ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

Efficient Bandwidth and Space Utilization Using Object Based Deduplication on the Cloud

[1] Ms. Uzma madeeha, [2] Smt. Shabana sultana
M.Tech - Information Technology
[1][2] Department of Computer science and Engineering National Institute of Engineering, Mysuru, India

***Abstract*— As cloud computing is mostly used domain now a days, keeping the cloud limitations in mind we try to efficiently utilize the cloud resources by applying basic optimization techniques. Since text data occupy very small space compared to images or multimedia files in general, we come up with some space optimizing and de-duplicating schemes. Instead of traditional techniques of block level de-duplication, we implement a technique where the images are divided into blocks and are given unique hash id and stored in ind ex table. Now instead of sending entire object onto cloud for comparison, we send only index table to check for duplicate entries and eliminate them. This way the bandwidth is very efficiently utilized compared to traditional schemes and also optimizes space utilized in cloud storage systems.**

***Index Terms*— Encryption, Decryption, Hash id, index table,**

I. INTRODUCTION

As Cloud computing is an emerging and trending technology which is ever growing. The most beneficial technology for internet based computing is cloud platform which is a “pay as you go” service. Cloud is nothing but the networks of internet. Cloud computing is nothing but the internet-based computing where the data and the software are present on the web. It creates virtual environment which provides real time computing of data anywhere, anytime globally. This feature makes it more attractive. Since it is such an intriguing field of technology and also globally popular, there is a need for coming up with efficient ways and cost effective faster ways of making this technology more advanced.

Since cloud computing is a platform where a user can utilize any amount of resources both hardware and software and pay only for those resources which the user makes use of, there is a demand for huge amount of data storage methods which are faster and efficient as well as secure. The beauty of the platform is that everything is stored on the internet and hence virtually available on the go across the globe. There can be all types of applications which can make use of the cloud computing services like the desktop applications, mobile based applications and browser based applications.

When it comes to cloud computing, storing data efficiently onto cloud is a major concern. Huge of data is stored onto cloud by multiple clients simultaneously. It is important to note that duplication of data has to be avoided in order to minimize of storage space as well as optimize bandwidth which in turn results in faster data exchange on cloud. There comes the concept of de-duplication. de-duplication is nothing but removing multiple copies of same data and retaining only single unique copy in order to avoid inefficient storage utilization or wastage of space. This concept ofde-duplication is entirely crucial onto cloud where tons of data are processed simultaneously innumerable times by multiple clients.

Now considering de-duplication technique based on granularity there exist 3 types of de-duplication techniques. First is the file level technique where the entire file is treated as a unit and stored onto cloud. This method is faster since it is treated as a file. But the drawback is it is inefficient. Minute changes to the text file is still treated as a new file and stored onto cloud. This results in space wastage. To overcome this there is second type technique that is fixed level block de-duplication. In this technique the entire file is divided into chunks called blocks. This block size is fixed for entire file. This method is efficient but is comparatively slower. Comparing blocks results in space utilization. But the drawback is selecting the block size. Third technique is the variable size de-duplication. Here the file may be divided into blocks of different sizes but the drawback is if the images in file are same and only text is changed it would still treat the block as new one and store it uniquely. This results in more space utilization and is still inconvenient.

II.LITERATURE SURVEY

This section describes the various existing schemes.

A. Secure de-duplication technique.

Here the paper explains the de-duplication of cloud files based on granularity into two main categories. One is file level and other is block level. The file level de-duplication the entire file is treated as a unit and is stored onto cloud as such. If the same file is tried to upload again then it eliminates the redundant copy and retains only one copy of file. The problem with file level is if same file is just modified a single letter it would still rewrite the entire file as a new one and store it incloud. This results in efficiency. In block level de-duplication [1], the entire file is divided into blocks and is stored onto cloud. Only the modified block is stored onto cloud. This way is more efficient than the first. Disadvantages are inefficient space utilization if small text is modified and if the block contains multimedia data. Deciding the block size is crucial. Not faster. Bandwidth utilization is more for comparing blocks.

B. Dynamic multidimensional index for large-scale cloud data.

The paper proposes a method called skip-octree structure which is a multidimensional index framework [2]. In this method the indexes are divided into hierarchical structure as in 3 subsets. Suppose 3 datasets q0, q1 and q3. Q0 stores all the data indexes, q1 stores half of q0 and q2 stores half of q1. When search is performed q2 is searched first and q1 and q0. Extended version of this to have multiple data indexes portioned to make the search easier. Disadvantages are since there are multiple partitions maintained require large storage space and hierarchical system. Large amount of data may lead

C. Secure file key sharing over single image on cloud computing.

In this paper, for encryption of images on cloud there are 2 keys used to encrypt. One key is hidden behind the image [3] itself. When file is accessed from server, primary key has to be entered only then the image key is to be used to display. Even if hacker hacks it, hacker will only be able to see the image not the key hidden behind. Disadvantages are if image is corrupted then it is not useful and key is lost. Hacker may be able to hack both the keys using advanced methods. Additional computation and processing for generating and maintaining 2 keys and hiding behind image file. Not efficient and secure.

D. A cryptography application using salt hash technique

Salt data [4] is a random data generated in order provide complexity to easy passwords to make hacking hard. In the paper, the random salt is added to the original password and salt password is generated using hash function. The salt is stored and salt hash in the database. This way it is expected to provide security from hacking. Disadvantages are if attacker gets access to systems, salts are easy to hack by hacking dictionaries. Brute force attack (same key for repetitive letters) used to system then salt fails to provide security.

III. PROPOSED SYSTEM

A. System model

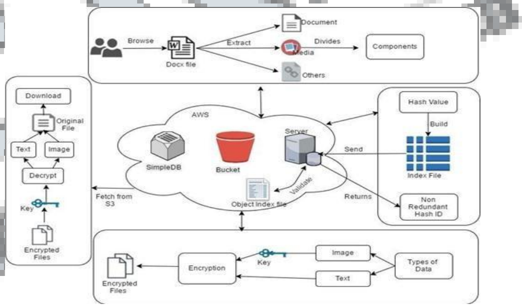

Fig 1. Proposed approach

As shown in fig 1 the system model clearly depicts the working architecture of de-duplication module.The 4 modules work together to provide the required functionality. First step is to sign up and create an account for each user. These details are added onto cloud in bucket along with files uploaded by user and components of those files. Upon registration and successful login user can make use of further services.

Initially the user has to browse the document file to extract the zip document which is to be uploaded. This extraction will result in 3 separate folders. One is document folder consisting of text data, media folder consisting of image data and other folder having Meta data. Only the image files are divided into smaller objects called components.

The next step is to build an index file. In order to do so, the components are assigned unique hash values which are generated using SHA algorithm. These hash values are calculated for each object and no 2 objects have same id. These hash value are used to build an index file which consists of all hash ids. In the de-duplication module, the client sends the index file onto cloud server. The server validates and indexes the file and returns only the hash values of non-redundant data. Only these are furthered processed for encryption.

After checking De-duplication, now the data has to be encrypted before uploading. The text data is encrypted according to user. The image file components are encrypted using keys present in cloud storage. These encrypted files are then uploaded onto cloud. For encryption of these components combination of hash id and primary key is used. These keys are maintained in the SimpleDB in cloud.

For user to download any file which is uploaded, the encrypted files are fetched from cloud storage and decrypted using same key used for encryption. After decryption the file components are merged as original file and fetched to user.

IV. ALGORITHMS

A. Rijndael algorithm

Rijndael algorithm is the advanced encryption standard algorithm and an iterated block cipher. It uses the same key for both encryption and decryption. Rijndael algorithm supports the block sizes of 128 bits and key sizes of 128, 192, 256 bits. It is the best among the combination of security, performance, efficiency, ease of implementation and flexibility.

Initially to encrypt the data master key is used, which is given by the user. These master keys are usually small in size and easy to hack. Hence, to become complex key, salt data are added to the master key. In the proposed approach, complex key generates an initialization vector. The first block of the data is encrypted by using the initialization vector along with the master key. The remaining blocks are encrypted with the help of previous encrypted block along with master key.

Steps:

***Step 1*: ByteSub Transformation**

Each byte of the block is replaced by its substitute in an S-box. Each byte is treated independently Single S-box is used for the entire state.

***Step 2**:* ShiftRow Transformation

Each row of the state is shifted cyclically a certain number of steps. The number a row is shifted can't be the same.

Step 3: Mix Column Transformation

State columns are treated as polynomials over GF(2^8). Each column is multiplied by modulo x4 + 1 by a fixed polynomial c(x) = `03` x3 + `01` x2 + `01`x + `02`

Step 4: Round Key Addition XOR round key with state.

V. CONCLUSION

Using the component oriented technique; the space is optimized to larger extent. Since the image files are optimized so efficiently even the file processing is faster compared to traditional methods. With the comparison of only hash ids, the bandwidth is utilized efficiently. Also the processing overhead is reduced and eases the complexity. Since the encryption of data is done it is secure. This is a highly secure and efficient de-duplication strategy to be implemented on cloud which can speed the processing and file related operations in faster way. Also it is cost effective since lesser storage space is used.

REFERENCES

[1] vruti satish radia,PIET, vadodra,gujrat, India and dheeraj kumar singh,PIET,vadodra,gujrat, Secure de-duplication technique techniques: A study. India.International journal of computer applications,volume 137-no 8, march 2016.

[2] Jing he, yue wu, yunyun dong, yunchan zhang and wei zhou. Dynamic multidimensional index for large-scale cloud data. . Journal of cloud computing: advances, systems and applications. 2016.

[3] M aruna, J nulyn punitha. Secure file key sharing over single image on cloud computing. IJCSMC, vol 5, issue 3, march 2016.

[4] Pritesh n patel, jigisha k patel and paresh v virparia. cryptography application using salt hash technique. International journal of application or innovation in engineering of management. Volume 2. Issue 6, june 2013.