LOAD BALANCING AND DEDUPLICATION

Mr.Chinmay Chikode
Dept.of Information Technology
Marathwada Mitra Mandal's College of Engineering,
Pune, India.

Mr.Mehadi Badri
Dept.of Information Technology
Marathwada Mitra Mandal's College of Engineering,
Pune, India.

Mr.Mohit Sarai
Dept.of Information Technology
Marathwada Mitra Mandal's College of Engineering,
Pune, India.

Ms.Kshitija Ubhe
Dept.of Information Technology
Marathwada Mitra Mandal's College of Engineering,
Pune, India.

ABSTRACT— Load Balancing is a method of distributing workload across multiple computing resources such as cluster of computers, network links. The goal of LB is to optimize the resource usage, evade overload, maximize throughput and to minimize the response time. This was identified as a major concern in Cloud Computing to scale up the increasing demands. Deduplication : In computing, data deduplication is a specialized data compression or data splitting into chunks technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage. This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the_match frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced. In Deduplication hash calculations can be done in real-time as data enters the target device.

1. INTRODUCTION

Cloud computing enables on-demand network access to a shared pool of configurable computing resources such as servers, storage and applications. These shared resources can be rapidly provisioned to the consumers on the basis of paying only for whatever they use. Cloud storage refers to the delivery of storage resources to the consumers over the Internet. Private cloud storage is restricted to a particular organization and data security risks are less compared to the public cloud storage. Hence, private cloud storage is built by exploiting the commodity machines within the organization and the important data is stored in it. When the utilization of such private cloud storage increases,

there will be an increase in the storage demand. It leads to the expansion of the cloud storage with additional storage nodes. During such expansion, storage node in the cloud storage need to be balanced in terms of load. In order to maintain the load across several storage nodes, the data needs to be migrated across the storage nodes. This data migration consumes more network bandwidth. The key idea behind this Application is to develop a dynamic load balancing algorithm based on deduplication to balance the load across the storage nodes during the expansion of private cloud storage.

2. MOTIVATION

Main motivation of the system is to remove a load on cloud based servers and avoiding data duplications using the some methodologies and algorithm. This system is basically used to function on Hash Code detection techniques which is used for avoiding multiple storage of the files on the Cloud Server.
For the load balancing techniques system split the file into three chunks and stored into the three different location and the access is only for the valid personnel or authorized person i.e only who has login credentials with the valid user key which is given by the admin. Even if the data is hacked original information is not leaked, it is kept safely.

3. OBJECTIVE

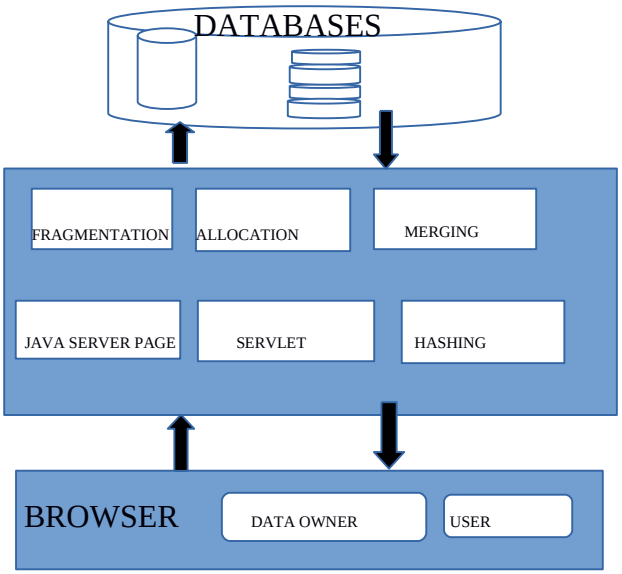
- The main purpose of this system is to maintain load on cloud by providing security and avoid duplication of the data using the same methodologies and algorithm.
- This system basically performs on Hash Code generation technique which avoids the multiple storage of the files on the Cloud Server.
- For the load balancing technique, we split the file into three chunks and store them into the three different random locations and the access is only for the valid personnel or authorized person, only who has login credentials with the valid user key (private key) which is given by the admin.

4. LITERATURE SURVEY

The proposed theory in[1] based on Reclaiming space from duplicate files in a server less distributed file system in the year 2004. In this concept, Cloud computing offers a new way of service provision by re-arranging various resources over the Internet. The most important and popular cloud service is data storage. In order to preserve the privacy of data holders, data are often stored in cloud in an encrypted form. However, encrypted data introduce new challenges for cloud data deduplication, which becomes crucial for big data storage and processing in cloud. In [2] he proposed a hybrid cloud approach for secure authorized deduplication and concluded that, Dedupli-cation has proved to achieve highcost savings, e.g., reducing up to 90-95 percent storage needs for backup applications. How to manage encrypted data storage with deduplication in an efficient way is a practical issue.
They studied [3] cloud computing vulnerabil-

ities where he found that Outsourcing data to a third-party administrative control gives rise to security concerns. The data compromise may occur due to attacks by other users and nodes within the cloud. Therefore, high security measures are required to protect data within the cloud. However, the employed security strategy must also take into account the optimization of the data retrieval time. The authors apporches in [4]Auditing algorithms for achieving integrity is given.

5. ARCHITECTURAL DIAGRAM



6. OVERALL DESCRIPTION

6.1 PRODUCT PERSPECTIVE:

This system developing perspective is to provide high-level security to every user from Hackers. An online shopping website is a form of electronic commerce which allows consumers to directly buy goods or services. So, data security plays an important role when it comes to managing all Internet shopper's data including personal information, banking information etc. Situation is when user registers & completes his/her shopping process & logouts. Attacker accesses database & modifies data. So, to overcome this problem, we use a 3-layer system such as Analyzer Engine, Secured Layer, Rollback Engine.

6.2 PRODUCT FUNCTION:

- This System has a functionality to ask information for the customer to the login and send the username, password with the help of the admin.
- Those have valid login credentials can easily perform upload, delete, and download operations using private key.
- Using the Advanced Encryption standards (AES) and MD-5 algorithm the data security and load balancing is managed.
- The Hash Code is used to create code according to the file data and stored into database, if the code is same then "Duplicate" file message will be displayed, otherwise the code is unique and then the file splits into three different chunks and is stored into three Different locations.
- If the user tries to Delete the file without Private Key and its login credentials fails.
- The Login credential on a match; all of the three chunks get merged into a single file and Delete/Download Operations are performed, this makes it faster and more secure.

7. IMPLEMENTATION

- **ALGORITHMS**

• T-coloring Algorithm:

T-colouring algorithm is use for storing data on different chunks. To place the chunks of the files on random server with no two chunks residing on the adjacent servers.

Suppose we have a graph G = (V;E) and a set T containing non-negative integers including 0. The Tcoloring is a mapping function f from the vertices of V to the set of non-negative integers, such that Sf(x)- f(y)S T, where (x;y) ¿ E. The mapping function f assigns a color to a vertex. In simple words, the distance between the colors of the adjacent vertices must not belong to T. Formulated by Hale the T-coloring problem for channel assignment assigns channels to the nodes, such that the channels are separated by a distance to avoid interference.

• Advanced Encryption Standard:

AES algorithm is used to encrypt the data. AES comprises three block ciphers, AES-128, AES-192 and AES-256. Each cipher encrypts and decrypts data in blocks of 128 bits using cryptographic keys of 128-, 192- and 256-bits respectively. (Rijndael was designed to handle additional block sizes and key lengths, but the functionality was not adopted in AES.) Symmetric or secret-key ciphers use the same key for encrypting and decrypting, so both the sender and the receiver must know and use the same secret key. There are 10 rounds for 128-bit keys, 12 rounds for 192-bit keys, and 14 rounds for 256-bit keys – a round consists of several processing steps that include substitution, transposition and mixing of the input plaint ext and transform it into the final output of cipher text.

• MD-5 Algorithm:

MD-5 algorithm generates a hash code on the basis of file content. Cryptographic hash functions are mathematical operations run on digital data; by comparing the computed ”hash” (the output from execution of the algorithm) to a known and expected hash value, a person can determine the data’s integrity. For example, computing the hash of a downloaded file and comparing the result to a previously published hash result can show whether the download has been modified or tampered with. A key aspect of cryptographic hash functions is their collision resistance: nobody should be able to find two different input values that result in the same hash output.

8.RESULTS

1.Registration of users with valid credentials:
The intent of adding valid user credentials check is to resolve implementation security by accepting a valid email address to which the access key is sent that permits the user to download any file by self authentication. Along side such an action the valid user credentials prove the authenticity of a specific user by default.

2.Uploading files:
While uploading files the user is allowed to send the file of redundant content that is then checked with a preexisting file(if any) and this allows the hashing algorithm to be put to use for redundant data check successfully.

3.Downloading files:
Each specific file needs the input of an access key sent to the user via a generally set email address. This process confirms individual data integrity balance by allowing to download one file at a time.

4.Deleting a file:
During the process of deleting any files, the data is not lost from the server but instead the user pointer linked to the data implicitly is invalidated thereby allowing simultaneous presence and safe execution of the function for the respective user.

5.Overall system functionality and utility:
The overall functioning of the system is to provide safety using the AES encryption algorithm, then the MD5 algorithm for progressive hash key generation to impart an attribute of uniqueness and finally applying the the T-coloring algorithm to inundate the server system performance by load balancing and division of data into a specific set of chunks. Hence, the system applies load balancing and de-duplication of data and is scalable up to the cloud systems.

9. CONCLUSION

This system proposes the architecture of deduplication system for cloud storage environment and gives the process of avoiding deduplication in each stage. In Server, system employs the file-level and chunk-level deduplication to avoid duplication. Load sharing algorithm which has a policy to partition the data into various partitions across various storage facilitations. These fragments are placed in different storage locations and hence it is difficult to trace the data. This fragmentation in turn also imparts the ease of load balancing. Even if the data is hacked original information is not leaked, it is kept safely. Most importantly the data is hashed encoded which allows unique identification of the data file alongside appropriate access optimization.

10. REFERENCES

[1] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer “Reclaiming space from duplicate files in a server less distributed file system” in Proc 2004
[2] Jin Li, Yan Kit Li, XiaofengChen, Patrick P.C. Lee, and Wenjing Lou. “A Hybrid Cloud Approach for Secure Authorized Deduplication’, IEEE transactions on parallel and distributed systems” in 2015.
[3] B. Grobauer, T.Walloschek, and E. Stocker, “Understanding cloud computing vulnerabilities” in 2011.
[4] Jian Liu, Kun Huang, Hong Rong, Huimei Wang, and Ming Xian “Privacy-Preserving

Public Auditing for Regenerating-Based Cloud Storage" in 2015.
[5] C. S. Pawar, and R. B. Wagh, Priority Based Dynamic Resource Allocation in Cloud Computing with Modified Waiting Queue, 2013 International Conference on Intelligent Systems and Signal Processing (ISSP), 2013.