Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 4, April 2016, pg.529 – 533

DATA DEDUPLICATION FOR NON SECURE TEXT AND DOC FILES IN CLOUD

Mallika D K G[1], Prof. Sudheer Shetty[2]

[1]Department of Computer Science & Engineering, Sahyadri College of Engineering & Management, Mangaluru, Karnataka, India

[2]Department of Computer Science & Engineering, Sahyadri College of Engineering & Management, Mangaluru, Karnataka, India

[1] mallika.mtechscs14@sahyadri.edu.in; [2] sudheer.cs@sahyadri.edu.in

Abstract— Information when interpreted we get data. Data about this data is called Metadata. For example: File consists of Information. Therefore File is data. And details like File name, file size etc. are Metadata. Data is stored in storage spaces. Currently the most dominantly used storage method is Cloud Storage. There has been a large increase in the usage of the Internet in the past few years. The Cloud has played a huge role in it. Traditional Data systems are slowly fading away and Cloud storage system is falling into place. One of the issues Cloud storage is facing is data redundancy. To solve this issue one should achieve Data Deduplication. Data Deduplication is a concept where unique data or a particular /similar pattern of data is identified and stored. This is then compared to other data available in the system. If a match is found, then the data is replaced by a link or a reference to the stored data.

Keywords— cloud computing, cloud storage, deduplication.

I. INTRODUCTION

Cloud computing is a platform that provides seemingly unlimited "virtualized "resources to users across the whole Internet, while hiding platform and implementation details. Highly available storage and massively parallel computing resources are provided at relatively low costs. Prevalent use of cloud, an increasing amount of data is being stored and shared by users with specified privileges, which define the access rights of the data stored. In short we can say that cloud computing is computing as a business model that provide computing resources as a service on demand to customers over the Internet. [3]

Cloud providers pool computing resources together to serve customers via a multi-tenant model. Computing resources are delivered over the Internet where customers can access them through various client platforms. Customers can access the resources on-demand at any time without human interaction with the cloud provider. From a customers' point of view, computing resources are infinite, and customer demands can rapidly change to meet business objectives. This is facilitated by the ability for cloud services to scale resources up and down on demand leveraging the power of virtualization. Moreover, cloud providers are able to monitor and control the usage of resources for each customer for billing purposes, optimization resources, capacity planning and other tasks. Cloud storage is one of the services in cloud computing which provides virtualized storage on demand to customers.

Cloud storage can be used in many different ways. [4] For example, customers can use cloud storage as a backup to store a backup copy of data, as opposed to maintaining their own storage disks. Establishments have moved their archival storage to the cloud which they can achieve more capacity at a lesser cost, rather than buying additional physical storage. Applications running in the cloud also require temporary or permanent data storage in order to support the applications.

As the amount of data in the cloud is rapidly increasing, customers expect to reach the on-demand cloud services at any time, while providers are required to maintain system availability and process a large amount of data. Providers need a way to dramatically reduce data volumes, so they can reduce costs while saving energy consumption for running large storage systems. [2] This is where the concept of data deduplication can plays a key role

In this paper we propose a novel approach that makes use of data deduplication. Logical pointers are created for other copies instead of storing redundant data. Deduplication can reduce both storage space and network bandwidth [5].

II. LITERATURE SURVEY

A. Existing Deduplication Architecture:

We are looking at system architectures of existing works of deduplication for cloud backup services such as SAM [11], AA-Dedupe [8], CABdedupe [9], and SHHC [10].

SAM [10] system architecture is composed of three subsystems: File Agent, Master Server and Storage Server. Clients subscribe to backup services, then File Agents are distribute and installed on their machines, while service provider provides Master Server and Storage Server in datacentre to serve the backup requests from clients. Most of existing solutions that use deduplication technology primarily focus on the reduction of backup time while ignoring the restoration time. The authors proposed CABdedupe [16], a performance booster for both cloud backup and cloud restore operations, which is a middleware that is orthogonal and can be integrated into any existing backup system. CABdedupe consists of CAB-Client and CAB-Server, which is placed on the original client and server modules in existing backup systems.

The main aim of these related works are the following: SAM aims to achieve an optimal trade-off between deduplication efficiency and deduplication overhead, CABdedupe reduces both backup time and restoration time. AA-Dedupe [15] aims to reduce the computational overhead, increase throughput and transfer efficiency, while SHHC [17] tries to improve fingerprint storage and lookup mechanism, however has a concern of scalability. SHHC is a novel Scalable Hybrid Hash Cluster designed for improving response times to fingerprint lookup process. Because of a large number of simultaneous requests are expected in cloud backup services.

In order to solve this problem, the hash cluster is designed for high load-balancing, scalability and minimizing the cost for each fingerprint lookup query. The hash cluster is designed as middleware between the clients and the cloud storage. It provides the fingerprint storage and lookup service.

There are other works on deduplication storages which their architectures are designed for scalability issue, for example; Extreme Binning [12], and Droplet [13].

Extreme Binning is used to build a distributed file backup system. The architecture of such system is composed of several backup nodes. Each backup node consists of a compute core and RAM along with a dedicated attached disk. The first task when a file arrives to the system for backup is, it must be chunked. The system can delegate this task to any one of the backup nodes by choosing one according to the system load at that time. After chunking, stateless routing algorithm is used to route the chunked file by using its chunk ID. The chunked file will be routed to a backup node where it will be deduplicated and stored.

Droplet, a distributed deduplication storage system designed for high throughput and scalability. It consists of three components: a single meta server that monitors the entire system status, multiple fingerprinting servers that run deduplication on input data stream, and multiple storage nodes that store fingerprint index and deduplicated data blocks. Meta server maintains information of fingerprinting and storage servers in the system. When new nodes are added into the system, they need to be registered on the meta server first. The meta server provides a routing service with this information.

The client first connects to the meta server and queries for list of fingerprinting servers, and then connects to one of them. After this, a raw data stream containing backup content will be sent to this fingerprinting server, which calculates data block fingerprints and replies results to the client. Fingerprint servers check duplicated fingerprint by querying storage servers.

The nature of data in cloud storage is dynamic [1], [14]. For example, data usage in cloud changes overtime, some data chunks may be read frequently in period of time, but may not be used in another time period. Some datasets may be frequently accessed or updated by multiple users at the same time, while others may need the high level of redundancy for reliability requirement. Therefore, it is crucial to support this dynamic feature in cloud storage. However, current approaches are mostly focused on static scheme, which limits their full applicability in dynamic characteristic of data in cloud storage.

B. Motivation for Proposed System

When a simple non secure .txt or .doc file has to be uploaded to the system, the existing deduplication methods are unnecessary since security aspect doesn't come into picture. Even if the existing systems are used, the complexity will exist which results in needless usage of resources. Hence we propose a system for non-secure files with reduced complexity in the system.

C. Related Works

Waraporn Leesakul, Paul Townend, Jie Xu used propose a dynamic deduplication scheme for cloud storage, which aiming to improve storage efficiency and maintaining redundancy for fault tolerance.

In order to improve availability while maintaining storage efficiency, Waraporn Leesakul, Paul Townend, Jie Xu [2] proposed a deduplication system which considers both the dynamicity and taking Quality of Service (QoS) of the Cloud environment into consideration. After identifying the duplication, the Redundancy Manager then calculates an optimal number of copies for the file based on number of references and level of QoS necessary. The numbers of copies are dynamically changed based on the changing number of references, level of QoS and demand for the files. The changes are monitored, for example, when a file is deleted by a user, or the level of QoS of the file has been updated, this will trigger the redundancy manager to re-calculate an optimal number of copies.

Authors Chun-I Fan, Shi-Yuan Huang, and Wen-Che Hsu [6] brief us about the following in their paper. In cloud environments, users store their data or files in cloud storage but it is not infinitely large. In order to reduce the requirement of storage and bandwidth, data deduplication has been applied. Users can share one copy of the duplicate files or data blocks to eliminate redundant data. Besides, considering the privacy of sensitive files, the users hope that the cloud server cannot know any information about those files. They often use certain encryption algorithms to protect the sensitive files before storing them in the cloud storage. Unfortunately, previous schemes have a security problem. These schemes did not satisfy semantic security. They propose a hybrid data deduplication mechanism which provides a practical solution with partial semantic security.

When studied a paper authored by Pasquale Puzio, Refik Molva, Melek O¨nen, Sergio Loureiro [7]., We learn that with the continuous and exponential increase of the number of users and the size of their data, data deduplication becomes more and more a necessity for cloud storage providers. By storing a unique copy of duplicate data, cloud providers greatly reduce their storage and data transfer costs. The advantages of deduplication unfortunately come with a high cost in terms of new security and privacy challenges.

They propose ClouDedup, a secure and efficient storage service which assures block-level deduplication and data confidentiality at the same time. Although based on convergent encryption, ClouDedup remains secure thanks to the definition of a component that implement an additional encryption operation and an access control

mechanism. Furthermore, as the requirement for deduplication at block-level raises an issue with respect to key management, we suggest including a new component in order to implement the key management for each block together with the actual deduplication operation. We show that the overhead introduced by these new components is minimal and does not impact the overall storage and computational costs.

III. MODEL AND ARCHITECTURE

A. Proposed System

We are in a need of a system that can upload non secure documents into the cloud removing redundancy thereby eliminating extra usage of storage space and unnecessary usage of bandwidth by eradicating uploading of redundant data.

B. System Architecture

The architectural design will show the conceptual model of the application. It shows the overall architecture of the system.

The system has an admin who is the owner of the system. The user logs in and uploads or downloads into the system. As per the option chosen, the alert is sent by the system using pop up messages.

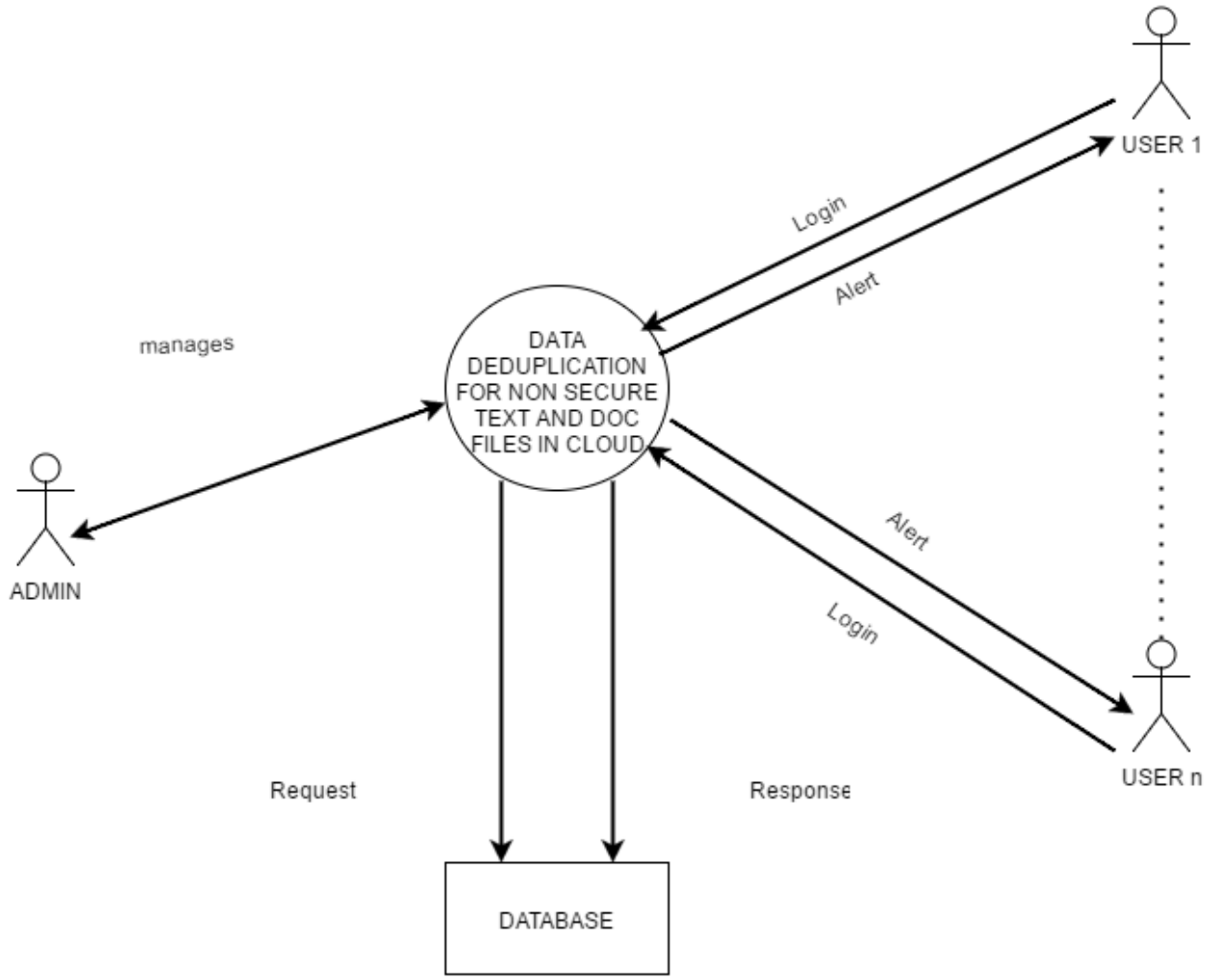

Fig. 1 System Architecture

C. System Implementation

We intend to perform three main tasks using the system i.e. inserting the data, updating the data and deleting the data. Basically we will be hashing techniques to extract the metadata. This data will be then compared with previously existing values of the data already existing in the system. If there is a match we can conclude that redundant data exists and thereby eliminate it.

Advantages:

•Obviously storage space can be saved.

•Because redundant flies will not be uploaded but only a reference will be given to the existing file, the energy consumed to store it and the bandwidth used for uploading also can be saved.

•Overall maintenance becomes easier, and also overall costs can also be reduced.

IV. CONCLUSIONS

In this work we present a method which is a data deduplication mechanism which provides a practical solution that is more secure than previous techniques in some cases.

The storage space will be saved; the energy consumed to store and maintain the data is reduced since redundancy is reduced. Also the bandwidth will be saved. The maintenance becomes smooth, thereby reducing overall cost

Acknowledgement

I would like to thank God Almighty for blessing me to complete this work.

I am profoundly indebted to my guide, Mr. Sudheer Shetty, Associate Professor, Department of Computer Science and Engineering, Sahyadri College of Engineering and Management, for innumerable acts of timely advice and encouragement.

I would like to express my sincere thanks to my beloved family members and friends for their wishes and encouragement throughout the work.

REFERENCES

[1] W. Cong, W. Qian, R. Kui, C. Ning, and L. Wenjing, "Toward Secure and Dependable Storage Services in Cloud Computing," Services Computing, IEEE Transactions on, vol. 5, pp. 220-232, 2012.

[2] Waraporn Leesakul, Paul Townend, Jie Xu,"Dynamic Data Deduplication in Cloud Storage", in 2014 IEEE 8th International Symposium on Service Oriented System Engineering,2014,pp 1-6.

[3] T. G. Peter Mell, "The NIST Definition of Cloud Computing," National Institute of Standards and Technology NIST Special Publication 800- 145, September 2011.

[4] SNIA Cloud Storage Initiative, "Implementing, Serving, and Using Cloud Storage,"Whitepaper 2010..

[5] SNIA, "Advanced Deduplication Concepts," 2011.

[6] Chun-I Fan, Shi-Yuan Huang, and Wen-Che Hsu,"Hybrid Data Deduplication in Cloud Environment", 2012 IEEE Journal, 978-1-4673-2588-2/12,pp 1-4..

[7] Pasquale Puzio, Refik Molva, Melek Onen, Sergio Loureiro,"ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage" in 2013 IEEE International Conference on Cloud Computing Technology and Science,2013, pp. 1-8..

[8] T. S F. Yinjin, J. Hong, X. Nong, T. Lei, and L. Fang, "AA-Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment," in Cluster Computing (CLUSTER), 2011 IEEE International Conference on, 2011, pp. 112-120.

[9] T. Yujuan, J. Hong, F. Dan, T. Lei, and Y. Zhichao, "CABdedupe: A Causality-Based Deduplication Performance Booster for Cloud Backup Services," in Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International, 2011, pp. 1266-1277.

[10] X. Lei, H. Jian, S. Mkandawire, and J. Hong, "SHHC: A Scalable Hybrid Hash Cluster for Cloud Backup Services in Data Centers," in Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on, 2011, pp. 61-65.

[11] T. Yujuan, J. Hong, F. Dan, T. Lei, Y. Zhichao, and Z. Guohui, "SAM: A Semantic-Aware Multi-tiered Source De-duplication Framework for Cloud Backup," in Parallel Processing (ICPP), 2010 39th International Conference on, 2010, pp. 614-623.

[12] D. Bhagwat, K. Eshghi, D. D. E. Long, and M. Lillibridge, "Extreme Binning: Scalable, parallel deduplication for chunk-based file backup," in Modeling, Analysis & Simulation of Computer and Telecommunication Systems, 2009. MASCOTS '09. IEEE International Symposium on, 2009, pp. 1-9.

[13] Z. Yang, W. Yongwei, and Y. Guangwen, "Droplet: A Distributed Solution of Data Deduplication," in Grid Computing (GRID), 2012 ACM/IEEE 13th International Conference on, 2012, pp. 114-121.

[14] K. Yang and X. Jia, "An Efficient and Secure Dynamic Auditing Protocol for Data Storage in Cloud Computing," Parallel and Distributed Systems, IEEE Transactions on, vol. PP, pp. 1-1, 2012.

