IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

Application-aware big-data de-duplication in cloud computing

Jagadeeshwari J[1*], Joshitha Gunreddy[2], Harshitha H G[3], Harshitha R[4], Spoorthi Rakesh[5]

[1,2,3,4,5]School of Computing and Information Technology, REVA University, Bangalore, India

Corresponding Author: jagadeshwarij6@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si14.121124 | Available online at: www.ijcseonline.org

Abstract— Cloud storage has been widely used because it can provide seemingly unlimited storage space and flexible access, while the storage resource is vulnerable to the cost issue since the data should be maintained for a long time. Data deduplication techniques make sure that only one distinctive instance of knowledge is maintained on storage media.In this paper, we discuss the benefit when a deduplication technique is adopted to the cloud storage, then we propose a deduplication framework for cloud environments. The deduplication application divides a given file into smaller chunks, then searches the index table that consists of hash values of chunks to judge duplicate data, finally stores non repeated chunks.

Keywords— Deduplication, MD5, Cloud, Hash

I. INTRODUCTION

Cloud storage is a finest way to store large amount of data, due to its high reliability and scalability. Cloud storage is a model in what data is being stored managed and maintained. When the storage in the cloud increases automatically the cost increases exponentially. It has become essential to make use of the efficient cloud storage, since earlier most of the companies were depending on in-house servers for storing the data, but now the best thing about the cloud is that the data can be accessed from any part of the world, and data de-duplication is another advantage of the cloud where it not only saves space but also removes the redundant data.

It is made easy to send and receive the data or files from the cloud. The data before being stored in the cloud will be divided into many blocks and sequences then they are being stored in the different cloud storage media, they are in turn being owned by storage providers, The users from the different parts can recollect the data using the sequence numbers the file is brought back to the original format and downloaded.

II. RELATED WORK

The related work of a study is an important part as it provides the context and purpose of the study. Hence there is a need for related work study that contributes to prepare different aspects of data de-duplication system.

A. Cloud data storage

In the most up-to-date decade, the request of outsourcing knowledge is considerably expanded. Cloud storage is a remote platform that uses an extremely virtualized, multi-tenant infrastructure to supply enterprises with ascendable storage resources which will be provisioned dynamically by the organization. Information storage and smart performance are the basic wants which must be satisfied.

These services are provided by varied distributed computing specialist organizations like Drop box, Google App Engine, Microsoft, AmazonS3 and lots of additional. Cloud based storage has many distinctive attributes that frame it enticing for enterprises making an attempt to vie in today's data-intensive business setting.

- The resources are distributed such that the availability is increased.
- In case of any disasters or faults, the resources are replicated for recovery.

B. Data Deduplication

In computing, information deduplication is a technique for eliminating duplicate copies of identical information. It is also called single -instance storage. This technique is employed to boost storage utilization and might even be applied to network data transfers to scale back the amount of bytes that has got to be sent. In de duplication process, unique data is identified and stored in the cloud. Further process continues with comparing the new chunk that has to be stored in the cloud with the chunks already present in the cloud. If it already exists the duplicate chunk is replaced with a small reference that points to the stored chunk.

Advantages of Data De-duplication in cloud are:

- Clears storage space
- Adept replication
- Effective use of network bandwidth
- Cost-effective

III. METHODOLOGY

This paper gives the idea of how to eliminate redundant data and store only unique copy of the data. Using MD5 algorithm (cryptographic hash function) the system is developed which does this function.

The work carried out involves 4 phases which discusses the system architecture and the algorithm used for this project. Further, in the next phase the different techniques will be discussed and finally, in the last phase the work flow of the project is explained in detail. The system is designed to be user friendly by developing an interactive GUI (Graphical User Interface).

3.1 SYSTEM ARCHITECTURE

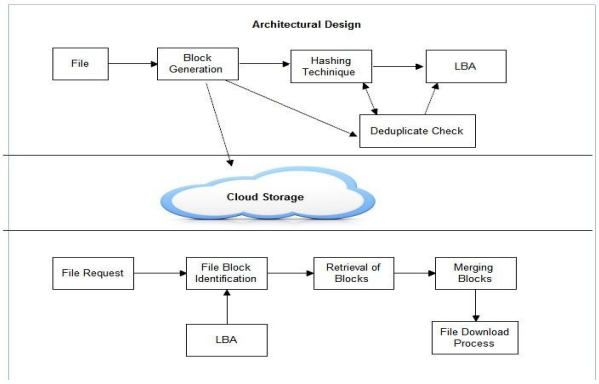

Figure 1: System Architecture

- LBA: Logical block addressing.
- Hashing technique: MD5 algorithm.

The above fig 1 deplicts the architectural design of the proposing system. New users should get users ID and password to upload their files when the file is uploads. The first phase is dividing the file into blocks. Then these blocks are processed through a hashing technique (MDS) and then checked if the same redundant data is already present in the cloud. If it is a different block of data. Then logical block addressing in cloud else a pointes will be initiated to store the address of similar block that is existing in the cloud.

In the download process when the file is being requested the block are retrieved through the help of logical block addressing then the entire file which is the result of merging of all these blocks is downloaded.

3.2 ALGORITHM USED

A. MD5- Message Digest 5

The MD5 hashing algorithmic rule is a unidirectional cryptographic operation which suggests the inverse of the operation cannot be computed. Message of any length can be given as input to this algorithm and after the hash digest is generated, it outputs a fixed length digest value of 128 bits which is used to authenticate the original message.

Working of the algorithm:

Step 1: Receive the Message and Transform it into bits.
Step 2: Append the required Padding Bits (Make the message bit length to be the exact multiple of 512 bits as well as 16 wordBlocks). A single "1" followed by required number of "0"s are added.
Step 3: Split the total bits thus obtained into 128 bits blocks each
Step 4: Initialize a four word Buffer (A, B, C, D) which is used to compute the message digest total 128 bits.
Step 5: Perform AND, OR, NOT and XOR operations on A, B, C and D by taking 3 of them as inputs and get one as output.
Four auxiliary functions that take as input 3 32-bit words and turn out as output one 32-bit word.
Step 6: Step 6 is repeated until 128 bits hash (16 bytes) is obtained. Output starts with the low-order computer memory unit of A, and finish with the high-order computer memory unit of D.
Step 7: Stop

Benefits of MD5:

- Its faster approach is to look at the exact sizes of the files (chunk) and then only compare the ones that are of the same size.
- It is faster, has less number of steps and also produces a small length of output when compared to other algorithms.

3.3 TECHNIQUES USED

There are four techniques used:

- **Chunk splitting technique**
  In this technique, the image file or text file you uploading will be divided into number of blocks based on the packet size for the duplication check.

- **Hashing technique**
  In this technique, the hash code will be generated for the each block using MD5 algorithm which is divided based on the packet size and the hash code generated is unique for each block.

- **De-duplication Checking Process**
  In this technique, the blocks which are divided based on the packet size which are given a unique code will be uploaded to cloud. Before uploading to cloud, duplication of blocks are checked based on the code generated whether the block already present or it is different.

- **Chunk Merging Process**
  When the user downloads file, this component will consult the hash server component to obtain all chunks' address of this file, and then sequentially access the cloud to get data blocks according to these chunks address, eventually this component merges these data blocks into a complete file.

3.4 WORK FLOW

The following flow charts clearly explain how the file upload and file download takes place.

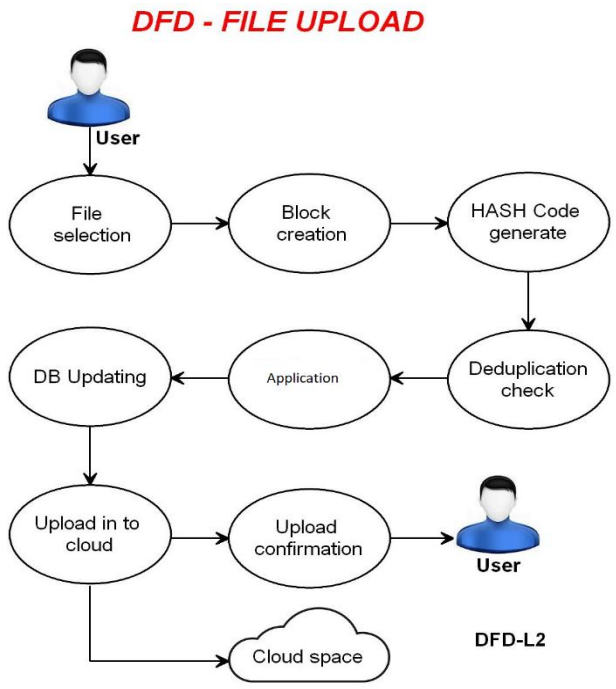

Figure 2: File upload process

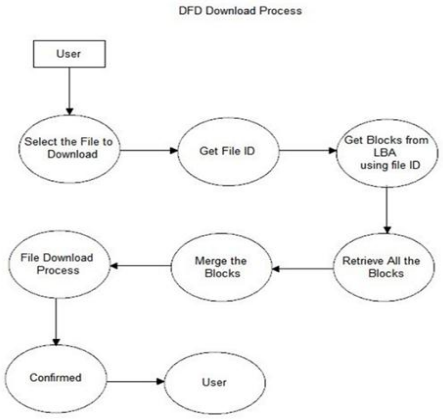

Figure 3: File downloads process

IV RESULTS AND DISCUSSION

The result of proposing system tells how effective and how it is applicable to check duplicate storage required. This system not only reduces the storage capacity required but also it improves the speed of duplication check for the file uploaded.

De-duplication has become widely effective when applied across multiple users and it also plays a vital role in cloud backup that reduces the storage space in cloud and can be used for further use and not only reduces the storage space but also saves the network bandwidth.

V CONCLUSION AND FUTURE SCOPE

The paper mainly centers around the information de-duplication philosophy and wordings as for the capacity system. Data deduplication is the new information computation modernization which removes duplication in data. When any picture or text is uploaded, the framework advices the staggered hash file to determine whether the piece is rehashed before sending the chunk and it just sends the unrepeated lump to the fundamental stockpiling system, consequently it can diminish the measure of correspondence and improve the extra room utilization. It varies from the pressure procedures by taking a shot at the information at sub-document level where as pressure encodes the information in the document to lessen its stockpiling requirement.

However pressure can be utilized to increase information de-duplication to give higher dedupe ratio. Small to huge organizations are undertaking this embracing new innovation as it gives huge rate of profitability by decreasing the capacity limit required to store the information and lessening system transmission capacity required to exchange the information.

REFERENCES

[1] Akanksha Upadhyay, Abha Sharma, DIFFERENT SECURE DATA DEDUPLICATION APPROACHES FOR CLOUD STORAGE: A REVIEW. IJARCS Volume 9, No. 3, May-June 2018
[2] Sarah Prithvika P.C , Ramani S , Jakkulin Joshi J and Sindhu K, Data Deduplication in Cloud Environment – A Survey. www.ijlemr.com || Volume 03 - Issue 01 || January 2018 || PP. 44-4
[3] Sumedh Deshpande, Anupama Murkute, Saurabh Patil, Sachin Kamble, Prof. Swati Shekapure, Deduplication on Encrypted Big data in Cloud. IJIRSET Vol. 7, Issue 4, April 2018 Vol. 7, Issue 4, April 2018
[4] Priyanka G. Masal, B.M. Patil , Encrypted Big Data with Data Deduplication in Cloud. International Journal of Computer Applications (0975 – 8887) Volume 174 – No.6, September 2017
[5] Bhairavi Kesalkar , Dipali Bagade , Manjusha Barsagade , Namita Jakulwar, Prof. Shrikant Zade , IMPLEMENTATION

OF DATA DEDUPLICATION USING CLOUD COMPUTING. IJARIIT(ISSN: 2454-132X) KITE/NCISRDC/IJARIIT/2018/CSE/105

[6] Sabale Nikita C, Prof.N.G.Pardeshi, A Survey Paper on Deduplication on Encrypted Big Data Using HDFS Framework. IJIRCCE Vol. 5, Issue 6, June 2017

[7] Priyadharsini.P, Dhamodran.P, Kavitha.M.S, A SURVEY ON DE-DUPLICATION IN CLOUD COMPUTING. IJCSMC, Vol. 3, Issue. 11, November 2014, pg.149 – 155

[8] Kinzal Patel, Prof. Kapildev Naina, Review on Data Deduplication In Cloud Computing. IJAERD Volume 4, Issue 11, November -2017

[9] Vaishnavi Moorthy, Arpit Parwal and Udit Rout, DE-DUPLICATION IN CLOUD STORAGE USING HASHING TECHNIQUE FOR ENCRYPTED DATA. ARPN VOL. 13, NO. 5, MARCH 2018

[10] Sadhana Poornachandra Rao, M.Kusuma, Application-Aware Big Data Deduplication in Cloud Environment. April 2018 | IJIRT | Volume 4 Issue 11 | ISSN: 2349-6002

[11] Dastagir Shaikh, Pratik Sen , Zubair Inamdar, Avoidance of Duplication of Encrypted Big-data in Cloud Storage IJARCCE Vol. 6, Issue 4, April 2017

[12] Supriya Milind More, Kailas Devadkar, A Comparative Survey on Big Data Deduplication Techniques for Efficient Storage System. IJIACS ISSN 2347 – 8616 Volume 7, Issue 3 March 2018.

[13] Zheng Yan, Deduplicatrd on encrypted big data in cloud. Volume 02 no 02 april-june 2016

