ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Cloud Storage Data Deduplication with Encryption

Aruna A S and Deepthi K
Assistant Professor,
Dept. of Computer Science,
CEV, Vadakara, Kerala, India

Abstract: Cloud storage is a remote storage service, where users can upload and download their data anytime and anywhere. Data duplication, data leakage, space consumption are the main issues of cloud storage. During the time of uploading data is converted into binary data, using AES algorithm cipher text is formed it is stored in cloud .Using MD5 algorithm we create a hash value and it is stored in hash table. Parallel to this plagiarism is running, it involves content checking. From the different methods of plagiarism, syntactic based method is used here. Syntactic-based methods do not consider the meaning of words, phrases, or sentence. Content checking is performed to eliminate duplicate cipher text, here a threshold value is set, if contents similarity is smaller than threshold value user opinion is asked to upload the data. Here data compression is performed at the time of uploading data to cloud and this is done to reduce the amount of storage space. The plagiarism and compression techniques avoids the unwanted usage of memory. This method provides better security for data. User can download data as needed.

Keywords: Deduplication, cloud storage, plagiarism, port stemmer

I. INTRODUCTION

A. Overview:

In the current digital world, data have priority for persons as well as for establishments. As the amount of data being generated increases exponentially time, duplicate data contents being stored cannot be with endured. Thus, employing storage optimization techniques is an essential requisite to large storage like cloud storage. Deduplication is a one such storage optimization technique that avoids storing spare copies of data. Currently, to ensure security, data stored in cloud as well as other large storage areas are in a scrambled format and one problem with that is, we cannot apply deduplication technique over such an encrypted data. Thus, performing deduplication securely over the encoded data in cloud appears to be a puzzling task.[1]

Cloud computing has become a very important topic and brings many advantages through various services. The complex hardware, database, and operating system can be handled by a cloud server. Users only need some simple devices, which can connect to the cloud server. However, in the environment, the cloud server can obtain and control all the uploaded data because all the data are stored or operated in the cloud. The security and privacy issues are very important in cloud computing. In order to protect privacy, users encrypt their data by some encryption algorithms and encrypted data to the cloud. Users can store their data in the cloud storage and download the stored data anywhere. Even if users exhaust their own storage spaces, the cloud storage server can expand the storage spaces without destroying the stored data. However, the fast growth of storage requirements burdens the cloud storage, which is not infinite. The cloud storage server typically applies the data de-duplication technique to reduce the consumption of storage space.

At the same time, the goal of encryptions is to keep information secret and make it difficult to distinguish the encrypted data (i.e., cipher texts) from random values. If an encryption is secure, it would be hard to obtain information from cipher texts. Hence, encrypted data de-duplication becomes a challenge because the first step of data de duplication is to search for duplicate data.

B. Scope:

Cloud computing increases the speed and dexterity which alludes to accessing the internet in a specific data centre of different hardware and software. It is a used to describe a class of network based computing that takes place over the internet. It comprises the procurement of dynamically adaptable and virtualized reserves as a indulgence over the internet. This technology allows more efficient computation by centralizing storage memory processing and bandwidth.

A hypercritical fight for cloud storage is the management of aggregate volume of accumulating data. In order to manipulate the data management, data compression or data deduplication technique has been proposed and intrigues more attention. Since the amount of data storage is larger, there may be large amount of duplicate copies. In order to avoid those unwanted data and to save the storage space, a peculiar data compression technique has been enabled to remove the redundant data .This helps to reduce the byte storage in cloud. Only one copy of the tautological data is kept and the remaining data are excluded.

Redundant data are replaced with pointers, so that only eminent data can be retrieved. Pointers are provided to users with same file so that there is no requirement to upload the file. Though there are many privileges, security threats may occur anytime. So some encryption systems are used and are with the help of cipher texts. Deduplication can be made possible by formatting contrast cipher texts for different users.

The users download the file that is encrypted and then AES keys are used to decrypt the file. Authorization is provided to guide the user while uploading the file in the cloud. Users without proper authentication are not allowed to perform checks on data. These checks are compassed in public cloud. After transmitting the file, checking is done for any existing privileges that correspond to match the honour of newly uploaded data. Hence for capable storage of uploaded data, Storage Service Provider is imported.

C. Deduplication:

According to the data granularity, deduplication strategies can be categorized [2] into two main categories: file-level deduplication and block-level deduplication which is nowadays the most common strategy. In block-based deduplication, the block size can either be fixed or variable. Another categorization criteria is the location at which deduplication is performed: if data are deduplicated at the client, then it is called source-based deduplication, otherwise target-based. In source-based deduplication, the client first hashes each data segment he wishes to upload and sends these results to the storage provider to check whether such data are already stored: thus only ”undeduplicated” data segments will be actually uploaded by the user. While deduplication at the client side can achieve bandwidth savings, it unfortunately can make the system vulnerable to side-channel attacks whereby attackers can immediately discover whether a certain data is stored or not. On the other hand, by deduplicating data at the storage provider, the system is protected against side-channel attacks but such solution does not decrease the communication overhead

D. Existing System Analysis:

Cloud storage is a remote storage service, where users can upload and download their data anytime and anywhere. It raises issues regarding privacy and data confidentiality because all the data are stored in the cloud storage. This is a subject of concern for users, and it affects their willingness to use cloud storage services. On the other hand, a cloud storage server typically performs a specialized data compression technique (data deduplication) to eliminate duplicate data because the storage space is not infinite.

Data deduplication, which makes it possible for data owners to share a copy of the same data, can be performed to reduce the consumption of storage space. In the existing system encrypted data deduplication mechanism which makes the cloud storage server be able to eliminate duplicate cipher texts and improves the privacy protection. Limitations: (1) There is no content checking in this method. (2) Only exact duplications of files are eliminated. (3) There is a chance of file with similar contents are uploaded

II.RELATED WORKS

Cloud computing is the 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. 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. But for protecting data privacy, sensitive data has to be encrypted before outsourcing, which obsoletes traditional data utilization based on plaintext keyword search. privacy-preserving multi-keyword ranked search over encrypted cloud data[3] is possible.

ALG-Dedupe is an application aware local-global source-deduplication scheme for cloud backup in the personal computing environment to improve deduplication efficiency [4]. There is a cloud-based file system named SEARS-Space Efficient [5] And Reliable Storage system that exploits the deduplication technique to reduce storage and traffic cost as well as the erasure coding technique to increase both the data reliability and the file retrieval speed.

III.PROPOSED SOLUTION

A. System Architecture:

A method termed plagiarism is introduced in the proposed system, it involves content checking. From the different methods of plagiarism, syntactic based method is used here. Syntactic-based methods do not consider the meaning of words, phrases, or sentence. Thus the two words “exactly” and “equally” are considered different. Nevertheless they can provide significant speedup gain comparing to semantic-based methods especially for large data sets since the comparison does not this involve deeper analysis of the structure and/or the semantics of terms. Benefits: This model avoids deduplication in cloud storage. It provides better security for data. The plagiarism and compression technique avoids the unwanted usage of memory.

Modules:

File Management: This module manages the downloading, uploading and searching of files. To implement the uploading process, the file data converted into binary is encrypted using AES. Using Plagiarism the content checking is done and the file is uploaded or discarded.

The downloading process request for the particular data and the searching is done; if it is available the data is downloaded.

Encryption: Advanced Encryption Standard (AES) was published by the national institute of standards and technology in 2001. AES is a symmetric block cipher. Cipher takes a plane text block size of 128 bits or 16 bytes. The key length can be 16, 24 or 32 bytes (128,192,256 bits). The algorithm is referred to as AES-128, AES-192 or AES-256 depending on the key length.

Each word is 4 bytes, and the total key schedule is 44 words for the 128-bit key. The ordering of bytes within a matrix is by column. The cipher consist of N rounds, where the number of rounds depends on the key length: 10 rounds for a 16-byte key , 12 rounds for a 24-byte key , and 14 round for a 32-byte key.

There are four different stages used:

1.Substitute bytes: Uses an S-box to perform a byte-by-byte substitution of the block.2.Shift Rows: A simple permutation.3.Mix Columns: A substitution that makes use of arithmetic over.4.AddRoundKey: A simple bitwise XOR of the current block with a portion of the expanded key.

Deduplication checking: The MD5 message-digest algorithm is a widely used cryptographic hash function producing a 128-bit (16-byte) hash value, typically expressed in text format as a 32 digit hexadecimal number. MD5 has been utilized in a wide variety of cryptographic applications, and is also commonly used to verify data integrity.

MD5 processes a variable-length message into a fixed-length output of 128 bits. The input message is broken up into chunks of 512-bit blocks (sixteen 32-bit words); the message is padded so that its length is divisible by 512. The padding works as follows: first a single bit, 1, is appended to the end of the message. This is followed by as many zeros as are required to bring the length of the message up to 64 bits fewer than a multiple of 512. The remaining bits are filled up with 64 bits representing the length of the original message, modulo 264.

MD5 digests have been widely used in the software world to provide some assurance that a transferred file has arrived intact.

Plagiarism: Document Plagiarism Opposed to other types of plagiarism (such as music, graphs, etc.), document plagiarism falls in two categories; source code plagiarism and free text plagiarism. Given the constraints and keywords of programming languages, detecting the former are easier than detecting the latter and hence source code plagiarism detection is not the focus of current research. Plagiarism takes several forms. Maurer et al stated that the followings are some of what considered practices of free text plagiarism:

• Copy-paste: or verbatim (word-for-word) plagiarism, in which the textual contents are copied from one or multiple sources. The copied contents might be modified slightly.

• Paraphrasing: changing grammar, using synonyms of words, re-ordering sentences in original work, or restating same contents in different semantics.

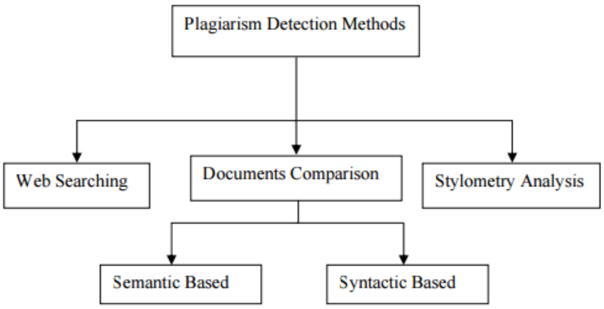
• No proper use of quotation marks: failing to identify exact parts of borrowed contents.

• Misinformation of references: adding references to incorrect or non-existing sources.

• Translated Plagiarism: also known as cross-language plagiarism, in which the contents are translated and used without reference to original work. Plagiarism detection methods can be broadly classified into three main categories

The first category tries to capture the author style of writing and find any inconsistent change in this style. This is known as stylometry analysis. The second category is more commonly used which is based on comparing multiple documents and identifying overlapping parts between these documents. The third category takes a document as input and then searches for plagiarism patterns over the Web either manually or in an automated manner.

Figure provides taxonomy of plagiarism detection methods.



Syntactic-Based Detection Unlike semantic-based, syntactic-based methods do not consider the meaning of words, phrases, or sentence. Thus the two words “exactly” and “equally” are considered different. This is of course a major limitation of these methods in detecting some kinds of plagiarism. Nevertheless they can provide significant speedup gain comparing to semantic-based methods especially for large data sets since the comparison does not involve deeper analysis of the structure and/or the semantics of terms. To quantify the similarity between chunks, usually a similarity measure is used.

Porter stemmer: Porter stemmer algorithm is a process for removing words from English. There are several types of stemming algorithms which differ in respect to performance and accuracy and how certain stemming obstacles are overcome.

Porter’s algorithm is important for two reasons. First, it provides a simple approach to conflation that seems to work well in practice and that is applicable to a range of languages. Second, it has spurred interest in stemming as a topic for research in its own right, rather than merely as a low-level component of an information retrieval system. The algorithm was first published in 1980; however, it and its descendants continue to be employed in a range of applications that stretch far beyond its original intended use.

Suffix-stripping algorithm: Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include:

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

Suffix stripping approaches enjoy the benefit of being much simpler to maintain than brute force algorithms, assuming the maintainer is sufficiently knowledgeable in the challenges of linguistics and morphology and encoding suffix stripping rules. Suffix stripping algorithms are sometimes regarded as crude given the poor performance when dealing with exceptional relations (like 'ran' and 'run'). The solutions produced by suffix stripping algorithms are limited to those which have well known suffixes with few exceptions. This, however, is a problem, as not all parts of speech have such a well formulated set of rules. Lemmatization attempts to improve upon this challenge.

Prefix stripping may also be implemented. Of course, not all languages use prefixing or suffixing.

File compression: GZIP power is format and a softer application used for file compression and de compression. The GZIP is based on the DEFLATE algorithm which is the combination of LZ77 and Huffman coding. GZIP is also used to refer to the GZIP file format which is a 10-byte header, containing a magic number, version number and a timestamp.

Although its file format also allows for multiple such stream to be concatenated. GZIP is normally used to compress just single files. Compress archives are typically created by assembling collections of file into a single tar archive, and then compressing that archives format. The ZIP format also used DEFLATE. The ZIP format can hold collection of file without an external archive, but is less compact than compressed tar balls holding the same data, because it compress file individually and cannot take advantages of redundancy between files.

III.ESULTS AND DISCUSSION

The main purpose of this research is to propose a reasonable data deduplication mechanism in which all data are stored as the encryption structure in the cloud storage. This research addressed an effective solution for the encrypted data deduplication cloud storage to achieve better results. As first step, The cipher structure in the cloud storage is formed, so the encryption of data will provide a high level of security. Here content checking is done as compared to existing ones and it is done using Plagiarism. Porter Stemmer enhances the searching of texts written in other languages. This avoids the deduplication in text files. For each file an independent hash value is generated using Message Digest algorithm (MD5).And also to lessen the storing space here compacting the data and it is done by using GZIP mechanism.

IV.CONCLUSION AND FUTURE WORK

Deduplication in cloud storage is an important current research area. We propose a feasible data deduplication mechanism in which all data are stored as the cipher structure in the cloud storage. The main two functional modules of this work is uploading and downloading. The two conditions provided for uploading are; making the file public and private. There is no special condition for downloading; it is just simply the downloading.

The cipher structure in the cloud storage is formed using Advanced Encryption Standard (AES) algorithm. The Encryption of data will provide a high level of security. Here sender and receiver use the same encryption/decryption key. The heart of this is content checking and it is done using Plagiarism. Porter Stemmer enhances the searching of texts written in other languages. This avoids the deduplication in text files.

For each file an independent hash value is generated using Message Digest algorithm (MD5). Inorder to reduce the storage space here compressing the data and it is done by using GZIP mechanism.

One main future work proposed here is to make this work more accurate.Furthermore, we will work on finding possible optimizations in terms of bandwidth and computation.

V.REFERENCES

[1] A Study on Deduplication Techniques over Encrypted Data Akhila K,Amal Ganesh,Sunitha C,2016

[2] ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage Pasquale Puzio SecludIT and EURECOM France ,Refik Molva,EURECOM ,France,Melek Onen ¨ EURECOM , France, Sergio Loureiro SecludIT

[3] Privacy-preserving multi-keyword ranked search over encrypted cloud, Ning Cao, Cong Wang, Ming Li. Department of ECE, Worcester Polytechnic Institute, USA,July 2015

[4] Yinjin Fu, Hong Jiang, “Application-Aware Local Global Source Deduplication for Cloud Backup Services of Personal Storage,” IEEE Transactions On Parallel And Distributed Systems, VOL. 25, NO. 5, MAY 2014.

[5] SEARS: Space Efficient And Reliable Storage System in the CloudYing Li, Katherine Guo, Xin Wang, Emina Soljanin, Thomas Woo,Dept of Electrical and Computer Engineering, Stony Brook University Bell Labs

[6] ”Authorized Deduplication:an approach for secure cloud Environment”,Vivek Waghmire,Smitha kapse,Science Direct 2015