Deduplication using Cloud Computing

Pratik R. Joshi , Akash A. Bagul , Avinash Y. Shelar , Pratik K. Dange
Department of Cloud Computing and Big Data , Padmashri dr.vithhalrao vikhe patil inst. of tech. and engi. polytechnic loni,rahata

***Abstract*: There has been decreased reliance on the deployment and maintenance of local storages as the cloud infrastructure has been consistently getting improved over time. The paradigm shift in the data storage department from local storages to cloud solutions has been aggressive. This is due to the fact that the cloud solution allows for a much convenient and hassle free experience that can allow for a much better handling and perseverance of data. There are a large number of organizations that utilize these services of which large number of employees access a singular file that leads to a lot of duplicates. The duplicates can take up unnecessary space which can be problematic with limited space and expensive online storage options. Therefore, there is a need for a de duplication mechanism where the duplicates on a shared cloud storage can be eliminated to improve experience and reduce storage consumption. The proposed approach utilizes the Amazon RDS approach which achieves highly satisfactory results.**

***Keywords*: Data Deduplication; Cloud Storage, Amazon RDS. Cloud computing, cloud solution, reliability, load balancing, encryption, secure de-duplication, data integrity.**

1. Introduction

Now-a-days, cloud computing is very important in the Information Technology. The data gathered through different sources and the Emergence of the Internet of Things in all aspects of applications increases data volume from petabytes to yottabytes, necessitating cloud computing paradigm and fog networks to process and store the data. Cloud computing (CC) produces a network-centered environment vision to users which provides access to the internet, to a collective pool of programmable grids, servers, software, storage, and amenities that could be quickly freed, with less supervision and communication to the cloud service provider. Data processing in all ways is carried out remotely in the cloud server with the help of internet connectivity. Cloud computing enables access to a shared pool of configurable computing resources like servers, storage and applications, etc. The storage services provided to users are though internet. There is chances of cloud disaster like problem in connection, performance, privacy and security, data management. To solve connection problem, we can implement offline storage & sync mechanism. To improve performance, load balancing is being an important task for doing operations in cloud and so as de-duplication also. As cloud computing has been growing and many clients all over the world are demanding more services and better results, so load balancing is necessary. Load balancing assure efficient resource utilization to customers on their demand and build up the overall performance of cloud. Every increasing volume of back up data in cloud storage may be a vital challenge so we can use de-duplication mechanism for eliminating the duplicate data. Many algorithms have been developed for allocating client's requests to available remote nodes. The key idea behind this project is to develop an offline store and sync mechanism, dynamic load balancing algorithm based on de-duplication to balance the load across the storage nodes during the expansion of private cloud storage.

2. Literature Survey

Because of rise in the costs, IT companies have started to externalize their IT services, which are maintained by specialized companies called service providers. This has made the cloud computing come up. Cloud Computing is a computing environment, where resources such as computing power, storage, network and software are abstracted and provided as services in a distributed network. Cloud Computing is a technology where the job is executed by sharing and using existing resources and applications of a distributed network environment. The resources can be allocated and de-allocated with ease by the service provider. A huge number of users request services to the cloud, which is run like large internet. Various companies use cloud computing due exponential growth in users and their needs. There are cloud computing data-centers all over the world to make cloud computing feasible. Different cloud services such as pay-per-use scheme which are offered at a lower price without intervention of owner and manager of these services.

Cloud computing consist of several characteristics such as:

- *On-demand:* Cloud services are given on-demand. Users can get there tasks done when they want.
- *Extensive Network Access*: In cloud computing resources are scattered over a network. These resources are accessed through various mechanisms.
- *Resource Pooling:* The resources are pooled accordingly. The resources are dynamically allocated and de-allocated accordingly.
- *Scalability:* Quantity of resources is increase at any time according to the customer’s requirements.

As it has been the norm, there are some issues with cloud computing as well. These issues come with huge number of requests that these clouds serve. Load Balancing, Redundancy and Fault tolerance are such issues.

Millions of service users across the globe constantly send service requests to the cloud for their storing or computing tasks.

The cloud computing needs to provide the abstraction that the user’s task is being done exclusively and provide the output without fail. When there is a surge in requests, the resource that serves these requests also needs upgrading and updating. The cloud computing has to function in such a way that it balances the load that is being out on it. A technique called load balancing is employed at this point. Cloud load balancing is the method of distributing services and computing resources in a cloud computing environment. Load balancing allows organizations to manage the workload demands by allocating resources to multiple computers, networks or servers within the cloud. By sharing the workload, the task is performed concurrently. It serves the basic idea that not all the burden should be forced on one server alone. All the servers and resources work in unison and the output is then generated in the end when all the resources have finished their assignment.

As cloud technology becomes prevalent along with it the data sharing and storage also became prevalent. The increasing volume of data needs to be managed because the less you store, the less will be the need of hardware resource. Service providers have to keep this in mind because adding hardware to store more data increases cost. To the user also it should be cheaper than actually storing the data at his end. Data de-duplication is one of the most popular technologies in storage right now because it allows companies to save a lot of money on storage costs to store the data and on the bandwidth costs. This is great news for cloud providers, because if you store less, you need less hardware. If you can de-duplicate what you store, you can better utilize your existing storage space, which can save money by using what you have more efficiently. If you store less, you also back up less, which again means less hardware and backup media. If you store less, you also send less data over the network in case of a disaster, which means you save money in hardware and network costs over time. Data de-dupliacting is really a game changer and cost saver.

The business advantages of data de-duplication include:

- Lowered backup costs
- Lowered hardware costs
- Lowered costs for business continuity and/or disaster recovery storage efficiency, network efficiency

In simple words, data de-duplication compares objects (usually files or blocks) and eradicates objects (copies) that already exist in the data set. The de-duplication process deletes block that are not unique.

A. *Existing Techniques in Clouds*

Following techniques are currently prevalent in clouds

In [1] the authors have introduced SRRS system which comprise of convergent algorithm to maintain data confidentiality and used role re-encryption algorithm to accomplish authorized data deduplication effectively. Management center is introduced to manage keys and user's roles. With the introduction of management center in the system, computational cost and overhead gets reduced on the client side. The SRRS system performs data deduplication and reduces storage space requirement and bandwidth consumption.

In [2] authors have proposed novel Attribute-Based Storage system which supports secure and efficient deduplication. It also explained drawback of standard Attribute-based encryption technique which does not support secure deduplication. The system works on hybrid cloud environment where private cloud is in charge of identical copies detection and public cloud opts for managing storage.

In [3] author explained ABE (Attribute Based Encryption) technique used to reduce storage space and share data efficiently. In this system if attributes of particular user is matched then the person is given right to compute and decipher the enciphered data.

In [4] authors have introduced convergent encryption technique to secure data in process of deduplication of data. The data outsourced is converted to cipher text before performing deduplication. The authors have also introduced different privileges to the users.

In [5], authors have introduced (MLE) which provide secure deduplication. This scheme is best for large files as this needs schema perpetuation at servers. As large files needs better maintenance scheme suits it. This scheme supports both file- level and block-level deduplication.

In [6] authors have introduced updatable block-level deduplication which provides deduplication on encrypted data and easy updation of data. The issue in file level deduplication of effective updating of data is overcome here. Some challenges are overcome by MLE and others are effectively dealt by UBLDe protocol. Dynamic Ownership management challenge is fulfilled here.

In [7] authors initiate idea to reduce the cost of updation of data. The existing MLE solution does not provide effective and secure updation of encrypted data to the user. The cost of updating single bit of data is quite high. So, the authors have introduced Updatable block-level message locked encryption technique which aims to reduce computation cost logarithm to file size. It has also introduced proof-of-ownership to users for access of files.

In [8], the author has introduced scheme which uses Symmetric Encryption algorithm, Hashing technique, Convergent encryption algorithm and token generation scheme to provide authorized duplication of data. Here the user data confidentiality and security is maintained. The data is protected both form passive and active attacks.

In [9] authors have introduced PoW (Proof-of-ownership) with data deduplication to support dynamic ownership management. This system support file-level, cross-user and block-level data deduplication. This scheme effectively carries out secure deduplication and maintains data confidentiality, consistency. It also reduces load of key management and storage space.

In [10] author has surveyed various methodologies and technologies for implementing data deduplication. They have also shown comparison of various technologies. The data confidentiality is compromised at different extent while performing data deduplication is depicted in the paper.

In [11] authors have introduced PoW (Proof-of-ownership) with data deduplication to support dynamic ownership management. This system support file-level, cross-user and block-level data deduplication. This scheme effectively carries out secure deduplication and maintains data confidentiality, consistency. It also reduces load of key management and storage space.

B. *Comparison of existing system with current system*

A lot of researchers have carried out their work in this section, we are discuss about previous work related to load balancing in cloud by using different technique .

In the current cloud server storage techniques there is a less security for the update, delete and download file. There is less load balancing techniques and no De-duplication. The current system has only provides the spaces on the server but not avoid the duplicate files.

In this project we are using hash code for the content of the file, if this code is find into the database then system generates a duplicate file message for the users else file will be divide into three chunks which is stored on the three different location so the load will be divide and automatically load balancing happened. We also using AES algorithm for encryption & private key mechanism to maintain privacy & security. Offline store & sync mechanism to overcome connection unavailability problem.

3. Data Deduplication Classification

Data deduplication can be classified based on following:

A. *Based on processing Unit*

Data Deduplication can be classified as into file-level data deduplication and block-level data deduplication. In file-level data deduplication complete file is processed from one end to another rather than dividing into multiple chunks. It involves whole file encryption and generation of single key for authorized access. In block-level data deduplication a single file is divided into multiple blocks. Each block is encrypted and multiple keys are generated. Each key share is provided to the user to avoid unauthorized access. Block size can be fixed-size or variable-size.

B. *Based on data implementation*

Data Deduplication can be divided into cross-user deduplication, server-side deduplication, and client-side deduplication.

In client-side deduplication known as source-based deduplication redundant data is eliminated before sending it to the target machine. Client-side deduplication transfers only single instance of data resulting in reduction of bandwidth consumption.

In server-side deduplication known as target-based deduplication, all data is send to the target machine where the redundant copies of data are removed. It increases bandwidth consumption resulting in cost hike but performance is better as compared to client-side deduplication.

The cross-user deduplication is widely used as it helps in better storage utilization and deduplication rate up to 90%- 95%.

4. Proposed System

1. This System has a functionality to ask information for the customer to the login and send the username, password and private key to the user with the help of the admin.
2. Those have a login credentials as well as private key for the login who can easily perform upload, delete, and download operations.
3. Using the Advanced Encryption standards (AES) and Secure Hash Code
4. (SHA) algorithm the data security and load balancing will be managing.
5. The Hash Code is created according to the file data and stored into database if the code is same then Duplicate file message will be arriving otherwise the code is unique then file split into three different chunk and stored it into three Different locations.
6. If the user tries to Delete or Download the file without Private Key and its login credential it gets fails.
7. The Login credential gets match then the all of three chunks gets merged into a single file and Delete/Download Operations performed this makes the faster and more secure. *Major Constraints:*

   1) User must enter his/her private key sent on his registered mail id.

Outcome:

1. Encrypted file will be uploaded.
2. Hash code of the file.
3. On download decrypted file will be downloaded.

Applications:

1. Allows enterprises to manage applications or workload demands by allocating resources among multiple computers or networks.
2. Provide single internet service from multiple servers, sometimes known as server farm.

Table 1
Hardware resources required

Sr. No.	Parameter	Minimum Required
1	RAM	2GB
2	Processor	Pentium 4 Processor and Above
3	HDD	160GB and above

Software Resources Required:

- Eclipse
- Apache Tomcat Server 7
- JDK 1.7
- MySQL 5
- MySQL Workbench 6

5. Algorithm/Methodologies Details

A. *Advanced Encryption Standard (AES)*

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.

All key lengths are deemed sufficient to protect classified information up to the "Secret" level with "Top Secret" information requiring either 192- or 256-bit key lengths. 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 plain text and transform it into the final output of cipher text.

B. Working and process

The system overview diagram provides an overview of the system with the important modules in the form of blocks. At first the users upload their files which are then utilized to achieve the hash keys for the de-duplication decision making that is used for labeling, indexing and label storage where the data can be accessed by the end user and the data is stored on the amazon RDS.

Architectural design:

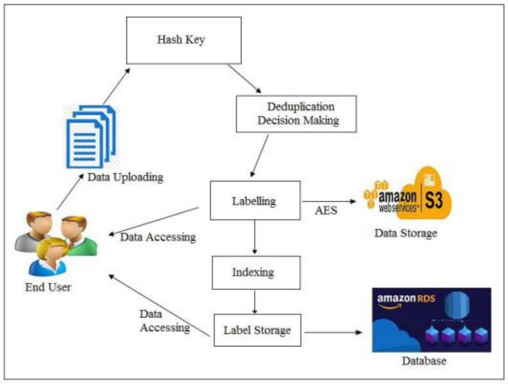

Fig. 1. Architectural diagram

Phase 1: This Phase Data is Be uploaded by authenticated user. De-duplication System applies Reverse circle Chiper Encryption Algorithm and Data is been sent to Next Phase.

Phase 2: All Encrypted Content is been Hashed Using MD5 Algorithmic procedure. And File to HashList is been maintained.

Phase 3: once the document hash is created it experienced through the procedure of deduplication. In this procedure sprout channel accumulated all the hash estimation of past records. In this arrangement of past qualities, new hash esteem is analyzed. When match is discovered, this hash esteem s bolstered to the similarity index relationship calculation for location of % connection. In the event that the esteem returned by similarity index is 1 then the record is copied and it won't spare to the cloud.

Phase 4: when duplication is recognized, every one of the references of this record with past documents are kept up for the future utilize.

Phase 5: In above case if File is not duplicate system flow comes to this step.

Phase 6: All Information of File is been saved using Inverted Index . Cloud plugin have been used to reduplicate cloud data

Phase7: De-duplication System provides feature to download file which was upload by the user.

6. Conclusion

Cloud storage space program offered by cloud computing has grown in recognition because of contemporary IT Industry development, digitalized development and also planet in deep Internet, Social Media, Smartphone, etc., It gives idea to discuss and make a methodological survey on secure Cloud Storage, Chunking algorithms and data De-duplication techniques, mainly growing Cloud Computing for efficient utilization in data storage to speed up the backup process, reduce the network delay within data centers, reduce server hit rate, reduce the power Efficiency. There are many businesses that make use of these services, and many of their employees have access to a single file, which results in several duplicates. With limited capacity and pricey online storage choices, the duplicates may take up extra space, which can be problematic. As a result, a de duplication technique is required so that duplicate files on a shared cloud storage system can be removed, enhancing user experience and consuming less space. The suggested approach makes use of the Amazon RDS strategy, which produces outcomes that are more than satisfactory.

In the future this system can be enhanced to work as the API or plugin software through which we can provide the services for other applications running in cloud

References

[1] Jinbo Xiong, Yuanyuan Zhang, Shaohua tang, Ximengl liu and Zhiqiang Yao, "Secure encrypted data with authorized deduplication in cloud," IEEE Access, vol. 7, pp. 75090–75104, Jun.2019.

[2] Hui cui, Robert H. Deng, Yingjiu Li, and Guowei Wu, "Attribute-based storage supporting secure deduplication of encrypted data in cloud," IEEE transactions on big data, vol. 5, no. 3, July-September 2019.

[3] Hua Ma, Ying Xie, Jianfeng Wang, Guohua Tian, and Zhenhua Liu, "Revocable attribute-based encryption scheme with efficient deduplication for e-health systems, " Volume 7, 2019.

[4] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou, "A hybrid cloud approach for secure authorized deduplication," IEEE transactions on Parallel and Distributed systems, 2015.

[5] Chen, R., Mu, Y., Yang, G., and Guo, F., "BL-MLE: Block-level message-locked encryption for secure large file deduplication," IEEE Transactions on Security, 2015.

[6] Yongjun Zhao and Sherman S. M. Chow, "Updatable block-level Message-locked encryption," Proc. IEEE Transaction on Dependable and secure computing, MAY 2019.

[7] Maozhen Liu, Chao Yang, Qi Jiang, Xiaofeng Chen, Jianfeng Ma, Jian Ren, School of Cyber Engineering, Xidian University, Xi'an, Shaanxi, " Updatable block-level deduplication with dynamic ownership management on encrypted data".

[8] Waghmare, V., and Kapse, S., "Authorized deduplication: An approach for secure cloud environment, " 2016.

[9] Hyungjune Shin, Dongyoung Koo, Youngjoo shin, and Junbeom Hur, "Privacy-preserving and updatable block-level data deduplication in cloud storage services," Proc. 2018 IEEE 11th International Conference on Cloud Computing.