AUGMENTED APPROACH OF DATA DEDUPLICATION AGAINST RESOURCE PLANNING IN BIG DATA ENVIRONMENT

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

SHAMSHER SINGH

41400022

Supervisor

MR. RAVINDER SINGH

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

May, 2017

PAC Form

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : 1792::M. Tech- CSE(Computer Science and Engineering)(Part Time)

COURSE CODE : CSEP546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0257

Supervisor Name : Ravinder Singh **UID :** 17750 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Shamsher Singh	41400022	2014	K1418	08872002736

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** _____

PROPOSED TOPIC : DATA DEDUPLICATION

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	7.25
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.25
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.50
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.75
5	Social Applicability: Project work intends to solve a practical problem.	7.50
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	8.00

PAC Committee Members		
PAC Member 1 Name: Janpreet Singh	UID: 11266	Recommended (Y/N): Yes
PAC Member 2 Name: Harjeet Kaur	UID: 12427	Recommended (Y/N): Yes
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): Yes
PAC Member 4 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Augement approach of deduplication against resource planning in BIG DATA environment**

Overall Remarks: Approved (with major changes)

PAC CHAIRPERSON Name: 11011::Dr. Rajeev Sobti **Approval Date:** 22 Nov 2016

5/20/2017 10:29:16 AM

ABSTRACT

In storage management systems it is the most difficult task to identify the duplicate data which resides on the storage disks. This duplicated data acquires high amount of storage space which sounds meaning less. On storage disk two or more copies of same data reduce the availability of space and increase the storage cost of data. Some researchers have been conducted researches to overcome this issue but still there are improvements required in this area. Our proposed work is based on an augmented approach to locate the similar data stored on disks in big data environment. As the result duplicated date will be removed from the storage media and free up the space, increase the system performance in terms of operational speed, and reduce the time for deduplication process.

Proposed approach is basically merging to different concepts, one is data deduplication which detects duplicate data and store only single instance of data, second is big data which is one of the hot topic now days because of its characteristics. Big Data is a group of bulky data sets which is almost difficult to handle or process by using traditional resources. Proposed work will eliminate the duplicate data by using inline data deduplication and post process data deduplication both, based on size of the data. Furthermore, in post process deduplication, file identification techniques will be implemented and then accordingly deduplication strategy will be applied on data sets. After post process deduplication an acknowledgement ticket will also be provided to application server in order to remove duplicate data from online storage media also.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "**Augmented Approach of Data Deduplication Against Resource Planning in Big Data Environment**" in fulfillment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Ravinder Singh. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Shamsher Singh

41400022

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**Augmented approach of Data Deduplication against Resource Planning in Big Data Environment"**, submitted by **Shamsher Singh** at **Lovely Professional University, Phagwara, India** is a bonafide record of his/her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Mr. Ravinder Singh)

Date:

Counter Signed by:

1) **HOD's Signature: _____**

   HOD Name: _____

   Date: _____

2) **Neutral Examiners:**

   (i) Examiner 1

   Signature: _____

   Name: _____

   Date: _____

   (ii) Examiner 2

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

I would like to take this noble opportunity to extend my deep-sense of gratitude to all who helped me a lot directly or indirectly during the development of this dissertation proposal.

Fore-mostly I want to express wholehearted thank to my mentor**, Mr. Ravinder Singh** for being such a worthy mentor and best ever adviser. His precious advice, motivation and critics proved the sources of innovative ideology, encouragement and main cause behind the successful completion of this dissertation. I am very much obliged to all the lecturers of computer science and engineering dept. for their heartfelt encouragement and support.

I also extend my sincerest thanks and gratitude towards all mates for their consistent support and invaluable suggestions provided at that time when I required the most. I am very grateful to my lovable family for their support, love and prayers.

Shamsher Singh

TABLE OF CONTENTS

LIST OF ACRONYMS / ABBREVIATIONS

IDC	International Data Corporation
ZB	Zeta Byte
PB	Peta Byte
SHA	Secure Hash Algorithm
PDA	Personal Digital Assistant
HAR	History Aware Rewriting
CAF	Cache Aware Filter
HDFS	Hadoop Distributed File System
RAID	Redundant Array of Inexpensive Disks
SSD	Solid State Drive
RAM	Random Access Memory
FD	Functional Dependencies
CDC	Content Defined Chunking
AE	Asymmetric Extremum
FSC	Fixed Sized Chunking
AGDO	Asymmetrical Grouping Data Organization
VM	Virtual Machine
LA	Learning Analytics
JVM	Java Virtual Machine
MB	Mega Byte
GB	Gega Byte

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

The first chapter begins with the background of the study. It entails the scope of Data Deduplication and Big Data in present scenario. Furthermore, the chapter sheds light on the data quality factors, algorithms, value chain of Big Data etc. The main focus of the chapter is to present deeper insight for research topic.

1.1 Introduction

In today's scenario data acts as an important asset for every organization. It decides the future of organization because it is the mine of information which supports the decision making process for future strategies. As indicated by the description by International Data Corporation (IDC) in 2011, the whole world carried around 1.8 Zeta Byte (ZB) of data which was increased nine times within past 5 years [1]. After this blasting growth in data IDC predicted that this amount of data will be double in every two years in future. From this we can easily understand the growth rate of data in our digital world. Now everyone carries digital assistants like smartphones, cameras, laptops and many more gadgets, these plays a big role behind the generation of this much amount data. Social networking sites like Facebook, LinkedIn, Twitter etc. are used to post millions of posts including photographs and videos which are the main cause of higher data generation rate. According to records Facebook generates 10 PB+ log data per month, Google processes data of 100's of PB [1].

As the quick growth in production of data, it becomes a matter of tension for every company to store data and perform analysis operations on this huge amount of data. Analysis makes any organization grow in this competitive world. To make any business a success data mining and analysis plays a vital role but here comes an obstacle that is quality of collected data on which analysis operations to be performed. A good quality data provides better results which leads the organization to better decision making stage [2].

Data quality assurance is a plus point for performing certain operations. It is a process of assuring that the collected data is free from anomalies or not. It guarantees that data is consistent, complete, accurate and free from any other errors. For this purpose, data cleansing, data aggregation, data transfer etc. activities are performed on data [2].

1.1.1 Data Quality Factors

Keeping in mind the end goal to guarantee the quality factor of collected data following are some data quality factors:

1. **Data Accuracy:** It ensures that data collected is correct in all the dimensions. It is assessed in two perspectives a) accuracy of data attributes, which has concern with the accuracy of data value sets, b) operational mistakes while collection of data.
2. **Data Currency and Timeliness:** It has concern with the updation of the data with respect to time.
3. **Data Correctness:** This ensures that data is correct in terms of formats, data types, data profiles etc.
4. **Data Usability:** This factor ensures the degree of usability of data in the process of data analytics.
5. **Data Completeness:** As data pool is collection of heterogeneous data; data completeness ensures that acquired data is complete in all manners or not.
6. **Data Security and Privacy:** This quality parameter make sure that data is safe and no privacy threat is present around.
7. **Data Accessibility:** This parameter indicates that how easily data is accessed by the users.
8. **Data Scalability:** It ensures that data should be easily scalable, transferable and accessible [2].

1.2 Data Deduplication

After generation and storage of huge amount of data there is a problem exist which is concern with the storage utilization of storage media. When any organization have to

make analysis or do any kind of survey for new strategies and decision making process, they try to collect higher amount of data so that they can conclude better and produce a quality product. It is quite obvious when large amount of data is to be collected then many data sets will be there which are exactly or somewhat similar to each other. Here comes the problem of data duplication. It consumes large amount of storage space to store data but in actual storage resources gone waste because of redundant data storage.

With the high rate of increment in the amount of data, it has become a problem to store it in such an efficient manner so that we can reduce the cost of storage and get enough space at low cost. To overcome this problem at the time of data storage a technique is used to eliminate the redundant data so that only a single unique copy of data will be stored and save space and cost, that technique is known as ***Data Deduplication***.

In today's era of digital world storage becomes a very expensive need for the software companies as well as the home users. Every day millions of users create PB's (Peta Bytes) of data in the form of videos, photographs and other documents. Every user wants its data to be safe from every aspect that's why they store multiple copies of their data at different locations. For this purpose, online storage is provided by many vendors. Another reason is incremental backups that are used for security and consistency means. To overcome this duplication problem data deduplication an effective technique used to free up the storage space.

In de-duplication cryptographic hash is used to locate and delete the redundant data from the backup taken. Hash value is a fixed length output of any data. In deduplication when data documents come for storage, its hash signature is made by utilizing secure hash algorithm (SHA). That hash signature is checked by the server from hash index that has all the hash signatures stored. If that hash signature matches with any other hash signature it means that data is duplicate and need not to store again, then data will be deleted but a reference will be generated to the original data. If hash signature does not match with any other signature, then that data will be stored in the disk and new signature entry will be done in index.

In order to stay updated about the term data deduplication following are some key points about it.

1.2.1 Two Ways to Perform Data Deduplication

1. **Inline Data Deduplication:** In this data undergoes the deduplication process before storing in the storage disk. When data comes for the storage in the disk the deduplication algorithm is applied on it and only the unique data blocks are stored.
2. **Post Process Data Deduplication:** In this deduplication is performed on data after storing it into the storage disk. After storing data, data fetched for deduplication process and after that only unique data blocks stored back into the memory and redundant data blocks are deleted [3].

1.2.2 Place to Perform Deduplication

1. **Source or Client side:** Deduplication process to done at the origin side of the data i.e. source or client side.
2. **Target side:** Deduplication process to done at the target side of the data i.e. where client wants to store data blocks [3].

1.2.3 Chunking Algorithms

1. **Hash Based Chunking:** In this hashing algorithm is used to find hash value of data chunks. Hash value of data chunk is called as fingerprint of that data chunk.
    a. **Fixed Length Chunking:** In this fixed reference window is used for deduplication process. It achieves less deduplication amount as compared to the variable length chunking.
    b. **Variable Length Chunking:** In this variable length window is used which is much effective than the fixed length chunking.
2. **Content Based Chunking:** In this data is considered as an object. One object is compared with all other objects to locate the redundant object [3].

1.3 Big Data

Generation of high amount of data on daily basis has gave birth to a new concept named as Big Data. In recent years' concept of Big Data captured, the attention of almost every industry and government because in near future it is going to be a big issue in terms of data storage and data analysis. Many Scientists have given different definitions of big

data but common statement which comes out is that ***"Big Data is a group of bulky data sets which is almost difficult to handle or process by using traditional resources"*** [4].

According to the Apache Hadoop Big Data is defined as the dataset which cannot be processed, managed and captured by ordinary computers. It couldn't be stored and managed with the standard database management programs because it consists of large amount of data, large variety of data including most of the unstructured and semi-structured data. Cloud Computing, Internet of Things, Hadoop, Data Center are the main technologies which are concern with big data [1].

Big Data concept comes under consideration few years back when internet, smartphones, PDAs and sensors were used commonly by every person. Everyone is concern with the protection and safety of their data and want to be last long with it, so they used to take backup of data and store multiple copies of it so that it can be easily accessible. This thinking gives the birth to high data generation rate. After this global revolution arrives in the form of social networking, it connects millions of people to each other no matter how geographically far they are from each other. People used to share their events in the form of photographs, videos, text stories etc. which comes in front in the form of Big Data.

1.3.1 Value Chain of Big Data

Big Data can be easily visualized by the value chain of Big Data. It includes mainly four phases:

1. **Data Generation:** In this phase data generation takes place. Data could be generate using mobile phones, cameras, computers in the form of photographs, videos, text etc.
2. **Data Acquisition:** This phase has the concern with collection of the generated data from numerous sources.
3. **Data Storage:** After gaining data, data should be stored in memory which takes place in this phase.
4. **Data Analysis:** In this phase processing takes place in order to achieve or generate meaningful information from collected data [1].

As an estimate almost 90 per cent of data of this era is generated in past 5-7 years and generated by the sensors used for weather statistics, posts on social networking sites, digital cameras, mobile phones, news posts etc. This much amount of data is almost non-processable by using traditional resources.

CHAPTER-2

LITERATURE SURVEY

This chapter pertains to the discussion and analysis of various studies done on theoretical framework of Data Deduplication and Big Data. It reviews the views of different authors on the subject area. Deduplication of data becomes a key research area of the computer industry because of the explosive increase of need of the storage space. In big data scenario every organization need to store huge amount of data to provide better services and attract their customers to generate high revenue. In this segment various studies have been inspected and a portion of the papers have been viewed as that has been taken as inspiration towards the study. The chapter has been divided into two sections. First section contains review related with Data Deduplication and second part cover the review related with Big Data.

2.1 Review Related with Data Deduplication

2.2 Review Related with Big Data

2.1 Review Related with Data Deduplication

Vikraman et.al (2014) conducted study on various data deduplication systems and process. This study described about the two ways to perform data deduplication i.e. inline deduplication and post process deduplication, also examined the places where data deduplication can be applied i.e. source or client side and target side. In data deduplication, process of chunking played a vital role. This study provides brief knowledge about different types of chunking such as fixed length chunking, variable length chunking and content based chunking. After chunking, indexing process described which generates fingerprints of every chunk and store into the index table. This study concluded that still many challenges present in data deduplication technique which could be taken for future research [3].

Zhu et.al (2012) proposed a backup method based on intelligent data deduplication entitled as BackupDedup. The study described an intelligent deduplication system which

used four different strategies. BackupDedup system used to choose the appropriate strategy according to the file type and application context. The results of study showed that BackupDedup utilizes multi de-duplication procedures all the while to significantly dispose of redundant information in the backup process in order to achieve the objective of adequately sparing storage space and system data transmission [5].

Kurav et.al (2015) proposed a parallel architecture for inline data de-duplication using SHA-2 hash. This study considered past research on data deduplication and also proposed an answer which was Parallel architecture for inline data deduplication which utilizes the secure hash algorithm 256 for performing information deduplication job remembering the objective to conquer the issues of time and to diminish hash collisions. Study concluded that for productivity and time assessment, delete and write operations are performed and helpful for storage servers where a huge sum of files has been saved and software industries dependably searches for new advancements so they can stay up with the latest and free for effective usage of the server nodes [6].

Sun et.al (2010) analysed the information backup and recovery in light of data deduplication. This study contrasted data de-duplication and other information storage techniques, investigations attributes of data de-duplication and applies the innovation to information backup and recovery. The concentrate additionally highlighted the exceptional procedure of asynchronous backup and recovery in light of information deduplication and concluded that information de-duplication can lessen backup volume, then spare client's information storage space, and cut the cost of storage capacity, asynchronous backup makes it more solid and controllable utilized as a part of design process [7].

Fu et.al (2016) studied the reducing fragmentation for In-line Deduplication backup storage via exploiting backup history and cache knowledge. According to this study fragmentation analysed in sparse and out-of-order containers. Sparse container responsible for low data restore performance and efficiency for collection of garbage information and out-of-order container reduce performance of restoration process. For reduction of fragmentation, study proposed History-Aware Rewriting algorithm (HAR) and Cache-

Aware Filter (CAF). HAR exploited ancient data in backup frameworks to accurately recognize and decrease sparse containers, and CAF exploits restore cache information to recognize the out-of-order containers that upset the restore routine. And concluded that CAF proficiently supplements HAR in datasets where out-of-order containers were prevailing, to lessen the metadata overhead of the garbage collection. Results showed that from practical world datasets indicate HAR essentially enhances the restore execution by 2.84-175.36 at a cost of just modifying 0.5-2.03 percent information [8].

Kumar et.al (2016) proposed technique that analyzed data set using Hadoop tool. The study examined that when the chunks have been obtained then these chunks are given to the MD5 algorithm module to generate hash values for the chunks. The study compared the duplicate hash values of MapReduce model with MD5 algorithm module to generate hash values i.e. already stored in bucket storage. The concluded that if these hash values are already present in the bucket storage then these can be identified as duplicate and if the hash values are duplicated then do not store the data into the Hadoop Distributed File System (HDFS) else then store the data into the HDFS. The proposed technique is analyzed using real data set using Hadoop tool [25].

Yang et.al (2008) introduced the new Backup Scheme with Data Deduplication Ability i.e. FBBM. Study concluded that in FBBM anchor-based chunking used to partition the data files in a variable sized chunks of data with the end goal of duplication recognition. Write once RAID (Redundant Array of Inexpensive Disks) has been used to store the data chunks and for keeping track of their indices and addresses after taking hash of their substance, this prompts to characteristically single instance storage and old data backup techniques beats by FBBM in terms of storage capacity and data transmission sparing. Result demonstrated that this technique is a promising system for current ventures to store or file their speedy developing important information with low storage and data transfer cost [9].

Wang et.al (2015) examined the I-sieve technology for an Inline Elite Deduplication framework utilized as a part of cloud storage. Study demonstrated that the

goal of I-sieve is to realize a high performance data sieve system based on iSCSI in the cloud storage system and designed the corresponding index and mapping tables and used a multi-level cache using a SSD to reduce RAM consumption and to upgrade query execution performance. A prototype of I-sieve is implemented based on the open source iSCSI target. Results of study showed the excellent deduplication and foreground performance and more importantly, I-sieve can co-exist with the existing deduplication systems as long as they support the iSCSI protocol [10].

Zhu et.al (2014) studied on data de-duplication on similar file detection. The study revealed that there has been existence of many problems in data de-duplication on block level management of metadata and read/write ratio. To accomplish higher deduplication elimination ratio, the conventional scheme is to grow the range of information for data deduplication, but that would make metadata fields bigger and increment the quantity of metadata entries. The study concluded that the when identifying the duplicated pieces of data, metadata should be continually move in and out into the memory of server and access blockage has been created and required to spot related files to categorize valued data aimed at deduplication. Study concentrate on another technique for block level data deduplication joined with parallel file record detection at the time of ensuring the deduplication elimination ratio [11].

Al-Janabi et.al (2016) presented a study on A Density-based Data Cleaning Approach for Deduplication with Data Consistency and Accuracy. The study uncovered that cleaning of information is a basic part of the data conversion phase at high amount of data storage media i.e. data warehousing where the mined data from relational databases are generally impure and unclean. This study introduced a uniform system and algorithms to incorporate data deduplication with conflicting data repairing and finding of the correct values in information and used the embedded density data in information to correct mistakes in light of data density where tuples that are near each other are pressed together. The study considered that the inconsistent or dirty data in terms of damages with respect to a set of functional dependencies (FDs), as these damages are common in practice [12].

Zhang et.al (2016) pointed out study on fast asymmetric extremum content defined chunking algorithm for data deduplication in backup storage systems. The study exhibit that chunk level deduplication assumes a vital part in backup storage frameworks. Other algorithms which also used for Content-Defined Chunking (CDC), experience the key difficulties, for example, bottleneck problem caused by very low chunking throughput, low quality chunking because of huge chunk size and unable to find exact chunk boundaries in low-entropy strings. This study tended to these difficulties and proposed another CDC algorithm called the Asymmetric Extremum (AE) algorithm. The result of study depicted that proposed system has given higher chunking output, ready to discover appropriate chunk limits in low-entropy strings and accurate chunk size. Additionally, AE enhances the throughput execution of the cutting edge CDC calculations by greater than 2.3X, which proved itself as a perfect hammer for the bottleneck problem in data deduplication process, and output speed of system increased by more than half, while accomplishing practically identical deduplication efficiency [13].

Krishnaprasad et.al (2013) has given a proposal for improving data deduplication with dual side fixed size chunking algorithm. Deduplication is the method to free the storage space by saving just a single instance of data on the storage media and vanish rest all duplicated data. Address of that single instance of data stored in table of indexes for future reference. Hash values are calculated as the end goal to recognize duplicated data. Fixed sized chunking (FSC) breaks file into fixed length pieces by starting this process from starting of the file. The primary disadvantage in this strategy is identified when new data bytes added into the middle of file, as the result all other data chunks get shifted from their positions. This generates another hash value and results as the less deduplication ratio. To get rid of this disadvantage, calculate hash values from ending point of the file along with starting point and store these hash values in separate metadata table. This study proposed new algorithm named as 'Double Side Fixed Size Chunking' in order to achieve better deduplication ratio. Instead of using costly variable size chunking for videos and sound records this algorithm could be the better replacement. This data removing will provide higher amount of space to store data records and save network transmission cost [14].

Fang et.al (2016) has given the operational RAID data design for object based deduplication reinforcement framework. The study proposed a substitute Redundant Array of Independent Disks (RAID) data design, Asymmetrical grouping data organization (AGDO), for object based data deduplication backup framework. The execution of AGDO has been assessed and ended up being adequate for the nonstop storage application. The outcome demonstrated that disk accesses were focused in a part of the disk over quite a while period and lessens the power utilization to 25 per cent in a 10-disk configuration. Also, object based deduplication consolidated with AGDO has awesome potential in expanding data restoration speed for compound documents and this mix makes normal restoration speed enhanced 11 per cent. Study concluded that object based data deduplication given the powerful solution for recognizing redundant data for compound files [15].

Sobe (2016) investigated that data deduplication algorithm has extended to be executed on a group of computer nodes and that deduplication algorithm detects equal blocks on different storage nodes and uses this information for deduplication, but also to keep block replicas in the group. Study described the protocol and report on experiments that demonstrated the feasibility of the approach, and quantify the cost of coordination [16].

2.2 Review Related with Big Data

Zhou et.al (2013) observed that the interest for data storage and processing is expanding at a fast speed in the big data scenario. The study revealed that such an incredible amount of data marks the limit on storage capacity and on the storage arrange and a critical part of the dataset in big data workloads is duplicate. The study demonstrated that as an outcome of deduplication innovation, which removes redundancy, turns into a great solution for spare storage space and network bandwidth in big data environment. The study inspected that the redundancy of big data workloads to confirm the requirement for deduplication, furthermore broke down the performance and energy affect brought by deduplication under different big data situations. The study identified three birthplaces of duplication in big data workloads: 1) deploying more nodes, 2) expanding the dataset, and

3) using replication mechanisms and concluded that level of repetition has been important workload characteristic that has effect on the productivity of deduplication in big data environment. The advantage of deduplication lessens when more SSDs are installed into the Hadoop framework. In a pure SSD scenario, deduplication can support squeeze more VMs at the cost of performance and energy overhead [17].

Kalota (2015) analysed the applications of big data in education. The study depicted that Big Data and analytics has been picked up an immense focus in recent time. Big Data feeds into the area of Learning Analytics (LA) that permitted scholarly foundations to better know the learners' needs and proactively address them. The study uncovered that it was necessary to have an understanding of Big Data and its applications. The study gave technologies utilized as a part of Big Data, and a portion of applications of Big Data in education, furthermore talked about areas of Big Data and present research scenarios. While Big Data can give the big advantages, for institutions to know their own needs, framework, assets, resources and constraint before moving on the Big Data trend [18].

Munir et.al (2015) investigated the big challenges to privacy and data protection for big data. The study examined the advantages of big data and all the more importantly the difficulties that have concern to the subject of security and data privacy. The study concluded that first essential part of data has been the way of big data considered before showing the capability of big data in the present days. Thereafter, the issue of security and data privacy has been in lighten before talking about the difficulties of implementing this issue in big data. What's more, study concentrate additionally on put forward the discussion on the capability of the current legal structure in securing individual data in the environment of big data [19].

Vidhya et.al (2014) looked at the diverse collection of big data analytics tools. The study showed that as big data has been got a lot of attention for their great and good reasons. Because of the expansive and complex pool of datasets it is hard to handle on traditional data processing tools. Moreover, the fundamental point of big data analytics to use the innovative analytic tools other than very large, different datasets which consist diverse

sizes from tera bytes to zetta bytes and different types, for example, organized or unorganized and batch or streaming. The study determined that big data has been valuable for data sets where their size or type is far from the capacity of traditional relational databases for handling, managing and capturing the data with low latency [20].

Tang (2014) led a study on big data cleaning. Data cleaning idea, an enthusiastic subject that has played a crucial part in the historical background of data management and data analytics, and as yet experiencing quick advancement. The study portrayed that data cleaning measured as a fundamental challenge in the time of big data, because of the expanding volume, speed and variety of data in numerous applications. The study gave different parts of data cleaning, for example, error discovery techniques, data repairing systems, and a common data cleaning framework and consist some discussion about our struggles of data cleaning strategies from the point of view of big data, in terms of volume, velocity and variety [21].

Chen et.al (2014) surveyed on big data. Study presented technologies, for example, distributed cloud computing, Internet of Things, data centers, and Hadoop for taking care of big data and concentrated on the four levels of the value chain of big data, i.e., generation of data, acquisition of data, storage of data and analysis of data. The study analyzed the few representative uses of big data, including enterprise administration, Internet of Things, online social networks and smart grid and gave a complete review and big-idea to readers of this sensational area, likewise talked about the open issues and future tips on big data [1].

Kakhani et.al (2013) investigated the research concerns in Big Data. The study uncovered that Big Data has attracted a great consideration from the educational world, industry and government and very difficult research zone in present situation. The study concluded that Academia and industry needs to cooperate to plan and grow new techniques and tools which excellently handle the preparing of big data and quick need of new machine learning and data mining systems to analyzed huge amount of data in near future that

require new technologies and methods to process, manage, store and investigate Big Data [4].

Gao et.al (2016) examined the issues, difficulties, and necessities of big data validation and quality assurance. The study analyzed that the quick progress of big data and analytics, big data is turning into an exceptionally great research and application area in scholastic research, industry, and government and expanding data quality issues bringing about incorrect data costs in enterprises and organizations. Current research sometimes talks about how to adequately approve big data to guarantee data quality. The study concentrate on big data validation and quality assurance, including the fundamental ideas, motivations, and validation process. The study finished up to guarantee data quality and to adapt to the risk issue, urban communities need big data quality certification done before data storing. The last group comprises big data clients, including researchers, business decision makers, and big data application merchants [2].

Verma et.al (2016) has made evaluation between Big Data Management Processing with Hadoop MapReduce and Spark Technology. The study observed that analysis of huge size of data from various resources has unproductive execution in Hadoop MapReduce. The study demonstrated that correlation for Spark has great utilizing as a part of memory preparing and quick, also concluded the Spark's capacity to perform batch processing and machine learning, additionally it has system for a preparing huge number data [22].

Kanchi et.al (2015) discussed the difficulties and solutions in big data management. This study observed a blast of data in numerous associations on the planet. Industry analysts and organizations looking towards big data as the future big thing to give openings, experiences, solutions and another approach to expand benefits in business. From public networks to records in a health care centers, big data has assumed a critical part to enhance businesses and development. Organizations struggle to get quality data and recover them for data analysis and business purposes. The study clarified that appropriate data management practices, systems, techniques and framework vanish these difficulties, issues and problems and concluded that through open source solutions, for example, Hadoop that

obtained the ability of adaptation to fault tolerance, adaptability and giving the advantage of running on group hardware. Another significant part of taking care of master data management gives a 360^0 perspective of the data over the whole organization. With appropriate technical and analytical practices of big data management, organizations created business values from the data. Great state of data quality and availability of business knowledge is the fundamental point of big data management [23].

Gadepally et.al (2015) has given the sampling operations on big data. This study states that the 3Vs i.e. Volume, Velocity and Variety of Big Data keeps on being a great challenge for frameworks and algorithms intended to store, handle and scatter data for discovery and investigation under real time restrictions. Study uncovered that basic signal processing operations, for example, sampling and filtering, which have been utilized for quite a long time to compress signals are regularly undefined in data that is described by heterogeneity, high dimensionality, and absence of known structure. The study depicted that the way to deal with sample large datasets, for example, online networking data furthermore evaluated the impact of sampling on a common predictive analytic: link prediction. Aftereffects of study showed that extraordinarily sampling a dataset can still yield important link [24].

CHAPTER-3

PRESENT WORK

This chapter portrays the reason justifying the selection of topic for Research. Additionally, the chapter throw a light on the research gap, problem definition, objectives of the study, research methodology and tools used for implementation.

3.1 Research Gap

As the quick growth in production of data, it becomes a matter of tension for every organization to store data and perform analysis operations on this huge amount of data. Analysis makes any organization grow in this competitive world. To make any business a success data mining and analysis plays a vital role but here comes obstacle in the form of large volume of collected data on which analysis operations to be performed. As huge volume of data contains almost 80 per cent of data in unstructured and duplicate form. Duplicate data in big data makes numerous confusions regarding storage of data in physical storage devices. As cleaned data is important for analysis, in same manner timely availability of data equally important. If data will be available, then analysis operations becomes easy to carry out.

i. High utilization of time and space in storage process of high amount of data.
ii. Less reduction in duplicated data in order to free up the storage memory and provide all time ready data.
iii. Deployment of HDFS at data storage as non-efficient manner.

In this study various data deduplication techniques and technologies being studied to find out optimal solution to reduce the duplication in big data in demand to free up the storage memory and provide all time ready data. Duplicated data put high pressure on the other factors of system such as operational speed, read/write operation, memory utilization, data transmission, power consumption, throughput and multi-tasking etc.

3.2 Problem Definition

In this era of Big Data there is a need of proper system which can handle high amount of data which is produced by the digital assistants, sensors, cameras and any other digital devices. Generation of data at such high rapidity arise a problem of storage and as well as reduction of duplication in storage media. Many researchers conducted studies to reduce the size of data by applying different methodologies and technologies but till now we are here without any stable system which can solve this problem.

This dissertation work has proposed a very effective system to manage the large sized data files in such an efficient way that it requires less amount of storage space and take less time to perform read/write operations. By using Hadoop HDFS as storage system and applying data deduplication at client and server level both storage problem can be reduced at maximum extend.

3.3 Objectives of the Study

The objectives of this study are multifold. This study aims to explore and analyze various data deduplication tools and techniques and develop an augmented technique in order to improve the data deduplication process in big data environment. New improved technique primarily focused on the large volume data in big data scenario. It will be developed in such a manner that replicated data will be removed in efficient manner and free up the storage space and improve the system performance with respect of time and read-write operations.

The research is focused on following objectives:

1. To reduce the storage size for files saved in big data scenario by removing duplicate data.
2. To deploy an augmented inline data deduplication technique in big data environment.
3. To compare the proposed approach with the existing by using the following parameters:-
    a) Data size after deduplication
    b) Deduplication ratio

c) Hash Time (MB/Sec)

d) Chunk Time (MB/Sec)

As, in today's information driven age, more and more unstructured data received again and again from different data sources for both transactional and analytical purposes, it is very important to improve storage capacity of systems to achieve higher performance by effectively utilizing the existing infrastructure as well as by adding new infrastructure.

3.4 Research Methodology

After review of approaches and techniques in the area of data deduplication it has been observed that there is more improvement required to achieve higher ratio of deduplication in the vast area of big data. In order to handle unstructured data generated by different sources and to free up storage space by removing duplicated data the following methodology will be implemented.

3.4.1 Connection Establishment Phase

1. Backup server sends message to application server about the scheduled backup for particular files, file groups, databases.
2. Backup server sends message to Name Node/Primary Storage Node to prepare disks for storage of incoming data [17].
3. Application server prepare files to take backup and calculate size of files.
   - 3.1. If file size is < 1GB then host itself applies data deduplication on files.
   - 3.2. If file size is > 1GB then send message to backup server about the size of files and unable to apply deduplication on this huge size.
     a. Backup server intimate to name node about the size of files and get ready to apply deduplication process and store data.
     b. Name node send a ticket to backup server having number of data nodes/secondary storage nodes and their addresses which will participate in deduplication process [17].
     c. Backup server send ticket to application server having number of data nodes and their addresses which will participate in deduplication process.
4. Hence connection is established [Figure 1].

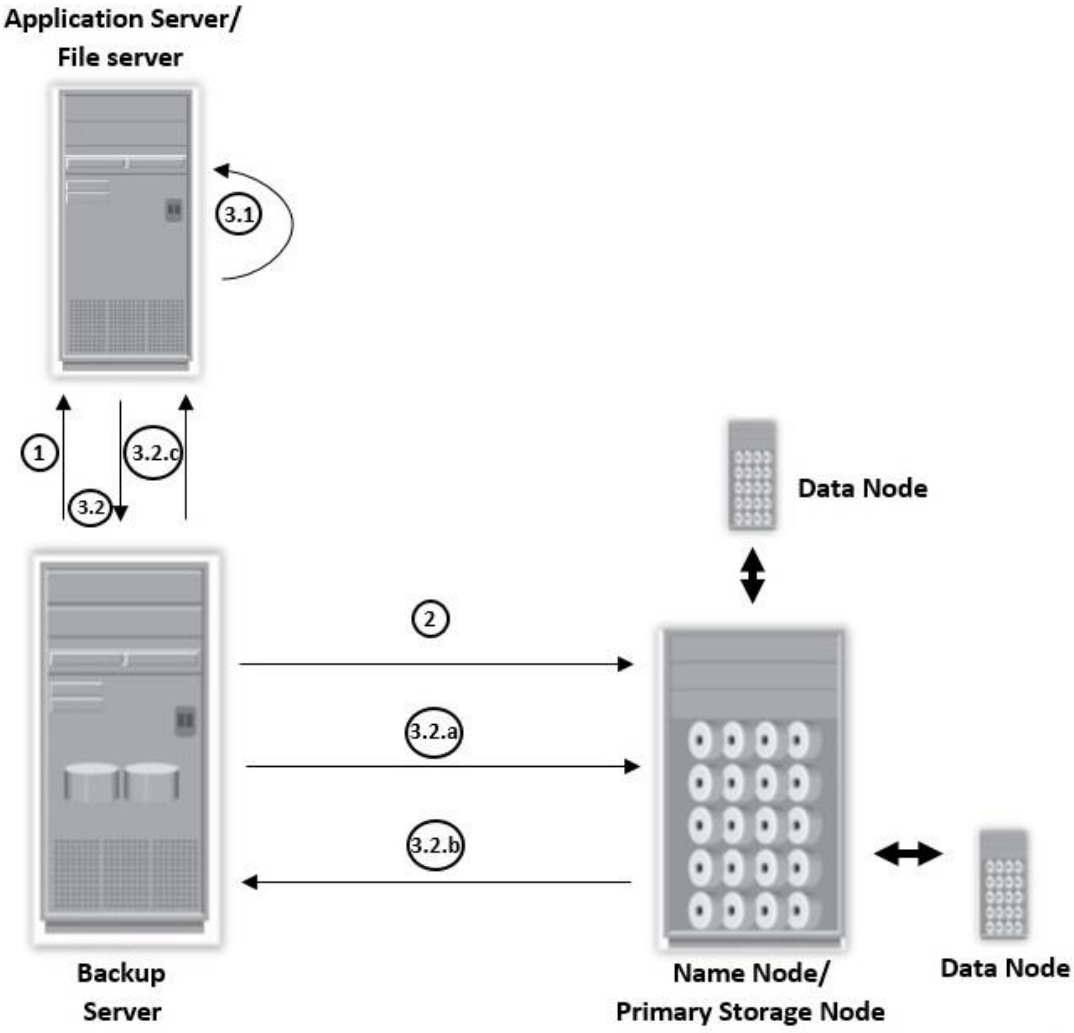


Figure 3.1 Connection Establishment between application server, backup server and primary name node

3.4.2 Chunking Phase

1. Application server divides files into fixed length chunks according to the number of data nodes which will participate in deduplication process.

$$No.of\ chunks = \frac{Total\ size\ of\ files}{No.of\ data\ nodes}$$

3.4.3 Data Transfer and Data Deduplication Phase

In this step, data deduplication and data transfer will be applied. A global hash index table will be shared by application server and secondary data nodes in order to apply deduplication process and to match hash signatures of file chunks. Global hash index table will be stored and managed by the primary data node.

For Pre-Process Data Deduplication and Data Transfer

1. If file size is < 1GB, then data deduplication process applied by application server.
2. Application server send deduplicated data to primary data node which stores that preprocessed deduplicated data into a dedicated secondary data node [Figure 2].

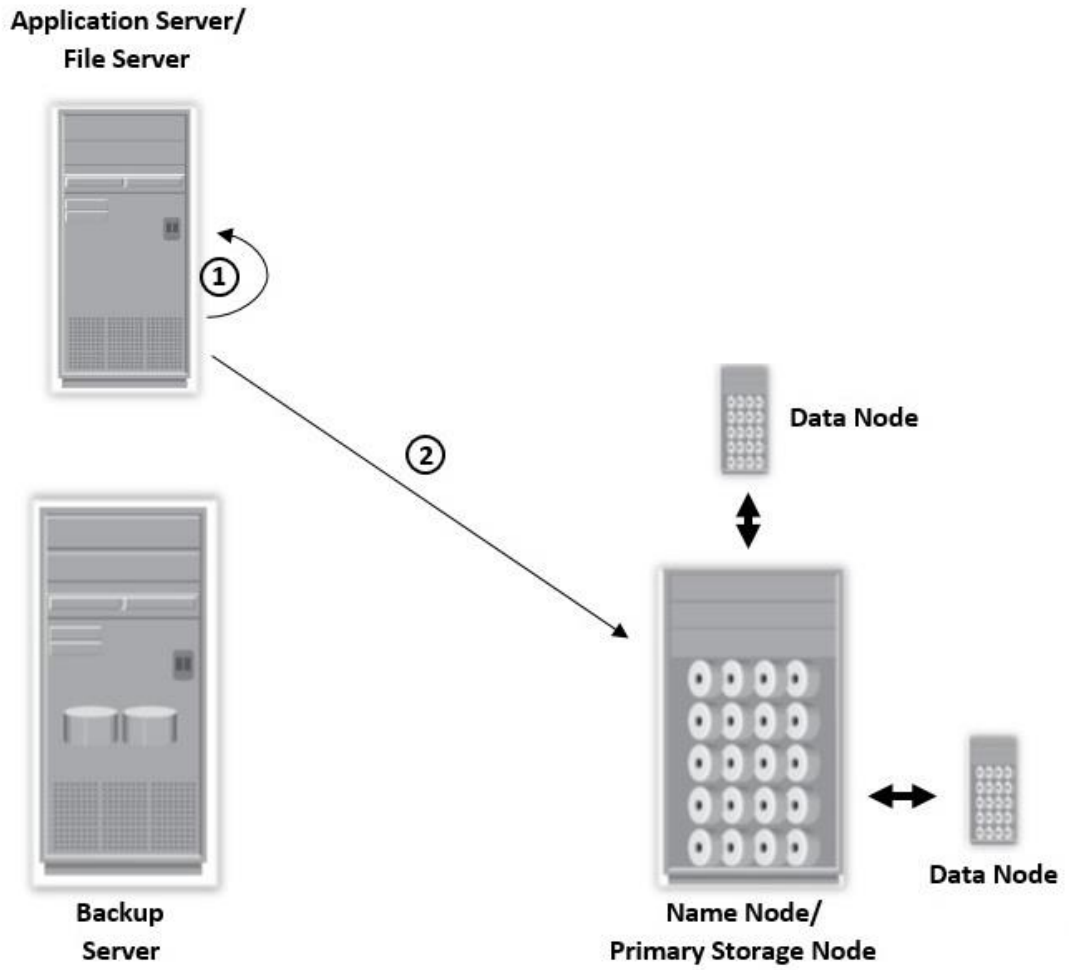


Figure 3.2 Pre-Process Data Deduplication and Data Transfer

For Post-Process Data Transfer and Data Deduplication

In this file type based data deduplication will be applied on data chunk.

1. If file size is > 1GB, then application server applies chunking process on files and transfer chunks to the secondary data nodes (which addresses were given by backup server to application server) [Figure 3].

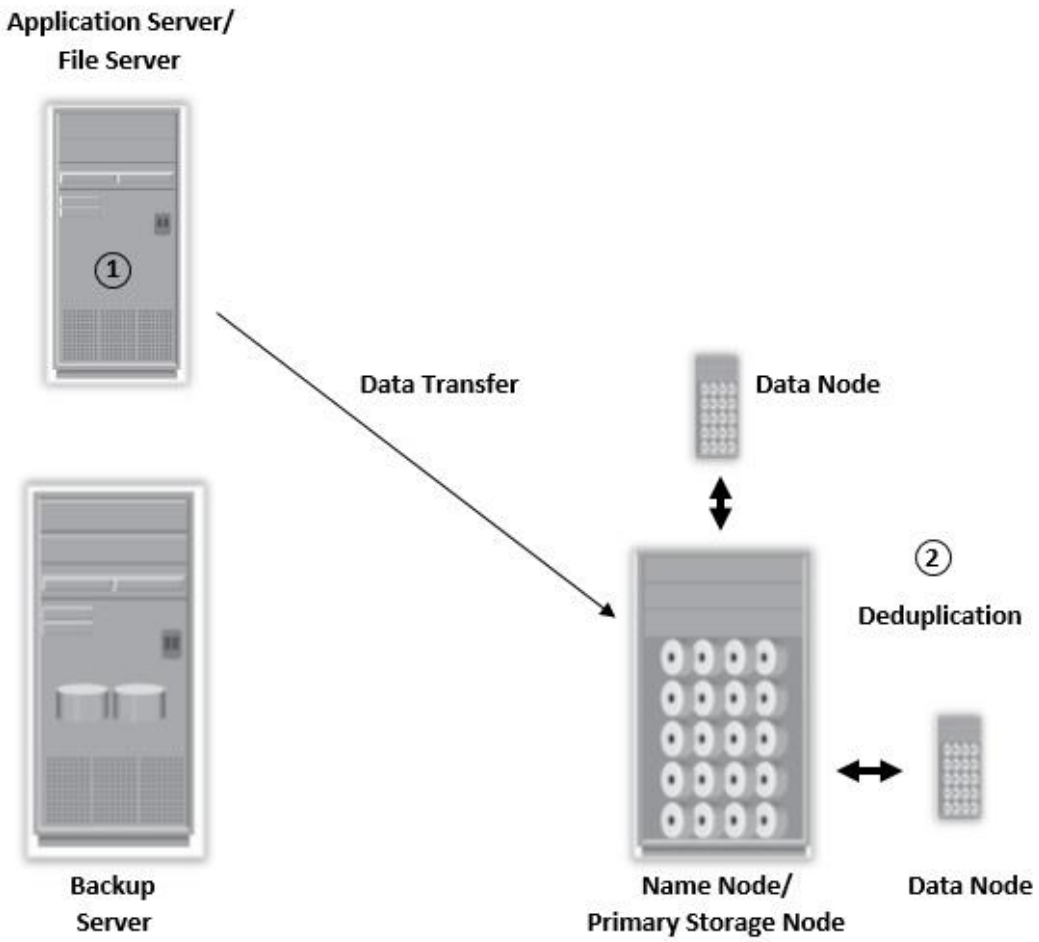


Figure 3.3 Post-Process Data Transfer and Data Deduplication

3.4.4 Acknowledgement Phase

1. After apply post-process data deduplication, secondary data nodes send acknowledgement to the primary storage node.
2. Primary node will pass acknowledgement to backup server.

3. Backup server will update its catalog and send acknowledgement about the duplicated data to application server, which is currently reside on it.
4. Application server applies flushing process on duplicated data and free up their storage space [Figure 4].

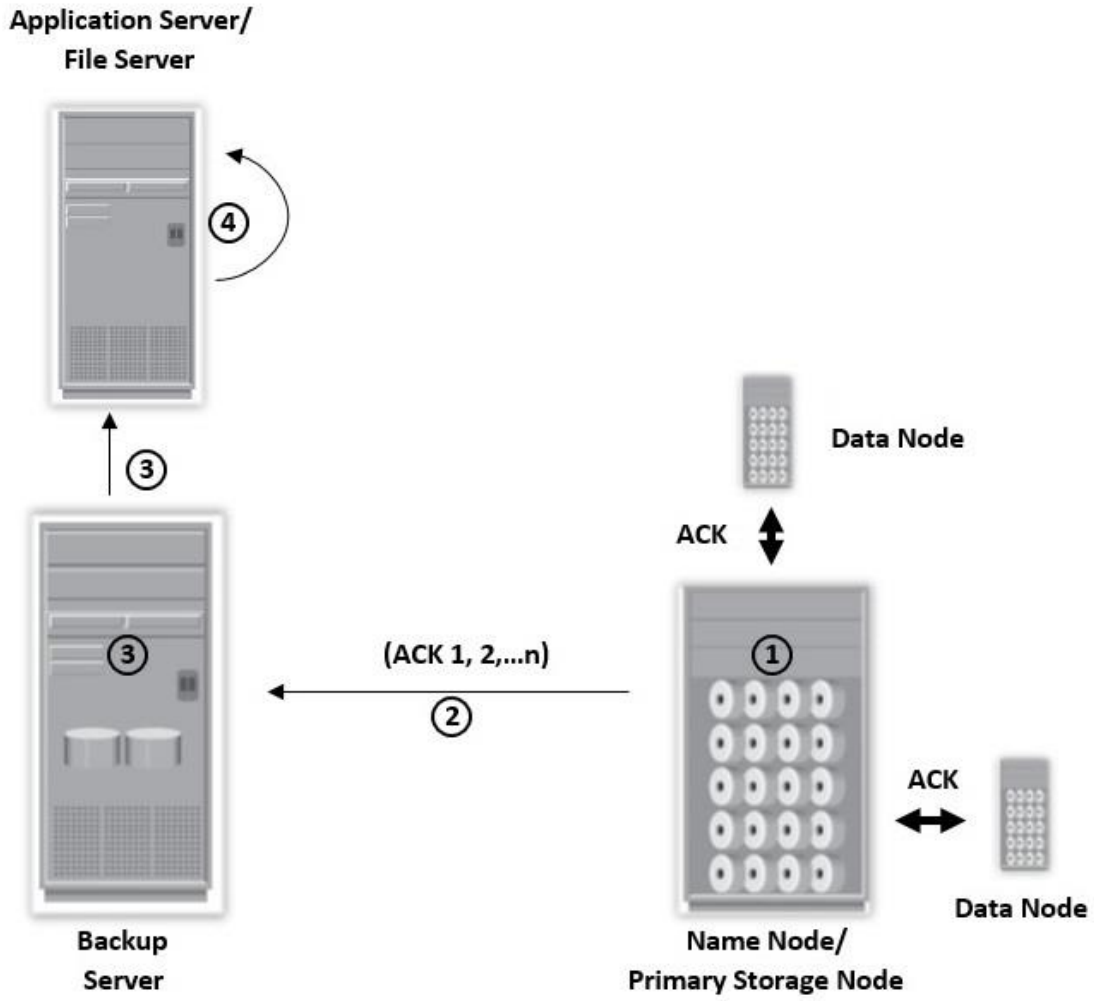
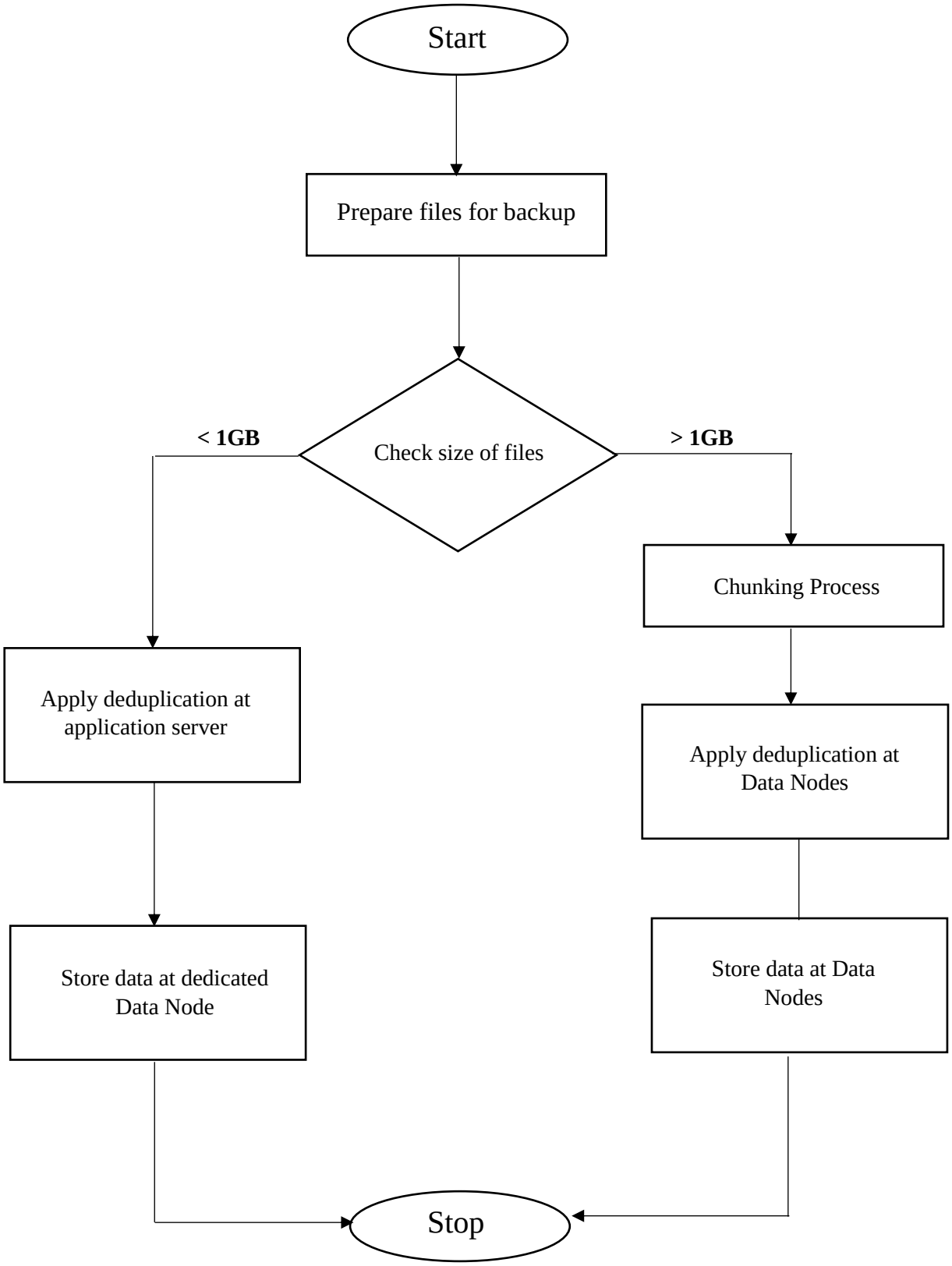


Figure 3.4 Acknowledgement process

This acknowledgement process will also be applicable for pre-process data deduplication, in order to update catalog of backup server.

3.4.5 Flowchart of Process



3.5 Tools Used

In implementation of proposed system following tools are used:

3.5.1 Java Programming Language

Java is an object oriented programing language which enables the programmers code the program at one system and use it anywhere. It is platform independent programing language, means after compilation of written code that code can be run at any platform without recompilation. It is developed by Sun Microsystems and released in 1995. Java applications can be deployed on any machine which runs JVM (Java Virtual Machine).

3.5.2 Hadoop

Apache Hadoop is an open source software framework used to store high amount of data in an efficient manner. It core is known as HDFS (Hadoop Distributed File System). In order to manage large data Hadoop split data files into file blocks and spread them across the nodes in a cluster.

3.5.3 Shell Script

Shell Script is a scripting language in which computer programs are designed to be run by Unix Shell. It group large and repetitive code into a small and simple script which can be executed at any point of time.

CHAPTER-4

RESULTS AND DISCUSSION

This chapter assesses the result obtained after implementation of proposed methodology and discussion on performance enhancement of data deduplication in big data environment. Two data set have been used i.e. greater than 1GB (>1GB) and less than 1GB (<1GB) for testing the performance of proposed system. Performance of data deduplication in big data environment have been measured on the ground of four parameters i.e. Data size after Deduplication process, Deduplication Ratio, Hashing (MB/Sec) and Chunking (MB/Sec).

Deduplication process is used for reducing the size of data stored in database. On the other hand deduplication ratio shows the percentage of reduced data after applying deduplication. Hashing implies data in terms of MB per sec. for which hash signature is generated and chunking measures data in terms of MB per sec. which is divided into smaller chunks. Table 4.1 made a comparison among the above mentioned parameters for two different data sets.

Data size before Deduplication (GB)	Techniques	Data size after Deduplication (GB)	Deduplication Ratio	Hashing (MB/Sec)	Chunking (MB/Sec)
0.840	Fixed sized	0.804	0.044	7.173	3.586
	Bucket Based	0.799	0.050	286.925	286.925
	Proposed	0.502	0.403	5.518	11.326
1.700	Fixed sized	1.470	0.136	14.506	7.253
	Bucket Based	1.020	0.400	870.400	870.400
	Proposed	1.020	0.400	1740.800	1740.800

Table 4.1 Results of Existing and Proposed System

4.1 Data Size after Deduplication (GB)

The table 4.2 depicts the results obtained after the implementation of proposed idea for two data set. In case of <1GB data set, fixed sized technique reduced the amount of data from 0.840GB to 0.804GB. On the other hand, bucket based technique reduced data set from 0.840GB to 0.799GB, but proposed system reduced more data size as compared to fixed sized and bucket based techniques (see table 4.2).

Data Size	Initial Data Size	Fixed sized	Bucket Based	Proposed
< 1GB	0.840	0.804	0.799	0.502
>1GB	1.700	1.470	1.020	1.020

Table 4.2 Data Size after Deduplication (GB)

It is clear from table 4.2 that for first data set i.e. >1GB data size reduce from 1.7GB to 1.470GB for fixed sized technique. For bucket based technique data size reduce to 1.020GB. This amount is same for proposed technique because both the techniques used same tool for analysis i.e. HDFC. In >1GB data set hashing and chunking (MB/Sec) is more as compared to previous one i.e. it increase from 14.506 to 1740.800 in proposed system.

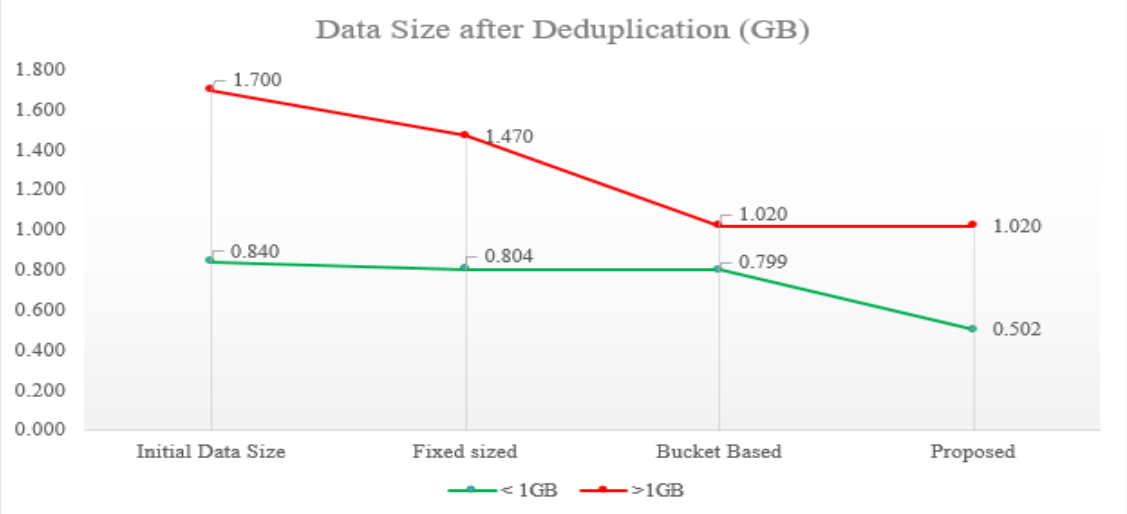


Figure 4.1 Data Size after Deduplication (GB)

4.2 Deduplication Ratio

The table 4.3 compares the deduplication ratio of proposed system with fixed sized and bucket based system for two different data sets. In case of <1GB data set, deduplication ratio for fixed sized technique was 0.044 per cent and deduplication ratio for bucket based technique was 0.050 per cent. It has been clear from table 4.3 proposed system having a more deduplication ratio than fixed sized and bucket based system i.e. 0.403 per cent.

Data Size	Fixed sized	Bucket Based	Proposed
< 1GB	0.044	0.050	0.403
>1GB	0.136	0.400	0.400

Table 4.3 Deduplication Ratio (in per cent)

On the other hand, for second data set i.e. >1GB, fixed sized system's deduplication ratio was 0.136 per cent. This ratio was less than bucket based and proposed system.

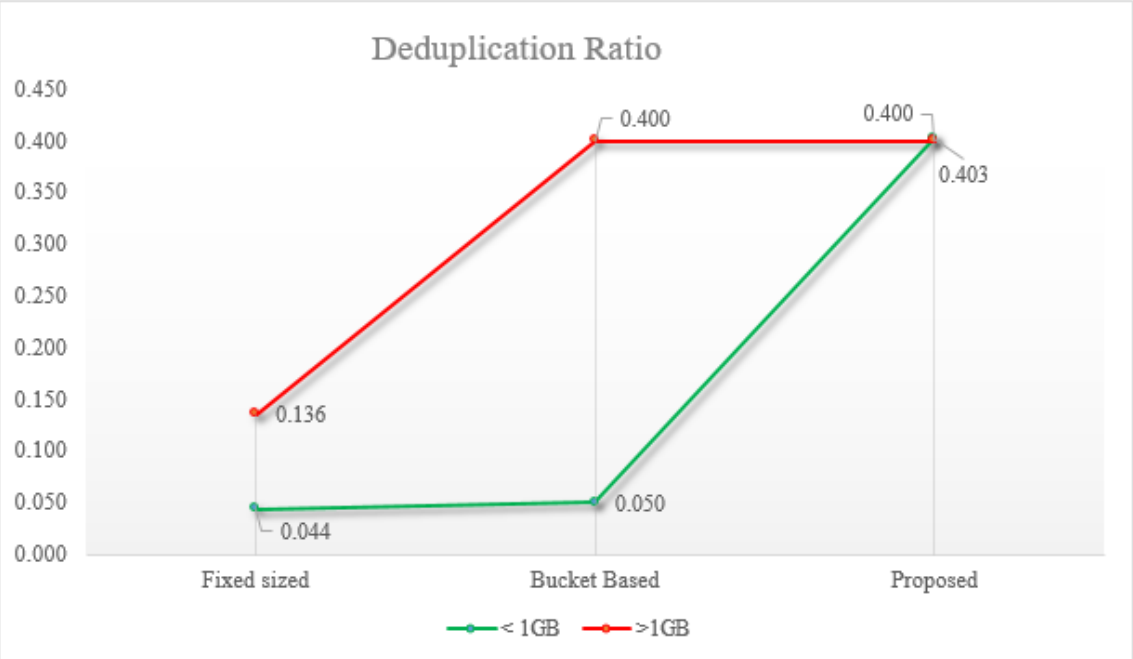


Figure 4.2 Deduplication Ratio (in per cent)

4.3 Hashing (MB/Sec)

The table 4.4 shows the hashing for all the systems. In case of <1GB data set, fixed size system generates hash signature of 7.173 MB/Sec and bucket based generate 286.925 MB/Sec. In contrast with >1GB data set, proposed system generating more hash signatures as compare to fixed size and bucket based systems i.e. 1740.800 MB/Sec. (see table 4.4).

Data Size	Fixed sized	Bucket Based	Proposed
< 1GB	7.173	286.925	5.518
>1GB	14.506	870.400	1740.800

Table 4.4 Hashing (MB/Sec)

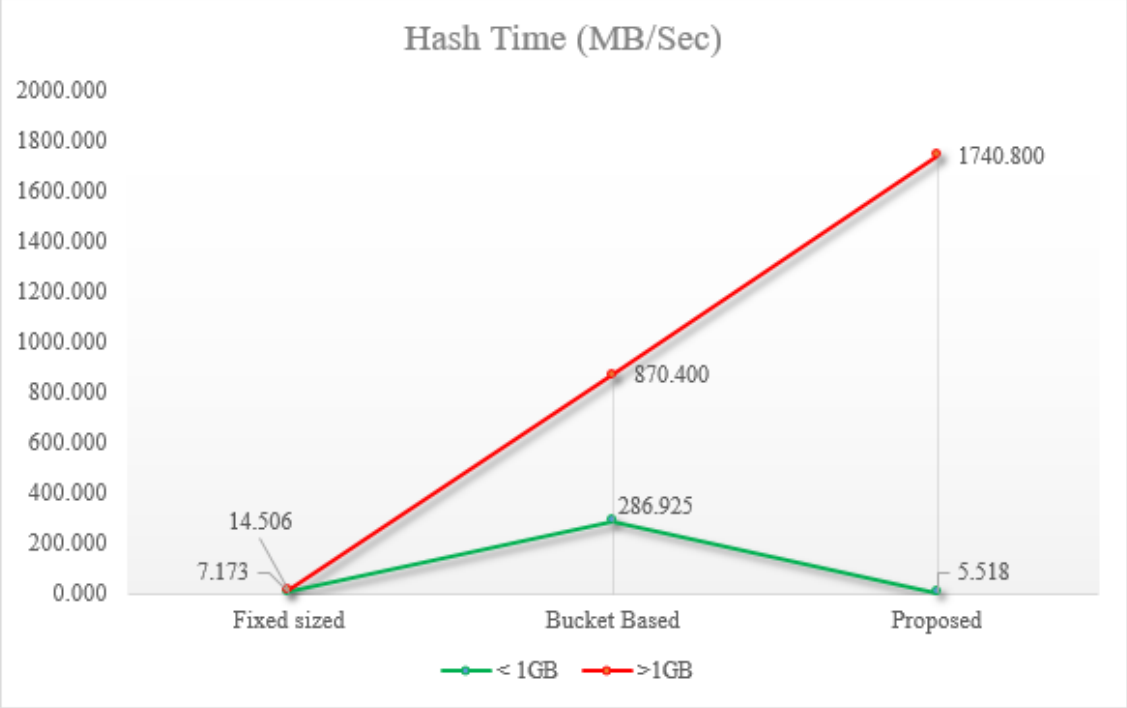


Figure 4.3 Hashing (MB/Sec)

4.4 Chunking (MB/Sec)

The table 4.5 illustrates that the chunking for all the systems. In case of <1GB data set, fixed size system generates 3.586 MB/Sec, bucket based generate 286.925 MB/Sec and proposed system generates the 11.326 MB/Sec.

Data Size	Fixed sized	Bucket Based	Proposed
< 1GB	3.586	286.925	11.326
>1GB	7.253	870.400	1740.800

Table 4.5 Chunking (MB/Sec)

In case of >1GB data set, proposed system generating more hash signatures as compare to fixed size and bucket based systems i.e. 1740.800 MB/Sec.

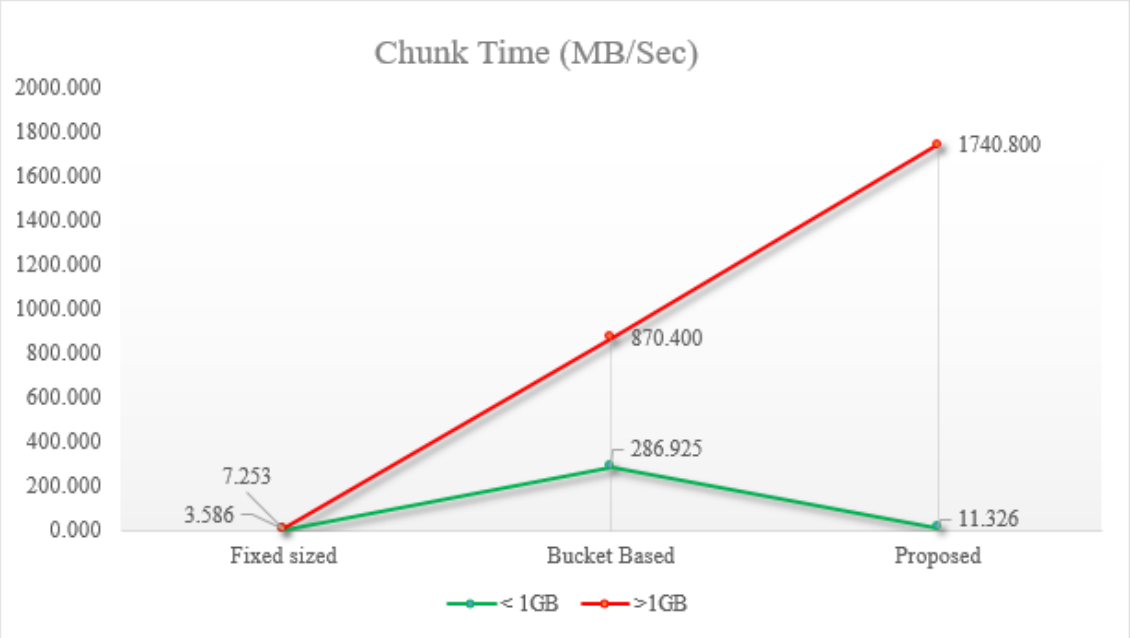


Figure 4.4 Chunking (MB/Sec)

Performance of data deduplication in big data environment have been measured on the ground of four parameters i.e. Data size after Deduplication process, Deduplication Ratio,

Hashing time (MB/Sec) and Chunking time (MB/Sec). Deduplication process is used for reducing the size of data stored in database. Hence, with implementation of prosed system, it help to reduced storage space consumption, enhance the system performance and also reduce cost of data storage in big data environment.

CHAPTER-5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

Duplication storage of data give security against data loss but it pushes high pressure on the storage systems. Storing multiple copies of same data at different places has concern with data security but in single storage instance presence of duplicated data makes no sense. As the solution to problem, present work proposed a technique which is useful for reduced storage space consumption, enhance the system performance and also reduce cost of data storage in big data environment. It is clear from the results that performing data deduplication on big data increase the performance of system in terms of deduplication ratio, hashing, chunking and storage space utilization. The study also provides efficient usage of computing resources with cutting the storage cost of data.

5.2. Future Scope

In present study transmission cost and network consumption are not taking into consideration while implementation of proposed system. But these are important factors for transferring large amount of data over network from one place to another place. In addition, the system can be implement in such a manner where costly hardware should not be required. Various existing algorithms might be proved one the best approaches among evolutionary algorithms to attain fruitful optimized results.

REFERENCES

[1] M. Chen, S. Mao and Y. Liu, "Big Data: A Survey," in *Mobile Networks and Applications*, 2014.

[2] J. Gao, C. Xie and C. Tao, "Big Data Validation and Quality Assurance-Issuses, Challenges, and Needs," in *IEEE Symposium on Service-Oriented System Engineering (SOSE)*, 2016.

[3] R. Vikraman and A. S, "A Study on Various Data De-duplication Systems," *International Journal of Computer Applications*, vol. 94, no. 4, pp. 35-40, 2014.

[4] M. K. Kakhani, S. Kakhani and S. Biradar, "Research Issues in Big Data Analytics," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. 2, no. 8, pp. 228-232, 2013.

[5] G. Zhu, X. Zhang, L. Wang, Y. Zhu and X. Dong, "An intelligent data de-duplication based backup system," in *15th International Conference on Network-Based Information Systems*, 2012.

[6] N. Kurav and P. Jain, "A Parallel Architecture for Inline Data De-Duplication Using SHA-2 Hash," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 4, pp. 1306-1312, 2015.

[7] G.-Z. Sun, Y. Dong, D.-W. Chen and J. Wei, "Data backup and recovery based on data de-duplication," in *International Conference on Artificial Intelligence and Computational Intelligence*, 2010.

[8] M. Fu, D. Feng, Y. Hua, X. He, Z. Chen, J. Liu, W. Xia, F. Huang and Q. Liu, "Reducing Fragmentation for In-line Deduplication Backup Storage via Exploiting Backup History and Cache Knowledge," *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS*, vol. 27, no. 3, pp. 855-868, 2016.

[9] T. Yang, D. Feng, J. Liu and Y. Wan, "FBBM: A new Backup Method with Data De-duplication Capability," in *International Conference on Multimedia and Ubiquitous Engineering*, 2008.

[10] J. Wang, Z. Zhao, Z. Xu, H. Zhang, L. Li and Y. Guo, "I-sieve: an inline high performance deduplication system used in cloud storage," *TSINGHUA SCIENCE AND TECHNOLOGY*, vol. 20, no. 1, pp. 17-27, 2015.

[11] Y. Zhu, X. Zhang, R. Zhao and X. Dong, "Data De-duplication on Similar File Detection," in *Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 2014.

[12] S. Al-janabi and R. Janicki, "A Density-based Data Cleaning Approach for," in *SAI Computing Conference*, London, 2016.

[13] Y. Zhang, D. Feng, H. Jiang, W. Xia, M. Fu, F. Huang and Y. Zhou, "A Fast Asymmetric Extremum Content Defined Chunking Algorithm for Data Deduplication in Backup Storage Systems," *IEEE Transactions on Computers*, pp. 1-14, 2016.

[14] K. P.K and B. A. Narayamparambil, "A Proposal for Improving Data Deduplication with Dual Side Fixed Size Chunking Algorithm," in *Third International Conference on Advances in Computing and Communications (ICACC)*, 2013.

[15] Y. Fang, T. Yu'an, Z. Quanxin, W. Fei, C. Zijing and Z. Jun, "An Effective RAID Data Layout for Object-Based De-duplication Backup System," *Chinese Journal of Electronics*, vol. 25, no. 5, pp. 832-840, 2016.

[16] P. Sobe, "Combination of Data Deduplication and Redundancy Techniques in Distributed Systems," in *29th International Conference on Architecture of Computing Systems (ARCS)*, 2016.

[17] R. Zhou, M. Liu and T. Li, "Characterizing the efficiency of data deduplication for big data storage management," in *IEEE International Symposium on Workload Characterization (IISWC)*, 2013.

[18] F. Kalota, "Applications of Big Data in Education," *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, vol. 9, no. 5, pp. 1602-1607, 2015.

[19] A. B. Munir, S. H. M. Yasin and F. Muhammad-Sukki, "Big Data: Big Challenges to Privacy and Data Protection," *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering,* vol. 9, no. 1, pp. 355-363, 2015.

[20] S. Vidhya, S. Sarumathi and N. Shanthi, "Comparative Analysis of Diverse Collection of Big Data Analytics Tools," *International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 8, no. 9, pp. 1646-1652, 2014.

[21] N. Tang, "Big Data Cleaning," in *Asia-Pacific Web Conference*, 2014.

[22] A. Verma, A. H. Mansuri and N. Jain, "Big data management processing with Hadoop MapReduce and spark technology: A comparison," in *Symposium on Colossal Data Analysis and Networking (CDAN)*, 2016.

[23] S. Kanchi, S. Sandilya, S. Ramkrishna, S. Manjrekar and A. Vhadgar, "Challenges and Solutions in Big Data Management--An Overview," in *3rd International Conference on Future Internet of Things and Cloud*, 2015.

[24] V. Gadepally, T. Herr, L. Johnson, L. Milechin, M. Milosavljevic and B. A. Miller, "Sampling operations on big data," in *49th Asilomar Conference on Signals, Systems and Computers*, Asilomar, 2015.

[25] R. R. S. C. J. Naresh Kumar, "Bucket Based Data Deduplication Technique," in *5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, Noida, 2016.