JCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Issue-8 E-ISSN: 2347-2693

A Review: Various Data De-Duplication Techniques to reduce memory consumption over Cloud Storage

Faheem Ahmad Reegu

Jazan University KSA, Saudi Arabia

Available online at: www.ijcseonline.org

Received: 22/Jul/2016 Revised: 30/Jul/2016 Accepted: 18/Aug/2016 Published: 31/Aug/2016

Abstract— The growth amount of data, it has converted an important problem to minimize the storing space that form information conquers and the bandwidth ingestion through system broadcast. To attain advanced de-duplication elimination proportion, the out-dated way is to extend the variety of information for data de-duplication, but that would make meta-data areas larger and increase the several of meta-data objects. When identifying the redundant information meta-data needs to continuously introduced and transferred into the recollection and access block will be process. Data de-duplication is most significant methods for removing identical copies of frequent records. It has been mostly used in cloud storing to minimize the quantity of storing space. To defend the confidential approved identical check is used. Dissimilar from the out-dated duplicate scheme, dissimilar rights of consumer are measured also the records itself. We study this approved duplicate check in mixture cloud architecture. The crossbreed architecture suggests about together the Private and Public Cloud. The private cloud plays a significant role in this system, moreover the security of this system is less and the private cloud of this apparatus is not secure, unofficial data subsequent in fail in security. Basically, it provides more safety, the isolated cloud is providing with multilevel verification. We have showed that our system is providing privacy to data and more secure. This paper, brief study of existing scheme has been described along with the construction and algorithms.

Keywords—Data De-duplication, Hybrid Cloud,Private and Public Cloud,Secure Data Architecture and Algorithms.

I. Introduction

Cloud Computing is used in favor of the Internet, *so* cloud computing revenue kind of web based computing everywhere dissimilar military such as Servers, storeroom as well as submission are used by an organization's throughout the Internet. It provides online information storage, arrangement & application. Cloud storage is that model where information can be located, controlled, backed up, stored & adapted. Cloud storage makes obtainable information to clients in any time, with high storage space & also makes it user friendly so that availability of information increases. Cloud storage is of three kinds: Public, Private & hybrid [1]. Information de-duplication is one of significant information density techniques for eliminate duplicate copies of repeat information & has been generally used in cloud storage in order to reduce the quantity of storage space & save bandwidth. For defense of information security, this paper makes a challenge to primarily attend to the predicament of accepted information de-duplication. Data De-duplication in the cloud is a novel technology that quickly decrease amount of digital data in information storage. It is basically the process of verifying the redundancy within his or her data and deleting it, but one client data de-duplication is not very cost saving. To maximize the benefits of data Deduplication, cross user De-duplication is exercised [2].

II. Method of Deduplications

There are some methods of de-duplication are discussed. To find the information de-duplication hashing algorithm & application aware local global de-duplication is used which will perform better results [3].

A. *Folder Level De-duplication*

In folder whole box file is checked for de-duplication files**.** When clients want to upload the file then by using hash algorithm like Rabin fingerprint, MD5, SHA-1, etc. a fingerprint of that file is generated. This fingerprint is unique for all different files. This fingerprint is used to store in place of whole file which reduces the storage space in information storage. An indicator is used to point the original file for the following copy.

B. *Block or Sub file De-duplication*

In block level whole file is divided into number of blocks or sub blocks [4]. Then mix up worth of these blocks is considered & then evaluate to hash value of each block. If the hash value of the block is unique then it is consider as a unique & stored in the information base, otherwise a pointer is stored. Only pointers are stored in the storage not the block of file which saves storage.

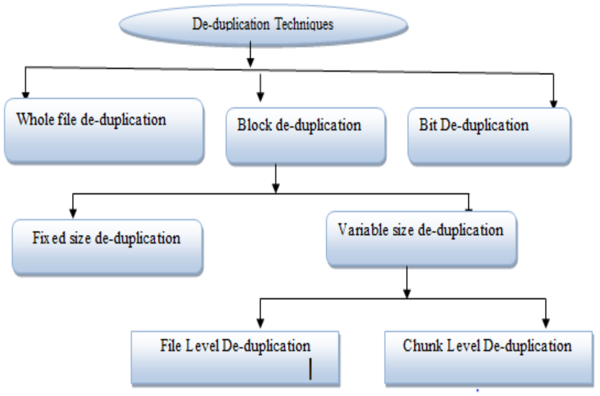
De-duplication Techniques
Whole file de-duplication
Block de-duplication
Bit De-duplication
Fixed size de-duplication
Variable size de-duplication
File Level De-duplication
Chunk Level De-duplication

Figure no: 1 Methods of Deduplication

According to the amount of chunk there are two processes in block De-duplication.

a) Fixed-Length block: In Fixed length blocks are divided into fixed length which is defined by the user. After that its duplication is checked. This method is speedy, easy & lowest amount CPU overhead.

b) Changeable Length block: In which the blocks are divided into variable sizes. Blocked are stored according to the size & then its de-duplication is checked by using de-duplication techniques [5].

Table no: 1 Techniques of De-duplication

Serial no.	**Techniques**	**Advantages**
1.	Folder Level De-duplication	Simple & fast
2.	Block or Sub file De-duplication	To reduce capacity in the 20:1 to 50:1 range for stored data

III. RELATED WORK

Vasilios et al [6] presents a migration support network, in which fundamental elements are cost effective system. They proposed a three level framework that satisfies al the necessity in view of cost assumption. They utilized the windows azure policy as a part of creating prototyping model. Besides, the ability to consolidate necessities for numerous administration sorts, e.g., information stockpiling & systems administration, is imagined to be given, encouraging the choice making in relocation sorts past the off-stacking of the application stack on a VM. **Haitao et al.** [7] proposed relocation methods taking into account (dynamic, receptive & shrewd procedures), albeit basically in light of the present information , can make the mixture cloud-helped VoD organization set aside to 30% transmission capacity cost contrasted & the Clients/Server mode. They can likewise handle unpredicted the glimmer group activity with little cost. It likewise demonstrates that the cloud cost & server transmission capacity picked assume the most essential parts in sparing expense, while the distributed storage size & cloud substance upgrade system assume the key parts in the client experience change. **C. Ward et al**. [8] acquainted the augmentations with a coordinated mechanization capacity called the Darwin structure that empowers on load movement for this situation & talk about the effect that computerized relocation has on the expense & dangers ordinarily connected with relocation to cloud. **Kang et al.** [9] proposed the migration algorithm .The VM to its best PM specifically, with the proviso that it has adequate capability. Then, if the relocation restriction is gratified, we transfer a different VM after this PM to oblige the novel VM. In addition, we learn a mixture system where a lot is working to recognize forthcoming VMs for the on-line expansion. Assessment upshots establish the great competence of our method.

IV. HASHING ALGORITHMS

Hashing the information means creating a mathematical form of specific information set that is unique for all information sets. For this hashing value we need algorithm which is known as hashing algorithm [10]. These hashing algorithms generate hash value or fingerprint by using some steps. Some of the hashing algorithms are disused below which have their own properties:-

1. MD 5 Algorithm
2. SHA -1 Algorithm
3. SHA-2 Algorithm
4. Havel Algorithm
5. Whirlpool Algorithm

A. MD-5 Algorithm

It is usually used for cryptographic purposes with a 128bit mix up worth. MD5 has been working in an extensive diversity of safety submission, & in addition[11] usually worn to make sure the dependability of accounts. An MD5 mix up is in universal articulated as a 32-digit Hexadecimal digit.

B. SHA Algorithm

Different kinds of SHA are SHA0, SHA 1, SHA2,SHA3, SHA 256 , SHA512. The safe Hash Algorithm is 1 of a figure of cryptographic hash purpose. There are at present 3 production of safe Hash method:

- SHA-1 is the unique 160-bit hash purpose and is alike to the MD5 method.
- SHA-2 is a relative of 2 alike hash functions, with dissimilar obstruct sizes, recognized as SHA-256 & SHA-512. SHA-256 uses 32-bit words where SHA-512 uses 64-bit words.
- SHA-3 is a prospect hash purpose criterion still in expansion.

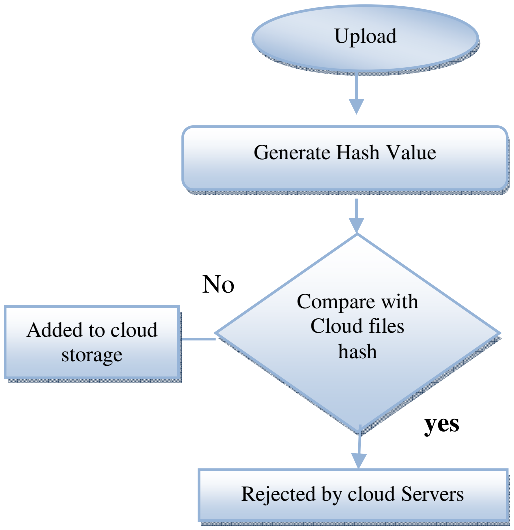

Figure no: 2 Normal Scenario of the Hash Algorithm

Table 1 Comparison of hashing algorithms

Algorithms	O/P(bits)	Internal state size	Rounds
WHIRLPOOL	512	512	10
PANAMA	256	8736	64
Havel	160	256	160
MD2	128	384	864
MD5	128	128	64
SHA-0	160	160	80
SHA-1	160	160	80
SHA-2	256	256	64

Pseudo Code SHA-1 Algorithm

```
Some Addition data to the last of the input
Set the start sha-1 values

For each 64-ute chunk do
  Extend the chunk to 320 bytes of data
Achieve 1st set of operations on chunk [j]  (x20)
Achieve 2nd set of operations on chunk [j] (x20)
Achieve 3rd set of operations on chunk [j] (x20)
Achieve 4th set of operations on chunk [j] (x20)

end
          return sha-1 values as a hash
```

Table 2 Comparison between Hashing Algorithms

SHA2	MD5	SHA1
0.5	0.5	1.0
0.5	1	2.1
0.8	1.5	3.3
1	1.8	4
1.1	2	5

In hashing technique some hashing algorithms are used. These hashing algorithms have their own properties like their output size, block size, rounds & performance. Hashing technique is used after uploading the file. When fingerprint of the file is generated then it is stored in the metadata & used for the comparison purpose. From the above block diagram the file[12] to be uploaded is fed to the hashing method which generates the hash worth. The hash value is compared with already existing hash values. If a matching hash value is found the particular file will not be added to the cloud storage, else server will store the file.

Pseudo Code of De-duplication

```
Step1: Calculate the two convergent key values

Step 2: Compare the two keys and files gets accessed.

Step3: Apply Deduplication to eradicate the duplicate values.

Step4: If any other than the duplicates it will be checked once again and make the data unique.

 Step5: That data will be unique and also more confidential the authorized can access and data is stored.
```

V. PROBLEM FORMULATION

Deduplication makes system a network efficient and storage optimization systems. At present, in the background of customer information allocation proposal the contests for great scales, extremely unnecessary internet information storing space is great. Due to this idleness storing space charge is decreases. Loading for this gradually central network information can be receiving by its de-duplication. The problems with existing information storage system. [13]

1. If we consider a case in which user update one same file on multiple time, it take space a lot on server memory.
2. If server have large amount of information then searching technique become slow.
3. Unwanted space consumption is a very costly when user are in billion.
4. Current hashing function or searching technique is not much better It is a more time consuming process to search any of records or de- duplicate any new content.

Data is most important in the system so we need accurate verification generator algorithm which finds files fast and accurately and current systems have this type of functions but the accuracy is less.

VI. ENCRYPTION ALGORITHM

A Public encryption key scheme is additively holomorphic if certain two cipher texts d1 = enc(pk,n1;r1) and d2=enc(pk,n2;r2) , it is possible to effectively calculate end(pk,n1+n2;r) with a novel random value r which depend upon r1 and r2 but can't be[14] resolute by anyone who knows only r1 and r2, even without information of the corresponding private key. Several algorithms come into the Deduplication.

- RSA Algorithm
- DES Algorithm

Pseudo Code of RSA algorithm

```
a) Generate keys:
    1) Generate tow larger prime number (P and Q).
    2) Solve N=PQ.
    3) Solve M=pi(N)=(p-1)(Q-1) // Euler method)
    4) Select any integer E, the rules to add E are:
        a) E is real int
        b) GCD(M,E)=1….Greater common divisor.
    5) Evaluate the Euler Theorem.

b)Encode/Identification:
Real Plain text (a block val)=X..X<N
Chiphertext = C..C(X^E)mod N
c)Decoder/Signing
Chiphertext=C
Dechiphertext=Y
```

A. *Password Authenticated Key Exchange*

Password based protocols are normally used for user authentication. However, such rules are susceptible to offline brute force attacks since client tend to select pwd with respect the low entropy that are hence guessable[15].

1. The model functionality for pwd authenticated key exchange.

Inputs:

- Alice's initial value is a password pwda;
- Bob's initial value is a password pwda;

Outputs:

- Alice's destination input is ka;

- Bob’s sink is kb;

Where id pwda=pwdb then ka = kb, and if pwda ≠ pwdb then bob w.r.t. Alice can’t differentiate ka and kb from the random character of the similar length.

2. The ideal functionality Fsame-Input-pake

Inputs:

- Alice initial value is a pwd pwda;
- There are m parties P1,…..Pn with Pwds Pwd1,…..Pwd,. respectively.

Outputs:

- Alice sink value is ka,1,….,Ka,n;
- Pi’s output is kb,I;

Where if pwda = pwdi then ka,i – kb,I, and otherwise pi can’t distinguish ka,I from the random character of the same length.

VII. Performance parameters

A. Memory Consumption: - De-duplication technique will affect the storage area of database. The sum of all file size in bytes will be evaluated before de-duplication and same procedure will be done after applying de-duplication technique. The difference of the above two will evaluate saved storage space or memory consumed. From this parameter we can assure that storage size is greatly reduced and memory is used efficiently.

B. Detection Time: **-** It is the time required to search for a duplicate file in a database. The inbuilt timer classes are responsible for their calculations**.**

C. Hash time: - It is the time required to upload the file according to its size and the time it takes for an algorithm to calculate its hash value. It is the obvious fact that hashed value of the file will take less time to get stored in database than the whole original file itself. Thus de-duplication will consume less uploading time.

D. Detection Accuracy: - This parameter will evaluate the constraint of how well our system is working accurately i.e. whether de-duplicator is working correctly or not.

VIII. Conclusion

Cloud is the costly storage provider, so the motivation is to use its storage area efficiently De-duplication has been proved to reduce memory consumption by removing the useless duplicate files. So far from the previous studies file level de-duplication is the better approach to be used, the focus of the proposed work will be on file level de-duplication. In this work, we propose a dynamic data Deduplication scheme for cloud storage, in order to fulfill a balance between changing storage efficiency and fault tolerance requirements, and also to improve performance in cloud storage systems. A lot of research has been carried out over this by means on hashing algorithm. From the previous hashing algorithms SHA2 will perform better than MD5 and SHA1.Our aim is to choose a well-built algorithm which will generate a good hash value in turn reducing cloud storage.

In this proposed work the use of Microsoft azure provides the replica of the cloud computing environment which is used by many companies. Thus the work can easily be accomplished by the use of cloud framework without any cost consumption usage.

References

[1] Kaushik, Vandna Dixit, Amit Bendale, Aditya Nigam, and Phalguni Gupta. "Certain Reduction Rules Useful for De-Duplication Algorithm of Indian Demographic Data." In *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, pp. 79-84. IEEE, 2014.

[2] Li, Min, Shravan Gaonkar, Ali R. Butt, Deepak Kenchammana, and Kaladhar Voruganti. "Cooperative storage-level de-duplication for I/O reduction in virtualized data centers." In *2012 IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pp. 209-218. IEEE, 2012.

[3] Backialakshmi, N., and M. Manikandan. "Data de duplication using N0SQL Databases in Cloud." In *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, pp. 1-4. IEEE, 2015.

[4] Nagarajaiah, Harsha, Shambhu Upadhyaya, and Vinodh Gopal. "Data De-duplication and Event Processing for Security Applications on an Embedded Processor." In *Reliable Distributed Systems (SRDS), 2012 IEEE 31st Symposium on*, pp. 418-423. IEEE, 2012.

[5] Saritha, K., and S. Subasree. "Analysis of hybrid cloud approach for private cloud in the de-duplication mechanism." In *Engineering and Technology (ICETECH), 2015 IEEE International Conference on*, pp. 1-3. IEEE, 2015.

[6] VasiliosAndrikopoulos, Zhe Song, Frank Leymann, “Supporting the Migration of Applications to the Cloud through a Decision Support System”, Institute of Architecture of Application Systems, IEEE, pp. 565-672, 2013.

[7] Haitao Li, LiliZhong, Jiangchuan Li, , Bo Li, KeXu, “ Cost-effective Partial Migration of VoD Services toContent Clouds”, 2011 IEEE 4th International Conference on Cloud Computing, pp. 203-110, 2011.

[8] C. Ward, N. Aravamudan, K. Bhattacharya, K. Cheng, R. Filepp, R. Kearney, B. Peterson, L. Shwartz, C. C. Young, “Workload Migration into Clouds – Challenges, Experiences, Opportunities”, 2010 IEEE 3rd International Conference on Cloud Computing, pp. 164-171, 2010.

[9] Kangkang Li, HuanyangZheng, & JieWu . “Migration-based Virtual Machine Placement in Cloud Systems”, 2013 IEEE 2nd International Conference on Cloud Networking (CloudNet, IEEE, pp. 83-90, 2013.

[10] Andoni, Alexandr, and Piotr Indyk. "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions." In *2006 47th Annual IEEE Symposium on*

Foundations of Computer Science (FOCS'06), pp. 459-468. IEEE, 2006.

[11] Venkatesan, Ramarathnam, S-M. Koon, Mariusz H. Jakubowski, and Pierre Moulin. "Robust image hashing." In *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3, pp. 664-666. IEEE, 2000.

[12] Swaminathan, Ashwin, Yinian Mao, and Min Wu. "Robust and secure image hashing." *IEEE Transactions on Information Forensics and Security* 1, no. 2 (2006): 215-230.

[13] Sengar, Seetendra Singh, and Manoj Mishra. "E-DAID: an efficient distributed architecture for in-line data de-duplication." In *Communication Systems and Network Technologies (CSNT), 2012 International Conference on*, pp. 438-442. IEEE, 2012.

[14] Rivest, Ronald L. "The RC5 encryption algorithm." In *International Workshop on Fast Software Encryption*, pp. 86-96. Springer Berlin Heidelberg, 1994.

[15] Barrett, Paul. "Implementing the Rivest Shamir and Adleman public key encryption algorithm on a standard digital signal processor." In *Conference on the Theory and Application of Cryptographic Techniques*, pp. 311-323. Springer Berlin Heidelberg, 1986.