International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Data Deduplication for Efficient Storage on Cloud using Fog Computing Paradigm

Shubham Sharma, Richa Jain, Pronika

Abstract*: Cloud services have taken the IT world by storm by making its services available to everyone over large geographic area. With the increasing amount of data generate every minute it has become increasing difficult to manage resources and the storage. Thus, data compression techniques like data de duplication that aims at executing the redundancy of data and forming chunks of data that can be stored on a distributed system can be proved to a logistic solution. But when it comes to cloud problems like security has always been a major issue. In order to eliminate these challenges, we need to implement a layer of fog computing they would deal with the shortcomings of cloud computing and at the same time present a filtration front before the incoming data.

***Keywords*: cloud computing, distributed-system, deduplication, fog computing.**

I. INTRODUCTION

Data deduplication has been a ranging trend in the technical market. It is a process that helps in filtering duplicated data and removing this duplicated data that might be encountered into the system. This technique has enabled the system to save storage and even the bandwidth of the network over which these data packets are transported as the repeated data is filtered and only one cop of that data is kept in the system or on the network [1]. Data de duplication is not a sacred technique but has been observed by the computer systems since long ago. When we try to copy the same file over similar storage space, we are given a choice of either replace the old file with the new one or else stop the copying command. This practice has been made possible by numerous algorithms that perform the functionality of data de duplication and thereby generate the result as discussed above. De duplication is not limited to system and file applications but has expanded its horizons over to the new technology that is taking over the IT sector that is cloud. [2]

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
Shubham Sharma*, Student, B. tech in CSE, Manav Rachna International Institute of Research & Studies, Faridabad.
Richa Jain, Student, B. tech in CSE, Manav Rachna International Institute of Research & Studies, Faridabad
Mrs. Pronika , Assistant Professor, CSE department of Manav Rachna International Institute of Research and Studies, Faridabad

The concept of data de duplication work with compression of data b ensuring that redundancy of data is eliminated, and that the data is stored only once over the cloud or file rather than being stored repeatedly. De duplication of data has allowed the systems to form a more cost effective and operational effective solutions to the problems of redundancy in the data. The user data that is being generated is very large and cannot be controlled in any form. The need of data de duplication has arose to deal with this large data and provide an efficient solution that will break the data into chunks and store then into distributed systems. When this technique is deployed to deal with data, the redundancy checks are made that allows the systems to store only one copy of the file. The retrieval of the old file is also possible when the redundant data is added back to its pace in the stored file. This retrieval is easier and takes on lesser resource involvement than the actual file would take given that it wouldn't have been de duplicated and stored 'n' number of times by the user every time any changes take place. When the file system is divided into chunks, it poses a threat to be accessed by a vicious third-party source and can be used to determine the owner of the master file. Although with advancement of technology, it has been made possible to store chunks of redundant data more secure environment. Data deduplication has its application in the distributed environment systems that involvecertain restrictions in order to protect the privacy of the data. The concept of data de duplication not only helps in helping in a storing the data in a redundantway but also helps in optimizing the bandwidth of the network [3]. The goal of data de duplication is to store a file into a compressed form and ensure that the resources needed to do so are minimal at their best. This also ensures that the file takes up the defined storage and remains prominent in its form.

II. PROBLEM

Cloud computing is a business model which is responsible for providing various services. On demand of the service requester and this service requester can belong to B2B (Business to Business) or B2C (Business to customer. Customer represent the individual end user which request for a cloud service for their personal purpose. More than 72 services are provided on cloud platform and among these services, Cloud storage is the one which is used by most of the people around the world. There are around 3.6 billion cloud consumers worldwide and to store their data we need a large amount of storage space in datacenter. If every user store same data multiple time, then the quantity of data increased and to store and manage unnecessary data we need more storage space in datacenter.

Data-deduplication helps in eradicating unnecessary data. Since data can be in various forms like image, text, video and audio.

File type	Total Size (GB)	Deduplicated Size (GB)	Savings%
Image	1.1	0.269	75.5%
Text Document	0.5	0.305	39%
Video	17.2	16.9	1.7%
Audio	1.1	0.42	61.8%
	19.9	17.89	10.01

Fig 1: Result analysis of various types of file data

III. REVIEW CRITERIA

There are multiple server are installed in a datacenter an and if data deduplication is not followed then multiple server may store the same file for example: There is a group picture of college classmates and there are around 30 friends in one image and all these 30 people store this picture on cloud to remember it as a memory . Since the picture is taken from a high quality camera therefore we can assume it size as 1GB but the same image is stored on cloud by 30 people therefore cloud require 1GB*30GB=30GB of storage to store this image when data deduplication is not used but when data deduplication is followed then it require only 1GB of space which saves 29GB of cloud storage for other work.It is one of the most simple example to explain the importance of data deduplication for the efficient cloud storage[4].

File Level Data Deduplication

In file level data deduplication if there are n user and these n user wants to store the same file x in the datacenter and instead of storing n*x files we store only one x file in datacenter [5].

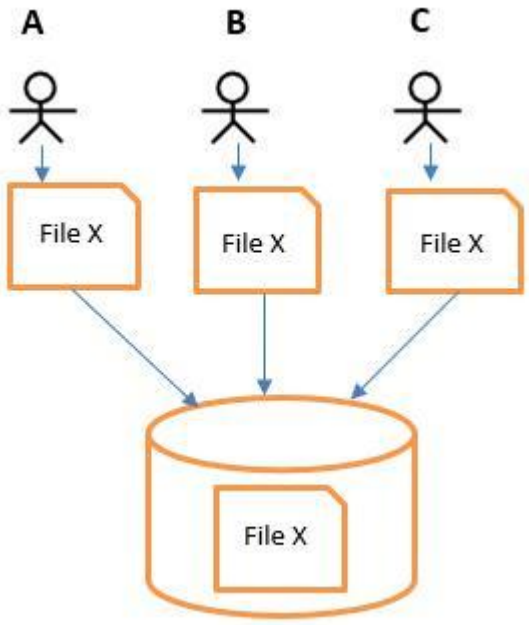

Fig 2: File level data deduplication

Block Level Data Deduplication

In block level data deduplication data is logically divided into blocks for example as shown in given fig there are three user A,B and C. Data file of every user is divided into logical blocks for example there are 12 logical blocks of 3 users and instead of these 12 blocks only 5 unique logical blocks are stored in file.

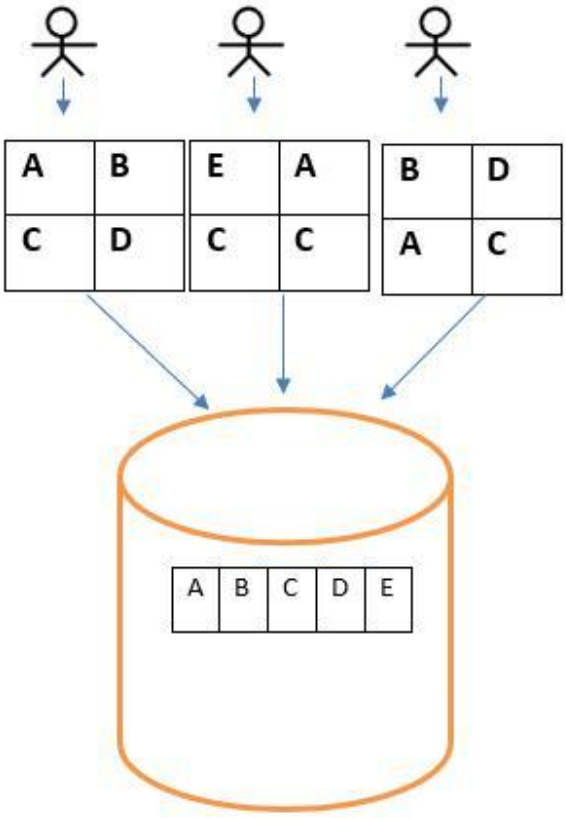

Fig 3: Block level data deduplication

Byte level De duplication:

When taking a bit stream into consideration, it has been observed that these bits contain a similar or dissimilar pattern for each data file. The data centers contain these repeating data that can be stored into databases in the form of patterns. This type of de duplication cares to study and remember the semantics of the data. Byte level de duplication focuses on defining and identifying the pattern in the incoming stream. Once the pattern is identified, it can be stored in the system it does not need to be saved again and again and thus reduce the redundancy. The data is always backed up on the disk ensuring fast recovery in times of need. Byte level de duplication does not hold challenges to be managed by the single system. This technique can be useful in many environments like file systems, virtualization and SAN and LAN [6].

Target Deduplication

Target de duplication is useful to remove redundancy when the data kept at backhand flows between the application and the backhand. This technology is supported by intelligent disk and virtual tab libraries.

Target de duplication reduces the amount the storage that is being used but cannot reduce the data that is to be sent over the network LAN or WAN when backup occurs [8].

IV. PROPOSED WORK

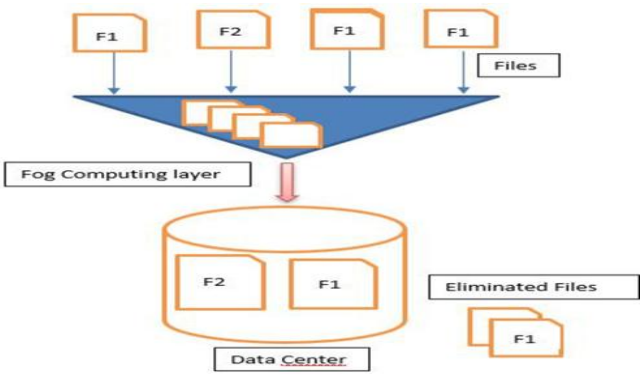

Fig 4: Proposed Model

The methodology we proposed in this paper is shown in Fig 5.

We add another layer which is known as fog computing [9] between data center and edge devices and all the data deduplication algorithms are implemented on this layer which act as a filter for the redundant data .All the redundant data is filter out by this layer and only an original data can pass out to data center which saves the computation power of datacenter and time which it can take to filter out the redundant data. As we represent in the Fig. There are 4 files namely (F1, F2, F1 and F1) and such that 3 of them carry a same name. Before storing data on cloud, it passes through a fog computing layer where all the data deduplication algorithm is implemented and this layer helps in filtering out the redundant data and out of three files of same type only one of them is stored and all the similar files are eradicating from this fog computing layer. This saves bandwidth of a network and storage space in cloud or heads unless they are unavoidable. Fog computing is a just another layer between cloud and end user which increase the capabilities of cloud [11].

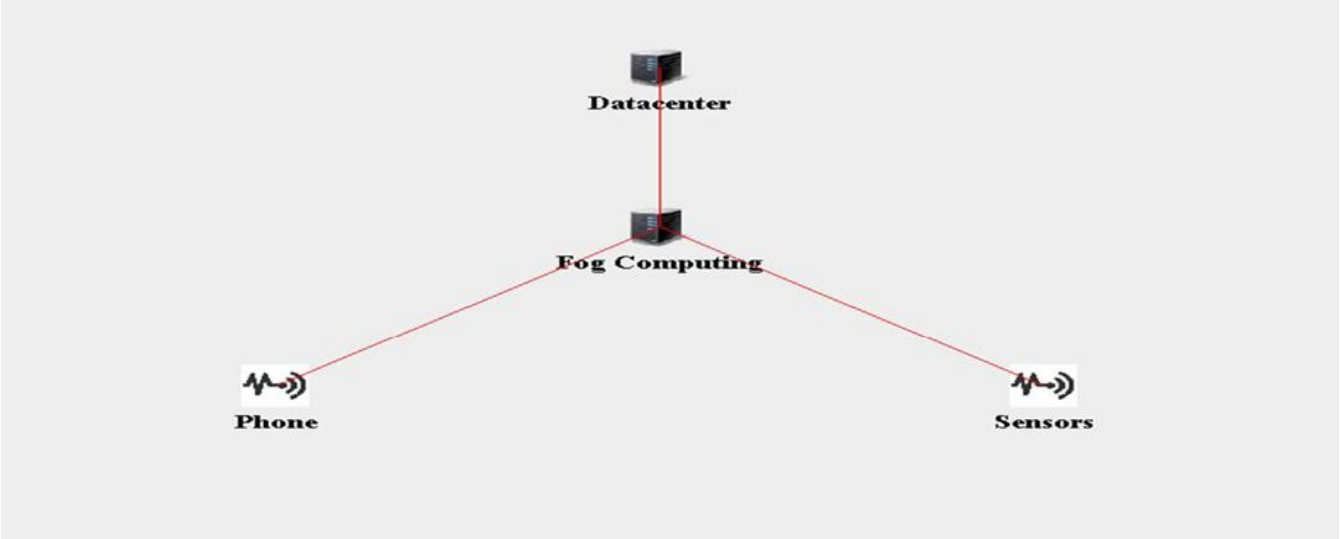

Fig 5: Fog computing simulation for data deduplication using iFog sim

V. CONCLUSION

The result shows that implementing a layer of fog computing can help in the better performance of the de duplication technique. This will enable the fog layer to filter the incoming data and reduce the workload of the de duplication algorithms by limiting them to check and approve the redundancy of the data forwarded by the fog layer. This also helps the system to make a smart decision about the data that will be sent to the data center and acquire the storage presented. Using fog computing along with de duplication techniques allow lesser usage of sacred resources and hence leads to optimized bandwidth. When dealing with cloud the major challenges that are faced includes: latency and security. The cloud faces a major problem with data stream delays when stored or retrieved. Not only this but one of the major concerns when it comes to cloud is security and privacy of the data. These issues can be mitigated using fog computing. This technique can be integrated over the edge devices and suppress the challenges presented by cloud computing

REFERENCES

1. Storer, M. W., Greenan, K., Long, D. D., & Miller, E. L. (2008, October). Secure data deduplication. In Proceedings of the 4th ACM international workshop on Storage security and survivability (pp. 1-10).
2. Prahlad, A., Muller, M. S., Kottomtharayil, R., Kavuri, S., Gokhale, P., & Vijayan, M. (2012). U.S. Patent No. 8,285,681. Washington, DC: U.S. Patent and Trademark Office.H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
3. Riteau, P., Morin, C., & Priol, T. (2011, August). Shrinker: Improving live migration of virtual clusters over wans with distributed data deduplication and content-based addressing. In *European Conference on Parallel Processing* (pp. 431-442). Springer, Berlin, Heidelberg.
4. He, Q., Li, Z., & Zhang, X. (2010, October). Data deduplication techniques. In *2010 International Conference on Future Information Technology and Management Engineering* (Vol. 1, pp. 430-433). IEEE.

5. Dutch, M. (2008, June). Understanding data deduplication ratios. In *SNIA Data Management Forum* (p. 7).
6. Venish, A., & Sankar, K. S. (2016). Study of chunking algorithm in data deduplication. In Proceedings of the International Conference on Soft Computing Systems (pp. 13-20). Springer, New Delhi.
7. Klose, M. F. (2013). U.S. Patent No. 8,412,677. Washington, DC: U.S. Patent and Trademark Office.J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.
8. Mandagere, N., Zhou, P., Smith, M. A., & Uttamchandani, S. (2008, December). Demystifying data deduplication. In Proceedings of the ACM/IFIP/USENIX Middleware'08 Conference Companion (pp. 12-17).Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].
9. Koo, D., & Hur, J. (2018). Privacy-preserving deduplication of encrypted data with dynamic ownership management in fog computing. Future Generation Computer Systems, 78, 739-752. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). *Title* (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
10. https://www.ijrte.org/wp-content/uploads/papers/v8i5/E6199018520.pdf
11. J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com

AUTHORS PROFILE

Shubham Sharma is pursuing B. tech in CSE with specialization in Cloud Computing with IBM at Manav Rachna International Institute of Research & Studies, Faridabad. He has participated in many workshops and conferences and published research papers in International Conferences. He has been researching in the area of cloud computing, machine learning, deep learning, and cybersecurity.

Richa Jain is pursuing B. Tech in Computer Science with specialization in cloud computing from Manav Rachna International Institute of Research and Studies, Faridabad. She is a pioneer and hopes to come up with the solutions of existing business challenges in society. She is placed with Tata Consultancy Services.

Mrs. Pronika is a 2007 B. Tech graduate from KUK University, India. She pursued her MTech from Banasthali Vidyalaya and was awardee degree in year 2009. She is working as an Assistant professor in the CSE department of Manav Rachna International Institute of Research and Studies with an experience of 11 years at hand and also subsequently doing her PhD. She has been researching in the area of cloud computing, database, computer network and security.