ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 5, March 2017

National Conference on Advanced Computing, Communication and Electrical Systems - (NCACCES'17)

24th - 25th March 2017

Organized by

C. H. Mohammed Koya
KMEA Engineering College, Kerala- 683561, India

Deduplication in Hybrid Cloud

Akhila M. [1] Arya R.[2], Revathy O.R[3], Sharon K.A.[4]

U.G. Student, Department of Computer Engineering, KMEA Engineering College, Edathala, Kerala, India[1]

U.G. Student, Department of Computer Engineering, KMEA Engineering College, Edathala, Kerala, India[2]

U.G. Student, Department of Computer Engineering, KMEA Engineering College, Edathala, Kerala, India[3]

U.G. Student, Department of Computer Engineering, KMEA Engineering College, Edathala, Kerala, India[4]

ABSTRACT: Deduplication is a process of eliminating duplicate copies of files. Deduplication in hybrid cloud hence will avoid the duplicate copies of file in the cloud which are uploaded by the users. The uploaded files are checked both file level and block level to avoid duplicate copies in two files and in two blocks. There are basically three modules and they are public cloud, private cloud and users. The users with privileges can read or write in the uploaded file. Duplicate copies of file are not uploaded into the cloud but users are provided with the pointer to the existing file in the cloud. We propose a new way of deduplicating videos and providing a backup for files when they are in case lost. We also try to provide a help desk which provides the users with necessary information regarding hybrid cloud deduplication and an ever note option where they can note their own data's.

KEYWORDS: Deduplication, Hybrid cloud, Cipher text, Evernote, Cloudinary.

I.INTRODUCTION

Cloud computing is most widely used method now a days as it provides necessary storage area. Since there are large number of data, we need to increase the storage space by eliminating duplicate copies of data.Traditional system may not be able to check for duplication and it reliably needs to manage huge number of convergent keys. To eliminate the problem of duplication hybrid cloud is used. An encryption method is used to reduce the number of keys to be managed. Only one copy of data is kept in cloud and the remaining is excluded. The duplicate data's are replaced by pointers, so that the data can be retrieved by the users. Pointers are provided to the same file user uploaded, whose privilege is set by the private cloud.

Cloud environment provides scalable and elastic storage capabilities to the users. Since only one copy of file is stored a large amount of space in the cloud becomes free and can be used to store new useful data's. Duplicate copies are detected by either file level or block level. File level eliminate the duplicate copies of same files whereas block level can eliminate copies at blocks.

A proof of ownership protocol is used to provide security since the uploading file may contain confidential data. Therefore, the data's are encrypted using keys created from cryptographic hash value and the ciphertext related to the files can be uploaded to cloud. Since identical files form same convergent key and ciphertext they can be easily identified.

The identical copies here can only be detected for images and data files, so here we propose a new concept of eliminating duplicate videos in the cloud by converting the prescribed number of frames in the videos into images and then finding out whether these converted images are identical. We also propose an idea of help desk that providesservice and support for users in the cloud. An 'evernote' option is provided for the users where they can note the necessary information. Backups are also provided for the data that are uploaded into the cloud incase of data loss. Users can also share their data files to others when needed.

II.RELATED WORKS

1. A redup:Here deduplication is done in image virtual machine image storage. A fragmentation technique is used here but it will degrade the read performance of the system. In fragmentation some block of files refer to another identical blocks. In this case file access will not be sequential. A revdedup method is proposed which deletes similarities in the virtual machine and removes any duplicates in the write path. The duplicate copies in old backup method can be removed by a method of hole punching and the segment compaction help[s to compact then segment to avoid duplicate copies reclaim the continuous space in the cloud storage. The revdedup is implemented on the client side and hence multiple user can make change in the server virtual machine's images.[2]
2. Message Locked Encryption Scheme: A message locked encryption scheme called as MLE provide a way of encrypting message using a key which is derived from the message. The encryption and decryption is done using the key formed from the message. These scheme doesn't rely on permanent secret keys but changes their secret key depending on the message. A key generation algorithm is used that generates a key from message. Message locked encryption provides secure deduplication. The ciphertexts are formed from the key generated from the message. These ciphertext can be tagged using tagging algorithm and can be used to avoid duplications in the cloud. The message can be decrypted from ciphertext using the key. Letting the tag equal to the ciphertext can make deduplication secure and can help in restricting the attacks.[3]
3. Leakage: Cloud services are the most widely used services nowadays. Security and storage must be provided for the clouds to make it more efficient way of storage. The deduplication of data can be done at the client side of cloud storage and is one of the efficient techniques. But this can lead to leakage of data at the client side itself hence a proof of ownership protocol can be used to make sure that a particular file is owned a particular valid user. To make it some more secure an encryption method can be used for the data. The encryption is done depending on the hash value but the direct combination of proof of ownership protocol and convergent encryption is not possible so a hash-as-a-proof method is proposed, here the hash gives the index of the file among huge number of files in the cloud. It is also treated as a proof for the ownership of file.[4]
4. Venti: Venti is an archival server that provides a write once archival repository such that it may be distributed among a number of client machines and applications. The write once policy is implemented in order to prevent accidental or malicious destruction of data. The storage technology used is the magnetic disc. In venti system, a unique hash table is used to store the blocks' contents. Thus it acts as an identifier for read and write operations. These unique hash is 'fingerprint', is used to address blocks. Such that, a block cannot be modified without changing its address; the behaviour is intrinsically write-once. In addition to the block level network storage system, duplicate copies of blocks are coalesced to reduce the storage consumption.[5]
5. The main problem in cloud computing is deduplication. Data Deduplication is a technique that is mainly used for reducing the redundant data in the storage system which will unnecessarily use more bandwidth and network. So here some common technique is being defined which finds the hash for the particular file and with that the process of deduplication can be simplified, David Geer. [6]

III.DEDUPLICATION

Data deduplication is an important technique used for eliminating duplicate copies of data and has been widely used in cloud storage to increase storage space. To protect the confidentiality of data whilesupporting deduplication, the convergent encryption technique has been proposed to encrypt the data. To better protect data security,this paper makes an attempt to formally address the problem of authorised data deduplication. Different from traditional deduplication system, the differential privileges of users are further considered in duplicate check besides the data itself. The analysis demonstrated that our scheme is secure in terms of the definitions specified in the proposed model. As a proof of concept, we implement a prototype of our proposed authorised duplicate checkscheme. we show that our proposed authorised duplicate checkscheme incurs minimal overhead.

IV.CLOUDINARY

Use cloudinary's image service instead managing your images in-house, it reduces IT costs. With cloudinary, all your images are automatically uploaded, normalized, optimized and backed-up in the cloud. We can securely upload images and files directly from our server. Cloud Store as many images as needed and can be dynamically managed to fit any graphics design. Apply effects, resizing, cropping, face detection, watermarks and tons of image processing capabilities.

On-the-fly real-time videomanipulation and web optimizationUpload all your video clips to thecloud. With a tweak of a URL,Cloudinary will transcode yourvideos for web view across allbrowsers and mobile devices.Scale, crop and further manipulateyour videos on-the-fly according toyour graphic design. Your videoswill be optimized and streamed viaa fast CDN.

Cloud backup, also known as onlinebackup, is a strategy for backing up datathat involves sending a copy of the dataover a proprietary or public network to anoff-site server. The server is usuallyhosted by a third-party service

provider,who charges the backup customer a feebased on capacity, bandwidth or numberof users. In the enterprise, the off-siteserver might be owned by the company,but the chargeback method would besimilar.
We can also protect & restore cloud servers files with cloud backup. Cloud backup safeguards your business by helping to protect the important files your website or application needs and quickly get back to normal operations by rapidly restoring files after a system failure or file loss. Block-level compression anddeduplication reduce storage costs byup to 20x compared to uncompressedbackups. After your first backup, additional backups are incremental tolower your total costs. Easily create, schedule, and managefile-level backups through our ControlPanel or API. Because it's integratedwith Cloud Servers on our high-capacitynetwork, protecting files and restoringbackups takes just minutes. Enterprise-grade encryption helps ensure data is securedwith a passphrase known only to you.After creating the key, data isencrypted before it leaves the server,and remains encrypted while stored.

Cloud backup solutions enable enterprises orindividuals to store their data and computer fileson the Internet using a storage service provider,rather than storing the data locally on a physicaldisk, such as a hard drive or tape backup.Cloud backup providers enable customers toremotely access the provider's services using asecure client log in application to back up filesfrom the customer ' s computers or data centerto the online storage server using an encryptedconnection.

Implementing cloud data backup can helpbolster an organization's data protectionstrategy without increasing the workloadon information technology staff.Online backup systems are typically builtaround a client software application thatruns on a schedule determined by thepurchased level of service. If thecustomer has contracted for dailybackups, for instance, the applicationcollects, compresses, encrypts andtransfers data to the service provider'sservers every 24 hours. To reduce theamount of bandwidth consumed and thetime it takes to transfer files, the serviceprovider might only provide incrementalbackups after the initial full backup.Cloud backups often include the softwareand hardware necessary to protect anorganization's data, including applicationsfor Exchange and SQL Server.Most subscriptions run on a monthly oryearly basis. While initially used mainlyby consumers and home offices, onlinebackup is now used by small andmedium-sized businesses (SMBs) andlarger enterprises to back up some formsof data. For larger companies, cloud databackup may serve as a supplementaryform of backup.

Easy to UseSchedule, manage, andrestore your files usingCloud ControlPanel or API. Quicklyrestore backups to theoriginating Cloud Server(or any Cloud Serverrunning the backupagent), and choosespecific files and foldersto back up.Flexibility is another benefit of the cloudbecause no additional hardware isneeded.Third-party cloud backup has gainedpopularity with SMBs and home usersbecause of its convenience. Thetechnology has an initial upfront cost andthen lower monthly or yearly paymentplans that appeal to many smalleroperations. Capital expenditures foradditional hardware are not required andbackups can be run dark . However, thecost of keeping data in the cloud foryears adds up. In addition, costs rise asthe amount of data to be backed up tothe cloud increases.Pricing models vary by vendor, but it'simportant to be on the lookout for hiddencosts. While most products for backingup to the cloud are sold using a price-per-gigabyte-per-month payment model,providers can also use a sliding scalemodel, set usage minimums and addtransaction costs .

V.HOW TO RESTORE A CLOUD BACKUP

To update or restore cloud backup, customers need to use the service provider's specificclient application or a Web browser interface.Files and data can be automatically saved tothe cloud backup service on a regular,scheduled basis, or the information can beautomatically backed up anytime changes aremade (also known as a "cloud sync").Enterprise-Grade Cloud BackupFor enterprises, enterprise-grade cloud backupsolutions are available that typically addessential features such as archiving anddisaster recovery. Archiving features help tosatisfy an enterprise’s legal requirements fordata retention, and as part of a company'sdisaster recovery plan, the remote, off-sitestorage provided by cloud backup helps ensurethe data remains safe should the enterprise’slocal data be jeopardized by a disaster such asa fire, flood, hacker attack or employee theft.

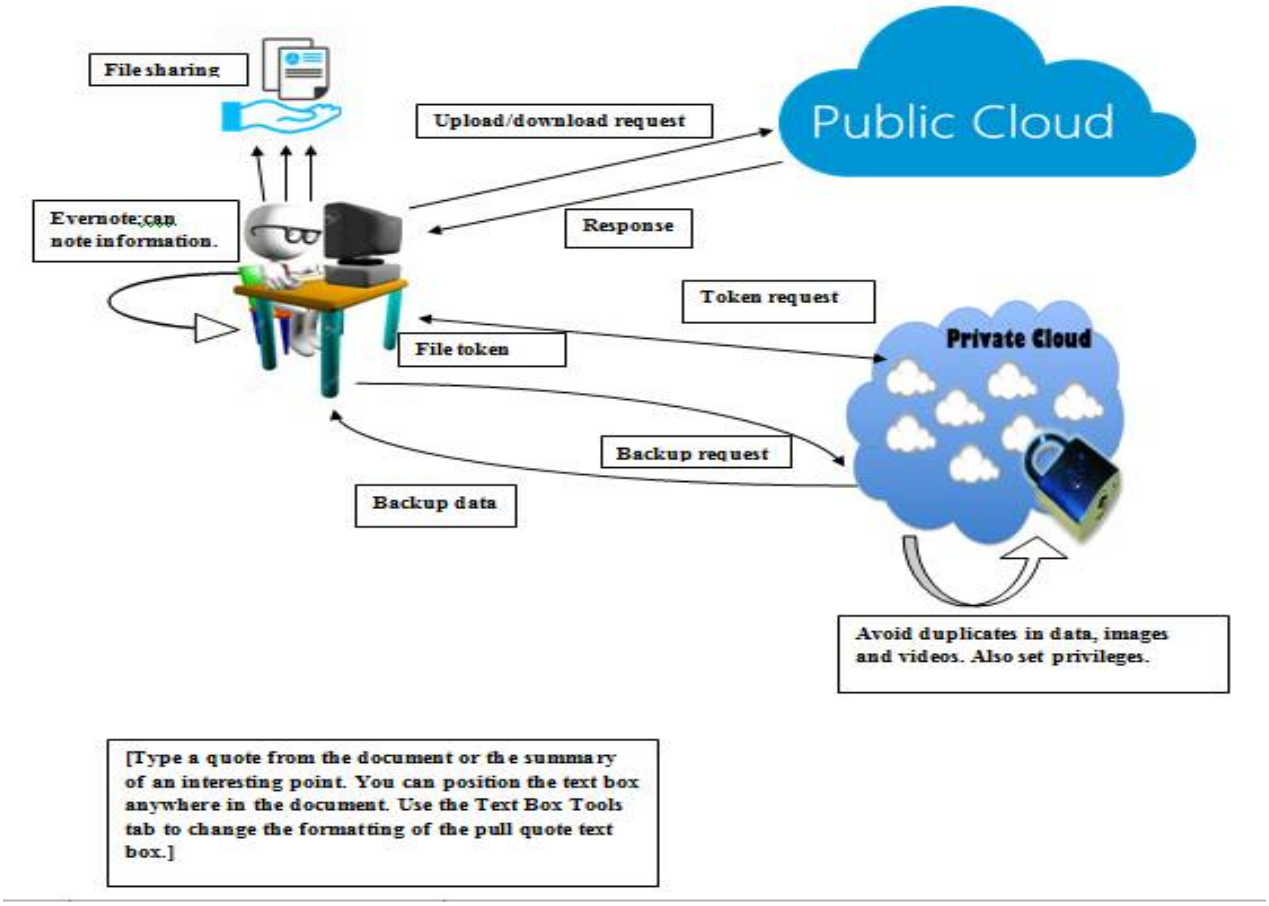

Fig. 1. Architecture for Hybrid cloud for deduplication with back up.

VI.PROPOSED TECHNIQUE TO SOLVE THE PROBLEM OF DUPLICATE VIDEO DETECTION

Refering the paper 'Efficient and Robust Detection of Duplicate Videos in a Large Database: A Survey', by Soni R. Ragho and C. S. Biradar, we get into a suggestion that duplicate videos and infringement in videos can be detected using two techniques. They are Digital watermarking and Content-based copy detection (CBCD) techniques respectively. In order to accumulate the features of images, which are obtained from videos to create fingerprints are extracted using robust or compact frame based descriptor and color layout descriptor. The fingerprints are encoded by vector quantization. The video is checked for copyright by determining the existence of watermark using the Digital watermarking technique. The CBCD technique is that, it compares the fingerprint of the query video with the fingerprint of the original videos from the database. The signatures/fingerprints from text media stream are compared to the original media signature/fingerprint to check if there is a copy of original media in test stream.

VII.CONCLUSION

In this paper, it was shown that the hybrid cloud increases the storage space by eliminating duplicate copies of data. We demonstrated that our paper meets the current demands. It was shown that a proof of ownership protocol is used to provide security. It was also shown that the identical copies here can only be detected for images and data files. The detection of identical copies for video, evernote, sharing process, backup and help desk are kept as future work.

REFERENCES

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication",IEEE Transactions On Parallel And Distributed System Vol:PP NO:99 Year 2014
[2] Chun-Ho Ng, Patrick P. C. Lee, The Chinese University of Hong Kong, Hong Kong, "RevDedup: A Reverse Deduplication Storage System Optimized for Reads to Latest Backups", June 28, 2013
[3] MihirBellare, SriramKeelveedhi, Thomas Ristenpart , "Message-Locked Encryption and Secure Deduplication", March 2013
[4] Jia Xu Institute for Infocomm Research, Ee-Chien Chang National University of Singapore, Jianying Zhou Institute for Infocomm Research, "Leakage-Resilient Client-side Deduplication of Encrypted Data in Cloud Storage"
[5] Sean Quinlan and Sean Dorward Bell Labs, Lucent Technologies, "Venti: a new approach to archival storage"
[6] David Geer, "Reducing the Storage Burden via Data Deduplication.computer.org|, December 2008.
[7] Soni R. Ragho1, C. S. Biradar2, "Efficient and Robust Detection of Duplicate Videos in a Large Database: A Survey", Volume 4 Issue 6, June 2015