ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

A Survey on Secure Deduplication on Hybrid Cloud

Priyanka Damkondwar

M.Tech. Student, Department of Computer Science and Engineering, SGGSIE&T Vishnupuri, Nanded, India

ABSTRACT: A survey on industry trends is been noted where the usage of hybrid cloud architecture can be used which supports, the upcoming industry challenges by providing the efficient way of storing their data in the cloud environment by using the combination of both public and private clouds, So that it provides the facility to store sensitive data in private cloud and less critical or nonsensitive data on public cloudavings can be made. Since the demand for data storage is increasing day by day and by the industry analysis we can say that digital data is increasing day by day, but the storage of redundant data is excess which results in most of the storage used unnecessary to keep identical copies. So the technology de-duplication is introduced to efficiently utilize the cloud storage system.Deduplication mainly used in all cloud storage services, which eliminates redundant data by storing only a single copy of each file or block and it also reduces the space and bandwidth requirements of data storage services, Across multiple users deduplication is most effective.

KEYWORDS:Hybrid cloud, data deduplication, authorized duplicate check, confidentiality.

I. INTRODUCTION

Cloud computing generate huge amount of data , require to be store in an efficient and secure fashion. And today, cloud computing have taking a lot of research interest and industrial implementation. The reason why this is being so popular goes back to organizational needs and adaptation for cloud services including storage, platforms and other services that are offers by cloud providers as shown in fig. 1. So Small, medium and even large organizations are not buying their own storage, they uses cloud services to store their data. The service is quit simple. It consists of a Cloud storage space assigned to a user for free or with reasonable fee. In addition, clouds can provide full functioning platforms to organizations allowing them to build their own specific platform and share it with others with worrying about their communication as soon as they are subscribed to the cloud. Hence cloud computing is being very popular today.

A hybrid cloud is an integrated cloud service. It utilises both private and public clouds to perform distinct functions within the same organisation. All cloud computing services should offer certain capabilities to differing degrees but public cloud services more cost efficient and scalable than private clouds. Thus, an organisation can maximise their capabilities by using public cloud services for all non-sensitive operations, only relying on a private cloud where they require it and ensuring that all of their platforms are seamlessly integrated.For dynamic or highly changeable workloads hybrid cloud is important. In private cloud application run, but it uses additional computing resources from a public cloud it uses computing resources which are needed. In Hybrid cloud model public cloud and private cloud resources are connected.Hybrid cloud used in big data processing.Suppose, a company could use hybrid cloud as storage to retain its accumulated business, sales, test and other data, and then run analytical queries in the public cloud, which can scale to support demanding distributed computing tasks.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

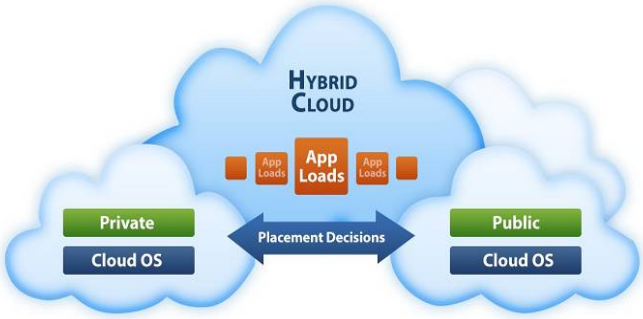


Figure 1: Hybrid Cloud Model

Public, private, and hybrid cloud technologies will also present opportunities to develop new architectures and processes that will advance security and management capabilities in ways that were not possible with physical computing restrictions. Therefore, while the hybrid cloud concept introduces new architecture considerations such as data migration, multi-cloud management, and distributed security models, it also presents new possibilities where security is concerned. Another key benefit of the hybrid cloud approach is the flexibility it offers. Companies wanting to capitalize on the benefits of both the private and public cloud approach are turning to a flexible hybrid cloud model. The hybrid approach allows businesses to take advantage of the scalability and cost-effectiveness that a public cloud computing environment offers, in combination with privateIn this paper, we present multi-cloud storage security challenges and the current adopted solutions in achieving security with their advantages and disadvantages.cloud computing and a strategic decision to keep some server operations on premise.

II. DE-DUPLICATION

The data deduplication is a technique that storeonly a single copy of redundant data, and provide pointers to thatcopy instead of storing actual copies of data. Withthe transition of services from tape to disk, In the backup process data duplication has become key component. Bystoring and transmitting only a single copy of redundant data,deduplication provides savings of both disk space and networkbandwidth. According to recent statistics, the most-impactful storage technology isconsidered to be deduplication and it is estimated to be applied to 75% of all backups in the nextfew years Data deduplication strategies can be categorized according to the basic data units they handle.Now a days data deduplication is popular technologies in storage because it offers services to companies to save money on storage costs to store the data and also on the costs of bandwidth to move the data when replicating it . It require less storage space because it stores only one copy of data , thus it requires less backup storage, that means it require less hardware and also backup of media. If storeage is less, means send less data over the network in case of a disaster, that means save money in hardware as well as network costs overtime. The benefits of data deduplication include:

•It Reduces hardware costs and backup costs

•Reduced costs for business continuity / disaster recovery

•Increases storage efficiency and network efficiency

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

There are two types of data deduplication :-

A. File-leveldeduplication:- In which only a single copy of each file is stored. Two or more files are identified as identical if they have the same hash value. This is a very popular type of service offered in multiple products.

B. Block-leveldeduplication:- In which segments files into blocks and stores only a single copy of each block. The system could either use fixed-sized blocks or variable-sized chunks. The discussion in this paper may be applied to both strategies.

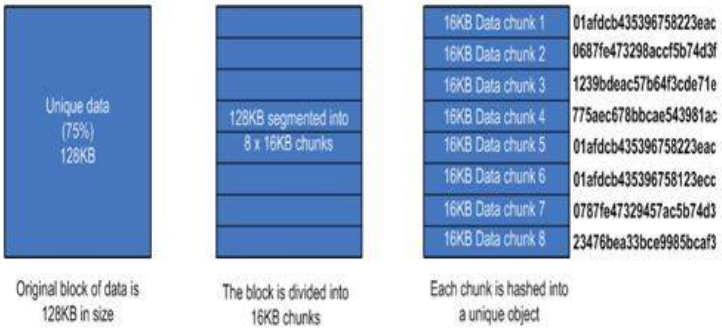
In easy way to understand data deduplication term ,is it smiply compares usually files or blocks are also known as objects and elimante copies that are already available in the data set. Deduplication process removes duplicate blocks and store only unique blocks. To happen this process it requires following four steps:

1. First it takes input data and then divide it into blocks or chunks.

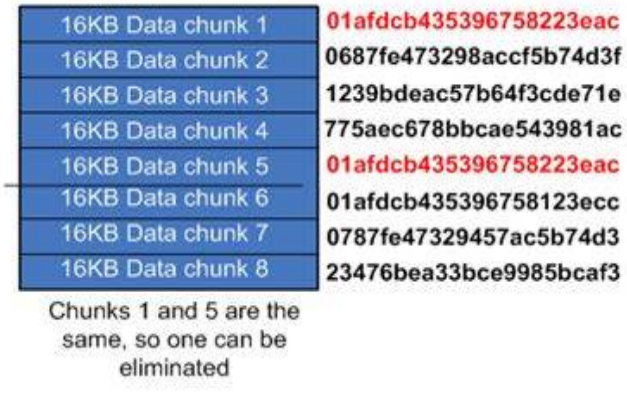
2. For each block or chunk of data it determine hash value using hashfunction.

3.Using these hashvaleus check that two blocks having same hashvalue that means block already been stored in database.

4.Replace the duplicate block of data with a reference to the chunk of data that already in the database.



Once the data is chunked, an index can be created from the results, and the duplicates can be found and eliminated. Only a single instance of every chunk is stored.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

III. LITERATURE SURVEY

Differential authorized duplicate check can not support previous deduplication systems which is most important in many application . In authorized deduplication system each user is assigned with set of privilegs during system initialization. Overview of cloud deduplication as follows.

A. Post-process Dedulpication:-

In this deduplication process first data is stored on storage device then perform duplication check . The main importance of this deduplication process is that there is no requirement to wait for the hash calculations and before storing the data ,analyzing of it completed and thus utilization of storage performance incrases. Implementations provides policy-based operation to users , Thus they have ability to defer optimization on active files, or to process files based on type and location. The main drawback of this process is that you may unnecessarily store redundant data for a temporary time which is problem if the storage of system is near fullcapacity. A block of data comes into the appliance and is written to storage completely .Then a separate process reads the block and checks it for redundancy. If it has been processed earlier (is considered to be a duplicate), it is deleted and replaced with a reference. If the block does not have duplicates, no changes are made. This method shortens time of data transfer from the source to storage but it needs more free space on a server disk and a lot more I/O than in-line deduplication.

B. In-line Dedulpication:-

In this process where deduplication hash calculations are created on the target device as the data enters the device in real time. Then device check a block that it already stored on the system it does not store the new block, it provides references to the existing block. The important benefit of in-line deduplication over post-process deduplication is that it requires less storage as redundant data has been elimnated, it stored only unique data. The disadvantage is, it is offenly contended that because hash calculations and searches takes more time, it means that the data ingestion can be slower thereby reduces the backup performance of the device.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

C.Source versus Target Deduplication:-

Another way to consider about data deduplication is where it happens. Deduplication occurs near to data where data is created commonly referrd as a source deduplication. Deduplication occurs where the data is stored this process is called target deduplication.Source deduplication guarantee that data on the data source is deduplicated. This deduplication usually occurs directly within a file system. The file system will offenly scan new files creating hashes and compare with the hashes of existing files. Files with same hashes found means they are duplicate files thus the duplicate file is removed and the new file refer to the old file. However, duplicated files are refered as to be different entities and any duplicated files are later modified, it uses system call Copy-on-write a means changed block is created. Users and backup applications are aware or transparent about this deduplication process. Storing a deduplicated file often cause duplication of file resulting the backups need more capacity source data. Target deduplication is the process of eliminating duplicate copies of the file from the secondary storage. It store such as data repository or a virtual tape library.

One of the most common forms of data deduplication is ,comparing chunks of data to find duplicates. This occur, when each chunk of data is assigned an identification, this calculated using hash function know as cryptographic hash functions. In implementations of many process, the assumption is that if identification means it is calculated hash value is same then data or content is identical this cannot be valid for all cases because principle of pigeonhole .other implementations process firest verify that the data with same hash value are duplicate files. Then software check that given hash value already exists then it replace that duplicate data with reference pointer .Whenever data has been deduplicated and user want to read file, then system provides pointer to that data chunk.

IV. PROPOSED SYSTEM

In this approach ,deduplication of data is achived by providing the proof of data from the owner of the data and this proof is helpful when uploading a file on cloud. Before uploading file to cloud each file is assigned with set of privileges to specify which type of users are allowed to perform duplicate check of the file and access of the files. Because of this users needs to provide his privileges as input before submitting duplicate check request for a file. Thus user is able to perform duplication check iff its matched privileges and copy of that file stored in cloud.

A. Encryption of Files:-

In this system common secret key k usedfor encryption and decryption data. It also used to convert the plain text to cipher text and viceversa. Following three basic functions are used,

KeyGenSE: k is the key generation algorithm that generates κ using security parameter 1.

EncSE (k, M):C is the symmetric encryption algorithm that takes the secret κ and message M and then outputs the ciphertext C

DecSE (k, C):M is the symmetric decryption algorithm that takes the secret κ and ciphertext C and then outputs the original message M.

B. Confidential Encryption:-

It provides data confidentiality in deduplication. A user derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derives a tag for the data copy, such that the tag will be used to detect duplicates.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

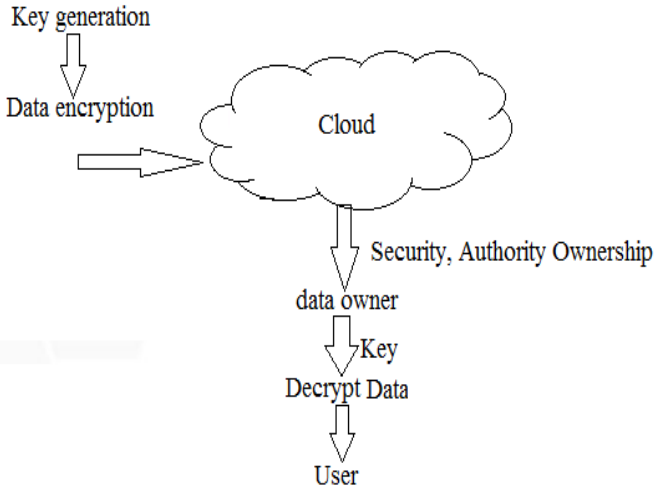
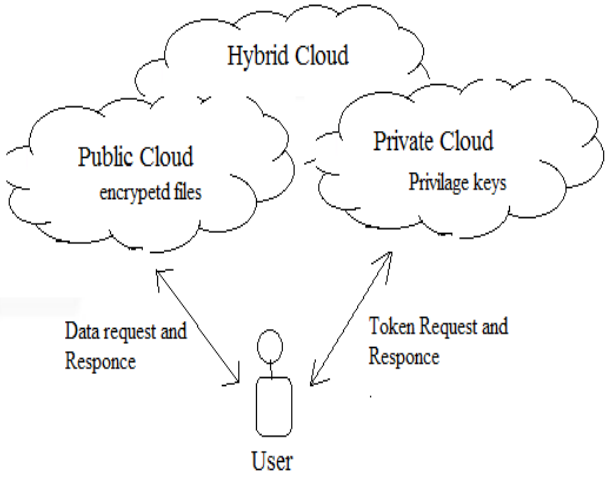


Figure : Confidential Data Encryption

C. Proof of Data:-

In proof of ownership protocol user needs to give his proof to server before downloading and uploading a file from or to the server, how it gives proof to server as follows:

User have to provide the convergent key and verifying that to prove his ownership at server ,if there is a duplicate of file and match privilege are stored in server.



V. Conclusion and Future Work

Cloud computing has reached at a higher level that leads it into a productive phase. This means main issues with cloud computing have been addressed to a degree that clouds have become interesting and benefical for full commercial exploitation. This does not mean that all the issues listed above have actually been solved, only that the according risks can be tolerated to a certain degree. Cloud computing is therefore still as much a research topic, as it is a market offering. For better confidentiality and security in cloud computing we have proposed new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Proposed system includes proof of data owner so it will help to implement better security issues in cloud computing.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart. "Dupless: Serveraided Encryption for Deduplicated Storage", In USENIX Security Symposium, 2013.
[2] P. Anderson and L. Zhang. "Fast and Secure Laptop Backups with Encrypted De-duplication", In *Proc. of USENIX LISA*, 2010.
[3] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. "Secure Deduplication with Efficient and Reliable Convergent Key Management", In IEEE Transactions on Parallel and Distributed Systems, 2013.
[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg., "Proofs of Ownership in Remote Storage Systems" In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.
[5] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou.,"Secure Deduplication With Efficient and Reliable Convergent Key Management", In IEEE Transactions on Parallel and Distributed Systems, 2013.
[6] C. Ng and P. Lee.,"Revdedup: A Reverse Deduplication Storage System Optimized for Reads to Latest Backups", In Proc. of APSYS, Apr 2013.
[7] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant TermSuggestion in Interactive Web Search Based on ContextualInformation in Query Session Logs", J. Am. Soc. for Information science and Technology, vol. 54, no. 7, pp. 638-649, 2003.
[8] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider,"Twin clouds: An architecture for secure cloud computing", In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.
[9] W. K. Ng, Y. Wen, and H. Zhu. "Private Data Deduplication Protocols in Cloud Storage", In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.
[10] R. D. Pietro and A. Sorniotti "Boosting Efficiency and Security in Proof of Ownership for Deduplication", In H. Y. Youm and Y. Won, editors, ACM Symposium on Information, Computer and communications Security, pages 81–82. ACM.
[11] S. Quinlan and S. Dorward"Venti: A New Approach to Archival Storage", In Proc. USENIX FAST, Jan 2002.
[12] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui.,"A Secure Cloud Backup System with Assured Deletion and Version Control", In 3rd International Workshop on Security in Cloud Computing, 2011.