INTERNATIONAL RESEARCH JOURNAL OF INDIA

A REVIEW PAPER ON HYBRID CLOUD APPROACH FOR SECURE AUTHORIZED DATA DEDUPLICATION

[1]Roshani Hukare & [2]Jagdish Pimple

roshanihukare@gmail.com

Computer Science Engineering Department, Nagpur Institute of Technology, Nagpur India

ABSTRACT:

Cloud computing is best concept to handle big database as the world is moving towards digitization. The amount of digital data in the world is growing exponentially with time. Thus, employing storage optimization techniques is an essential requirement to large storage areas like cloud storage. Cloud computing is best concept to handle big datasets. Data de-duplication is one of the best storage optimization techniques for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and save bandwidth. However, there is only one copy for each file stored in cloud even if such a file is owned by a huge number of users. As a result, Data Deduplication reduces bandwidth requirements, speeds up the data transfers, and improves storage utilization. Thus here in this paper we would be discussing data Deduplication techniques along with securing techniques thus forming secure Deduplication.

KEYWORDS: Data Deduplication, Cloud Storage, Data Security.

1. INTRODUCTION:

Cloud computing is an emerging technology, which can organize enormous resource of computing, storage and applications, and enable users to enjoy ubiquitous, convenient and on demand network access with great efficiency and minimum economic overhead. By these interesting features, both individuals and enterprises are motivated to outsource their data to the cloud, instead of purchasing software and hardware to manage the data themselves. Today world is moving on digitization and cloud computing is best concept to handle big datasets. Storage issue can be easily handled using cloud, but storing large and confidential data on cloud is not that easy. Using cloud can not only be expensive but also could be a threat to the confidentiality and security of the data [1]. Data Deduplication with Hybrid cloud is the best approach because hybrid cloud combines both public and private clouds, bound together by technology that allows data and applications to be shared between them. Hybrid cloud provides companies greater litheness and more deployment options by allowing data and applications to shift between private and public clouds. It provides the facility to store sensitive data in private cloud and less critical data on to the public cloud where huge savings can be made. Non- important actions are performed using public cloud while important actions are performed using private cloud. Adapting hybrid cloud is very easy for many companies as they will be having in-house cloud and will require only leverage the existing public cloud capabilities.

INTERNATIONAL RESEARCH JOURNAL OF INDIA

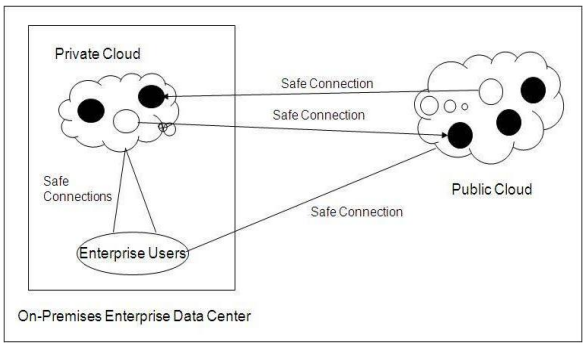

Fig 1: Example of Hybrid cloud Environment

Different techniques are used by Cloud storage providers to improve storage efficiency and one of leading technique employed by them is data deduplication. Instead of saving everything repeatedly, the ideal scenario is one where only the new or unique content is saved. Data de-duplication provides this basic capability. Data Deduplication offers the ability to find out and remove redundant data from within a dataset. Now, data deduplication is widely used by various cloud storage providers like Drop box, Amazon S3, Google Drive, etc. Data Deduplication is one of the best data compression techniques. Data de-duplication is necessary for reorganizing backup operations, enhancing information protection, cutting backup windows, removing load from information networks, and reducing backup infrastructure. De-duplication is a method to reduce the required Storage capacity since only the unique data is stored (refer Figure 2).

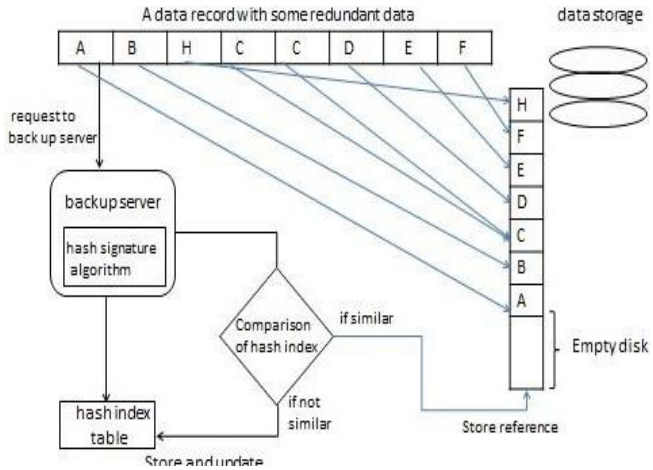

Fig 2: Deduplication Process

2. LITERATURE SURVEY:

The Hybrid Cloud is the architecture that provides the Organization to efficiently work on both the private and public cloud architecture in combination by providing the scalability to adopt. The author Neal Leavitt proposed here some of the basic concepts and idea about how best and easy to adopt this environment, Neal Leavitt [2].

Cloud provides infrastructure-as-a-service (IaaS) system, in which IT infrastructure is deployed in a provider's data center as virtual machines. With IaaS clouds' growing popularity, tools

INTERNATIONAL RESEARCH JOURNAL OF INDIA

and technologies are emerging that can transform an organization's existing infrastructure into a private or hybrid cloud. OpenNebula is a effective infrastructure manager which deploys virtualized services on both a local pool of resources and external Infrastructure as a Service clouds. It provides features not found in other cloud software or virtualization-based data center management software, Borja Sotomayor, Rubén S. Montero and Ignacio M. Llorente, Ian Foster [3].

Deduplication is a data compression 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 value for the particular file and with that the process of deduplication can be simplified, David Geer [4].

Data Deduplication is the widely used technique. This paper implements both File Level and Block Level Deduplication on Hybrid Cloud so as to avoid storage of same data uploaded by the any user . Here, data is encrypted before deduplication using AES (Advanced Encryption Standard) algorithm. The Security of the file is ensured by providing OTP's to the user that checks to see if the user is authenticated. This papers offer additional security by tracing and providing the MAC and the IP address of the hacker's machines, also introducing features like user friendly and easy to understand, Taranpreet Bhatti, Ashish James, Siddhi Narvekar, Prof. Varsha Wangikar [5].

In the data deduplication process, the first and the most important step is the granular division and subdivision of data. For deduplication of data different algorithms and methods are summed in this paper. An efficient and effective chunking algorithm is a must for proper granularity. If the data is chunked accurately, it increases the throughput and the net deduplication performance. The file-level chunking method is efficient for small files deduplication, but it is not relevant for a big file environment. The problem with the fixed-size chunking method is that it fails to detect redundant data if some bytes are altered. The variable-size chunking methods overcome this fixed file size chunking problem by creating boundaries based on the content of the files. The content-based chunking methods provide good throughput as well as decrease the space utilization, A. Venish and K. Siva Sankar [6].

To improve the consistency of data while achieving the confidentiality of the user's outsourced data without an encryption mechanism, distributed deduplication system is used. Four constructions were proposed in this paper to support file-level and fine-grained block-level data Deduplication. The security of tag consistency and integrity were achieved. This Deduplication system has been implemented using the Ramp secret sharing scheme and demonstrated that it incurs small encoding/decoding overhead compared to the network transmission overhead in regular upload/download operations, Meghana Vijay Kakde, N.B.Kadu [7].

To shield the data security by including differential rights of users in the duplicate check, authorized data deduplication was proposed. Here, they implemented several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, the duplicate-check tokens of files are generated by the private cloud server with private keys. Proposed security model in this paper is

INTERNATIONAL RESEARCH JOURNAL OF INDIA

secure in terms of insider and outsider attacks. The proposed authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer, Jagadish, Dr.Suvarna Nandyal [8].

An intelligent workload factoring, service for organization customers which makes the best use of the present public cloud services including their private owned data centers. It allows the organization to work between the off-premises and the on-premises infrastructure. The efficient core technology that is used for intelligent workload factoring is a fast redundant data element detection algorithm, that helps us factoring all the incoming requests based on the data content and not only on volume of data, Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and Akhilesh Saxena [9].

For secure deduplicated storage, two models are proposed in this paper. These two models are authenticated and anonymous. These two designs demonstrate that security can be combined with deduplication in a way that provides a diverse range of security characteristics. A plan is created for each file that describes how to reconstruct a file from chunks in both the authenticated and unknown models. This file is itself encrypted using a unique key. In the authenticated model, sharing of this key is managed through the use of asymmetric key pairs. In the anonymous model, storage is immutable, and file sharing is conducted by sharing the map key offline and creating a map reference for each authorized user . This paper implemented that the security of each model with regard to a number of security compromises and found that the system is mostly secure against external attackers, Mark W. Storer, Kevin Greenan, Darrell D. E. Long, Ethan L. Miller [10].

Data De-duplication is the data compression technique that is most effective and most widely used but when it is applied across the multiple users the cross-user deduplication tend to have to many serious privacy implications. In this paper, they discussed simple mechanisms which can enable the cross-user deduplication which will reduce the risks of the data leakage and also some of the security miss uses are discussed with how exactly to identify the files and to encrypt them while sending, DannyHarnik, Benny Pinkas, Alexandra Shulman- Peleg [11].

In order to optimize upload bandwidth and storage space over cloud, Source Based Deduplication is one of the best options. Deduplication is applied when data is on the source i.e. when data is created. Then the non-duplicate data is back up to the cloud. It helps in better and optimized utilization of resources. It is also helpful in incremental backup of new blocks in the user's instances. Distributed deduplication systems helps to achieve security, confidentiality and reliability if data. In addition distributed Deduplication system supports block-level Deduplication and file-level Deduplication. Hence if both the approaches are united then better deduplication ratio and reliability of data can be achieved, Vruti Satish Radia, Dheeraj Kumar Singh [12].

Data de-duplication technology is a process of removing duplicate data, identification of redundant files and to decrease the requirement to store data in order to utilize the overall storage capacity. Duplication of data or record can occur within a data block; within a file or in a specific data byte. Three levels of data de-duplication is available- File level de-duplication, Block-level de-

INTERNATIONAL RESEARCH JOURNAL OF INDIA

duplication, Byte-level de-duplication. In this paper, they implemented these three levels Deduplication. They thoroughly studied about various Data De-Duplication methods widely used in storage servers worldwide. These strategies can be optimized for utilizing storage Capacity, Neha Kaurav [13].

To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed here to encrypt the data before outsourcing. First attempt was made by this paper to better protect data security, which formally addressed the problem of authorized data deduplication. Different from traditional deduplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. They also present several new deduplication constructions supporting authorized duplicate check in a hybrid cloud architecture. Security analysis demonstrates that their scheme is secure in terms of the definitions specified in the proposed security model. Implemented a model of proposed authorized duplicate check system and conducted test bed experiments using their model, to check the genuineness of the concept. Proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations, Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee Wenjing Lou[14].

3. CONCLUSION:

Hybrid cloud approach with secure data Deduplication offers a great flexibility to the organizations. Data Deduplication is data compression technique for eliminating duplicate copies of data and reduces the storage space and bandwidth requirement. In this paper we have studied various techniques for data deduplication. Deduplication can be done on file level as well as block level. Most of the method studied here, work on the basis of convergent encryption, which is a simple approach that makes data deduplication compatible with encrypted data. To better protect data security, several new deduplication constructions supporting authorized duplicate check in a hybrid cloud architecture discussed in many techniques. Authorized duplicate check scheme incurs minimal overhead compared to normal operations.

INTERNATIONAL RESEARCH JOURNAL OF INDIA

4. REFERENCES:

[1] Zhihua Xia, Member, IEEE, Xinhui Wang, Xingming Sun, Senior Member, IEEE, and Qian Wang, Member, IEEE, "A Secure and Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data", IEEE Transactions on Parallel and Distributed Systems, Vol. 27, No. 2, February 2016.

[2] Neal Leavitt, "Hybrid Clouds Move to the Forefront", Published by the IEEE Computer Society, May 2013.

[3] Borja Sotomayor, Rubén S. Montero and Ignacio M. Llorente, Ian Foster, "Virtual Infrastructure Management in Private and Hybrid Clouds", Published by the IEEE Computer Society, 2009.

[4] David Geer, "Reducing the Storage Burden via Data Deduplication.computer.org", December 2008.

[5] Taranpreet Bhatti, Ashish James, Siddhi Narvekar,Prof. Varsha Wangikar, "Secure Authorized Deduplication on Hybrid Cloud", International Research Journal of Engineeringand Technology, April 2016.

[6] A. Venish and K. Siva Sankar , "Study of Chunking Algorithm in Data Deduplication", proceeding of the international conference on soft computing system ICSCS 2015, Vol. 2.

[7] Meghana Vijay Kakde, Prof. N.B.Kadu, "SurveyPaper on Deduplicating Data and Secure Auditing in Cloud", International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016.

[8] Jagadish , Dr.Suvarna Nandyal, "A Hybrid Cloud Approach for Secure Authorized Deduplication", International Journal of Science and Research 2013.

[9] Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and Akhilesh Saxena, "Intelligent Workload Factoring for A Hybrid Cloud Computing Model", Published by the IEEE Computer Society, 2009.

[10] Mark W. Storer, Kevin Greenan, Darrell D. E. Long, Ethan L. Miller, "Secure Data Deduplication", ACM 978-1-60558-299-3/08/10, October 31, 2008.

[11] Danny Harnik, Benny Pinkas, Alexandra Shulman- Peleg "Side Channels in Cloud Services Deduplication in Cloud Storage. Copublished by the IEEE Computer and Reliability Societies, November/December 2010.

[12] Vruti Satish Radia, Dheeraj Kumar, "Secure Deduplication Techniques: A Study", International Journal of Computer Applications (0975 – 8887) Vol. 137 – No.8, March 2016.

[13] Neha Kaurav, "An Investigation on Data De-duplication Methods And it's Recent Advancements", Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014.

[14] 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,Volume: PP, Issue:99, Date of Publication: 18.April.2014.