Security of Various Data De- Duplication Approaches

Divya Jain
Samrat Ashok Technological Institute
Vidisha, India

Vivek Sharma
Samrat Ashok Technological Institute
Vidisha, India

ABSTRACT

In this paper, the cloud computing has develop into more and more admired. It brings revolutionary innovation with regards to cost, reserve organization and utilization. Cloud computing offer nearly unlimited resources, highly reliable on-demand services with minimal infrastructure and operational cost and storage system. In the cloud storage system with data de-duplication, un-trusted entities including a cloud server and users may cause security threats to the storage system. These services offers to end-users rather than a product by sharing resources, software and information, hence economic benefit and data de-duplication is the key for Cloud terms of capital and operational expenditure.

Keywords

Cloud Computing, Data De-duplication, style, styling, insert.

1. INTRODUCTION

Cloud computing (CC) is a promising and emerging technology for the next generation of IT applications. The difficulty and problems in the direction of the quick development of cloud computing are data security and privacy issues. Cloud computing is a capable tool that cost-effectively allows data outsourcing as an examination using Internet tools with elastic provisioning and usage-based pricing [1]. Cloud computing provides a low-cost, scalable, location independent infrastructure for data management and storage that is available anyplace and anytime over the Internet application on mist stowage amenities such as Drop-Box, Mozy and Memopal are increasing recognition.

Cloud computing has raised the delivery of IT services to a novel stage that carries the console of conventional utilities such as water and electricity to its users. The advantages of Cloud computing, such as cost effectiveness, scalability, and ease of management, encourage more and more friendship and service providers to become accustomed it and present their explanations passing through Cloud computing models. According to a modern review of IT decision makers of huge companies, 68% of the respondents expect that by 2014, more than 50% of their company’s IT services will be migrated to Cloud platforms [2]. Cloud computing has become a scalable service consumption and delivery platform. Figure 1.1 shows the system architecture in cloud computing. In a cloud environment, the cloud provider grips a huge number of distributed examines (e.g. databases, servers, Web services, etc.), which can be offered to expensive for increasing a range of cloud applications. Expensive of cloud applications can prefer from an extensive collection of distributed services when creating cloud applications. These examinations are frequently bring into play distantly through communication links and are enthusiastically put together into the applications. The cloud application designers are located in different geographic and network environments.

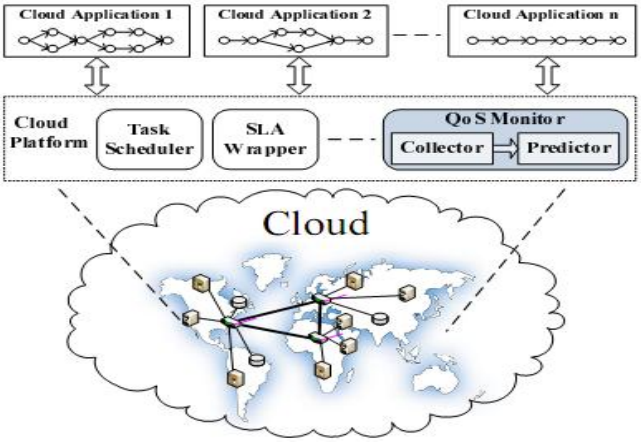

Figure 1: System Architecture of cloud computing

2. DATA DEDUPLICATION

Data deduplication is the procedure of recognizing and removing unnecessary copies of varying information in the cloud or storage .This technique is currently widely used to improve the storage utilization. Data deduplication can be done at file level and block level. File level deduplication identifies and removes same files. This method is also called as Single instance storage. Once the file is stored other same copy of file point to this pointer. Block level detects redundant data within and across files. This method is also called as sub file deduplication For the most part in cloud remote storage context, deduplication fulfills one of the most common requirement as reducing the overall costs of providing the same service that client could do themselves in their own data centers. Not only the process utilizes storage hardware costs it also decreases backup and recovery costs while improving network efficiency. Cloud services like Dropbox employs deduplication across multiple user client, as the user base grows, service cost per user decreases subsequently.

From end-users' point of view, one of the most notable advantage of deduplication is the storage service cost. Basically, they pay less for the same amount of storage. For example, Dropbox offers 1TB storage for only $9.99 while an external hard drive would cost more than a few hundred dollars for the same capacity. Another noteworthy gain is the upload time/bandwidth for duplicated files. Large files such as .iso or movies could be uploaded almost instantly if they are duplicated, providing significant boost to user experience. In general, deduplication process involves 4 steps:

1. Split data into small units, (ex: files, blocks).
2. Calculate unique hash value for each unit.
3. Detect if there is any file on the storage having the same hash value.
4. Put reference to the duplicated file.

Data can be compared at file level where the scheme generates hashes of files and compares. Meanwhile, block/bye level units allow deduplication in a finer grained manner with the cost of more operations. Hashing mechanisms are also varying among schemes but the general thought is to produce a unique, mathematical representation of unit that can be located and compared. Data Deduplication is a method to prevent duplication of repeating data. During deduplication processes, unique chunks of data such as files or block of bytes are analyzed followed by discarding the duplication. The result is only one instance of repeat data are stored/transferred hence greatly improve storage utilization and transfer time/bandwidth. Initially, deduplication is heavily used on duplication prone situations such as backups or virtual desktops [5].

3. DATA REPLICATION AND STORAGE ON CLOUD COMPUTING

A Data Grid is a geographically-distributed teamwork in which all participants necessitate admission to the datasets yield within the association. Replication of the datasets is consequently a key requirement to ensure scalability of the cooperation, dependability of data access and to reservation bandwidth. Replication is constrained by the size of storage existing at altered positions inside the Data Grid and the bandwidth amongst these sites. An imitation organization organization therefore confirms admittance to the necessitated data while management the essential storage.

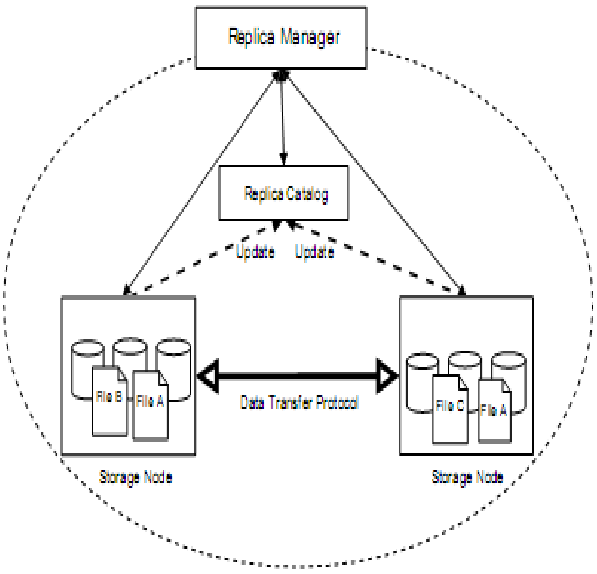

Figure 2: Replica Management Architecture.

An imitation organization organization, shown in Figure 2, consists of cargo space nodes which are connected to each other through high-performance data transport protocols. The replica manager uninterrupted the formation and supervision of reproductions allowing to the requests of the customers and the accessibility of storage, and a collection or a directory keeps pathway of the duplications and their sites. The collection can be demanded by presentations to determine the number and the positions of existing duplications of a specific dataset. The client libraries permit querying of the catalog to determine datasets and to apply for replication of a exacting dataset [9].

4. LITERATURE SURVEY

In this paper [6], author has to solving efficiently the predicament of deduplication with discrepancy rights in obscure computing, here they think about a mixture obscure architecture consisting of a public cloud and a private cloud. As using existing approach for data deduplication the confidential obscure is concerned as a alternative to authorize information owner/users to strongly achieve photocopy ensure with discrepancy benefits. Such architecture is convenient and has concerned much awareness from make inquiries from data owners only outsource their data storage by utilizing public cloud while the data process is deal with in private cloud. A new method sustaining differential duplicate ensure is planned beneath this mixture cloud building where the S-CSP reside in the community obscure. The user is only permitted to execute the duplicate check for files marked with the parallel privileges.

The main goal of this paper is to provide stronger security by encrypting the file with differential privilege keys. In this approach, the users without corresponding privileges cannot achieve the duplicate check. In addition, such unofficial Owners cannot decode the ciphertext even join together with the S-CSP.

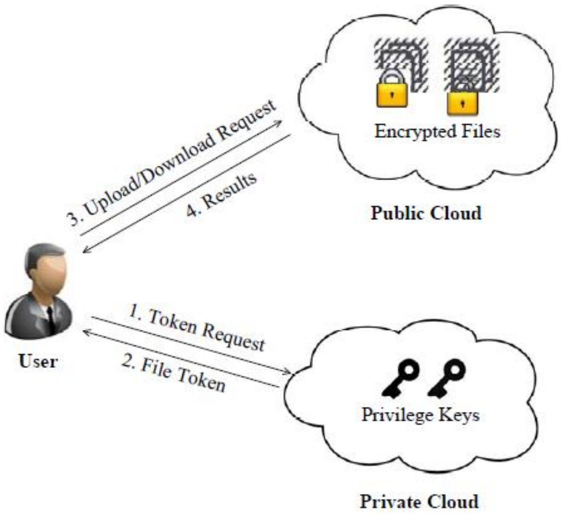

Figure 3. Architecture for authorized deduplication[6].

As their proposed method has to authorize duplicate check and conduct test-bed experiments to calculate the overhead of the prototype. Security examination shows that their organization is protected in stipulations of the definition particular in the planned sanctuary mock-up. Here they show that the transparency is negligible compared to the standard convergent encoding and sleeve upload operation.

To guard the discretion novelist has been planned [7] to encode the information before outsourcing. To enhanced protect data security this paper makes the initial attempt to officially concentrate on the difficulty of authorized data deduplication. Unusual from conventional deduplication systems the degree of difference rights of owners are added measured in photocopy make sure as well the data itself. Here they also present common new deduplication growing sustaining official duplicate patterned in mixture haze building. Security analysis shows that their method is secure in expressions of the descriptions particular in the anticipated security representation. As a proof of idea, they put into practice a example of our planned approved duplicate check method and behavior testbed experiments using our prototype. We demonstrate that our suggested authorized replica check method bring upon yourself negligible transparency evaluated to normal operations. It keeps the memory by deduplicating the data and thus makes available us with enough memory. It provides authorization to the private firms and guards the discretion of the significant data.

In this broadside, author [8] offerings a new confidentiality preservative sanctuary explanation for mist amenities. Here in this method deal with user indefinite access to cloud amenities and shared storage servers using non-bilinear group signatures to ensure anonymous authentication of cloud service client's user. Users use tamper resistant devices during the generation and storing of user explanations to defend alongside collusion attacks. Here the solution delivers registered users with anonymous access to cloud services and also offers anonymous authentication. This signifies that user's personal attributes (age, valid registration, successful payment) can be proven without make knowing user's identity. Consequently, users can use amenities without any danger of profiling their performance. On the other hand, if users break provider's rules, their access rights are withdrawn. Here we analyze modern confidentiality conserving explanations for cloud amenities and summarize our explanation based on progressive cryptographic mechanisms it also suggestions unidentified access, unlink ability and the confidentiality of transmitted data. Due to this fact, cloud service providers using our solution can authenticate more clients in the same time. Additionally, there method gives output the experimental results and measure up to the performance with related solutions.

In this paper author [10] has try to evaluate how can mist breadwinners earn their purchaser's conviction and provide the security, privacy and reliability, when a third party is meting out sensitive data in a remote machine established in various countries. A thought of utility cloud has been characterized to provide a variety of services to the users. Various technologies can help to concentrate on the challenges of security, privacy and trust in cloud computing. Unfortunately, the implementation of cloud computing came before the suitable technologies become visible to deal with the supplementary confronts of trust. This opening between implementation and improvement is so extensive that cloud computing consumers don't fully expectation this innovative way of computing. To close this opening, we require identifying with the trust issues join together with cloud computing from both a technology and business perception. Then we'll be able to establish which up-and-coming technologies could best address these problems. Here the author [10] has analyzed the trusted computing in the cloud calculating situation and the function of trusted computing platform in cloud computing. The advantages of this move toward are to make bigger the trusted computing technology into the cloud figuring situation to accomplish the trusted computing prerequisites for the cloud computing and then accomplish the trusted cloud computing.

S. No.	Paper	Author	Issues
1.	Secure De-duplication And Statistics Sanctuary With Effectual And Reliable CEKM	N.O.AGRAWAL , S.S.KULKARNI	The problem of impairment of pilfered records if we diminution the assessment of that stolen evidence to the invader to completing effectual and consistent key administration in secure de-duplication.
2.	Private Data De-duplication Protocols in Cloud Storage.	Wee Keong Ng, Yonggang Wen, Huafei Zhu	how to define the security of private data de-duplication protocols how to formalize the functionality of private data de-duplication protocols, and how to construct private data de-duplication protocols if exist.
3.	A Secured and Authorized Data Deduplication in Hybrid Cloud with Public Auditing.	Sharma Bharat, Mandre B.R.	Issue is that duplicate check do not support differential privileges from convergent encoding uniform nevertheless afford confidentiality.
4.	A Study on Authorized De-duplication Techniques in Cloud Computing.	Bhushan Choudhary, Amit Dravid	The security issue is to estimate the resourceful consumption of raincloud band width and floppy convention.

5. PROPOSED METHODOLOGY

Setup: Here in this phase first of all the Elliptic Curve Parameters are set and public and private key pairs are generated using KeyGen(.). Suppose the General Elliptic Curve Equation is defined by:

$$y^2 = ax^3 + bx + c$$

Where, $4a$

Client chooses any random point over elliptic Curve E(F) that would be the chosen Secrete key of the client sk using secrete key and Common Base Point B public key is generated.

Pk=Sk.P

SigGen:The Shared Data File F={m1,m2…mn}, first of all choose a random integer 'u' and hence generate Tag for the Shared Data File F using

$$T_m = name||Nd||u||Sig$$

Client Starts generating Signatures Sg for each of the block mi,

$$S_g = (H(mi).u^{mi)^{\alpha}}$$

The Client Generating of Linked List based on the signatures and create a First Node of the Linked List and the other Nodes are constructed using H(mi).

Client Signs the Generated Started Linked List Root Node using secrete key sk

$$sig_{sk}(H(R)) \leftarrow (H(R))^{\alpha}$$

Client Sends {F,Tm,Sg,) to Third Party Auditor (TPA).

Data Deduplication: When the Block is received to the TTP will checks the Data is Already stored to the Storage Panel or not. If already stored then discarded, otherwise stores in Storage Panel.

6. CONCLUSION

Data privacy has been a major concern in cloud storage since users have to trust the cloud service providers for security and privacy. Here we also review on various existing method for new de-duplication buildings backup authorized matching checkered in hybrid mist construction to analyzed almost every security threat various de-duplication method for both the cloud models, in which the duplicate-check tokens of files are generated by the private cloud server with private keys.

7. REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al., “Above the clouds: A berkeley view of cloud computing," Technical Report UCB/EECS-2009-28, Dept. EECS, UC Berkerely, 2009.

[2] B. Narasimhan and R. Nichols, “State of cloud applications and platforms: The cloud adopters’ view,” Computer, vol. 44, no. 3, pp. 24–28, 2011.

[3] Pearson, S., & Benameur, A. Privacy, security and trust issues arising from cloud computing. In Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on (pp. 693-702).

[4] Tweney, A., & Crane, S. (2007). Trustguide2: An exploration of privacy preferences in an online world. Expanding the Knowledge Economy: Issues, Applications, Case Studies.

[5] Meyer, D. T., and Bolosky, W. J. A study of practical deduplication. ACM Transactions on Storage (TOS), 2012.

[6] 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 Systems, 2014.

[7] N.B. Kadu, Mr. Amit Tickoo, Mr.Saurabh I. Patil, Mr. Nilesh B. Bhagat , Mr. Ganesh B. Divte, “A Hybrid Cloud Approach for Secure Authorized Deduplication” International Journal of Scientific and Research Publications, Volume 5, Issue 4, April 2015.

[8] Gade Swati,Prof.Prashant Kumbharkar, “Cryptosystem For Secure Data Sharing In Cloud Storage” IJIRT Volume 1 Issue 6 2014.

[9] Lukas Malina and Jan Hajny, “Efficient Security Solution for Privacy-Preserving Cloud Services” 6th International Conference On Telecommunications Signal Processing Year 2013.

[10] Pardeep Kumar, Vivek Kumar Sehgal, Durg Singh Chauhan, P. K. Gupta and Manoj Diwakar, “Effective Ways of Secure, Private and Trusted Cloud Computing “JCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011.

[11] Wee Keong Ng, Yonggang Wen, Huafei Zhu, “Private Data De-duplication Protocols in Cloud Storage” ACM 978-1-4503-0857-1/12/03, 2011.

[12] Sharma Bharat, Mandre B.R. “A Secured and Authorized Data Deduplication in Hybrid Cloud with Public Auditing” International Journal of Computer Applications (0975 – 8887) Volume 120 – No.16, June 2015