IJCTA, 9(26) 2016, pp. 7-13

Improving De-Duplication Efficiency with multilevel Hashing in Cloud Storage Backup

S. Pavai Madheswari[1], X. Ignatious Viola[2] and S. Radhika[3]

ABSTRACT

Cloud Storage is a model in which data is stored and maintained on third party server rather than on the dedicated servers used and made available to users over a network. An average of 60% of digital data storage in cloud is redundant. To eliminate this problem data de-duplication has been proposed, which is a technique that identifies the duplicate data and removes the redundancies in order to effectively utilize the storage space. This technology is used to optimize the storage system by reducing the amount of redundant data and thereby maximizing the utilization of available energy. The proposed methodology performs data de-duplication in a private cloud storage backup which improves data de-duplication efficiency by looking up fingerprints concurrently in small indices classified by file type rather than a single full unclassified index.

Keywords: De-duplication, Cloud computing, Fingerprints, Private cloud.

1. INTRODUCTION

Cloud computing is a remarkable technology that delivers flexible applications, web services and IT infrastructure as a service over the internet using usefulness pricing model. The Cloud is a cost-effective approach to technology as there is no need to make usage predictions, upfront capital investments or over purchase hardware or software to meet the demands of peak periods. Cloud computing incorporates virtualization, data and application on-demand deployment, internet delivery of services, and open source software. The different forms of cloud models are Public cloud, Private cloud and Hybrid cloud. Public clouds services are rendered by third party service providers and applications from different customers are likely to be mixed together on the cloud's servers, storage systems, and networks. Here the computing infrastructure is provided by the cloud vendor at the vendor's premises. Private clouds are built for the exclusive use of single client. Private clouds can also be built and managed by the organization's own administrator. Here the computing infrastructure is dedicated to a particular organization and not shared with other organizations.

1.1. Cloud Storage

Cloud storage is a service model in which data is maintained, managed and backed up remotely and made available to users over a network. Cloud storage is a model of data storage in which the digital data is stored in logical pools, the physical storage spans multiple servers, and the physical environment is typically owned and managed by a hosting company. These cloud storage providers are responsible for keeping the data available and accessible, and the physical environment protected and running. People and organizations

[1] Professor, Department of Science and Humanities, R.M.K. Engineering College, Kavaraipettai - 601 206, Tamil Nadu, India, *E-mail: ac@rmkec.ac.in.*

[2] Assistant Professor, Department of Science and Humanities, R.M.K. Engineering College, Kavaraipettai - 601 206, Tamil Nadu, India, *E-mail: willsviola@gmail.com*

[3] Associate Professor, Department of Science and Humanities, R.M.K. Engineering College, Kavaraipettai - 601 206, Tamil Nadu, India, *E-mail: srp.mca@rmkec.ac.in*

buy or lease storage capacity from the providers to store user, organization, or application data. The storage cloud provides Storage-as-a-Service. The organization providing storage cloud uses online interface to upload or download files from a user's desktop to the servers on the cloud. Typical usage of these sites is to take a backup of files and data. Cloud storage is a service-level agreement (SLA) between a cloud storage service provider and a client that specifies the details of the service, usually in quantifiable terms. The different forms of cloud storage are private cloud storage, public cloud storage and hybrid cloud storage.

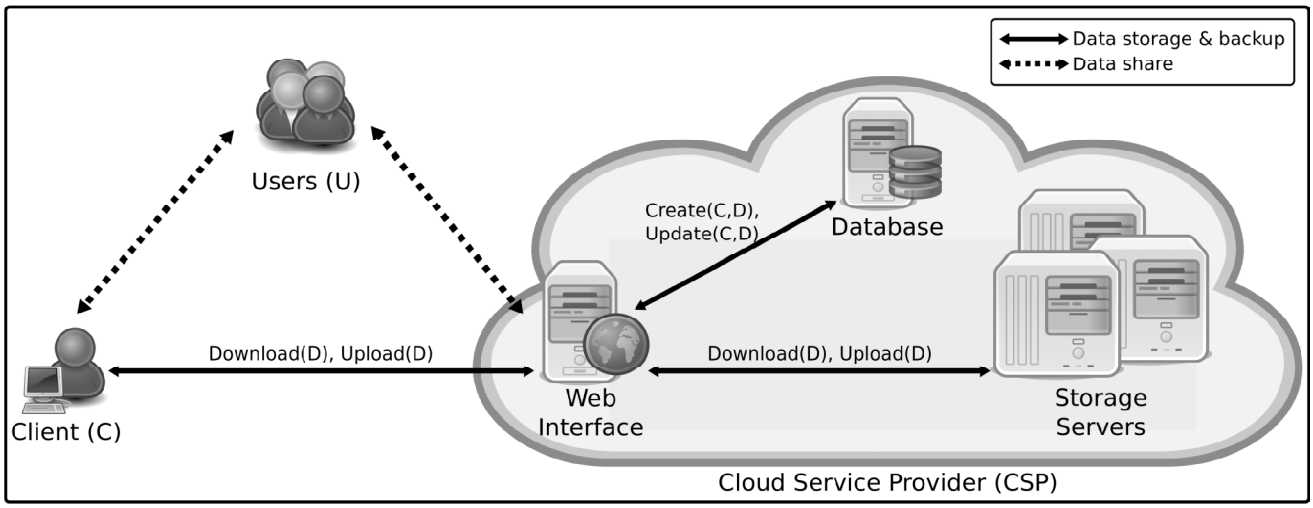

Figure 1: Cloud computing working process

1.2. Advantages of Cloud Storage

Cloud storage has quite a lot of advantages over traditional data storage. In Cloud Storage stored data can be accessed from any location through internet connection. Businesses and organizations can trim down annual operating expenses by using cloud storage. There is no need to carry around a physical storage device to save and retrieve information. Users of cloud storage can drag and drop files between the cloud storage and the local system.

1.3. Private Cloud Storage

Private cloud is a model of cloud computing that is operated within the corporate firewall. Private cloud is best for businesses with dynamic or changeable computing needs that require direct control over their environments.

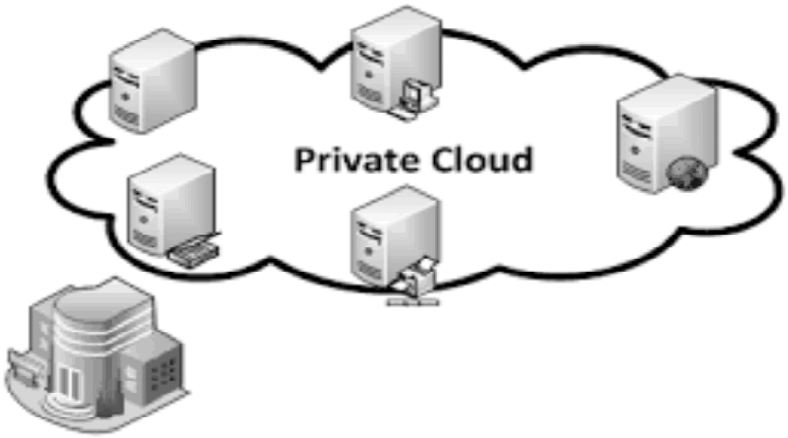

Figure 2: Private Cloud

1.4. Public Cloud Storage

A public cloud is a standard cloud computing model, in which services are rendered by third party service providers that is open for public use. Public cloud services may be free or obtainable on a pay-per-usage model.

1.5. Hybrid Cloud Storage

Hybrid cloud storage is an integration of at least one private cloud and one public cloud infrastructure. An organization can store dynamically used and structured data in private cloud and unstructured and archival data in a public cloud.

2. PRIVATE CLOUD STORAGE BACKUP

Cloud storage backup [1] is an approach for backing up data that involves removing data offsite to a managed service provider for protection. A major gain of using cloud backup is that it can make managing a backup system easier. Data moved offsite should be de-duplicated to avoid the redundancy and it is done by Cloud Storage Controller (CSC). This controller provides data protection, security, advanced virtualization features, and performance for an array of locally attached disk drives. The benefits of CSC are as follows. It creates a seamless and highly robust connection to cloud storage, while requiring no changes to applications running in the data center. Applications are able to access the cloud using standard block and file access protocols. Also, it accelerates the performance of applications using cloud storage through advanced WAN techniques including caching, de-duplication, compression, and protocol optimization. And, the Cloud Storage Controller provides the same features and capabilities expected of local storage arrays, such as thin provisioning, automated storage tier and volume management.

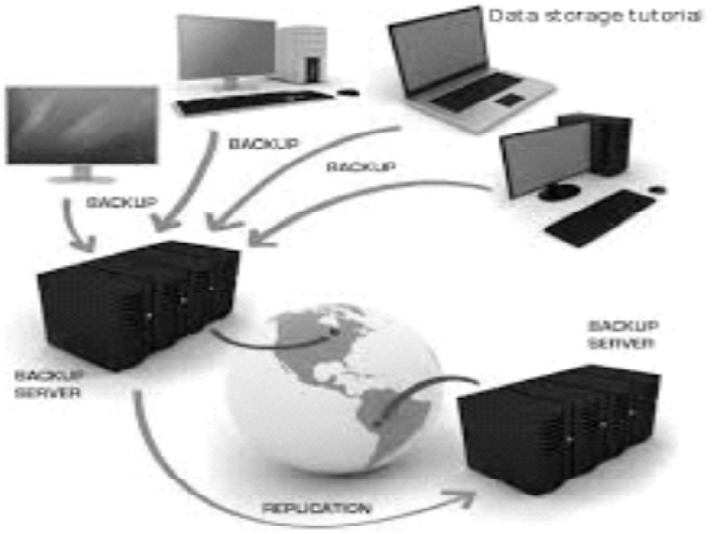

Figure 3: Cloud Storage backup

3. OVERVIEW OF DE-DUPLICATION

Data de-duplication has been proposed, which is a technique that identifies the duplicate data and removes the redundancies in order to effectively utilize the storage space. This technology is used to optimize the storage system by reducing the amount of redundant data and thereby maximizing the utilization of available energy. Benefits of data de-duplication are, it requires less storage space which will save money on storage device expenditures, efficient use of disk space allows for longer disk maintenance periods and reduces the need for tape backups, it also reduces the data sent across a WAN for remote backups and replication. Data

de-duplication is becoming popular in today's IT landscape - lowering infrastructure costs and saving money on backups for organizations who have successfully implemented de-duplication solutions. Duplication lets an organization keep 20 times more data in a given amount of storage. Award Jim Gray predicted that the amount of information across the world would double every 18 months [1].Globalization increased high demands on backup and also on how efficiently to store information. This requires companies to buy more storage, consume more power, energy and spend more time. To cope with this, organizations are increasingly using Data de-duplication. Data de-duplication is a data compression technique for eliminating cross grained redundant data typically to improve storage utilization [2].

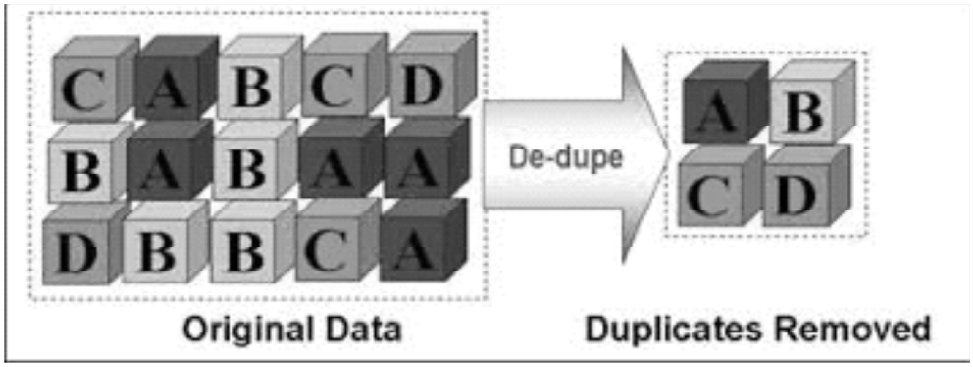

Figure 4: Data de-duplication technology

3.1. Benefits of data De-duplication

- Optimizes storage space by eliminating duplicate data
- Reduces storage device expenditures
- Frees up network bandwidth

3.2. Types of De-duplication

(a) In-line de-duplication: The data is de-duplicated before it is written to a backup device. Duplication is removed as the data enters a System. This is the process where the hash calculations are created on the target devices 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 rather references the existing block [3].

(b) Post-process de-duplication: The data is de-duplicated after it is written to a backup device. Data de-duplication process begins only after the entire back up is over.

3.3. Data De-duplication strategies

(a) File level data de-duplication: File-level data de-duplication compares a file to be archived with those already stored by checking its attributes against an index. If the file is distinctive, 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 copy of the file is saved and redundant copies are replaced with a pointer that points to the original file.

(b) Block level data de-duplication: In block-level data de-duplication the file -file level. As its the file is typically broken down into chunks that are checked for redundancy with the previously stored data.

4. REVIEW OF LITERATURE

According to [1], an average of 60% of data can be de-duplicated for individual users by using cross-user de-duplication techniques. Thus data de-duplication offers a unique opportunity to cloud storage providers to provide their users with more space at a low cost. [1].

Guo-Zi Sun, yu Dong, Dan-Wei Chen compares data de-duplication with other data storage methods, analyses characteristics of data de-duplication and applies the technology to data backup and recovery. In the process of backup based on data de-duplication user's data is divided into data blocks. Fingerprints are calculated for each data block, the newly created fingerprints are compared with the fingerprints of back up data. If a match occurs the data block corresponding to the fingerprints identified as duplicate data and a pointer and Meta data of the duplicate block is stored in the backup. Otherwise the new block is added to the backup. In the process of recovery based on data de-duplication with the help of user information such as file name, backup date, the data to be recovered is extracted from backup by reading the meta data blocks.[2].

Backup de-dupe supports the online source-side de-duplication and is capable of selecting different de-duplication algorithm according to the corresponding data types. Backup de-dup employs multi de-duplication strategies simultaneously to substantially eliminate redundant data in the backup process. [3].

Seungkwag et al, proposed a new cross-user source based de-duplication providing enhanced security in cloud storage. [4].

ALG-Dedupe improves de-duplication awareness and further combines local and global duplication detection to strike good balance between cloud storage capacity saving and de-duplication time reduction. [5].

5. METHODOLOGY

Our design consists of two Modules

i) Cloud Controller Module

ii) Cloud Backup Module

Cloud Controller Module

The users of the private cloud are provided with separate username. Whenever a user wants to upload a file, a unique 128 bit hash value of the file generated by md5 and file type is sent to the Cloud Controller. In Cloud Controller a fingerprint index is created to store the hash value of the each file uploaded by the user along with username.

Cloud Backup Module

This module performs the function of de-duplication detection by comparing the incoming fingerprint index with the backup node fingerprint index. In backup storage the fingerprint index is categorized by the file type in order to improve the efficiency of the de-duplication process.

The algorithm is as follows

1. The file index from cloud storage controller is compared with index in backup which is categorized by the file type.
2. Comparison is done by multithreading.
3. It starts checking by file type and then hash value.
4. If the match is found with the hash value along with file type, then the file is a duplicate one.
5. If the file is identified as duplicate, then it is not saved into the disk.
6. If the match is a not found then the file is assumed as new file and it is updated into backup node.

Our proposed method can achieve high de-duplication throughput by looking up fingerprints concurrently in small indices classified by file type rather than a single full unclassified index.

6. EXPERIMENTAL SETUP AND RESULTS

Cloud storage was built with Hostinger. Hostinger provides a simple web service interface that can be used to store and retrieve any amount of data at anytime from anywhere on the web. File level de-duplication was implemented. Different types of files were uploaded to the cloud storage and checked for de-duplication.

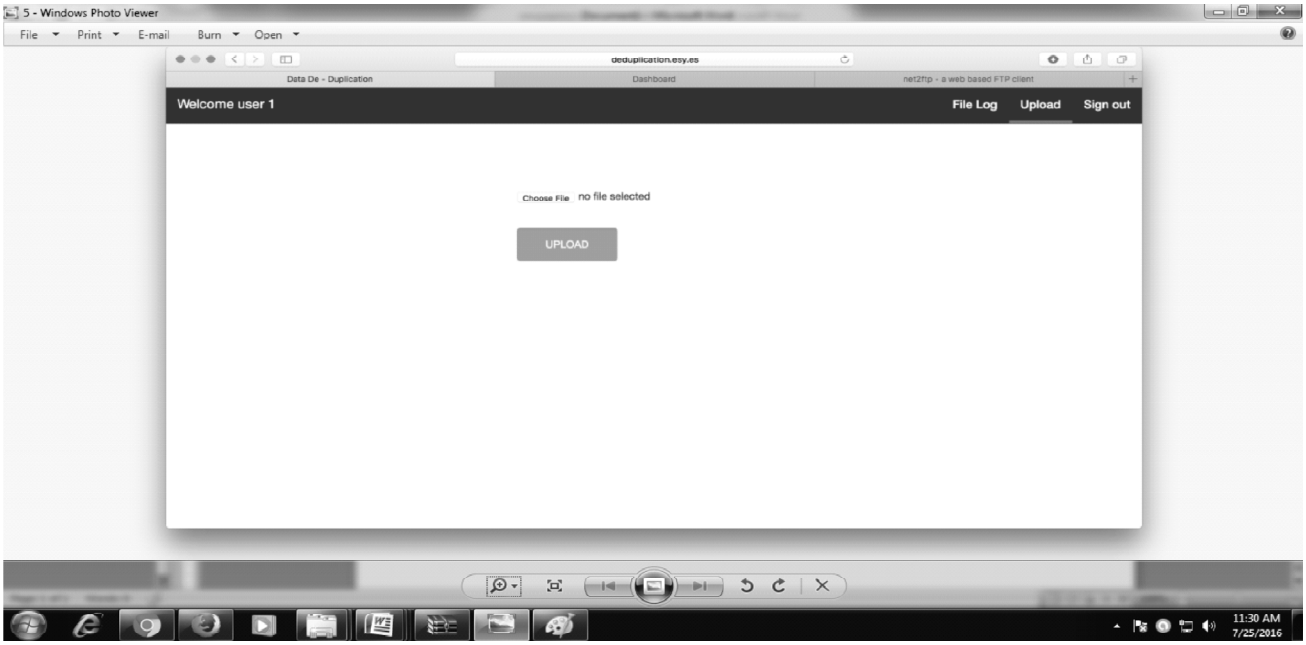

Figure 5: This window allows users to upload any type of file to cloud storage.

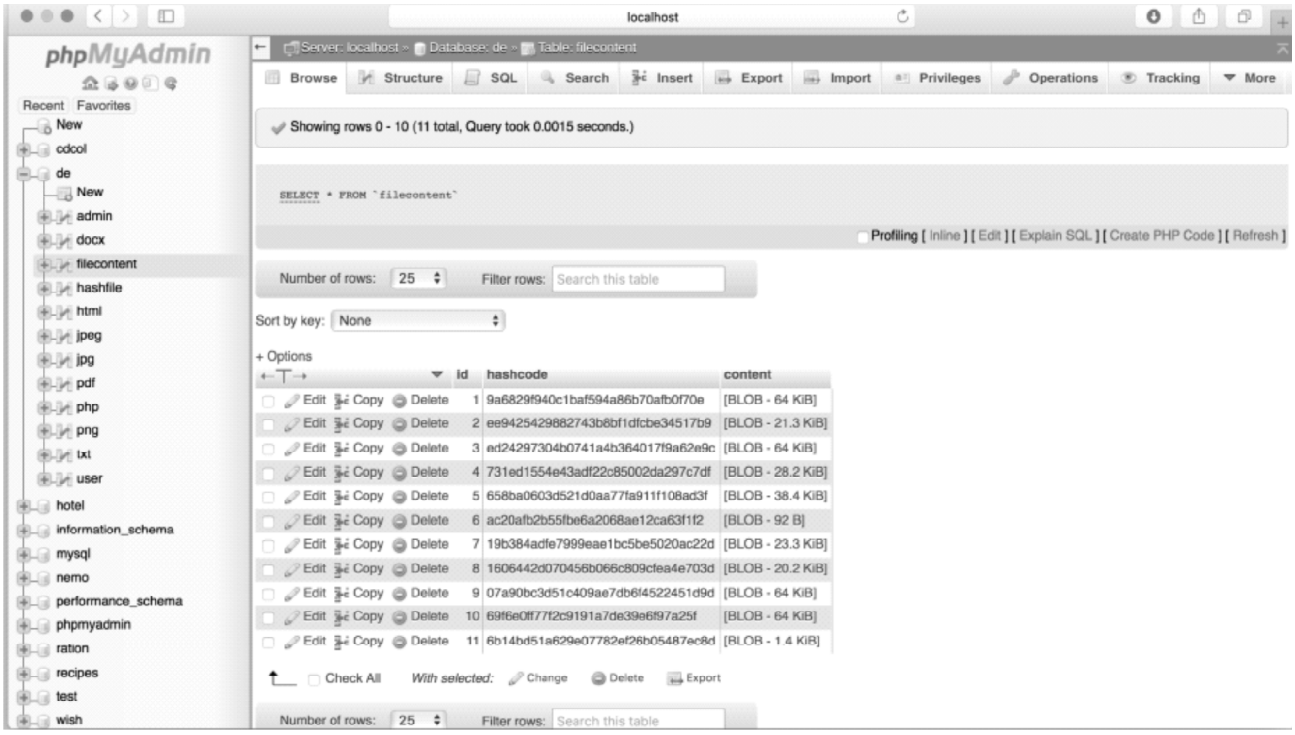

Figure 6: Shows the classification of files based on file type

Figure 7: Daily report of duplicate file in the backup.

7. CONCLUSION AND FUTURE WORK

Data de-duplication technology is a new direction of storage technology and has a wide range of application. Data de-duplication can be applied to cloud storage, which leads to a mass storage [9][10]. Data de-duplication reduces storage cost and network bandwidth, saves times and also provides a stable data storage for users. In this paper a novel approach is proposed for cloud backup deduplication using hashing algorithm. Our future work is to implement this same for block level storage in order to avoid duplication and also to analyze the performance of the same.

REFERENCES

[1] C.Soghoian "How dropbox scarifies user privacy for cost saving", http://paranoia.dubfire.nte/2011/04/how-dropbox-sacrifies-user-privacy-for.html.

[2] Guo-Zi Sun, yu Dong, Dan-Wei "Data backup and recovery based on data de-duplication," International Conference on Artificial Intelligence and Computational Intelligence, pp. 379-382, 2010.

[3] Guofeng Zhu "An Intelligent Data de-duplication based backup System," 15th International Conference on Network-Based Information Systems, pp.771-776, 2012.

[4] Seungkwang Lee "Privacy-Preserving Cross-User Source based data de-duplication in cloud Storage", International Conference on ICT Convergence, pp. 329-330, 2012.

[5] Yinjin Fu, Hong Jiang "Application-Aware Local-Global Source de-duplication for Cloud Backup Services of Personal Storage" IEEE transactions on Parallel and distributed Systems, pp.1155-1165, 2014.

[6] Guo-Zi Sun, yu Dong,Dan-Wei Chen,Jei Wei "Data backup and recovery based on data de-duplication" International Conference on Artificial Intelligence and Computational Intelligence, 379-382,2010.

[7] Forman G, Eshgh K, Ch Iocchett IS. "Finding similar files in a large document repositories", The *11th ACM SIGKDD* International Conference on Knowledge Discovery and Data Mining (KDD'05), pp. 394-400, 2005.

[8] Jain-ming Lin School of Buisness Administration Zhejianf Gongshang,Dongsheng Liu, Zhejiang Gongshang University Hangzhou, China, Shi-wen Gao, China, WeiChen, "A Web Page De-duplication Algorithm Based on Data Cleaning" International Joint Conference on Artificial Intelligence, pp. 544-547, 2009.

[9] D. Harnik, B. Pinkas, and A.Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud sstorage," IEEE Security Privacy, pp. 40-47, 2010.

[10] The Pros and Cons of File-Level Vs. Block-Level Data Deduplication Technology. [Online]. Available:*http:// searchdatabackup.techtarget.com/tip/The-pros-and-cons-of-file-level-vs-block-level-data-deduplication-technology,* accessed Jan. 2, 2015.