ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# Optimized Storage Approaches in Cloud Environment

**Sri M.Tanooj kumar, A.Radhika**
Department of Computer Science and Engineering, Andhra Loyola Institute of Engineering and Technology, Andhra Pradesh, India

## Abstract

The cloud computing is a technology using internet and remote servers the users can run and deploy their applications .Different services are provided by cloud service providers(CSP).They are software-as a service , plat-form as a service, infra-structure - as a service etc. Storage as a service is also provided by CSP to back up the users data in to cloud .The storage service provided by the CSP is cost effective, reliable. These services are called "pay-as-you-go". That is depending on the usage of service consumer is charged. So for maximum utilization of cloud storage optimization is a task .In this paper we propose a technique to optimize the usage of cloud storage space .The work in this paper mainly focuses on Deduplication and Compression techniques. Deduplication is a technique used to identifying the redundancy of the data and eliminating the duplication of data there by maintaining only one copy the original data. For optimum utilization of cloud storage the file is compressed after Deduplication before storing into cloud storage. So the bandwidth and time required to transfer file over the network is reduced.

**Keywords**: Cloud storage, Deduplication, Compression, Optimization,CSP.

## Introduction

Cloud computing is a new form of distributed computing mode after grid computing. This technology allows many businesses and users to use the data and application without an installation. Cloud storage is a kind of cloud computing[1-3].since 2006 there have been some of the more successful cloud facilities such as Amazon's Elastic compute cloud, IBM's blue cloud, Nimbus, Open Nebula and Google's Google App engine and so on. Best Buy's Gift tag applet uses Google appending to let users create and share wish lists from web page's they visit .Different services are provided by CSP among them IaaS is to start, stop, access and configure their virtual servers and storage. The main advantage of cloud storage is to provide space for data storage and enables users to access data at any time. Many big internet based companies have come to realise that only a small amount of their data storage capacity is being used. This has led to the renting out of space and storage of information on remote servers or clouds. Information is temporarily cached on desktop computers, mobile phones or other internet linked devices.

Cloud storage is a computing model where data is stored on multiple dedicated storage servers. The storage servers may be hosted by third parties. In computing environment large volumes of duplicate copies of data exist. Deduplication technique reduces these redundant copies of data.For example when a backup application creates a backup, depending on the data criticality it creates a series of individual files or a big file. Back up of these files has to be taken, every back up creates a redundant copy of data .A lot of n/w resources are required
to transmit these large volumes of data over the n/w. Data deduplication is a process where duplicate copies of data are eliminated and maintaining only one copy of the original data over the server [4].This reduces the storage space, cost of storage space and n/w bandwidth. The existing system send all the blocks of data, without eliminating the duplicate copies .So we propose a system, where only the deduplicated unique blocks are send over the n/w.

The rest of the paper is organized as follows.Chapter2 describes different deduplication techniques.Chapter3 describes about compression techniques.Chapeter4 describes system design.Chapter5 describes results.Chapter6 describes conclusion.

## Background and motivation

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data .This is used to improve storage utilization and also can be applied to n/w data transfers to reduce the number of bytes that must be

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

sent .In deduplication process unique chunks of data Or bytes 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 point to the stored chunk .Given that the same byte pattern may occur dozens, hundreds or even thousand times, the amount of data that must be stored or transferred can be greatly reduced.

**Detection of similar data**:
The process of detection of identical files is mainly based on two levels and they are file level and the data block level or data at a sub file. The results for both will vary .The difference lies in the amount of reduction each produces and the time each approach takes to determine what is unique.

**File level deduplication**
*:*It is also called as single instance storage (sis),file level data deduplication compares a file with those already stored by checking it's attributes against an index.If the file is unique ,it is stored and the index is updated, if not only a pointer to the existing file is stored. The result is that only one instance of the file is saved and subsequent copies are replaced with a stub that points to the original file. This approach provides a scalable solution with the division of chunk index into two tiers namely Primary and Secondary index [5].In this approach, all the chunk_ IDs that constitute a file and the minimum chunk_ID among them are found.

This minimum Chunk_ID is termed as representative Chunk_ID ..According's to boarder's Theorem, two files are said to be nearly similar, when the representative Chunk_IDs of both files are same. Primary Chunk index consists of representative Chunk_ID,whole file hash address of the secondary index or bin.Bin is made up of three fields namely, Chunk_D,Chunk size and the storage address of the chunk[6].

When ever a file has to be backed up,it is chunked and both the representative Chunk_ID and the hash value for the entire file are found.Representative chunk_ID is searched in the primary index and if it is not,present,the incoming file is considered as new.Hence a new bin is created and all chunk_ID's,their corresponding size and a pointer to the actual chunks are added to the disk. Representative Chunk-ID,hash value of new file and the pointer to the newly created bin are added to the primary index.If the Representative Chunk_ID of the incoming file is already present in the primary index but the hash value is not modified in the primary index and the updated bin is written back to the disk .If the whole file hash value of the incoming file is considered as a duplicate.Hence the bin need not be updated. Since every incoming chunk is checked only against the indices of similar files,this approach achieves better throughput compared to the chunk level deduplication.

**Block level deduplication:**
It operates on the sub file level .As it's name implies, the file is typically broken down in to segments, chunks [5] or blocks, that are examined for redundancy vs previously stored information. The most popular method for determine whole file detection is, using Hash technology. As for data block detection the fingerprint is checked through the fixed block size.

**The difference b/w file level and block based deduplication:**
A change within the file causes the whole file to be saved again. Block based deduplication would only save the changed blocks b/w the version of the file and the next.
File level is more efficient than Block based data deduplication. Indexes for file level deduplication are significantly smaller which takes less computation time when duplicates are being determined .Back up performance is less affected by the deduplication process.
File level requires less processing power due to the smaller index and reduced number of comparisions. The impact on the system performing the inspection is less.
The impact on recovery time is low .Block based deduplication will require "reassembly" of the chunks based on the master index that maps the unique segments and pointers to unique segments. Since file based approaches store unique files and pointers to existing unique files, there is less reassemble.

**Types of deduplication *:***
**In-line and post-processing deduplication**: It may occur "in-line" as data is flowing or "post-process" after it has been written.
**Post-process deduplication:** New data is first stored on storage device and then process at a later time will analyse the data looking for deduplication. The advantage is that there is no need to wait for the hash calculations and lookup to be completed before

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

storing the data there by ensuring that storage performance is not degraded. The disadvantage is that you may unnecessarily store duplicate data for a short time which is an issue if the storage system is near full capacity.

**In-line deduplication*:*** This is the process where the deduplication hash calculations are created on the target device as the data enters the device in real time .If the device spots a block that is already stored on the system it does not store the new block, just references to the existing block. The benefit of In-line over post process is that it requires less storage as data is not duplication. The disadvantage is that hash calculations and lookups take so long.

**Source vs Target deduplication*:*** It means where data deduplication occurs when the deduplication occurs close to where data is created, it is often reoffered to as "source deduplication".When it occurs near where the data is stored it is commonly called "target deduplication".Source deduplication ensures that data on the data source is deduplicated i.e it takes place directly within a file system. The file system will periodically scan new files creating hashes and compare them to hashes of existing files. When files with same hashes are found then the file copy is removed and the new file points to the old file. The deduplication process is transparent to users and back up applications .Backing up a deduplication file system will often cause duplication to occur resulting in the backups being bigger than the source data.

**Target deduplication:** It is the process of removing duplicates of data in the secondary storage. Generally this will be a backup store such as a data repository or a virtual tape library.

**Global deduplication**: Deduplication of files is performed across multiple systems.

**Hash functions**

Data integrity assurance and data origin authentication are essential security services in data storage .The broadest definition of authentication within computing systems encompasses identity verification, message origin authentication, ,message content authentication. In IPSEC (INTERNET PROTOCOL SECURITY) the technique of cryptographic hash functions utilized to achieve these security services

Hash functions compress a string of arbitrary length to a string of fixed length. they provide a unique relationship between the input and the hash value and hence replace the authenticity of a large amount of information by a much smaller hash value .In recent years there has been an increased interest in developing message authentication code(MAC) derived from hash code .Among the many reasons behind this are the cryptographic hash functions such as MD5 and SHA2.SHA2 is asset of cryptographic hash functions(SHA-224,SHA-256,SHA-384,and SHA - 512) DESIGENED BY THE NATIONAL SECURITY AGENCY.The method to implement the MAC IP sequrity has been choosen hash based HMAC,which uses an existing Hash function conjunction with a secret key .With minor modification, HMAC can easily replace one Hash Function with another[7].

MD5[8] is message digest algorithm takes as input a message of arbitrary length and produces as output a 128 message digest of the input. This is mainly intended in digital signature applications where a large file must compressed in a secure manner before being encrypted with a private key under a public key cryptosystem.

## Compression

Using Deduplication only duplicate copies of files are identified and eliminated there by maintaining only one copy of the original file but the file is not compressed.So using compression with deduplication we can save more amount of disk storage space.Data compression refers to the process of reducing the amount of data required to represent a given quantity of information.The aim of data compression is otreduce redundancy in stored or communicated data thus effectively increasing data density.it has important applications in the area of file storage ,data transmission, and distributed systems. The advantages of data compression are less disk space, faster reading and writing, faster file transfer, storing compressed data or transmitting it reduces the storage and transmission cost, compressing a file size to half of it's original size is equivalent to doubling the capacity of the storage medium.

Many algorithms have been developed to compress file sizes, two broad categories are Lossless data compression and lossy data compression .The compression technique used in this paper is entropy based encoding algorithm for Lossless data compression.

Huffman compression of files is based on the frequency of occurrence of a symbol in the file that is being compressed. The Huffman algorithm is based on statistical coding , which means that probability of a symbol has a direct bearing on the length of it's representation. The more probable the occurrence of a

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

symbol is, the shorter will be it's bit size representation. In any file, certain characters are used more than others. Using binary representation, the number of bits required to represent each character depends upon the number of characters that have to be represented. Using one bit we can represent two characters, i.e., 0 represents the first character and 1 represents the second character. Using two bits we can represent four characters, and so on. Unlike ASCII code , which is a fixed length code using seven bits per character, Huffman compression is a variable length coding system that assigns smaller codes for more frequently used characters and longer codes for less frequently used characters In order to reduce the size of files being compressed and transferred.

## System architecture

The Architecture of the optimized storage system is shown figure1.Entire architecture is divided into four layers.

**Interface layer**: using user interface the user can select the file for deduplication, type of deduplication and also can specify the split length .

**Chunk layer***:* Different segments of the file are created based on split length.The hash values are calculated for these segments using MD5 algorithm.

**Deduplication layer**: The duplicated chunks are detected by comparing hash values generated by the chunk layer. The chunks are compressed after eliminating the redundant copies.

**Storage layer**: After eliminating duplicated values the compressed file is stored in cloud storage.

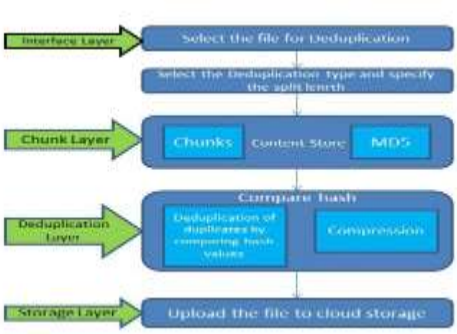


***Figure :1 System Architecture***

*Algorithm to segment file*
**Algorithm 1**: File segment
**Input:**File selected
Split size
**Output**:Files divided into segments based on split size
**Procedure:**SegmentFile
/*Enter the split size.*/
/*Read the input file and creat the bytes stream.*/
**While**(bytes!=-1)
{Divide the file into number of bytes specified by split length and creat the new file with .txt extension}
**end while**
**/***Different segments of the file is created.*/
**end**

*Algorithm to upload file*
**Algorithm 2:Upload File**
**Input:** File
**Output:** File uploaded to Cloud
**Procedure: Upload**
**begin** upload
/*Select file to upload*/
/*Select cloud storage*/
/*upload the file to cloud storage */
**end**

## Performance evaluation

Java language was used for implementing deduplication techniques. Only file level deduplication techniques implemented. The set of Sample files along with their duplicate copies are chosen for conducting test The set comprises of only text files.

| sno | file name | file size | compressed file | compressed file |
|---|---|---|---|---|
| | | bytes | size with out Dedup | size with Dedup |
| 1 | sree.txt | 6807 | 3947 | 3947 |
| 2 | sree01.txt | 6807 | 3947 | 0 |
| 3 | sunny.txt | 3183 | 1737 | 1737 |
| 4 | sunny01.tx | 3183 | 1737 | 0 |
| 5 | sunny.txt | 3183 | 1737 | 0 |

***Table 1:Output for Different text File***

Result are shown in above table 1.After Deduplication and compression the files are saved on the cloud storage and the storage space saved. Two runs are performed in this particular application . Testing was done by storing a file and then a duplicate copy of the same file.Result are shown in Table1 is the original file size,compressed file size with out deduplication and with deduplication.Five files are choosen for deduplication.By applying deduplication approach on these files we are able to save cloud storage. Results of file level Deduplication is shown in the above Table1 .Deduplicaton is performed on the set of text files and

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

it 's own copy. The graphs for the above results are shown below Figure2,by taking file name on X-axis, vs file size in bytes(y-axis).

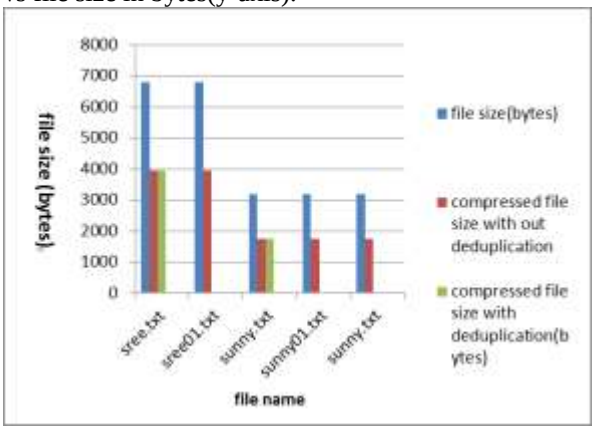


***Figure 2:File level Deduplication***

The space saving ratio is termed as Deduplication efficiency The efficiency is calculated for different test cases is shown in table 2.Another graph is plotted against between by taking space reduction ratio on (x-axis) and disk space saving ratio percentage on (y-axis) is shown in below figure3.

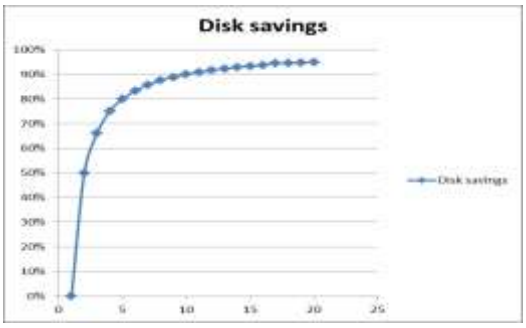


***Figure 3:Spce Reduction Perecentages***

File level deduplication achieves better througput as it compares every incoming chunks along with chunks of similar files.

## Conclusion

The optimized cloud storage technique (OCS) is Deduplication with compression. Deduplication helps in saving the disk storage space ,so that the cost of storage space and Band Width required is reduced.This application helps in easy maintenance of data on the cloud platform so that no duplicate files are saved in the cloud.

## References

1. Weiss. Computing in the clouds[J]. netWorker 2007,11(4):16-25.
2. Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal,et al.Cloud computing and emerging IT platforms :Vision, hype, and reality for delivering computing as the 5 th utility[J].Future Generation Computer Systems 2009,25:599-616.
3. Twenty experts define cloud computing[URL].http:// cloudcomputing.sys.con.com/read/612375_p.htm (18.07.08).
4. Century : of Cloud Computing Fabio Technical Report UBLCs-2011-03 May 2011.
5. Deduplication and Compression Techniques in Cloud Design" by Amrita Upadhyay, pratibha R Balihalli,Shashibhushan ivaturi and shrisha rao 2012 IEEE.
6. Amazon Inc., "Amazon Elastic Compute Cloud,"http://aws.amazon.com/.
7. National Institute of Strandards and Technolgy,The Key-Hash Message Authentication code (HMAC),Fedaral information processing standards publication # HMAC,2001.
8. R.Rivest,The Md5 Message Digest algorithm,RFC 1321,MIT LCS&RSA Data security,inc,APril 1992.
9. Extreme Binning:Scalable ,Parallel Deduplication for Chunk Based File Back up by Deepavali Bhagwat,Kave Eshghi,Darrell D.E.Long,Mark Lillibridge.
10. "AMazon Simple Storage Service API reference,Api Version 2006-03-01,"2006.

| Sno | test case name | Space reduction ratio | Space reduction percentage % |
|---|---|---|---|
| 1. | Case1 | 1.7296 | 42.18% |
| 2. | Case2 | 1.9353 | 48.32% |
| 3. | Case3 | 4.0751 | 75.46% |
| 4. | Case4 | 4.8346 | 79.31% |
| 5. | Case5 | 8.6230 | 88.40% |