Performance-oriented Data Deduplication Scheme to Improve Primary Storage System in Cloud

[1]Sinchana M.K, [2]R.M Savithramma

[1]PG Student, [2]Assistant Professor
Department of CSE, SIT, Tumkur, India

***Abstract*: Cloud computing has rapidly turned out to be a standout amongst the most huge field because of its developmental administrations gave model of figuring in the IT business as well as in the product and equipment industry. Although cloud computing has many benefits, it also comes with many drawbacks and these drawbacks impact on the performance of the system. Here, we have considered the issue called duplicate data that is present in the cloud that affects performance considerations such as storage processing speed. In-order to resolve this performance issue, we have done a project based on data dedupliction in which we have used two fragmenting process named recursive method and commonality find to show the performance variation between them with respect to processing speed and also the security of the data has been increased.**

***Index Terms*: Primary storage, I/O performance, data deduplication.**

I. INTRODUCTION

Cloud computing is a booming technology used for the virtual computation. Cloud computing is computing resource for the storage, application and many other services. Cloud provides services namely: Infrastructure as a Service, Software as a Service and Platform as a Service. Cloud computing is a combination of software and hardware which provides services over a internet.

Cloud maintains all the services provided for the user. Cloud provides virtual cloud for the user to store the data. Virtual clouds are the imaginary cloud which stores huge amount of data to save the physical data cost. Virtual cloud provides users virtual storage, virtual memory, virtual application and virtual operating system. Which users can access across the cloud over the internet. Virtual memory is created over an existing machine is called as virtual machine.

Data deduplication has been used as an effectual method in the cloud backup in-order to improve the storage space. In recent times, research shows that average to high data redundancy evidently present in the Virtual Machine and Storage Systems. By applying the data deduplication technique, we can save space savings up to 90 percent in the VM and 70 percent in the storage systems.

Present data deduplication method for the primary storage such as iDedupe, Offline Dedupe are capacity oriented and focus on the storage capacity reserves and these existing data deduplication scheme fail to address one of the most significant problem in the primary storage, that is performance.

Therefore, to solve the above deduplication–induced problems and performance issues of the primary storage in the cloud, a Performance-Oriented data Deduplication scheme called POD is used rather than Capacity Oriented data Deduplication to progress the performance of the primary storage in the cloud. In this project, in-order to provide optimal performance for primary storage, we have created two fragmenting algorithm in data deduplication named as recursive comparison and commonality find which are the two chunking method used to check which process increases the performance in terms of processing speed storage and comparison has been made in between them.

II. LITERATURE SURVEY

T. Clements, I. Ahmad, M. Vilayannur, and J. Li [1] proposed a decentralized deduplication in SAN cluster file system, where file system hosting virtual machine contains many duplication blocks of data which causes wastage of space in storage. This reduces the performance of the primary storage system of input/output requests. To avoid this duplication of data, author proposed decentralized deduplication in storage area network cluster file system. The decentralized cluster file system is applied to virtual machine to avoid data deduplication. The author proposed DEDE block level deduplication system where DEDE manipulated the file cluster system data.

K. Srinivasan, T. Bisson, G. Goodson, and K. Voruganti [2] proposed an algorithm which depends on two key in-sight obtained from real world workloads, which are spatial locality and temporal locality. By utilizing Spatial Locality, they specifically deduplicate just sequence of disk blocks, which decreases fragmentation and lessen the search occurred by deduplication. While Temporal Locality enable us to replace the costly, on-disk deduplication meta data by a lesser, in-memory cache. This technique facilitates them to exchange the capacity savings for the performance.

A. El-Shimi, R. Kalach, A. Kumar, A. Oltean, J. Li and S. Sengupta [3] have shown how to optimize deduplication of the primary file based server data in-order to get both high deduplication savings and minimum resource consumption by using novel fragmenting algorithm, fragment compression, splitting and a low RAM footprint fragment index. They designed an architecture to achieve high deduplication savings with less computational overhead. In the paper, they concentrated on parts of the framework

which address scaling deduplication processing resource usage with information size to such an extent that memory, CPU, and disk resource stay available to satisfy the primary workload of serving IO.

A. Meister, A. Brinkmann, J.Kaiser, T. Cortes, M. Kuhn and J. Kunkel [4] have proposed a gigantic potential for data deduplication in HPC storage systems that isn't encouraged by the present HPC file systems. They showed the main research on the capability of data deduplication in HPC centers, which have a place with the most requesting storage producers. By using data deduplicaton process we can decrease the online information that gets saved and this decrease can simply be accomplished by a subfile deduplication approach, while methodologies dependent on entire file comparision just escort to little capacity savings. If data is redundant, it could be evacuated by deduplication methods.

III. METHODOLOGY

The data deduplication is the strategy used to save the storage in the virtual machine disk image storage or any other storage area. Here, we are proposing the POD architecture which is the main design of the experiment.

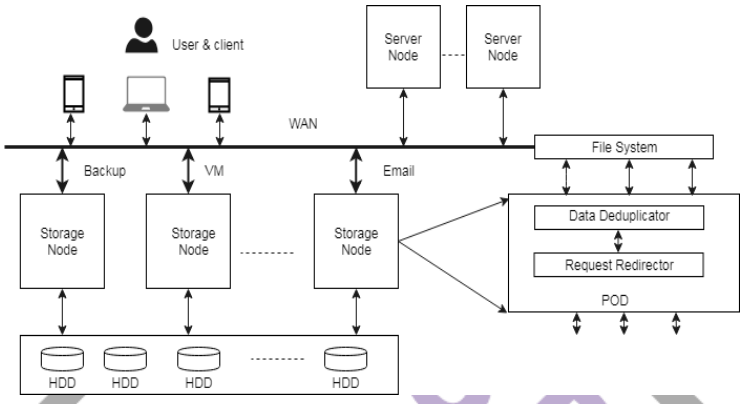

Figure 1: POD system architecture

Figure.1 demonstrates the system architecture outline of our proposed performance oriented deduplication method in perspective of the system I/O in the cloud. In above figure, POD exist in the storage node and communicate with the file system by means of standard read/write interface. The POD is integrated into any HDD based primary storage system to speed up the system performance. POD is autonomous of the higher file-system, which makes POD more adaptable and convenient than entire file deduplication and it can be set up in a variety of environments.

POD has two major components: Data De-duplicator and Request Redirector. The Data Deduplicator module is culpable for comparing upcoming data, checking whether the upcoming data is redundant or unique. depending on this information, the Request Redirector module determine whether the write request should be deduplicated, and maintains data reliability to avert the referenced data from being overwritten and updated.

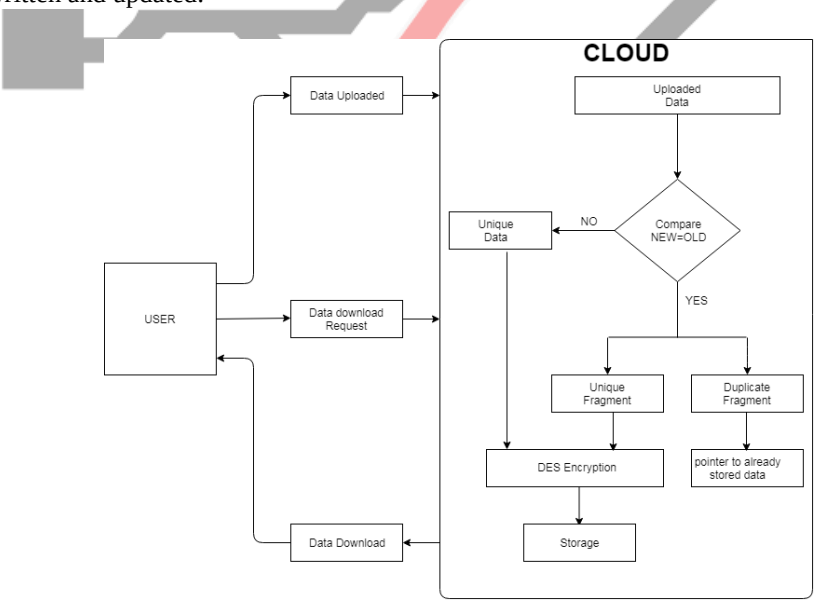

Figure 2: Proposed system

Here, when a user uploads a file, the uploaded data is compared with the existing data in the cloud. If the data is unique, then DES algorithm is applied to it and it will be stored in the storage. If the data contains duplicate data, then this duplicate data is discarded and its value is stored in the pointer of already stored.

During the decryption process, when a user requests for a particular file, then that data will be decrypted and fetched from the cloud storage.

IV. EXPERIMENTAL ANALYSIS

The experimental steps are as follows

- To not allow everyone to access the data we are proving register and login for new user and already existed one.
- After login is successful, they will get popup message logged in successfully.
- Next, data owner uploads a file by clicking upload button and selecting a file that he needs to upload.
- The uploaded data will go to two fragmenting process simultaneously, that is recursive method and commonality find and gets compared and fragmented. Next, we get a message which shows that data has been uploaded. The file uploaded is encrypted by using DES encryption method and it is stored in virtual storage of the user.
- Data owner can download the file by selecting his desired file from the list and by selecting the download button.
- The performance of recursive method and commonality find method has been noted for different sized data that have been uploaded and a graph has been plotted to represent it which is shown in figure 3.

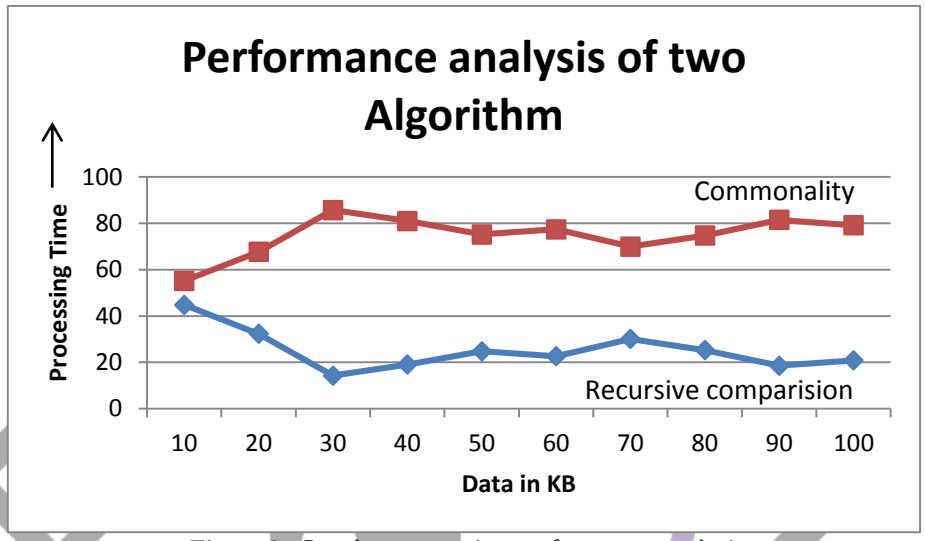

Figure 3: Graph representing performance analysis

V. CONCLUSION

Proposed framework provides a performance-oriented deduplication scheme to increase the performance of the primary storage systems in the cloud. Here the comparison between two fragmenting process that is recursive comparison and commonality find shows that how different method of fragmenting causes performance variation with respect to processing speed and storage of the primary storage.

We have also compared processing speed of two fragmenting process to represent which process works better. For security purpose, we have used DES algorithm. Existing plan could be upgraded to a variety of other objectives in cloud computing in terms of performance.

REFERENCES

[1] A. T. Clements, I. Ahmad, M. Vilayannur, and J. Li, "Decentralized deduplication in SAN cluster file systems," in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2009, pp. 101–114.

[2] K. Srinivasan, T. Bisson, G. Goodson, and K. Voruganti, "iDedup: Latency-aware, inline data deduplication for primary storage," in Proc. 10th USENIX Conf. File Storage Technol., Feb. 2012, pp. 299–312.

[3] A. El-Shimi, R. Kalach, A. Kumar, A. Oltean, J. Li, and S. Sengupta, "Primary data deduplication-large scale study and system design," in Proc. USENIX Conf. Annu. Tech. Conf., Jun. 2012, pp. 285–296.

[4] D. Meister, J. Kaiser, A. Brinkmann, T. Cortes, M. Kuhn, and J. Kunkel, "A study on data deduplication in HPC storage systems," in Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal., Nov. 2012, pp. 1–11.

[5] R. Koller and R.Rangaswami, "I/O Deduplication:Utilizing content similarity to improve I/O performance" in 8th UsenixConference on File and Storage Technology.,Feb 2014.

[6] Michal Kaczmarczyk, Marcin Barczynski, Wojciech Kilian, and Cezary Dubnicki, "Reducing impact of data fragmentation caused by in-line deduplication", Proceedings of the 5th Annual International Systems and Storage Conference., June 2015.

[7] Bo Mao, Member, Hong Jiang, Suzhen Wu and Lei Tian," Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud",IEEE Transactions On Computers, VOL. 65, NO. 6, JUNE 2016.