A NOVEL APPROACH TO SECURE AND ENCRYPT DATA DEDUPLICATION IN BIG DATA

Dr. C. Murugamani*
Professor, Information Technology, Bhoji Reddy Engineering College for Women, Hyderabad.
Dr. C. Berin Jones
Professor, Computer Science & Engineering, Bhoji Reddy Engineering College for Women, Hyderabad.
E-mail: cberinjones@gmail.com
*Corresponding author E-mail:murugananija@gmail.com.

Abstract
Recently, there is an increasing demand for storing large amount of data in digital form has become quite challenging task In big data storage, there will be a large amount of duplicate data are presented in the data base. Existing techniques do not improve the performance and efficiency of the system. In this paper the bucket based deduplication technique is introduced, where the big data stream is divided to create fixed size chunks using chucking algorithm. Then the generated chunks are given to the enhanced MD5 algorithm to form hash values for thongs chunks. In order to detect the duplicate hash values in the data base Map Reduce is used and the hash values are compared with already stored hash values in the bucket. The Experimental results conclude that the proposed technique outperforms than any other existing techniques and improve the efficiency the system by analyzing the real dataset using Hadoop tool.

Keywords: Big data, Deduplication, Chunking, Hadoop, Map Reduce.

1. Introduction

Big Data is a massive volume of both structured and unstructured data, this data is large so it is critical to process the database and software methods. Big Data is used everywhere in the world (i.e) online and offline.

Big data determines a very large volume of data which grows exponentially and makes sure that the availability of data in both structured and unstructured format. Big data is a wide term for data sets which is very large or complex that traditional data processing applications are inadequate. It has several challenges like analysis, capture, data curation, search, sharing, data storage, data transfer, data visualization and information privacy. It uses the predictive analytics or other certain advanced method to extract values from data to the particular size of data set. The accuracy in big data leads to confident decision making, which means great operational efficiency, cost reductions and reduced risk.

Big data has many uses related to industries. It has been used to access risk in insurance companies and to track replies to the products. It has also been used in monitoring the various things like wave movements, flight data, traffic data, financial transaction, health and crime.

Data deduplication is particularly used for compressing the repeated data and for eliminating the repeated data, mostly to develop the storage area. In the deduplication process, duplicate data is eliminated, leaving only one copy of the data is stored. All data are maintained and are used for further processing. Deduplication reduced the required storage capacity. So only different data is stored. Data Deduplication is of two types. They are,

1. File-level Deduplication
2. Block-level Deduplication

1. File-level Deduplication: File level deduplication is described by the multiple copies of the same file, and it stores the first copy and also links other references to the first file. Only one copy

stored on the disk. Finally the space is saved on the disk related to number of copies of the file.

2. Block-Level Deduplication: Block-level deduplication is also called as Variable block-level deduplication. Here data blocks are taken itself, if another copy of this block already exists.

The final copies are not stored on the disk, but a pointer has found to point the original copy.

In this technique, it is used to improve storage space and can be applied to network data, transfers to decrease the number of bytes sent. In this process, different chunks of data are detected and are stored in repository during the process of analysis. Analysis continues on other chunks are compared to stored copy data and whenever the same data occurs, the repeated chuck is replaced with the small repository chuck.

The same frequency is dependent on the chunk size, then the amount of data transferred can also be reduced.

The existing data deduplication methods are primary storages, such as iDedup and offline Dedup. On storage capacity saves and selects the large request for deduplication and sent the small requests (i.e , 4KB,8KB or less). The small Input/ Output requests only contain atomic fragment of the storage capacity required in making the deduplication. The existing workload studies displaying the small files are controlled in primary storage systems (more than 50%) and the roots the system performance. Data deduplication replaces same regions of data with reference to data already saved in the disk. By comparing with the compression techniques, the data deduplication can eliminate not only data repetitions within a single file, but also the data redundancy among multiple files. In order to find the redundancy in data blocks, deduplication has compared the contents with a large amount of data. Data deduplication has a negative impact on the performance of data servers which is both computation and I/O intensive. To avoid this constraint, an approach is implemented simultaneously by distributing the computational and I/O tasks to individual nodes in the storage. With the help of utilizing the computation capability and storage capacity of multiple nodes in cloud solved the bottleneck and improves the throughput of data storage.

Two Approaches are used to perform the data deduplication, such as Inline Deduplication and Post Process Deduplication .In inline deduplication the data receives the deduplication performances before storing the storage disk. If the data comes from storage area, then the deduplication algorithm is applied and the independent data blocks are stored. In Post Process Deduplication is implemented on data after storing into the storage disk, then the data extract for deduplication process. After that, independent data blocks are stored into the memory and delete the repeated data blocks. In section I introduction has been described , in section II related works and several literature papers were disused, in section III network design has been explained in detail, in section IV results and discussion has been analyzed in detail, in section V concludes the paper.

2. Related Work

Kumar et al. [1] proposed a novel data deduplication technique based on bucket storage. This proposed technique uses the fixed size chunking algorithm to divide the big data into fixed size chunks. And then MD5 algorithm is applied to generate hash values for chunks. Finally Map reduce is used to check whether the chunk is duplicate or not and it compares the values with the hash values already stores in the buckets to detect the duplication in the chunks. HDFS is a distributed file system stores the chunks which is identified as real.

Liu at al. [2] studied tremendous challenges on the storage because of the exponential growth of data.

They presented a scalable and reliable cluster data deduplication system Halodedu, which uses MapReduce and HDFS to manage data storage and to process data deduplication simultaneously. HBase was deployed to maintain the availability and reliability of metadata and also to store the backup files.

The comparative results displayed that it improves the speed and scalability of the system.

Luo et al. [3] investigated the challenges of saving memory space and the necessary abilities to move big data within desired time frame. Here they presented a cloud storage system called Boafft to achieve scalable throughput and capacity with help of multiple data servers. This

considerably reduces the loss of deduplication data. Initially the Boafft make use of an efficient routing algorithm to detect the storage location, which eventually minimizes the network overhead. Then it establishes an in-memory similarity index to remove the large number of random disk reads and writes thereby increasing the local data deduplication. The comparative results concludes that it improves the data deduplication ratio.

Mao et al. [4] explained the growth of data in the cloud and also showed how it causes space contention in memory and data fragmentation on disks. With the help of these observations they proposed method called POD, a performance-oriented I/O deduplication which improves the I/O performance of storage systems without any loss of storage space in the cloud. POD is a two-way approach to enhance the performance and also minimizes the performance overhead of data deduplication. The experimental results showed that POD outperforms and provides better capacity savings than iDedup. Tang et al. [5] proposed a data deduplication as an efficient technology to reduce storage cost for cloud storage systems, especially when massive volume of data has become normalcy in this era of Big Data. Primary storage. The direct interaction layer with service users secured the advantages of deduplication techniques because of its expensive manufacturing costs. Nevertheless, the primary storage is regularly accessed by users, workloads of primary storage systems are highly latency-sensitive. Such workload feature are highly challengeable to create both performance and space efficiency in deduplication schemes for primary storage systems. Existing deduplication schemes on primary storages do not pay any attention to achieve desired storage space while preventing the inherent performance faults.

Wen et al. [6] addressed the problem of managing the convergent keys in big data by presenting an efficient scheme known as Big data Outsourcing with Secure Deduplication, BDO-SD. The proposed is a convergent encryption technique encrypts the user query and also maintains the privacy. The simulation results conclude that this scheme achieved better efficiency than the existing data deduplication.

Yan et al. [7] explained the challenges of encrypted data and also the existing solutions of encrypted data deduplication which suffered from security breech. They proposed a new scheme to encrypt the deduplicate data and stored in cloud environment. The performance evaluation had been done on extensive analysis and simulations. And the results shows that it outperforms than other existing solutions for encrypting data deduplication. Zhou et al. [8] analysed about data deduplication technology that removes duplicate data, turns out to be one of the appealing solutions saving disk space and traffic in a big data environment. They analysed and characterised the impacts of performance and energy in bid data environments by identifying the causes of redundancy in big data workloads. Here they also unleash the relations between energy overhead and the degree of redundancy and also the efficiency of deduplication in an SSD environments. Ren et al. [9] proposed a secure data deduplication scheme which is based on differential privacy lies on constructing a hybrid cloud framework. The proposed method uses convergent encryption algorithm to encrypt original files and uses differential privacy mechanism to protect the system against side channel attack. The evaluated results shows that it can protect the system efficiently and also saves memory space and network bandwidth..

3. Network Design

Problem Statement

In big data their will a massive amount of data was gathered, stored, observed and analyzed for father processing. These database contains large number of files and it is very complicated to access those data. Because of big amount of data in the database, there will be large amount of duplication in data and it takes more memory space. The Proposed work removes the duplication in the data using MapReduce algorithm and reduces the memory space. The input given large number of files containing large number of data. The output obtained is data without duplication.

Proposed Work

In proposed work, first collect the dataset from DATA.CSV. The real data is divided into unique

chunks to perform this task by using fixed chunking algorithm. These algorithm start the number of chunks and generate the size of chunks (Eg. 64MB).It denote the file is divided into different chunks of size 64 MB. Find out the duplicate data by using chunks algorithm. Here use enhanced MD5 algorithm to create a hash values of this chunks. These hash values are hidden values so the data in chunks cannot be authorized by another person that may security breach of system. Now these hash value are inserted into the HDFS (Hadoop Distributed File System).Then start the different buckets the values are stored into the hash table, the corresponding buckets are stored in hash value. If duplicated data is identify then remove the duplicate files from the data and store the different data into HDFS. A new data is stored in HDFS, first use the chunking algorithm to generate the chunks. If the hash generated from the chunks and transfer these chunks for the verification process. To find out the hash value duplicated or not by using MapReduce approach, if hash value are identify that is duplicate so cannot be stored in HDFS otherwise store into buckets. This will minimized the duplicated data and reduce the memory space.

Steps in Proposed Work

Step 1: In this step the different data are stored in the disk and the repeated data is removed, for the recovery the data the metadata is used. It explains the chunking and record the deduplication process for each chunk.

Step 2: In this step when the two files have some duplicated data then single copy of the data is stored in the database. The references in the database is allowed for accessing by file,

Step 3: By using metadata file they receive the request of read and it given to the client.

Algorithm

Enhanced MD5 Algorithm

Input: data
Output: data without duplication
begin
Insert data
Separate data into the chunks
Bit values are assigned at the end of last chunk
if (last chunk size<other chunk size)
add extra bits to the last chunk

end
elserepeat first four steps

Chunking Algorithm

Input: dataset
Output: Fixed size chunks
begin

end
Initialize chunk size
Create chunks
Initialize the memory buffer size
Read the source file
Separate the bytes from the data

4. Results and Discussion

Performance Analysis

Large number of files are divided into fixed number of chunks according to the number of nodes in deduplication process .Chunks are calculated using the formula:

Performances analysis of the proposed method is measured using the following metrics: Data size after deduplication, Deduplication Ratio, Hash time, Chunk time. Table I explains the details of data size before deduplication using fixed size and Bucket based techniques.

Table I: Compression of Fixed Size and Bucket Based Techniques

Data size beforededuplication	Techniques	Data sizeafterdeduplication (GB)	DeduplicationRatio	Hash time(MB/s)	Chunktime(MB/s)
2.5(GB)	Fixedsize BuckettBased	1.23 1.05	0.4432 0.5431	190.0 50	170.0 60
1.5(GB)	Fixedsize BuckettBased	0.85 0.80	0.5462 0.6734	180.32 40	160.32 40

1. Data Size after Deduplication

Data size after deduplication is computed by dividing data size before deduplication and total number of data.

$$Number\ oooChunks = \frac{Totalssssseoffssle}{Numberofdata}$$

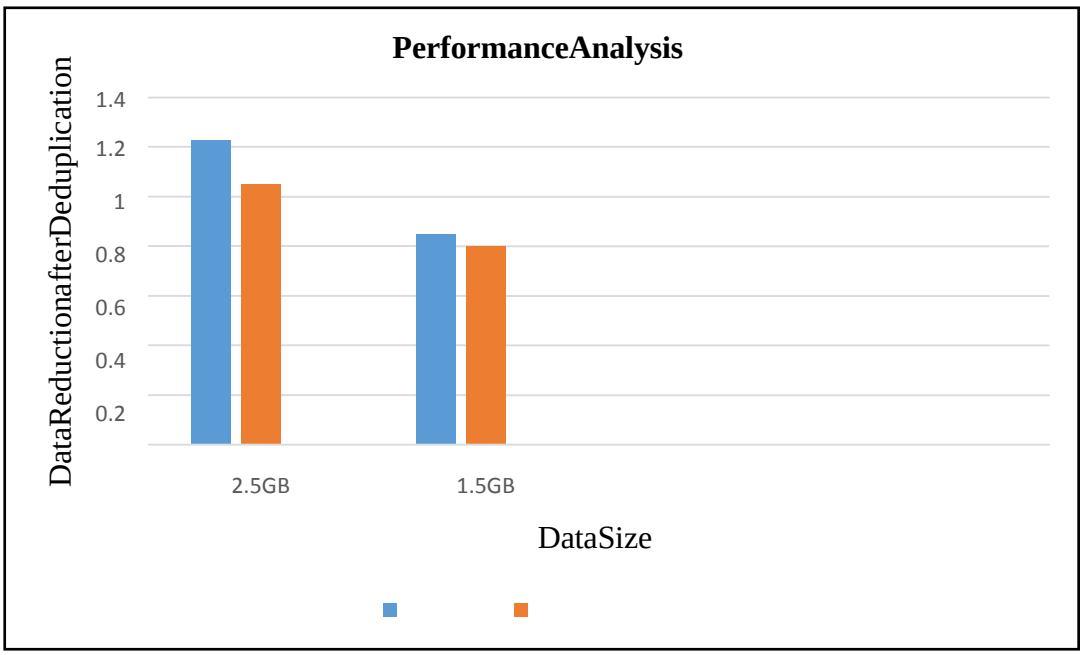

Fig.1:PerformanceAnalysis

2. Deduplication ratio

It is the ratio of calculating the output size of data into the input size of data. The formula of deduplication ratio is given below,

$$Datassszeaooter\ deduplscatsson = \frac{datassszebeooorededuplscatsson}{totalnumberooodata}$$

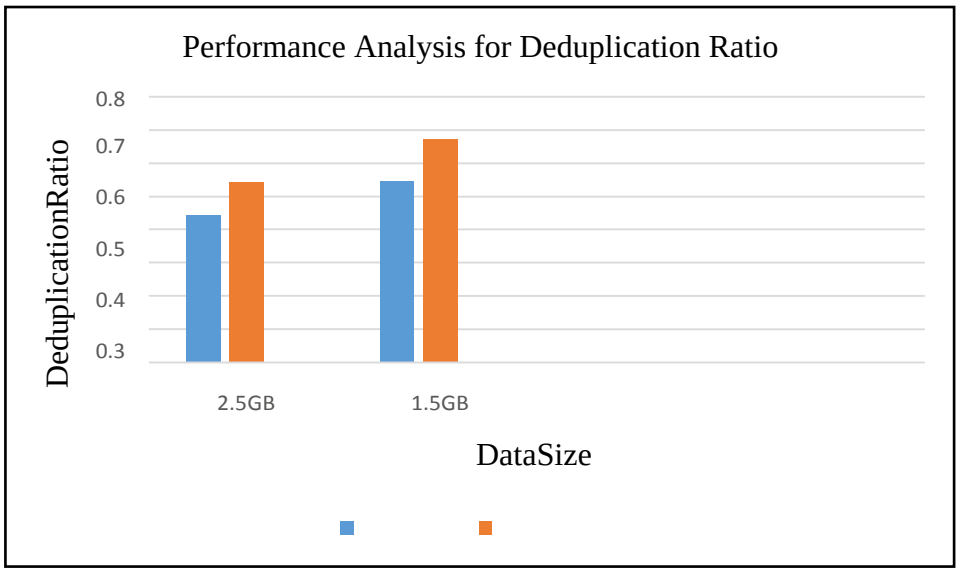

Fig.2:Performance Analysis for Deduplication Ratio

3. HashTime

The total time taken to perform hash operation is called hash time. The formula for hash time is given Below

$$Hashtssme = \frac{tssmetakentoperooormhash}{totaltssme}$$

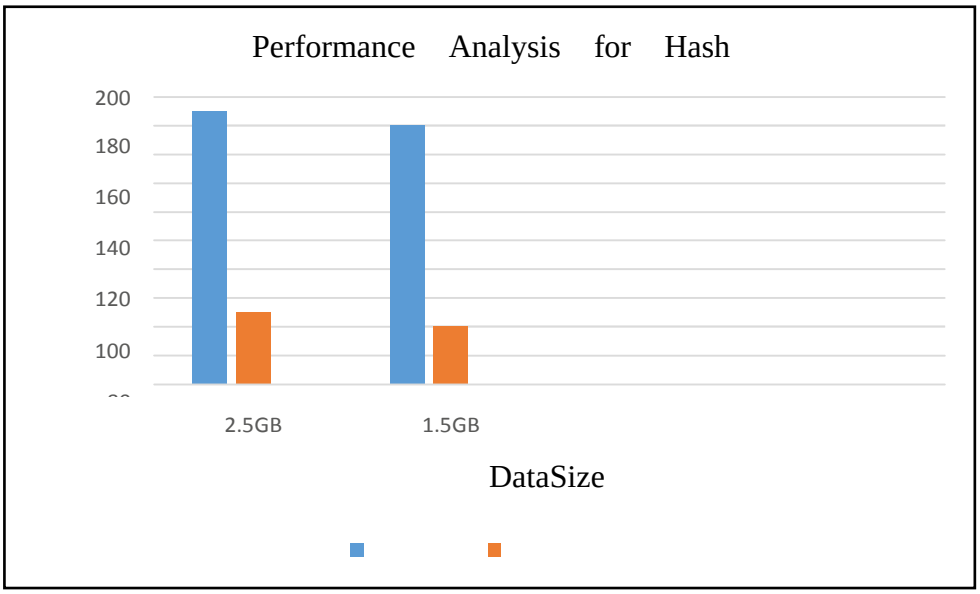

Fig.3:Performance Analysis for HashTime

4. Chunk Time

The total time taken to generate chunks is called chunk time.The formula for generating chunks is given below**.**

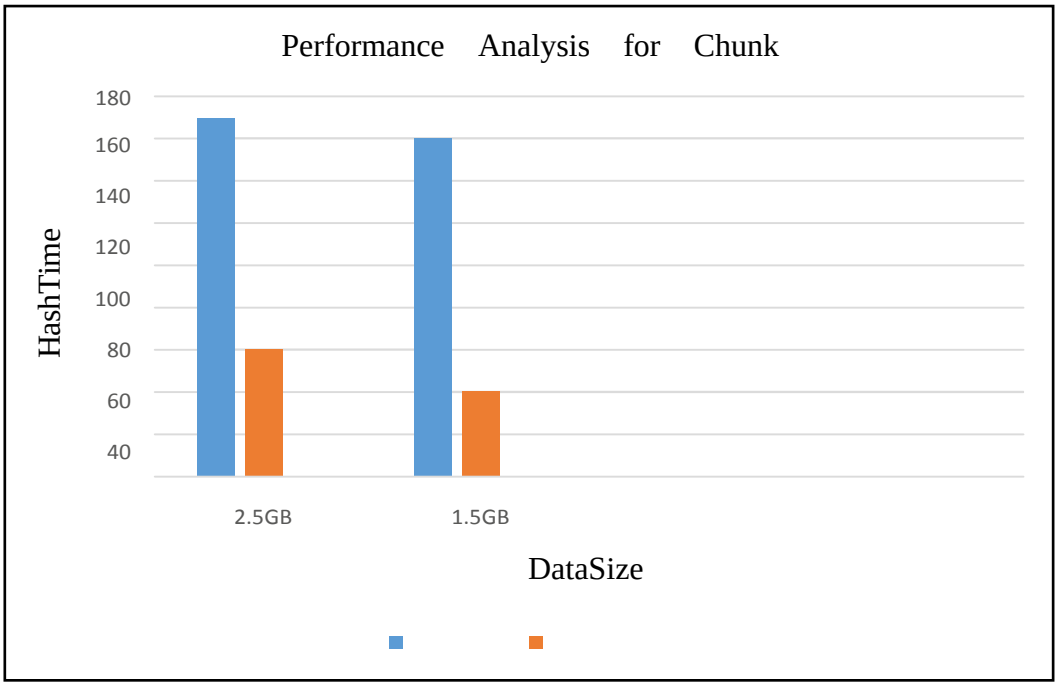

Fig.4:Performance Analysis for Chunk Time

5. Conclusion

Big data storage is a very difficult task and it is not very easy to manage such a large amount of data. To handle this issue, Hadoop tool must provide HDFS which manages data by removing the duplicated data in the database.

The proposed technique, which is based on bucket storage, efficiently stores data and deduplicate data using e-MD5 algorithm. With the help of e-MD5 algorithm the duplicated data are eliminated from the database and by using MapRedue algorithm the hash values generated for data are compared with already storedhash valuesin the bucket.Results showed that proposed technique's deduplication ratio, data size reduction, hash time and chunk time are observed and compared to existing fixed size chunking technique. The proposed technique provides better result an improves the performance of the system.

References

[1] N. Kumar, R. Rawat, S.C. Jain, Bucket based data deduplication technique for big data storage system. 5thInternationalConferenceon Reliability,InfocomTechnologiesandOptimization(TrendsandFutureDirections)(ICRITO), 2016, pp.267-271.

[2] Q. Liu, Y. Fu, G. Ni, R. Hou, Hadoop Based Scalable Cluster Deduplication for Big Data. IEEE 36thInternational Conferenceon DistributedComputing Systems Workshops (ICDCSW),2016,98-105.

[3]S.Luo,G.Zhang,C.Wu,S.Khan&K.Li,Boafft: distributeddeduplicationforbigdatastorageinthecloud,IEEEtransactionson cloud computing, 1,2015, 1-1.

[4] B.Mao,H.Jiang,S.Wu,L. Tian, Leveraging data deduplicationto improve the performance of primarystoragesystems inthecloud, IEEEtransactions on computers,65(6),2016,1775-1788.

[5] Y.Tang,J.Yin,W.Lo,Saud:Semantics-awareandutility-driven deduplication frameworkforprimarystorage,IEEE7thInternationalSymposiumonCyberspaceSafetyandSecurity(CSS),2015IEEE12thInternationalConferenceon EmbeddedSoftwareandSystems(ICESS),2015IEEE17thInternationalConferenceonHighPerformanceComputingandCommunications(HPCC),2015,190-197.

[6] M. Wen, K. Lu, J. Lei, F. Li, J. Li, BDO-SD: An efficient scheme for big data outsourcing with securededuplication, IEEE Conference on Computer Communications Workshops, 2015,214-219.

[7]Z.Yan,W.Ding,X.Yu,H.Zhu,R.H.Deng,Deduplicationonencryptedbigdataincloud,IEEEtransactionsonbigdata, 2(2), 2016138-150.

[8] R. Zhou, M. Liu, Li, T, Characterizing the efficiency of data deduplication for big data storage management, IEEEinternationalsymposiumonworkloadcharacterization(IISWC),2013,98-108.

[9] J. Ren, Z. Yao, J. Xiong, Y. Zhang & A. Ye, A secure data deduplication scheme based on differentialprivacy, IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS), 2016, 1241-1246.

[10]Y.Fu,N.Xiao,H.Jiang,G.Hu,W.Chen,Application-Aware BigData Deduplicationin Cloud Environment,IEEETransactionsonCloudComputing,1, 2017, 1-1.

[11] S. Singh, R. Singh, Next Level Approach of Data Deduplication in the Era of Big Data, International Journal of AdvancedResearchinComputerScience,8(4), 2017.