International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

Cross User Bigdata Deduplication

Yash Karanje[1], Ankita Jadhav[2], Nikhita Biradar[3], Ketaki Kadam[4], Prof. M.P. Navale[5]

[1,2,3,4]Student & [5]Asst. Professor
Dept. of Computer Engineering, NBN Sinhgad School of Engineering, Pune, Maharashtra, India

---

Abstract - *Today's world has been digitalized to a large extent. The total amount of data generated per day is more than 2.5 exabytes out of which social media fuels up with maximum contribution along with business transactional data, sensor-generated data. Such a huge amount of data must be managed properly to use it for certain business domain-specific decision-taking purposes. It is very challenging to store and handle such large amounts of data which is mostly redundant in nature and that too present over multiple cloud platforms for multiple users; it requires high resources including the storage cost, backup time, processing time; which in turn decreases the system throughput. So, Data Deduplication is the most preferable solution we propose here for the above issue. We propose a model that will perform deduplication of data for multiple users to achieve the uniqueness of textual data (only) uploaded by multiple users; data access must be efficient though, maintain the privacy of data against brute-force attacks. This purpose will be achieved by employing certain algorithms like a Fixed-size blocking algorithm & Hashing algorithm and effective data organization. It will not only preserve the space by means of reducing storage allocation but also effectively manage network bandwidth.*

Key Words: Hash Key, Secure Hashing Algorithm, Brute-Force Attack, Inter-User Deduplication, Intra-User Deduplication, Fixed Size Blocking Algorithm, Cryptokeys, Clustering.

1. INTRODUCTION

A large amount of data is getting generated from different resources per unit time; its proper handling, processing & storage must be done effectively. Here, we're concerned with the data privacy, size of that data and redundancy is the main cause behind that. This results in improper knowledge discovery, inefficient decision making & data may get exposed to brute-force attacks; redundancy also leads to inefficient storage space utilization. To deal with the above issues, we use data deduplication.

We propose a model that will perform deduplication of data for multiple users to achieve the uniqueness of textual data (only) uploaded by multiple users; data access must be efficient though, maintain the privacy of data against brute-force attacks. This purpose will be achieved through a certain hashing algorithm and effective data organization. This application mainly focuses on efficient data access, preserve the privacy of textual data owned by multiple users by avoiding redundancy of textual data.

Initially, we have a text input file and this file supposed to go under the Blocking/Chunking algorithm to create blocks of data. Then the individual block will go through a Hashing Algorithm (SHA-256) to generate hash keys. For the sake of efficient storage cost utilization, clusters will be created to store those generated hash keys so that the search operation becomes comparatively more easy and optimal to perform. Further checking will take place in two phases namely Intra User & Inter User deduplication checking.

Consider an email system where a single attachment (text document) sized 5kb is supposed to be sent to 20 different users. If sent individually, we will have 21 copies in total resulting in 105kb of data. As long as the document is supposed to be in read-only format, it is always a good choice to make it a shared entity among all users so that we will have only one copy of it sized 5kb only available for all users. This will preserve space because we have prevented data redundancy by keeping only a single & unique copy of it.

2. LITERATURE SURVEY

Nowadays due to the exponential growth in the use of emerging technology such as cloud computing and big data, the rate of data growth is also increasing rapidly, so data deduplication technique [10] has been widely developed in cloud storage because it can significantly reduce storage costs by storing only a single copy of redundant data.

In the current system, [10] initially private keys are generated for all the users available and it also generates corresponding system parameters. The data uploading phase consists of four parts, those are tag generation, intra-deduplication, inter-deduplication and data encryption / key recovery. In tag generation, the User wants to upload data that chooses a random intra-tag along with the private key. In the intra-deduplication part has only checked the duplicate among the outsourced data from the same domain. In the inter-deduplication phase, checks the duplicate from the root node by comparing the data length.

In the end, data encryption is performed so that information can only accessed and decrypted by that specific user having the correct encryption key.

Hadoop Based Scalable Cluster Deduplication for Big Data [7] by Qing Liu, Yinjin Fu, Guiqiang Ni Used techniques like Fixed-size blocking algorithm, Map reduce and HDFS. In this paper, they have used the Mapreduce technique for parallel deduplication framework. The index table is distributed in each node which is stored in lightweight local MySQL databases.

Bucket Based Data Deduplication Technique for Big Data Storage System [6] by Naresh Kumar, Rahul Rawat, S. C. Jain used Fixed Size blocking Algorithm and Bucket based technique. The file is divided into blocks using the Fixed Size blocking Algorithm and Buckets are used to store the Hash value of blocks. The map-reduce technique applied to compare hashes stored in a bucket with an incoming hash of the block.

Application- Driven Metadata Aware Deduplication Archival Storage System [1] by Chuanyi LIU, Yingping Lu, unhui Shi, Guanlin Lu, David H.C.Du, Dong- ShenWANG used Variable size Chunking Metadata information. In this paper, they used metadata information of different levels in the I/O path such that more Meaningful data Chunks can be generated in the process of file partitioning to achieve inter-file level deduplication.

Extreme Binning: Scalable, Parallel Deduplication for Chunk-based File Backup [2] by Deepavali Bhagwat, Kave Eshghi, Darrell D. E. Long used File-level Borders theorem, File similarity techniques. Extreme binning uses file similarity. First, it chooses the minimum hash index value of a particular file as its characteristic fingerprint using the border's Theory. Then it transfers the files to the same deduplication server to deduplicate.

Authorized Data Deduplication Using Hybrid Cloud Technique [5] by Mane VIdya Maruti, Mininath K.Nighot used Content level Deduplication, file-level Deduplication. Duplication is checked in an authenticated way. Proof of ownership is needed to file duplication check. The user is needed to submit the file and proof of ownership of the file before sending the request for a duplicate check to the cloud.

[Literature Survey Summary]

3. METHODOLOGY

- In this proposed system, deduplication is used for efficient use of storage space and a better way to handle the duplicate data. There are 2 phases of deduplication checking; first is intra-user deduplication and second one inter-user deduplication.

- Multiple users upload text files in their respective local space. For each file uploaded, the blocking algorithm will be applied to divide the file into multiple blocks. Every block of data i.e. object will go under the hashing algorithm (SHA-256) and generate the hash key. The generated hash key will be unique for that block of data.

- If intra-user checking fails to locate the record, then only that particular block will be stored in that local space & hence inter-user checking will take place. Clusters will be maintained on both local & global side to store hash keys & the same will be used while deduplication checking.
- Clusters are created to store hashkeys and every cluster will maintain certain unique & independent characteristics shown by hashkeys. Those clusters will be the file storage systems stored in persistent storage. The test hash key will be given to every individual function representing those clusters and will be compared in that cluster only whose function it satisfies. If no match found then only that record will be added to the current cluster and hence the data. Functions representing every cluster must be unique and independent.

That's how two level deduplication checking will take place.

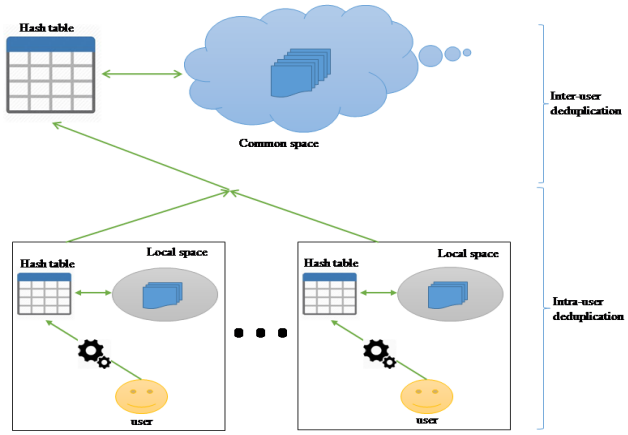

Fig – 1: Architecture diagram

Table – 1: Literature Survey Summary

Sr. No.	Title	Authors	Techniques Used	Methodology
1	Hadoop Based Scalable Cluster Deduplication for Big Data	Qing Liu, Yinjin Fu, Guiqiang Ni	Fixed size blocking Algorithm, Map reduce and HDFS.	They have used Mapreduce technique for parallel deduplication framework. Index table is distributed in each node which is stored in lightweight local MySQL databases.
2	Bucket Based Data Deduplication Technique for Big Data Storage System	Naresh Kumar, Rahul Rawat, S. C. Jain	Fixed Size blocking Algorithm, Bucket based technique	Buckets are used to store the Hash value of blocks. Map reduce technique applied to compare hashes stored in bucket with incoming hash of block.
3	Application- Driven Metadata Aware Deduplication Archival Storage System	Chuanyi LIU,Yingping Lu, unhui Shi, Guanlin Lu , David H.C. Du,Dong-ShenWANG.	Variable size Chunking metadata information	Used metadata information of different levels in the I/O path such that more meaningful data chunks can be generated in the process of file partitioning in order to achieve inter-file level deduplication.
4	Extreme Binning: Scalable, Parallel Deduplication for Chunk-based File Backup	Deepavali Bhagwat, Kave Eshghi,Darrell D. E. Long	File level Borders theorem, File similarity	Extreme binning uses file similarity. First, it chooses minimum hash index value of particular file as its characteristic fingerprint using border's Theory. Then it transfers the files to the same deduplication server to deduplicate.
5	Authorized Data Deduplication Using Hybrid Cloud Technique	Mane VIdya Maruti, Mininath K.Nighot	Content level Deduplication, file level Deduplication	Duplication is checked in authenticated way. Proof of ownership is needed to file duplication check. User is needed to submit the file and proof of ownership of the file before sending the request for duplicate check request to cloud.

4. CONCLUSION

Here, we've proposed a cross user bigdata deduplication system to maintain the uniqueness of textual data generated by multiple sources & owned by multiple users. Since the amount of data getting generated is tremendous; hence it becomes challenging to handle & process such huge data for knowledge discovery purposes. This system mainly focuses on redundancy issues; moreover, it also achieves efficient data organization & access, preserves privacy. For that, we've made use of the Fixed-size blocking algorithm and Hashing algorithm to generate crypto keys. This is how the paper summarizes existing technologies & proposed methodologies for bigdata deduplication.

REFERENCES

[1] C. Liu, Y. Lu, C. Shi, et al., "ADMAD: Application-driven metadata aware deduplication archival storage System", in Proc. 5th IEEE Int. Workshop Storage Netw. Archit. Parallel I/Os, 2008, pp. 29–35.7.

[2] Deepavali Bhagwat, Kave Eshghi, Darrell D. E. Long, "Extreme Binning: Scalable, Parallel Deduplication for Chunkbased File Backup", in Proc. IEEE Int. Symp. Modell. Anal. Simulation Comput. Telecommun. Syst., 2009, pp. 1–9.

[3] Hyungjune Shin, Dongyoung Koo†, Youngjoo Shin, and Junbeom Hur, "Privacy-preserving and Updatable Block-level Data Deduplication in Cloud Storage Services", IEEE 11th International Conference on Cloud Computing, 2018.

[4] M. Miao, J. Wang, H. Li, and X. Chen, "Secure multi-server-aided data deduplication in cloud computing," Pervasive and Mobile Computing, vol. 24, pp. 129–137, 2015

[5] Mane VIdya Maruti, Mininath K.Nighot, "Authorized Data Deduplication Using Hybrid Cloud Technique", 2015 International Conference on Energy Systems and Applications (ICESA 2015), 2015

[6] Naresh Kumar, R. Rawat, and S. C. Jain, "Bucket Based Data Deduplication Technique for Big Data Storage System",5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, 2016, pp. 267-271

[7] Q. Liu, Y. Fu, G. Ni, R. Hou, "Hadoop Based Scalable Cluster Deduplication for Big Data", IEEE 36th International Conference on Distributed Computing Systems Workshops, 2016.

[8] Supriya Milind More, Kailas Devadkar, "A Comparative Survey on Big Data Deduplication Techniques for Efficient Storage System", IJIACS, ISSN 2347-8616, Volume 7, Issue 3, 2018.

[9] Xingyu Zhang, Jian Zhang, "Data Deduplication Cluster Based on Similarity-Locality Approach", 2013.

[10] Xue Yang, Rongxing Lu, Ali A. Ghorbani, Jun Shao, Xiaohu Tang, "Achieving Efficient and Privacy-Preserving Multi-Domain Big Data Deduplication in Cloud", 8.2881147, IEEE Transactions on Services Computing, 2018.