International Journal of Advanced Technology in Engineering and Science
Vol. No.7, Issue No. 05, May 2019
www.ijates.com
ijates
ISSN 2348 - 7550

A survey on: Application-Aware Big Data Deduplication in Cloud Environment

Prof.Vani B[1]Associate Professor,Dept of CSE,

Santosh Lama[2],**Rajan Boudel**[3],**Hridesh Chaudhary**[4],**Sanjeev Kumar Raut**[5],

Dept of CSE,Sambhram Institute of Technology, Bangalore-560097

Abstract-*Deduplication process is currently the widely used process in the cloud storage to improve the IT resources and the efficiency.There are many techniques designed to solve the problem of deduplication but many of those techniques has failed to struck down the problem of efficient deduplication process. They have the problem of low duplicate elimination ratio and are less scalable. To eliminate this problem, we introduce the concept of AppDedupe. It is an Application-aware scalable inline distributed deduplication framework in cloud environment. We can meet this challenge by exploiting application awareness, data similarity and locality to optimize distributed deduplication with inter-node two-tiered data routing and intra-node application-aware deduplication.In application aware deduplication, the file is broken down into a small chunks. Each small chunks of a file is assigned with a hand print and the hand print is stored in the lookup table. The handprint is assigned to speed up the process of deduplication with the high efficiency. The AppDedupe process has the highest deduplication efficiency with the high deduplication effectiveness in comparison to other traditional deduplication process.*

KeyTerms-Bigdata deduplication, application awareness, data routing, hand printing, similarity index

1. Introduction

The ongoing technological development in the field of computing environment has led to flash flood of data from the various domains over the past few years. The data centres are flooded with the thousands of bytes of data and the entanglement of data management led us to the big data. According to the IDC, the 75% of our data in the world of computerized storage is duplicate data. The common example of bulge of duplicate data is Good morning message in WhatsApp. The same data is forwarded to same people for many numbers of time and the data will be stored in the user's system multiple time which leads to the data duplication.In big data there are many data which can be simply copy of the same data. The copy of same data will lead to the problem of storage capacity and the data retrieval. The deduplication technology helps us to overcome this problem by eliminating the duplicate data and decreases the Administrationtime, operational complexity, power consumption and human error.

In the deduplication process the large data is divided into small parts which is also known as chunks.The each chunk of data is assigned with the finger prints. The finger print will be stored in the lookup

table. When the duplicate chunks is intercepted by the application, it will eliminate the duplicate data with the finger prints and only the unique data will be stored in the storage to increase the communication and storage efficiency. The source inline deduplication iseffective because it eliminates the duplicate data before the data reaches the storage system which reduces the requirement of physical storage capacity. The deduplication in wide area network is more tougher than the deduplication within the same system or small network of system.The both inter node and intra node deduplication framework in large scale faces the problem of communication overhead in inter node and chunk index lookup disk bottleneck in intra-node. Because of this the parallel deduplication in high rate is not possible. So to eliminate this problem, The AppDedupe is introduced which isascalable source inline distributed deduplication framework by leveraging application awareness.It is a middleware application and they are deployed by the big data centre. The elimination of duplicate data takes place before it reaches the storage systemonly the link of the data will be stored in the storage system. It maintains the high data deduplication efficiency and performs deduplication in each node independently and in parallel.

2. Related work

Initially, "The View of Cloud Computing" has been proposed where the process and the mechanism which are used in the process of cloud computing are explained. "The Experiencing Data De-Duplication" has been carried out where the improvement of the efficiency and reducing capacity requirements had been explained. The next work related to the data de-duplication is "An Application-Aware Framework for Video De-Duplication", where the video to be stored in the cloud environment are prevented to be reoccurring in the cloud. "Multi-level Comparison of Data De-duplication in a Backup Scenario" is the next work that had been carried out. This deals with the comparison of the data in the different levels for the appropriate backup".

Extreme Binning: Scalable, Parallel De-duplication for the chunk Based File Backup" is the mechanism that had been carried out for the data de-duplication process. This helps to check the redundancy of the data by dividing them into the chunk before storing it in the cloud. The next work "A Framework for Analysing and Improving Content Based Chunking Algorithm" had been proposed which helps to analyse and improve the storing of unique data in the virtual environment by an algorithm.

The next work related to our project is "Avoiding the Disk Bottleneck in Data Domain De-duplication File System", which mainly deals with the parallel processing while the data is to be stored in the cloud. Due to the bottleneck, there occurs a data traffic at the last moment of the storage.Hence to avoid such bottleneck the task was proposed for the File System.The next work that had been related to our project is "Fast and Secure Laptop Backups with Encrypted De-Duplication". Generally, this task deals with the efficient and appropriate data de-duplication from the laptop with proper

speed and security. The data that are to be stored it in the cloud are encrypted which provide the secure Data De-Duplication. Similarly, there is also "Primary Data De-Duplication,Large Scale Study and System Design" which provides the primary steps that are required for the de-duplication process in the field of large-scale study. For the largescale study and system design, there is a vital role of data de-duplication to provide efficient storage by eradicating the redundant data, hence primary data de-duplication had been carried out.The next work is "Content-Aware Load Balancing for Distributed Backup" which helps to analyse the content of the data and distribute them accordingly so that the load to be stored in the cloud are balanced.

The next task which is related to our project is "AA-Dedupe: An Application-Aware Source De-duplication Approach for Cloud Backup Services in the Personal computing Environment". As we know that, in the present era, the implementation of personal computing Environment has been widely deployed. Hence there must be the provision of the De-duplication Approach in the data while storing it in the cloud for the backup purpose.

2.1 Existing System

Researchers have proposed many dynamic PoS schemes in single-user environments.In the single user environment, deduplication process will take place within the user's own data but not the with the other user's data.

In multi-user cloud storage system needs the secure client-side cross-user deduplication technique, which allows a user to skip the uploading process and obtain the ownership of the files immediately, when other owners of the same files have uploaded them to the cloud server.

2.2 Proposed system

In the pre-process phase, users intend to upload their local files. Files will be divided in to the blocks and for each and every block,hash code will be generated, and blocks will be sent to the deduplication phase.

In the deduplication phase, it will verify all the blocks of the file whether the block is already uploaded to the cloud or not based on the hash code generated to each of the block. If already the block is uploaded then the user will get the ownership of the block, new blocks will be sent to the upload phase.

3. AppDedupe Design

The basic things to understand about the word AppDedupe is the mechanism where an application is created which is used to prevent the data repeatability while storing it in the virtual environment.Basically there are three components like Capacity, Throughput and Scalability on which the AppDedupe has been configured. The Capacity deals with the passing of the similar data through the same deduplication node hence provide the adequate result in the prevention of the duplicate data to be stored multiple times.

The next is throughput, which defines the efficiency of the process of deduplication. And the last component is the Scalability, which basically demonstrate that the app that is used for the deduplication

should be scaled inorder to handle the huge amount of data to as per the requirement.

Our main moto is to achieve the high deduplication throughput with the excellent capacity and scalability. These can be achieved by the design of the inline distributed deduplication framework which means that the deduplication should be done for each and every small chunk of the data that are to be stored in the cloud.

4. Methodology

As to meet the challenge faced in de-duplication we implemented the two-tiered data routing scheme to obtain scalable performance with high ratio of deduplication efficiency including file-level application aware routing decision in director components for managing files information and keep track of files and other one is super-chunk level similarityaware data routing in client components to measure whether the chunk is duplicate or not before sending the data chunk and only the unique data chunks will transferred over the interconnection network.

Based on these two-tiered data routing two algorithms were implemented and exerted one is Application Aware Routing Algorithm which will implement in application aware routing decision module of director and other one is Hand-printing Based Stateful Data Routing which will improve load balance for dedupe storage node by avoiding or not altering the already stored data when we try to add or delete the node in storage cluster.

Algorithm 1. Application Aware Routing Algorithm

Input: the full name of a file, full name, and a list of all dedupe storage nodes {S1, S2, …, SN} **Output**: a ID list of application storage node, ID_list = {A1, A2 … ,Am}

1. Extract the filename extension as the application type from the file full name fullname, sent from client side;

2. Query the application route table in director, and find the dedupe storage node Ai that have stored the same type of application data; We get the corresponding application storage nodes ID_list = {A1, A2, … , Am}⊆{S1, S2, …, SN};

3. Check the node list: if ID_list= ∅ or all nodes in ID_list are overloaded, then add the dedupe storage node SL with lightest workload into the list ID_list={SL};

4. Return the result ID_list to the client.

Algorithm 2. Handprinting Based Stateful Data Routing

Input: a chunk fingerprint list of super-chunk S in a file, {fp1, fp2, … ,fpc}, and the corresponding application storage node ID list of the file, ID_list={A1, A2, … , Am}

Output: a target node ID, i

1. Select the k smallest chunk fingerprints {rfp1, rfp2, …, rfpk} as a handprint for the super-chunk S by sorting the chunk fingerprint list {fp1, fp2, …,fpc}, and sent the handprint to k candidate nodes with IDs mapped by consistent hashing in the m corresponding application storage nodes;

2. Obtain the count of the existing representative fingerprints of the super-chunk S in the k candidate nodes by comparing the representative fingerprints of the previously stored super-chunks in the application aware similarity index, are denoted as {r1, r2, …, rk};

3. Calculate the relative storage usage, which is a node storage usage value divided by the average storage usage value, to balance the capacity load in the k candidate nodes, are denoted as {w1, w2, …,wk};

4. Choose the dedupe storage node with ID i that satisfies ri/wi = max{r1/w1, r2/w2, …, rk/wk} as the target node.

5. Conclusion

In this paper, we describe App-Dedupe, an application-aware scalable inline distributed deduplication frame-work for big data management, which achieves a trade-off between scalable performance and distributed deduplication effectiveness by exploiting application awareness, data similarity and locality.

It implements a two-tiered data routing scheme to route data at the super-chunk granularity to reduce cross-node data redundancy providing good load balance and low communication overhead, and employs application-aware similarity index based optimization to improve deduplication efficiency in each node with very low RAM usage.

The main advantage of AppDedupe design framework is in the state-of-the-art distributed deduplication for large clusters of data by outperforming the stateful tight coupling scheme in cluster wide deduplication and improves stateless loose coupling schemes with high scalability and low overhead communication.

References

[1] K. Srinivasan, T. Bisson, G. Goodson, and K. Voruganti. "iDedup: Latency-aware, inline data deduplication for primary storage," Proc. of the 10th USENIX Conference on File and Storage Technologies (FAST'12). Feb. 2012.

[2] D. Bhagwat, K. Eshghi, D.D. Long, M. Lillibridge, "Extreme Binning: Scalable, Parallel Deduplication for Chunk-based File Backup," Proc. of the 17th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS'09), pp.1-9, Sep. 2009.

[3] W. Dong, F. Douglis, K. Li, H. Patterson, S. Reddy, P. Shilane, "Tradeoffs in Scalable Data Routing for Deduplication Clusters," Proc. of the 9th USENIX Conf. on File and Storage Technologies (FAST'11), pp. 15-29, Feb. 2011.

[4] T. Yang, H. Jiang, D. Feng, Z. Niu, K. Zhou, Y. Wan, "DEBAR: a Scalable High-Performance Deduplication Storage System for Backup and Archiving," Proc. of the 24th IEEE International Parallel and Distributed Processing Symposium (IPDPS'10), pp. 1-12, Apr. 2010.

[5] Y. Fu, H. Jiang, N. Xiao, "A Scalable Inline Cluster Deduplication Framework for Big Data Protection," Proc. of the 13th ACM/IFIP/ USENIX Conf. on Middleware (Middleware'12), pp. 354-373, Dec. 2012.

[6] M. Lillibridge, K. Eshghi, D. Bhagwat, "Improving Restore Speed for Backup Systems that Use Inline Chunk-Based Deduplication," Proc. of the 11th USNIX Conf. on File and Storage Technologies (FAST'13), Feb. 2013.

[7] Min Fu, Dan Feng, Yu Hua, Xubin He, Zuoning Chen, Wen Xia, Yucheng Zhang, Yujuan Tan. "Design Tradeoffs for Data Deduplication Performance in Backup Workloads," Proc. of the 13th USNIX Conf. on File and Storage Technologies (FAST'13), pp. 331-344, Feb. 2015.

[8] W. Xia, H. Jiang, D. Feng, Y. Hua, "Silo: a Similarity-locality based Near-exact Deduplication Scheme with Low RAM Overhead and High Throughput," Proc. of 2011 USENIX Annual Technical Conference (ATC'11), pp. 285-298, Jun. 2011

[9] Y. Fu, H. Jiang, N. Xiao, L. Tian, F. Liu, "AA-Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment," Proc. of the 13th IEEE Conf. on Cluster Computing (Cluster'11), pp. 112-120, Sep. 2011.

[10] B. Zhu, K. Li, H. Patterson, "Avoiding the Disk Bottleneck in the Data Domain Deduplication File System," Proc. of the 6th USENIX Conf. on File and Storage Technologies (FAST'08), pp. 269-282, Feb. 2008.