INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue4)

Available online at: www.Ijariit.com

Review Data De-Duplication By Encryption Method

Sonam Bhardwaj
M.Tech Student, UIET, Kurukshetra University
sbsonambhardwaj@gmail.com

Poonam Dabas
Assistant Professor, UIET, Kurukshetra University
poonamdabas.kuk@gmail.com

Abstract— Data deduplication is a technique to improve the storage utilization. De-duplication technologies can be designed to work on primary storage as well as on secondary storage. De-duplication with the use of chunking data that is passed through the de-duplication engine is chunked into smaller units and assigned identities using cryptographic hash functions. Thereafter, two chunks of data are compared to ascertain whether they have the same identity. Chunking for de-duplication can be frequency based or content based. Frequency based chunking identifies high frequencies of occurrences of data chunks. The algorithm uses this frequency information to enhance data duplication gain.

Keywords— deduplication, cloud, SHA1.

I. INTRODUCTION

Big data is an abstract concept. Apart from masses of data, it also has some other features, which determine the difference between itself and "massive data" or "very big data." Big Data as the next frontier for innovation, competition, and productivity. Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software. This definition includes two connotations: First, datasets volumes that conform to the standard of big data are changing, and may grow over time or with technological advances. Second, datasets volumes that conform to the standard of big data in different applications differ from each other[4].

In addition, NIST defines big data as "Big data shall mean the data of which the data volume acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies", which focuses on the technological aspect of big data. It indicates that efficient methods or technologies need to be developed and used to analyze and process big data [1].

Cloud computing has recently emerged as a popular business model for utility computing system. The concept of cloud is to provide computing resources as a utility or a service on demand to customers over the Internet.

Cloud storage is one of the services in cloud computing which provides virtualized storage on emend to customers. Cloud storage can be used in many different ways. For example, customers an use loud storage as a backup service, as opposed to maintaining their own storage disks. Organisations can move their archival storage to the cloud which they can achieve more capacity at the low-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 [2]

Data de-duplication, is a file system feature that only saves unique data segments to save space. It has been most popular and successful for secondary storage systems (backup and archival) Data de-duplication technology, the basic principle is to filter the data block to find the same data block, and the only instance of a pointer to point to replace. Data de-duplication technology to identify duplicate data, eliminate redundancy and reduce the need to transfer or store the data in the overall capacity. Duplication to detect duplicate data elements, to judge a file, block or bit and another file, block or bit the same. Data de-duplication technology to use

mathematics for each data element, algorithms to deal with, and get a unique code called a hash authentication number. Each number is compiled into a list, this list is often referred to as hash index.

Data deduplication is a technique to improve the storage utilization. De-duplication technologies can be designed to work on primary storage as well as on secondary storage. De-duplication with the use of chunking Data that is passed through the de-duplication engine is chunked into smaller units and assigned identities using cryptographic hash functions. Thereafter, two chunks of data are compared to ascertain whether they have the same identity. Chunking for de-duplication can be frequency based or content based. Frequency based chunking identifies high frequencies of occurrences of data chunks. The algorithm uses this frequency information to enhance data duplication gain. Content based chunking is a stateless chunking algorithm which partitions a long stream of data into smaller units or chunks and removes duplicate ones[6].

Hash Based De-duplication: Hash based data de-duplication methods use a hashing algorithm to identify "chunks" of data. Commonly used algorithms are Secure Hash Algorithm 1 (SHA-1) and Message-Digest Algorithm 5 (MD5). When data is processed by a hashing algorithm, a hash is created that represents the data. A hash is a bit string (128 bits for MD5 and 160 bits for SHA-1) that represents the data processed. If you processed the same data through the hashing algorithm multiple times, the same hash is created each time.

MD5: MD5, with the full name of the Message-digest Algorithm 5, is the fifth generation on behalf of the message digest algorithm. In August 1992, Ronald L. Rivest submitted a document to the IETF (The Internet Engineering Task Force) entitled "The MD5 Message-Digest Algorithm", which describes the theory of this algorithm. Message digest is a cryptographic hash function containing a string of digits. It's designed to protect the integrity of data & to detect the changes if made any to the data. MD5 takes the message of any length and computes a fixed 128 bit hash value by compressing the message.[2]

SHA-1: The design of SHA-1 algorithm imitates MD4 mostly, which accepts the bits of the maximum length of the message, to generate a 160-bit message digest. Similar with the MD5, this arithmetic operation is divided into 32-bit word of 512 bits' length block for processing units, including four loop operators, 20 rounds per loop, a total of 80 rounds.

Hadoop is the open source framework for storage and large scale processing of data-sets on the clusters of the commodity hardware. Hadoop is composed of following modules. Hadoop common Hadoop Distributed File System, Hadoop Yarn, Hadoop Map Reduce Hadoop[1] architecture consists of the Hadoop common package, which provide us to abstraction of the file system and OS level contact. Hadoop comes with the java files which are used to start the Hadoop and the other components of the Hadoop.

MapReduce divides workloads up into multiple tasks that can be executed in parallel. *MapReduce key attributes: a) Resource Manager:* Employs data locality and server resources to determine optimal computing operations. *b) Optimized Scheduling:* Completes jobs according to prioritization.
c) Flexibility: Procedures can be written in virtually any programming language. *d) Resiliency & High Availability:* Multiple job and task trackers ensure that jobs fail independently and restart automatically. *e) Scale-out Architecture:* Can add servers to increase processing power.

II. Literature Review

In the year 2010, Qinlu He, Zhanhuaiand Xiao [1] presented Data Deduplication Techniques for optimizing the storage system that greatly reduced the amount of data and energy consumed. Data Compression reduced the number of disks used in the operation for reducing disk energy consumption costs. Data deduplication strategies that were considered were file level, block level and Byte level, also the research show us the five stages of development that are: First phase, which is a data collection phase. Second is the phase that identifies data in Bytes. In the third phase data is re-assembled. Fourth phase removes all duplicate data. In fifth phase redundant storage of data is removed.

Amrita Upadhyay et al [2] deduplication and Compression Techniques in Cloud Design in the year 2012, aimed at reduction in storage space and bandwidth usage during file transfers. Existence of duplicate files was determined from the metadata. Files clustered into bins depending on their size, then segmented, deduplicated, compressed and stored. In this research, Deduplication was done in two situations: for existing files and for incoming files (new files). Binning is the process by which we can decide the size of each segment to be formed, based on the size of the parent file. This decreased the time to process and save the segments. Reduction

of 47.5% in processing time was achieved. After Segmentation, the Hash Value for each of the file segments calculated using Hashing algorithm SHA-1. The first time the deduplication was performed; hash values were calculated for every segment and recorded in metadata structure. For newly uploaded file, hash values calculated for its segments and then compared with the list of existing hash values. If there was a match in the values corresponding segment file is not saved.

P.Neelaveni [3]A Survey on Deduplication in Cloud storage, 2014 had various challenges: Bandwidth, Throughput, Computational overhead, Deduplication efficiency, Read and write efficiency, Backup window size, Transmission cost. Deduplication can be done to reduce the storage amount consumed by virtual machine images. There were issues like high duplicate tracking space, space overhead and high computation. Therefore, various deduplication strategies were adopted to benefit the cloud backup services.

Xing Yu-Xuan [4]: Traditional Deduplication systems based on convergent encryption even though provide Confidentiality but do not support the duplicate check on the basis of differential privileges. Paper presented idea of authorized data deduplication proposed to protect data security by including differential privileges of users in the duplicate check. To support stronger security the files were encrypted with differential privilege keys, users only allowed to perform the deduplication for the files marked with the corresponding privileges to access. Users can verify his/her presence of file after deduplication in cloud with the help of a third party auditor by auditing the data further auditor audits and verifies the uploaded file on time.

LiuFang et al[5]: AR-dedupe :An efficient deduplication approach for cluster deduplication in the year 2015 marked the following challenges as their research base. Decreasing data deduplication rate with the increasing dedupe server nodes. High communication overhead for data routing. Load balance for improving throughput of the system. Cluster Dedupe has three parts Backup client, Metadata and Deduplication server Nodes. First, it partitions large data objects into smaller parts called chunks and generates its fingerprint which can be uniquely represented in the backup client. Then, it transfers all chunks to deduplication server nodes according to its routing mechanism. Metadata management server keeps the information of all files for restoration. The algorithm used for chunking was Content Defined Chunking that forms chunks according to the content.

CONCLUSION

Decreasing data deduplication rate with the increasing dedupe server nodes. High communication overhead for data routing. Load balance for improving throughput of the system. Cluster Dedupe has three parts Backup client, Metadata and Deduplication server Nodes. First, it partitions large data objects into smaller parts called chunks and generates its fingerprint which can be uniquely represented in the backup client. Then, it transfers all chunks to deduplication server nodes according to its routing mechanism

REFERENCES

[1]Quinluhe, ZhanhuaiLi, XiaoZhang"DataDeduplicationTechniques", IEEEInternational Conference, 2010.

[2] Amrita Upadhyay,Pratibha R Balihalli,ShashibhushaIvaturi and Shisha Rao,"Deduplication and Compression Techniques in Cloud Design",IEEE 978-1-4673-0750-5/12/$31.00©2012 ,Pratibha R Balihalli,ShashibhushaIvaturi and Shisha Rao,"Deduplication and Compression Techniques in Cloud Design",IEEE 978-1-4673-0750-5/12/$31.00©2012

[3] P.Neelaveni,M.VijayaLakshmi,"A Survey On Deduplication in Cloud Storage",Asian journal Of Information Technology 13(6):320-330©Medwell journals,2014

[4] Xing Yu-Xuan,XlaoNOng,LiuFang,SunZhen,HeWan-Hui,"AR-Dedupe:An Efficient Deduplication Approach For Cluster Deduplication System",J.ShanghaiJiaotong Univ.(Sci.),2015,20(1):76-81

[5] Xing Yu-Xuan, XlaoNOng, LiuFang, SunZhen, HeWan-Hui,"AR-Dedupe:An Efficient Deduplication Approach For Cluster Deduplication System",J.ShanghaiJiaotong Univ.(Sci.),2015,20(1):76-81

[6] S. Kumar Bose, S. Brock, R. Skeoch, N. Shaikh, and S. Rao, "Optimizing live migration of virtual machines across wide area networks using integrated replication and scheduling," in Systems Conference (SysCon), 2011 IEEE International, 2011, pp. 97-102.

[7] S. K. Bose, S. Brock, R. Skeoch, and S. Rao, "CloudSpider: Combining Replication with Schedulingfor Optimizing Live Migration of Virtual Machines across Wide Area Networks," in Cluster, Cloud and Grid Computing (CCGrid), 2011 ,11th IEEE/ACM International Symposium on, 2011, pp. 13-22.

