of search, the user must provide the text file as input. From the user input query the set of keywords is extracted to search the relevant file from the server. This method promised to afford efficient results during the queries made.

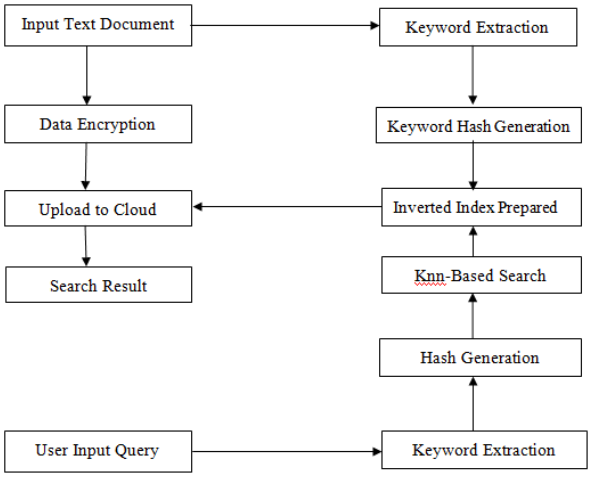

Figure3 Propose Model

V. CONCLUSION AND FUTURE WORK

The paper provides the knowledge and benefit of information retrieval from the cloud are explained through keywords. This is that all users typically having the same secure keys for trap-door generation by same-key encryption scheme. Using this cloud space, it improves precision, recall, time consumption and memory management by managing search space, which helps to reduce the amount of data from the specified index. Scale the Storage on Cloud Environment for sustaining the threshold scheme as the keyword extractor, by changing the policy of keyword extraction from raw data. In the future, we intend to expand the proposed word for expanding data such as forward moving data. And also search for easy and effective recovery of data, as well as improve encryption techniques.

VI. REFERENCES

[1]. Kahate, Atul, "Cryptography and Network Security", Tata McGraw-Hill ,India

[2]. Xia Zhihua, and Xinhui Wang, 2016 IEEE Transactions on Parallel and Distributed Systems, Pp- 340-352. ISSN No.- 1045-9219 DOI:10.1109/TPDS.2015.2401003

[3]. Zhangjie Fu and Kui Ren, 2016, IEEE Transactions on Information Forensics and Security.

[4]. W. Sun et al.,2013, in Proc. 8th ASIACCS ,Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking, Pp- 71–82.

[5]. Ning Cao et al., 2014, IEEE Transactions on parallel and distributed systems, Privacy-preserving multi-keyword ranked search over encrypted cloud data, Pp- 222-233.

[6]. Sumathi Sivaraj et al, 2014 International Journal of Computer Science and Mobile Computing, Vol.3 Issue.2, pg. 580-585

[7]. Mayank Kudale et al., 2014, International Journal of Scientific Engineering and Technology Research , Pp: 1142-1145.

[8]. Hongwei Li, Dongxiao Liu and Xuemin, 2014, IEEE transaction on emerging topics in computing, Enabling Efficient Multi-Keyword Ranked Search Over Encrypted Mobile Cloud Data Through Blind Storage, DOI:10.1109/TETC.2014.2371239

[9]. Li, Hongwei, et al. ,2015, IEEE Transactions on Emerging Topics in Computing, Enabling efficient multi-keyword ranked search over

encrypted mobile cloud data through blind storage, Pp-127-138.

[10]. B. Wang, W. Song, W. Lou, and Y. T. Hou, 2015, in Proc. IEEE INFOCOM , Inverted index based multi-keyword public-key searchable encryption with strong privacy guarantee, Pp. 2092–2100.

[11]. Li, Ruixuan, et al. 2014, Future Generation Computer Systems, Efficient multi-keyword ranked query over encrypted data in cloud computing, Pp-179-190.