IJESMR

International Journal OF Engineering Sciences &Management Research

A NOVEL TECHNIQUE FOR DATA DEDUPLICATION WITH SHA-1 IN HADOOP FRAMEWORK

Sonam Bhardwaj*1 & Preeti Malik2
*1&2CSE Department, Kurukshetra University,Kurukshetra,India-136119

Keywords: *Big Data, Deduplication, Sha-1, Hadoop ,HDFS*

ABSTRACT

Big Data a transpiring research matter in hand analyzing and processing which is a defiance for current systems leading to high processing costs and degraded performance and quality. The centralized architecture is unable to cope up with the challenge of massive data resulting in storage space issues and processing time conflicts. The proposed technique addresses the above problem by applying the deduplication technique on various dataset containing unstructured data and implementing SHA-1 algorithm for calculation of fixed size digests and only storing the unique values. The research work is favoured by Hadoop that contains Distributed MapReduce framework with Mapper and Reducer programs for processing and reduction of data respectively.By enforcing the proposed technique there is a gain in space saved, reduction in time consumed, increased deduplication ration as well as number of duplicate files are detected efficiently.

INTRODUCTION

Big data,not a theoretical concept but a practical eruption for the storage system[1]. The enormous increase in the data emphasize the need for secondary storage devices that are costly,huge and needs more management. Neverthless, the resource utilization decreases due to increase in the amount of data. Big data[2] has certain features that justifies its name that are vareity,velocity, value and volume. Vareity shows that big data extends beyond structured to unstructured data of any type. Velocity points to the speed at which data gets created in real-time continously. Value regards to economic function, productivity improvement, competition for enterprises and benefits for the customers. Volume the domain describing the huge masses of data making the scale increasingly big. To flinch with this storing of data problem there is a prominent solution of data deduplication[3].

Data deduplication is an adept approach for minimizing the desidratum of storage space by weeding out the duplicated contents hence saving the solitary data contents. There are four leading key ingredients that makes up this process of deduplication happens to be chunking, duplication detection ,enhancement of reliability and information life cycle management.

Deduplication technique can be performed on three levels file-level ,block level and byte level[4][5][6][7].

In file level, the whole file is fragmented into fixed or variable size chunks and the congruent chunks are ommited. Block level embrace the data to be divided into blocks and saving just the unique ones where as in byte level the data splits into streams of bytes and compared in the most primitive way ,that is, byte by byte standing the most accurate method for deduplication but taking a bit mor time than the rest ones[8].

In this paper,we toss around file level deduplication splitting the file into fixed size(64 MB) chunks calculating their hash values using Sha-1 algorithm, comparing those hash values and only acrue the unique hash valued data contents ,that is, any chunks having same hash value with the other then the unique ones are only shown and the duplicate ones are reported.

Sha-1 belonging to the MD4 family is the hashing algorithm used to calculate the Message digests(Hash values) to find for duplicate contents by going through 80 rounds and yielding 160 bits hash value. This 160 bits is the outcome of the process that makes sha-1 successful ,viz, padding, appending the length,buffer initialization, message processing in 16-word blocks and lastly giving the result of 160 bit hash value. The attracting feature of sha-1 is that it is free from collision(no two different inputs yields same hash values)[9].

The algorithm is implemented on Apache Hadoop[10][11] which is an open source framework having the capability of storing and large scale processing of data sets. The computation ,in this framework, takes place at the residing location of the data in lieu of moving data to compute location. The components strenghthening Hadoop[12]are its distributed file system(HDFS) and mapreduce framework. HDFS is a specially designed file system for storing huge datasets with cluster of commodity hardware and streaming access pattern while Mapreduce processes the data being stored in HDFS by dividing work load into multiple tasks those are executed parallely[13].

The paper is further divided into five more sections intimating Hadoop, Related works, Implementation, Results and Analysis, and last Conclusion.

HADOOP

The architecture of hadoop provides the abstracted file system and os level communication. It has all its nodes located in racks which makes it easy for it to determine where the nodes is and can easily process the data with results.

The architecture includes following components

- Name node and secondary name node
- Data node
- Task trackers
- job tracker

a) Architecture is a master-slave one of which Name node,secondary name node and job tracker are the master while data node and task tracker constitutes the slave. This internally has 2 layers one is the HDFS layer consisting of name node and data node for storing the data and the other is the MapReduce layer having task tracker and job tracker for processing the data. There is an intra-communication between masters and slaves.
HDFS[14] is written in java having hdfs cluster with Name node that administers the namespace(also the metadata stored in name node hard disk) along with controlling access to data by clients.

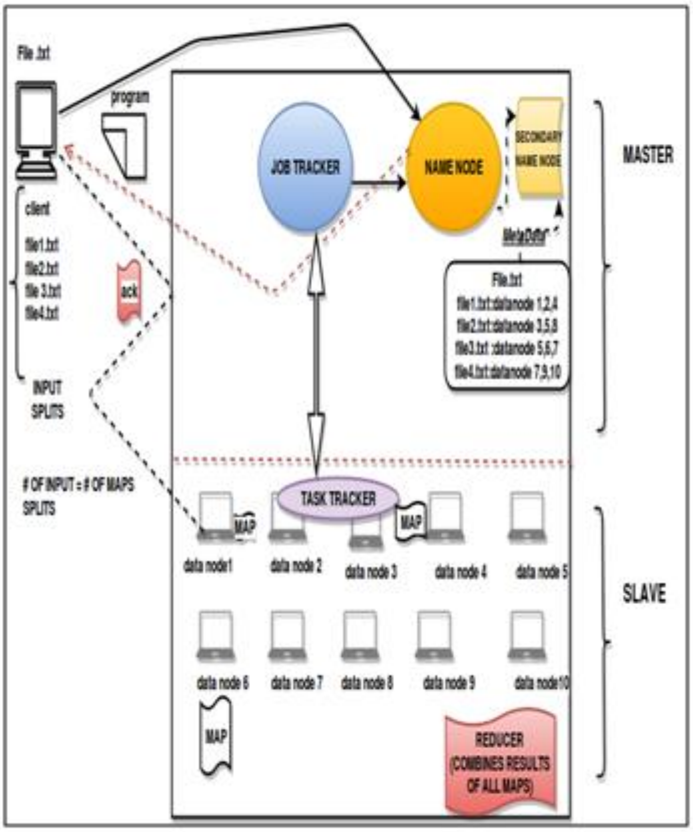


Figure 1: Hadoop Scaled Architecture

Secondary Name Nodemaintains a copy of NameNodedata to be used to restart the NameNode when failure occur. In addition,there are multiple data nodes that manages the memory attached to it. The input file is divided into various blocks which are then mapped to the datanodes supervised by namenode which keeps the track of what block is mapped to which datanode which are then accountable of i/o requests from clients. In hdfs, for high performance and protection of data, replication of data is made across multiple nodes.

b) Another basic component for hadoop is MapReduce that incurs computational framework. MapReduce is the programming fascimile and capable of implementing large datasets congenitally parallel befitting the distributed environment. Hadoop gains a cluster of nodes to run Map Reduce programs densely in parallel. Job tracker schedules all the jobs as well as the jobs split into multiple tasks over the cluster that executes on the worker nodes. Task tracker sends heart beat to the job tracker reporting its progress .

As by the name MapReduce consists of two major programs Map that processes input data and secondly the Reduce program that unites the proceedings into the final result. The usage of (key,value) pairs defined by users allows the output of one job to be used as input for another.

Data are broken in data blocks (generally 64 MB,as not too big for memory of data blocks and not too small for increased number of blocks) stored locally on different nodes and replicated for reliabilty, availaibilty, fault tolerance and data security. Hdfs is the constitution of the file system on which mapreduce programs run. Mapreduce program is not affected by the traditional hindrance of network bandwidth.

The below defined are the key attributes of MapReduce that intensify the performance of hadoop:

- Resource Manger: It bestow data locality and server resources to actuate optimized operations
- b)Optimized scheduling : Jobs are accomplished based on the prioroties given to the jobs that is the most critical will be processed first.
- Adaptability : The procedures can be written
- effectually in any programming language. It implies the flexibility of the MapReduce.
- High availability and bouyancy : The regular
- tracking of tasks through heart beat (block report) ensures that failure of jobs are independent and can be restarted automatically.
- Scaled – Architecture : This implies that for boosting performance servers may further be added.

RELATED WORKS

- ➢ In 2015,Big Data was analyzed using Computational Intelligence and Hadoop[15] to face the challenges to existing computer intelligence techniques Volume, Velocity ,Value and Veracity.

The work imparted computational intelligence in huge amount of data while using biologically inspired techniques and with big data analysis using Hadoop environment.
Big Data analysis using Hadoop was nature inspired and an effective method for mining and analyzing tons of data.

Computational Intelligence (CI) techniques are expected to provide powerful tools for addressing Big Data challenges. The main techniques in CI, such as evolutionary computation, neural computation and fuzzy systems are inherently capable of handling various amount of uncertainty, which makes CI techniques well suited for dealing with Variability and Variety of Big Data. On the other hand, the other two V's, Volume and Velocity may create serious challenges to existing CI techniques. The next two V's that is Value ad Veracity are equally important and yet challenging in dealing with big data. Consequently, new CI techniques needed to be developed to efficiently and effectively tackle huge amount of data, and to rapidly respond to changing situations.

The two approaches were compared taking few examples:

- Optimization Inspired by Evolution process of a Bacterial Colony and hadoop cluster-A new swarm intelligent technique called bacterial colony optimization (BCO) was considered such that the problem space was huge due to its evolutionary properties similar to the scalability of commodity hardware in hadoop in orderto provide availability and scalability properties to the system of computers.

- Support vector machines using nonlinear kernels on hadoop mahout and the kernel methods for trees and graphs through neural networks. The four major challenges of big data i.e. volume, velocity, variety and veracity targeted the big data mining. This was achieved via hadoop ecosystem and the swarm intelligent techniques. Here harmonic cryptosystem with secured multiparty computation of system matrix operation yielded high privacy preserving while data miners perform information retrieval from big data.The neural networks were applied on structured data for mining of useful data that uses a recurrent network for the analysis of data.
- Swarm intelligence was successfully applied in hosting research settings that focus on improving management andcontrol over large number of interacting entities thus, describing the collective behavior . It is primarily concerned with the design of multi agent systems by taking inspiration from collective behaviors of social insects and other animal societies .Swarm intelligence inspired Hadoop analysis of Big data .

- AR-dedupe[16] approach for cluster deduplication in the year 2015 marked the following challenges as their research base:

Decreasing data deduplication rate with the increasing dedupe server nodes.

1. High communication overhead for data routing.
2. Load balance for improving throughput of the system.

Cluster Dedupe has three parts Backup client, Metadata and Deduplication server Nodes. First, it partitions large data objects into smaller parts called chunks and generates its fingerprint which can be uniquely represented in the backup client. Then, it transfers all chunks to deduplication server nodes according to its routing mechanism. Metadata management server keeps the information of all files for restoration. The algorithm used for chunking was Content Defined Chunking that forms chunks according to the content.

There were two types of Data sets used to evaluate AR-Dedupe, ∑-dedupe[17] and Extreme binning[18] independently. They all partitioned data into chunks with static chunking of 4 KB size in Backup client. Extreme Binning does deduplication operations with a file size granularity, as the other with super chunk granularity.

The research succeeded in overcoming the challenges and improved 30% performance in terms Handprint index with application aware mechanism.

AR-Dedupe mainly consists of four parts: backup client, metadata management server, routing server and deduplication server nodes.Backup client first sorts data by application, then partitions large data into chunks and generates fingerprints. Finally, the backup client groups chunks into super-chunk and sends its handprint to routing server with its application type. After getting the s_id from routing server, it begins to backup.

Metadata management server stores the chunk information of all files and it manages files in different groups by application. Routing server is responsible for choosing the optimals id for each super-chunk and keeping system load balancing.

Deduplication server nodes consist of n deduplication servers, and each stores unique data chunk.

AR-Dedupe acquired a high data deduplication rate with low communication overhead, and at the same time kept the system’s load balancing well for the following reasons:

- Handprints were chosen to represent the similarity among super-chunks that could effectively reduce the communication overhead between backup client and routing server. Moreover, handprints decreased the amount of information needed to be stored for super chunks in routing server.
- By adding routing server to cluster deduplication system, AR-Dedupe could select the optimal deduplication server node which only needed a communication with routing server. The load information stored in the routing server which contributed to the cluster deduplication system to keep load balancing well.

- Besides, different handprint index tables were created according to different application types, which improved the efficiency of indexing and at the same time having little impact on the data deduplication rate.

ConvergentAR-Dedupe mainly consists of four parts: backup client, metadata management server, routing server and deduplication server nodes.Backup client first sorts data by application, then partitions large data into chunks and generates fingerprints. Finally, the backup client groups chunks into super-chunk and sends its handprint to routing server with its application type. After getting the s_idfrom routing server, it begins to backup.

Metadata management server stores the chunk information of all files and it manages files in different groups by application. Routing server is responsible for choosing the optimal s_ id for each super-chunk and keeping system load balancing.

Deduplication server nodes consist of n deduplication servers, and each stores unique data chunk.

- In June 2015, A Secured and Authorized Data deduplication in Hybrid cloud with public auditing benefited both storage provider and user by deduplication technique and auditing technique respectively.

Traditional Deduplication systems based on convergent encryption even though provide confidentiality but do not support the duplicate check on the basis of differential privileges. Paper presented idea of authorized data deduplication proposed to protect data security by including differential privileges of users in the duplicate check. To support stronger security the files were encrypted with differential privilege keys, users only allowed to perform the deduplication for the files marked with the corresponding privileges to access. Users can verify his/her presence of file after deduplication in cloud with the help of a third party auditor by auditing the data further auditor audits and verifies the uploaded file on time[19].

With traditional encryption different users encrypt data with their own key, which makes likeness data with different user key makes different ciphertext for that data which is unable for deduplication. The convergent encryption allows encrypt/decrypt data with convergent key on the data thus makes possible to apply to check duplicates. Thus with uploading user's data as ciphertext to cloud resolved security issues. To prevent from unauthorized access, proofs of ownership protocol can be used as privacy constraint. Proof of ownership denotes that user can download the decrypted and obtain particular data with convergent keys by specifying its ownership. Therefore, resolving security and privacy issues.

- A Survey on Deduplication in Cloud Storage[20] enlisted various challenges like Bnadwidth, throughput, Computational overhead, deduplication efficiency, read and write efficiency, Backup window size and transmission cost. Therfore many deduplication strategies were adopted to benefit the cloud backup services that were grouped into Application Based, GPU based, Hash Cluster based, Casualty based and SSD based(Solid state drives deduplication).
- The agenda of data deduplication is avoiding storage of multiple replicas of similar data block in physical storage medium. In [21,22,23], using deduplication there abide convincing reduction on image file storage. In 2009, Jin et al.[24] studied the deduplication potency on virtual machine disk images and observed that stored data cultivate slowly suceeding first few virtual images on disk. Takahashi et al. [25] achieved a rapid migration expeditive migration of virtual storage by adopting deduplication, that reduced the amount of data transmission among hosts by using reusable disk pages. Meister et al.[26], stressed on file recipe compression in data deduplication adopting a union of scalable and efficient compression technique for reducing file recipe size.

IMPLEMENTATION

Experimental setup

The research work has been carried out on operating system Linux Mint 17.3 Rosa Cinnamon 64-bit with Hadoop-2.7.2.tar.gz along with Eclipse 4.5.1[27] platform installed . The proposed technique is applied on various data-sets[28][29] , inferencing the applicability on Big Data.

Linux Mint is a virus free open source operating system that provides a feasible environment for hadoop to run. The MapReduce programs either runs in python or java. The concerned code is written in java fo which we have eclipse that is also a open source software certified by OSI that provides in-built java packages.

SHA-1 algorithm

Amongst the most trusted algorithms SHA-1 proves itself by calculating the hash values being free from collision restricting the input length to 264 that are processed in 512 bits blocks processing each block separately[30].

The 160 bits hash values being calculated are the outcome of 80 rounds of processing that involves the preprocessing of input as the initial steps consisting of padding, appending length and initialisation of buffers. The chronicle of the algorithm is as follows:

Step 1: Adding the padded bits at the end of message to make total length congruent to 448 mod 512. for padding first hit is for the '1' followed by '0' hits as per requirement.

Step 2 : 64-bit binary representation of the message is made for appending the length at the end of the message.

Step 3 : 160-bits SHA-1 buffer is initialized
Word A: 67 45 23 01
Word B: EFCDAB 89
Word C: 98BADCEF
Word D: 10 32 54 16
Word E: C3 D2 El FO

Step 4 :16- word Block processing
The inclination of the algorithm is a module that consists of four rounds of processing 20 steps each.

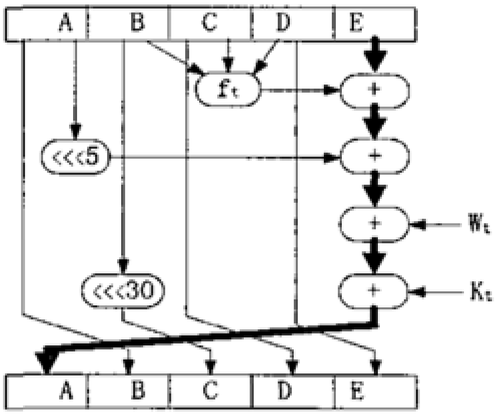


Figure 2: Sha-1 single step processing

Step 5: After all512-bits blocks have been processed, the output of the last block is the 160-bits message digest. Figure shows the operations involved in a single step.

Proposed Algorithm

File level Deduplication:

```
//Input:Data in the type of file from user
//Output: Several Blocks or no output
while
receive the data from users
do
Fetch the data(video,audio,text)
```

```
if
the size of data is greater than storage space
then
send message to users of denying the request of saved data
else
chunk the data into several blocks
update the data into saving space
compute the data digest by SHA-1 in distributed environment of map reduce
if
there are any collision in digests
then
print the chunk or filename
else
save the copy in data node of HDFS
analyse the time, saved storage space and deduplication ratio
end
```

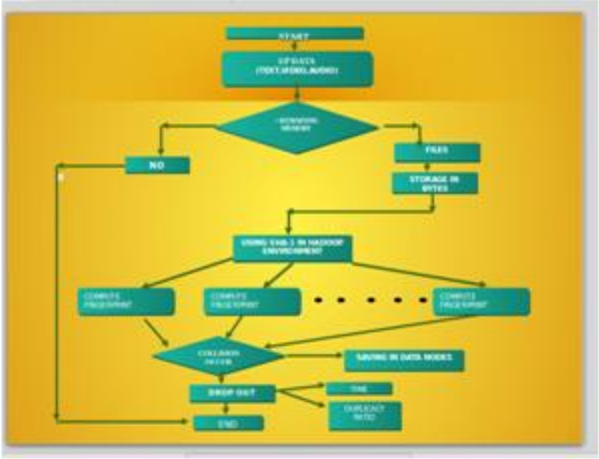


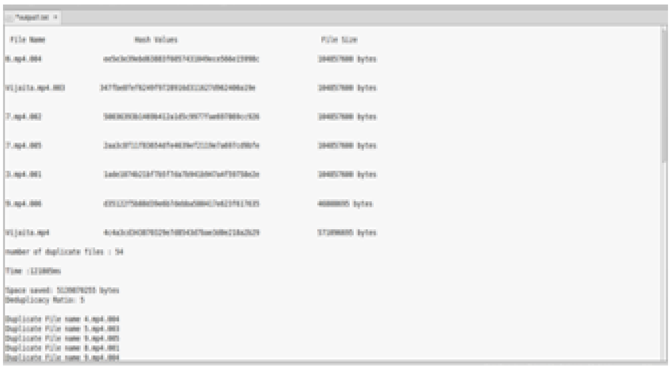
Figure 3: Proposed Methodology Flowchart

RESULTS AND ANALYSIS

The results calculated on Datasets of variable sizes on hadoop 2.7.2.tar.gz platform.

Table 1: Results showing various parameters

Datasets	Datasize before deduplication (GB)	Datasize after deduplication (GB)	Number of duplicate files	Time (ms)	Ratio	Space saved (GB)
Dataset 1	3.3	1.7	26	70677	2	1.6
Dataset 2	6.7	2.03	88	121660	3	4.67
Dataset 3	2.0	1.1	37	37666	2	0.9



```
File Name              Hash Values                        File Size
6.mp4.004          [illegible]                104857600 bytes

Vijaila.mp4.003    [illegible]                104857600 bytes

7.mp4.002          [illegible]                104857600 bytes

7.mp4.005          [illegible]                104857600 bytes

3.mp4.001          [illegible]                104857600 bytes

9.mp4.006          [illegible]                40000095 bytes

Vijaila.mp4        [illegible]                57109695 bytes
number of duplicate files : 54

Time :121809ms

Space saved: 5120870255 bytes
Deduplicacy Ratio: 5

Duplicate File name 4.mp4.004
Duplicate File name 5.mp4.003
Duplicate File name 9.mp4.005
Duplicate File name 8.mp4.001
Duplicate File name 5.mp4.004
```

Figure 4:Fingerprints and calculated parameters

Graph in Figure 5 and Figure 6 show analysis of the above output and results, portraying the sapce saved , time consumed , deduplication ration and duplicate files detected.

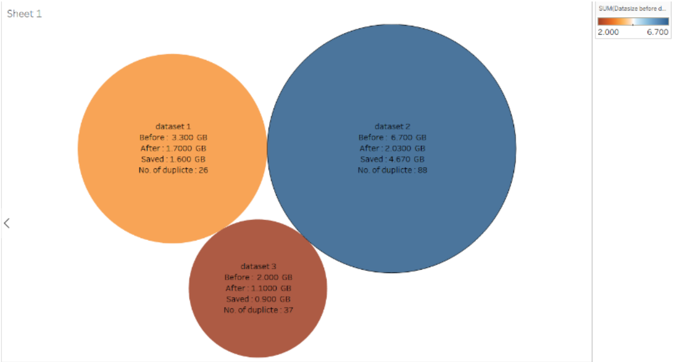


Figure 5: Graph showing the analysis of Space saved, time consumed, number of duplicate files detected for Dataset 1, Dataset2, Dataset3

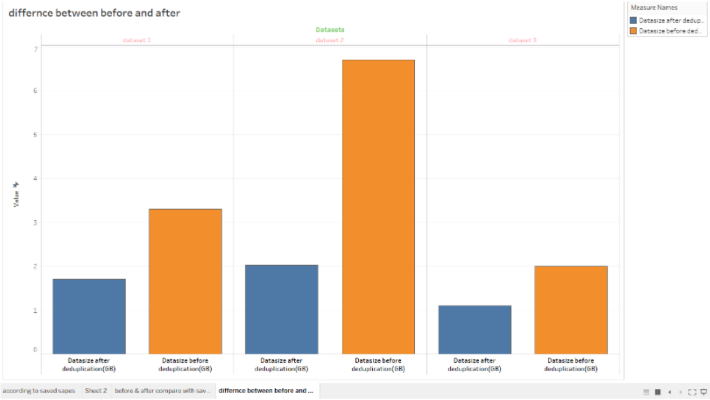


Figure 6: Graph showing comparison of space saved before and after deduplication for Dataset1, Dataset2, Dataset3

CONCLUSION

Duplication is not a new apparatus, in fact it is just data compression derivatives. With the advancement of information and network technology, gradual increase in the size of the data center and energy consumption due to IT spending increasingly in data deduplication for optimization of storage system can greatly reduce the amount of data.

This technique works on primary as well as distributed systems using Hash based SHA1 algorithm on Hadoop platform providing mapper and reducer to utilize space efficiently, removing deduplication and storing unique files by calculating and comparing the fingerprints. By this, time performance also improves as only the unique data is processed. The results show that on primary storage it takes less time running the program with Hadoop rather without hadoop it takes more time. The deduplication ratio is increased with hadoop platform and so is the space saved.

The results of the research can be improved by applying the technique on distributed storage like AWS cloud and recording other parameters like bandwidth, throughput and computational latency

REFERENCES

1. *A.K. Reshmy, D. Paulraj,"An Efficient Unstructured Big Data Analysis Method for Enhancing Performance using Machine Learning Algorithm", International Conference on Circuit, Power and Computing Technologies [ICCPCT], 978-1-4799-7075-9/15/ ©2015 IEEE.*
2. *Marco Pospiech, Carsten Felden, "A Descriptive Big Data Model Using Grounded Theory", 16th International Conference on Computational Science and Engineering, 978-0-7695-5096-1/13, IEEE,2013.*
3. *Min Chen, Shiwen Mao and Yunhaoliu, "Big Data-A Survey",©Springer.Science+ Business Media New York ,2014.*
4. *T. Yujuan, J. Hong, F. Dan, T. Lei, and Y. Zhichao, "CABdedupe: A Causality-Based DeduplicationPerformance Booster for Cloud Backup Services," in Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International, 2011, pp. 1266-1277.*
5. *X. Lei, H. Jian, S. Mkandawire, and J. Hong, "SHHC: A Scalable Hybrid Hash Cluster for Cloud Backup Services in Data Centers," in Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on, 2011, pp. 61-65.*
6. *Jaehong Min, Daeyoung Yoon and Youjip Won, " Efficient Deduplication Techniques for Modern Backup operation" IEEE Transactions on Computers, Vol. 60, No. 6, June 2011*
7. *Z. Yang, W. Yongwei, and Y. Guangwen, "Droplet: A Distributed Solution of Data Deduplication," in Grid Computing (GRID), 2012 ACM/IEEE 13th International Conference on, 2012, pp. 114-121.*
8. *Ruijin Zhou, Ming Liu, Tao Li,"Characterizing the Efficiency of Data Deduplication for Big Data Storage Management", 978-1-4799-055-3/13 ©2013 IEEE.*
9. *Zhenqi Wang, Lisha Cao, "Implementation and Comparison of Two Hash Algorithms" 2013 International Conference on Computational and Information Sciences 978-0-7695-5004-6/13 © 2013 IEEE DOI 10.1109/ICCIS.2013.195.*
10. *Apache Hadoop,http://hadoop.apache.org*
11. *Apache Mahout, https://mahout.apache.org*
12. *Apoorva Gupta,"Big Data Analysis Using Computational Intelligence and Hadoop: A Study",2nd International Conference on Computing for Sustainable Global Development (INDIACom),978-9-3805-4416-8/15,IEEE,2015.*
13. *Sara B.Elagib, Atahur Rahman Najeeb, Aisha H. Hashim, Rashida F. Olanrewgi, "Big Data Analysis Solutions Using MapReduce Framework", 5th International Conference on Computer and Communication Engineering, 978-1-4799-7635-5/14, IEEE,2014.*
14. *Ramya A V, E Sivasankar, "Distributed Pattern Matching and Document Analysis in Big Data Using Hadoop Map Reduce Model", International Conference on Parallel, Distributed and Grid Computing, 978-1-4799-7683-6/14 ©2014 IEEE.*
15. *ApoorvaGupta,"Big Data Analysis using Computational Intelligence and Hadoop: A Study",IEEE 2nd International Conference on Computing On Sustainable Global Development (INDIAcom),2015.*
16. *Xing Yu-Xuan, XlaoNOng, LiuFang, SunZhen, HeWan-Hui, "AR-Dedupe:An Efficient Deduplication Approach For Cluster Deduplication System",J.ShanghaiJiaotong Univ.(Sci.),2015,20(1):76-81.*
17. *Fu Y J, Jiang H, Xiao N, "A scalable inline cluster deduplication framework for big data protection",The ACM/IFIP/USENIX 13th International Conference on Middleware (Middleware'12), ACM, 2012: 354-373.*
18. *D. Bhagwat, K. Eshghi, D. D. E. Long, and M. Lillibridge, "Extreme Binning: Scalable, parallel deduplication for chunk-based file backup," in Modeling, Analysis & Simulation of Computer and Telecommunication Systems, MASCOTS '09. IEEE International Symposium on, 2009, pp. 1-9.*
19. *Sharma Bharat, Mandre B.R., "A Secured and Authorized Data Deduplication in Hybrid Cloud with Public Auditing",International Journal of Computer Applications (0975 – 8887) Volume 120 – No.16, June 2015.*
20. *P.Neelaveni,M.VijayaLakshmi,"A Survey On Deduplication in Cloud Storage",Asian journal Of Information Technology 13(6):320-330©Medwell journals,2014.*

21. *K. R. Jayaram, C. Peng, Z. Zhang, M. Kim, H. Chen, and H. Lei. An Empirical Analysis of Similarity in Virtual Machine Images. In Proceedings of the Middleware 2011 Industry Track Workshop, Middleware'11, pages 6:1-6:6, New York, NY, USA, 2011. ACM.*
22. *Clements A, Ahmad I, Vilayannur M, et al. Decentralized deduplication in SAN file systems // Proc of the USENIX ATC09. Berkeley:USENIX,2009:98-111.*
23. *Jayaram, K. R., et al. "An empirical analysis of similarity in virtual machine images." Proceedings of the Middleware 2011 Industry Track Workshop. ACM, 2011.*
24. *Jin K, Miller E L. The effectiveness of deduplication on virtual machine disk images[C]//Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference. ACM, 2009: 7.*
25. *Takahashi, Kazushi, Koichi Sasada, and Takahiro Hirofuchi. "A fast virtual machine storage migration technique using data deduplication." CLOUD COMPUTING 2012, The Third International Conference on Cloud Computing, GRIDs, and Virtualization. 2012.*
26. *Meister, Dirk, André Brinkmann, and Tim Süß. "File recipe compression in data deduplication systems." Proceedings of 11th USENIX Conference on File and Storage Technologies (FAST). 2013.*
27. *Eclipse platform, http://www.eclipse.org.*
28. *https://erondata.readthedocs.io/en/latest/achieve.ics.uci.edu/ml/machine-learning- databases/*
29. *https://data.sunlightlabs.com/dataset/*
30. *Dan Cao, Jun Han, Xiao-yang Zeng, "A Reconfigurable and Ultra Low-cost VLSI Implementation of SHA-1 and MD5 functions", 1-4244-1132-7/07 © IEEE,2007.*