A Survey on Deduplication Techniques in Cloud Storage with Cryptographic Techniques

Aayushi Vernika Das, Delisha Clair Sequeira, Gulshan Damini Patel and V R Srividhya
Department of CSE
AMCEC
Bangalore, India

***Abstract*- Cloud computing is a paradigm shift in the Internet technology. Data deduplication apart from storing space also reduces bandwidth. Data deduplication brings high system overhead and as a reason there is a trade -off in deduplication efficiency and system performance. Latest studies on adopting data deduplication technique to cloud system are analyzed and compared with existing work. It is expect that our suggestions would be efficient enough to understand the studies.**

Keywords-Deduplication, ABE, Encryption

I. INTRODUCTION

Cloud computing is a used to store manage process the data on a remote server rather than a local server or a desktop computer.Applications and services are accessed via the Web, instead of the hard drive attached to the system. The services are delivered and used over the Internet and are paid for by cloud customer, typically on an "as-needed, pay-per-use" business model. The cloud substructure is preserved by the cloud provider, not the individual cloud customer.

Cloud computing provides three main services:

1) Software as a service (Saas)
2) Platform as a service (Paas)
3) Infrastructure as a service (Iaas)

A. Software as a Service:

Software as a Service (SaaS) is well-defined as a software which is positioned over the internet. Saas make available an application to customer as a service on demand, in a "pay-as-you-go" model, at no charge when there is chance to produce revenue from streams other than the operator, such as commercial or user list sales.

B. Platform as a Service

Platform as a Service (PaaS) can be well-defined as a figuring platform that permits the making of web applications speedily and effortlessly and without the complexity of buying and maintaining the software and infrastructure underneath it.
PaaS is a platform provided for the creation of software which is delivered over the web.

C. Infrastructure as a Service

Infrastructure as a Service (IaaS) is also known as Hardware as a Service (HaaS) is a way of delivering Cloud Computing infrastructure – servers, storage, network and operating systems – as an on-demand service. Rather than purchasing servers, software, data centre space or network equipment, clients instead buy those resources as a fully outsourced service on demand.

Cloud storage has become so popular now a days not because of its cost efficient on demand usage but because of provision of huge amount of storage on which user can easily outsource its data and all the services it provides which are mentioned above. Cloud computing is emerging day by day due to its ability to provide cost efficient and on demand use of huge storage as well as resources. As there is a huge amount of growth of contents and resources online, cloud storage looks forward for better utilization of storage resources. Most of the surveys indicate that nearly half of consumed storage on cloud is occupied by duplicate files or data.

Challenge in front of today's cloud services is the management of that huge increasing quantity of data. There is a need to avoid large amount of redundancy in storage because of duplicate data[1]. For that reason in cloud storage systems, there is a requirement, for reducing the amount of data that is to be transferred or stored. This is crucial for providing benefits for performance of applications, storage costs and administrative issues.

To make data management scalable, de-duplication is introduced. In deduplication, instead of storing multiple copies of data which are exactly identical, it keeps only one physical copy. In public cloud environment Security and privacy is a concern. This Data confidentiality for users is achieved with Convergent Encryption instead of Traditional encryption.

As a result, Data De-duplication plays important role as it saves cost for cloud storage, by storing only a single copy of redundant data, and provide pointers to clients of that duplicate or redundant copies instead of storing actual copies of that data. In cloud computing to make data management scalable, deduplication has attracted more and more attention recently in the backup process. Deduplication stores and transfers only a single physical copy of identical data, in this way de-duplication saves both disk space as well as network bandwidth.

Two major business benefits of data deduplication are:
(i) Reduced costs in hardware, back up and business continuity/Disaster Recovery;
(ii)Increased storage and network efficiency.

Deduplication can occur at two different levels. They are:
(i) Source based-Prior to sending to a back up target, duplicates are removed, Used for fewer Data Sets
(ii) Target based-Duplicates are removed at the storage end of the target, Used for Larger Data Set

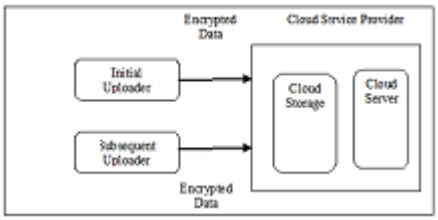

Fig. 1 Architecture of Data deduplication System

Fig. 1 represents the Architecture of a Data deduplication system[3]. It consists of the
(i) Data Owner that is a client who is the owner of the data, and wishes to upload it into the cloud storage to so as to reduce and save costs. A tag or Index is attached to it.
(ii) Cloud Service Provider that provides cloud storage services. It consists of a cloud server and cloud storage. The Cloud server de duplicates the data if essential and then stores it in the Cloud Storage. This takes care of all the requirements of a Cloud Service.

II. RELATED WORK

A. Secure data deduplication with dynamic ownership management in cloud storage[3]:
Here it is intended to achieve deduplication with dynamic ownership management. The scheme features a re-encryption technique that enables dynamic updates upon any ownership changes in cloud storage. The encryption technique used in it will allow the cloud server to control access to outsourced data even when the ownership changes dynamically by exploiting randomized convergent encryption and secure ownership group key distribution. Whenever an ownership change occurs in the ownership group of outsourced data the data are re-encrypted with an immediately updated ownership group key, which is securely delivered to valid owners.

It uses Elliptic Curve Cryptographic technique for encryption and Decryption and Hash function for key generation.

Advantages
1) Reduce space and bandwidth requirement
2) Computational cost is less almost negligible
3) High efficiency
4) Enhances data privacy and confidentiality
5) Tag consistency is guaranteed

B. Attribute based storage supporting secure deduplication of encrypted data in cloud [5]:
Data is stored in encrypted form with access control policies such that no one except the users with attributes of specific forms can decrypt the encrypted data. An encryption technique that meets the requirement is called attribute based encryption where a user's private key is associated with an attribute set. A message is encrypted under an access policy over a set of attributes and a user can decrypt a cipher text with his/her private key if his/her set of attributes satisfy the access policy.

The cloud is been designed in such a way that it has both the properties of ABE and deduplication; Here no matter how many times the file is been encrypted and have different access policies it will maintain only a single copy of the data in the cloud. ABE is widely used where the data providers outsource the data on the cloud and can share with different users who have some credentials.

Advantages

1) The system is the first that achieves the standard notion of semantic security for data confidentiality in attributes based deduplication.
2) Modifies a cipher text over one access policy into cipher texts of the same plaintext but under any other access without revealing the underlying plaintext.
3) Data consistency is achieved.

C. Secure enterprise Data deduplication in the cloud[4]:
Most enterprises are choosing to outsource the data on cloud for better management of IT resources, in terms of Security, control space and storage costs. This paper proposes a novel two level data deduplication-

1) At enterprise level.
2) At cloud storage provider level.

One of the major disadvantages in the existing system is that even though the space used is less and cost is low, Data deduplication will allow to maintain only a single copy of data due to this it will allow all authorized user to enterprise the data including confidential data. So data deduplication framework was introduced to preserve privacy of data while it still performs data deduplication. At enterprise level each individual enterprise will perform its own data deduplication among its users. At cloud storage provider level it will perform deduplication again.

It uses two level data deduplication framework that can be used in the cloud storage by enterprise who stores the data in a common cloud service provider. Here the cloud service provider is semi honest that is when many users are there it cannot be trusted to handle the data.

Advantages
1) Saves space
2) Cost is less

III. CONCLUSION

Study of Data deduplication and its different ways can be very helpful for cloud storage, which need very convenient and cost efficient storage system. Notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. Three different types of techniques are seen and compared and anyone can be used as per the need. Also, Focus is thrown on their working. Due to deduplication, client is not charged for duplicate storage also security analysis demonstrates that proposed schemes are secure in all terms.

IV. REFERENCES

[1] Bolosky, D. M. " A study of practical deduplication". in Proc. USENIX Conf. File Storage Technol, 1-1, 2011.
[2] W.K. Ng, W. W. ,"Private data deduplication protocols in cloud storage", 27th annual ACM Symp Appl comput, 441-446,2012..
[3] Junbeom Hur, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang(2016), "Secure data deduplication with dynamic ownership management in cloud storage",in Ieee Transactions On Knowledge And Data Engineering, Vol. 28, No. 11, November 2016.
[4] Fatema Rashid, Ali Miri, Isaac Woungang,"Secure enterprise Data deduplication in the cloud",IEEE Sixth International Conference on Cloud Computing,2013.
[5] ," Hui Cui, Robert H. Deng, Yingjiu Li, and Guowei Wu "Attribute based storage supporting secure deduplication of encrypted data in cloud", IEEE Transactions on Big Data,2017.