ISSN No. 0976-5697

Volume 8, No. 4, May 2017 (Special Issue)

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Next Level Approach of Data Deduplication in the Era of Big Data

Shamsher Singh
Research Scholar, Department of Computer Science Engineering
Lovely Professional University, Phagwara, (Punjab)
iamshamsher1989@gmail.com

Ravinder Singh
Department of Computer Science Engineering
Lovely Professional University, Phagwara, (Punjab)
ravinder.17750@lpu.co.in

Abstract**:** Word DATABASE itself have a strong and deep meaning which indicates that high amount of raw facts and figures are stored at a place in electronic form. Data gives chances to its owner organization to compete in challenging world and stand aside from the crowd. But to store high amount of data it requires high amount of storage space. Almost 55% data stored in computer memory present in duplicated form which costs very high to organization to manage it. In this paper we proposed a new strategy to locate similar data stored on disks in big data environment. As the result duplicated data will be removed from the storage media and free up the space, increase the system performance in terms of operational speed, and reduce the time for deduplication process.

Keywords: Backup, Big Data, Data Deduplication, Data Node, Name Node

I. INTRODUCTION

In today's era of digital world storage becomes a very expensive need for the software companies as well as the home users. Every day millions of users create PB's (Peta Bytes) of data in the form of videos, photographs and other documents. Every user wants its data to be safe from every aspect that's why they store multiple copies of their data at different locations. For this purpose, online storage is provided by many vendors. Another reason is incremental backups that are used for security and consistency means. To overcome this duplication problem data deduplication an effective technique used to free up the storage space [2] [5].
In de-duplication cryptographic hash is used to locate and delete the redundant data from the backup taken. Hash value is a fixed length output of any data. In de-duplication when any data comes for storage , its hash signature is created using secure hash algorithm (SHA) [8].
That hash signature is verified by the server in hash index which has all the hash signatures stored. If that hash signature matches with any other hash signature it means that data is duplicated and need not to store again, then data will be deleted but a reference will be generated to the stored original data. If hash signature does not match with any other signature, then that data will be stored in the disk and new signature entry will be done in index.

Mr. Vikraman conducted a study on various data deduplication systems and process. This study described about the two ways to perform data deduplication i.e. inline deduplication and post process deduplication, also examined the places where data deduplication can be applied i.e. source or client side and target side [6].

Two ways to perform data deduplication on backup:

Inline Deduplication: In this data undergoes the deduplication process before storing in the storage disk. When data comes for the storage in the disk the deduplication algorithm is applied on it and only the unique data blocks are stored [6].

Post Process Deduplication: In this deduplication is performed on data after storing it into the storage disk. After storing data, data fetched for deduplication process and after that only unique data blocks stored back into the memory and redundant data blocks are deleted [6].

II. REASON BEHIND THE STUDY

Generation of high amount of data on daily basis has gave birth to a new concept named as Big Data. In recent years' concept of Big Data captured, the attention of almost every industry and government because in near future it is going to be a big issue in terms of data storage and data analysis. Many Scientists have given different definitions of big data but common statement which comes out is that "Big Data is a group of bulky data sets which is almost difficult to handle or process by using traditional resources" [3] [4] [9][10].

According to the Apache Hadoop Big Data is defined as the dataset which cannot be processed, managed and captured by ordinary computers. It couldn't be stored and managed with the standard database management programs because it consists of large amount of data, large variety of data including most of the unstructured and semi-structured data. Cloud Computing, Internet of Things, Hadoop, Data Centre are the main technologies which are concern with big data [11] [12] [13].

Big Data concept comes under consideration few years back when internet, smartphones, PDAs and sensors were used commonly by every person. Everyone is concern with the protection and safety of their data and want to be last long with it, so they used to take backup of data and store multiple copies of it so that it can be easily accessible. This thinking gives the birth to high data generation rate.

After this global revolution arrives in the form of social networking, it connects millions of people to each other no matter how geographically far they are from each other. People used to share their events in the form of photographs, videos, text stories etc. which comes in front in the form of Big Data. As an estimate almost 90 per cent of data of this era is generated in past 5-7 years and generated by the sensors used for weather statistics, posts on social networking sites, digital cameras, mobile phones, news posts etc. This much amount of data is almost non-processable by using traditional resources [9] [11].

In order to reduce duplication in the data, data deduplication process must apply on storage to free up space which reduces the cost of storage of data and increase the performance of the system.

Mr. Chen surveyed on big data and presented various technologies, for example, distributed cloud computing, Internet of Things, data centres', and Hadoop for taking care of big data and concentrated on the four levels of the value chain of big data, i.e., generation of data, acquisition of data, storage of data and analysis of data [1].

III PROCESS OF DATA DEDUPLICATION

Following figures shows the process of data deduplication over the data files.

a) Figure 1 shows the chunking process of the file and deduplication checking process of each chunk. After that unique data chunks are stored in the disk and duplicated chunks are discarded. Metadata of each chunk is also stored for recovery purpose.

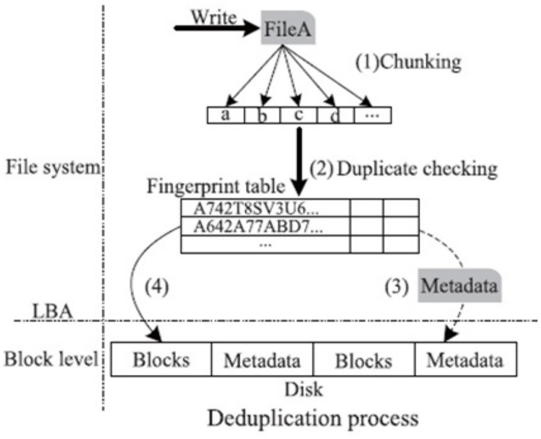

Figure 1. Chunking and Deduplication process

b) Figure 2 shows that after deduplication process if two files are having some duplicated data then only single copy of data is going to be stored in the disk and references to that data is made for accessing reasons by File B.

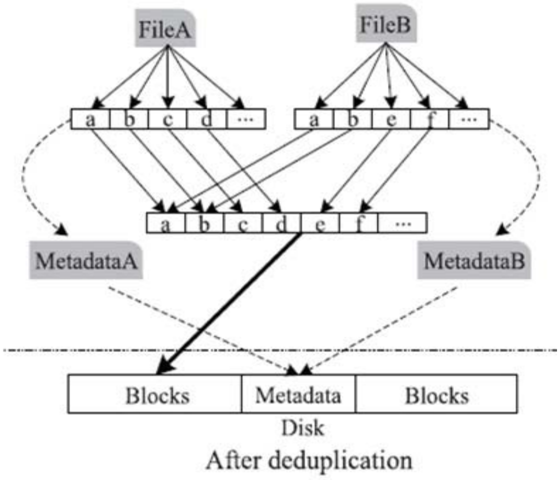

Figure 2. Storage of only one copy of data

c) Figure 3 shows the read operation of File A. When read request received then by using metadata file is being given to the client who made that request [7].

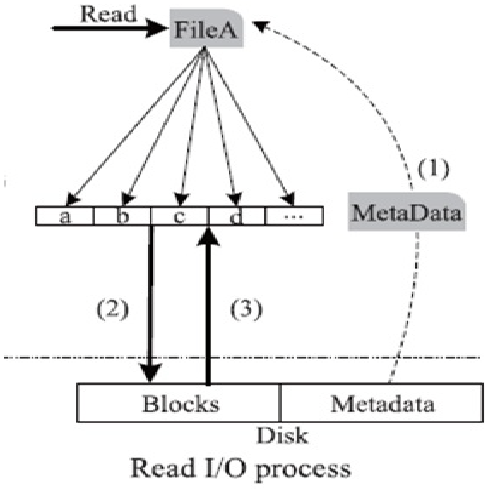

Figure 3. Read operation on files

IV) PROPOSED SYSTEM

In order to overcome the problem of duplicated data in storage media we propose following strategy to apply data deduplication process in Big Data environment. This system will work in following phases:

A) CONNECTION ESTABLISHMENT PHASE

1. Application server contacts to the backup server that it needs to store data in storage server.
2. Backup server sends message to Name Node/Primary Storage Node to prepare disks for storage of incoming data [4].
3. Application server prepare files to take backup and contact to primary storage node to get ready status. When gets ready status it sends data for storage as shown in figure 4.

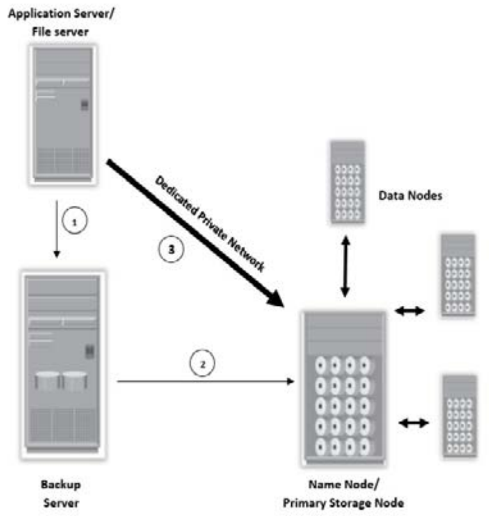

Figure 4. Connection Establishment between application server, backup server and primary name node

B) CHUNKING PHASE

Name divides files into fixed length chunks according to the number of data nodes participating in deduplication process.

$$No.of\ chunks = \frac{Total\ size\ of\ files}{No.of\ data\ nodes}$$

C) DATA TRANSFER AND DATA DEDUPLICATION PHASE

In this step, data deduplication and data transfer will be applied. A global hash index table will be shared by primary name node and secondary data nodes in order to apply deduplication process and to match hash signatures of file chunks. Global hash index table will be stored and managed by the primary data node.

1) FOR PRE-PROCESS DATA DEDUPLICATION AND DATA TRANSFER:

1. If file size is < 1GB, then data deduplication process applied by primary name node.
2. After deduplication name node send data to data nodes for storage along with updations in hash index table as shown in figure 5.

D) FOR POST-PROCESS DATA TRANSFER AND DATA DEDUPLICATION:

1. If file size is > 1GB, then name node applies chunking process on files and transfer chunks to the secondary data nodes.
2.
3. Secondary data node apply data deduplication process on data chunks and store unique copy of data into the memory. After this synchronization of hash index table takes place as shown in figure 5.

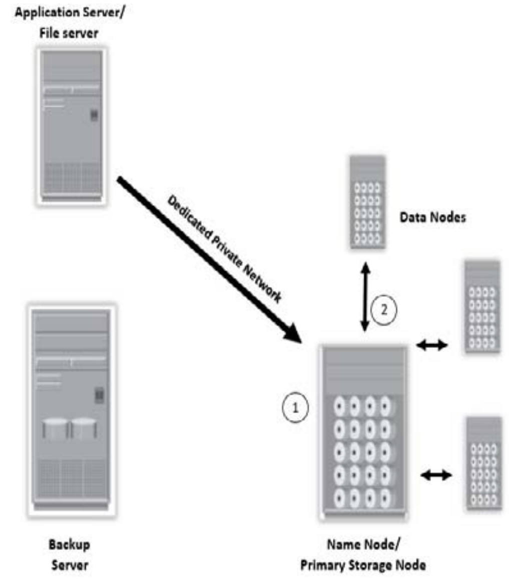

Figure 5. Process of Data Deduplication and Data Transfer

V) EXPECTED RESULT

The expected outcomes after implementation of this methodology could be summarized as follows:

- More storage space will be available to save data on disk in big data environment.
- Performance of deduplication will be increased in terms of speed and time.
- Unstructured data will be managed and deduplicated in efficient manner.

VI) CONCLUSION

Duplication storage of data may give security against data loss but it pushes high pressure on the storage systems. Storing multiple copies of same data at different places has concern with data security but in single storage instance presence of duplicated data makes no sense. As the solution to this issue data deduplication technique is perfect hammer. In big data environment a huge amount of data collected or received in order to process. Duplicated data in big data occupy massive storage space and processing power of system. As the result performing data deduplication on big data increase the performance of system in terms of deduplication ratio, speed of read/write operations will be increased due to less overhead, storage space utilization will be higher, data transmission speed will be increased and power consumption will be low. Also provide efficient usage of computing resources with cutting the storage cost of data.

REFERENCES

[1] M. Chen, S. Mao and Y. Liu, "Big Data: A Survey," in *Mobile Networks and Applications*, 2014.

[2] S. Shamsher, S. Ravinder, “A Viewpoint on different Data Deduplication Systems and allied issues” in *International*

Conference on Advanced Computing and Intelligent Engineering, 2016.

[3] M. K. Kakhani, S. Kakhani and S. Biradar, "Research Issues in Big Data Analytics," *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* vol. 2, no. 8, pp. 228-232, 2013.

[4] R. Zhou, M. Liu and T. Li, "Characterizing the efficiency of data deduplication for big data storage management," in *IEEE International Symposium on Workload Characterization (IISWC),* 2013.

[5] G. Zhu, X. Zhang, L. Wang, Y. Zhu and X. Dong, "An intelligent data de-duplication based backup system," in *15th International Conference on Network-Based Information Systems*, 2012.

[6] R. Vikraman and A. S, "A Study on Various Data De-duplication Systems," *International Journal of Computer Applications,* vol. 94, no. 4, pp. 35-40, 2014.

[7] Wang, J., Zhao, Z., Xu, Z., Zhang, H., Li, L., and Guo, Y. (2015). I-sieve: an inline high performance deduplication system used in cloud storage. *Tsinghua Science and Technology*, Vol. *20,* No. 1. pp. 17-27.

[8] G.-Z. Sun, Y. Dong, D.-W. Chen and J. Wei, "Data backup and recovery based on data de-duplication," in *International Conference on Artificial Intelligence and Computational Intelligence,* 2010.

[9] F. Kalota, "Applications of Big Data in Education," *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering,* vol. 9, no. 5, pp. 1602-1607, 2015.

[10] A. B. Munir, S. H. M. Yasin and F. Muhammad-Sukki, "Big Data: Big Challenges to Privacy and Data Protection," *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering,* vol. 9, no. 1, pp. 355-363, 2015.

[11] N. Tang, "Big Data Cleaning," in *Asia-Pacific Web Conference*, 2014.

[12] S. Vidhya, S. Sarumathi and N. Shanthi, "Comparative Analysis of Diverse Collection of Big Data Analytics Tools," *International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 8, no. 9, pp. 1646-1652, 2014.

[13] A. Verma, A. H. Mansuri and N. Jain, "Big data management processing with Hadoop MapReduce and spark technology: A comparison," in *Symposium on Colossal Data Analysis and Networking (CDAN)*, 2016.