International Journal Of Scientific & Engineering Research, Volume 7, Issue 10, October-2016
ISSN 2229-5518

Performance Evaluation of BigData Analysis with Hadoop in Various Processing Systems

Ms. Preeti Narooka[1] (Ph. D Student, Computer Science),
Banasthali University,
Jaipur, India
preeti.narooka@gmail.com

Dr. Sunita Choudhary[2] (Former Associate Professor, Computer Science),
Banasthali University,
Jaipur, India
sunitaburdak@yahoo.co.in

Abstract—In recent years BigData has become most popular area for research and development. In Business, Organizations need to focus towards their data-driven approach for gaining the competitive advantage. When more data is processing, interacting and integrating it provides meaningful data for making good decision. All this happened with the advent of advanced computational and storage system required for BigData [1]. On one side BigData help in giving a panoramic view on decision making, on other side it increases the performance complexity and expenditure of the process. This paper presents a study of data analysis using Hadoop on various capacities of systems. The data is being created through an algorithm and analysis is done in multiple systems to understand the limitations and constraints of systems. The experimental results helped in concluding that how the system performance can be affected using different memories & processors. The study could be further extrapolated to optimize the system performance by reducing the time complexity by controlling the RAM size in multiple systems.

Keywords—BigData; MapReduce; Hadoop Cluster; Memory; System performance.

I. INTRODUCTION

"BigData" appears to have become a buzzword overnight. It has been expected that the growth of BigData is going to increase rapidly in coming years. The media industry's requirement to store data for long time is the reason for producing big size of data on everyday basis [8]. As a result, these large data generators brought many challenges and issues [4] which also affects business intelligence technologies and that couldn't handle storing, analyzing, preparing and processing of such large volume data. With traditional database management system, it was impossible to handle petabyte size [9]. To overcome from this, market introduced new technologies which are known as BigData techniques [10].

Basically, BigData term was described initially in 1980s in both academia and industry –"handling large groups of datasets". But it was too early stage to define BigData because still people are just trying to understand its nature. More comprehensive definitions and descriptions have emerged.

There are too many definitions and literature available about BigData today. In this paper one of the definition among all which was offered by the BigData Commission at the TechAmerica Foundation in its report, "Demystifying BigData" is discussed.

A. *BigData*

"BigData is a term that describes large volumes of high-velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information"[2].

BigData concept is basically used for handling large and complex data to process and store [7]. BigData is completely different from traditional database system. It coordinates and support Cloud Computing and other emerging technologies .It uses MapReduce technology for processing the big analytical data.

B. *MapReduce*

MapReduce technology is vitally used in analysis of the BigData. It basically divides the task into number of key-value pairs. This process is termed as mapping, and it results into generation of intermediate key-value pairs. These intermediate results are further processed by the

International Journal Of Scientific & Engineering Research, Volume 7, Issue 10, October-2016
ISSN 2229-5518

reduce function to achieve the final results of BigData analysis [8].

The MapReduce definition as per BigData analytics is given below:

"The input is formed by a set of key-value pairs, which are processed using the user-defined map function to generate a second set of intermediate key-value pairs. Intermediate results are then processed by the reduce function."

MapReduce was originally developed by Google but has now been adapted by many BigData tools, among others Hadoop.

C. Hadoop

Apache Hadoop is an open source framework for MapReduce process. The use of Hadoop framework [3] focuses on computational problems and it gives bandwidth to developer to write parallel processing programs. Hadoop includes 1). Hadoop distributed file system (HDFS). 2). Hadoop MapReduce.

In this paper I section focus on the BigData usability, related technologies/ processes and the prevailing challenges, section I also guides to select the problem area in BigData analysis. Section II is about experimental set up of Hadoop framework to process the designed algorithm on single and multiple systems and finding the system parameters affecting the performance of the process. Section III compares the single system and multiple systems & justifies the usability of multiple systems. Different experiments are set by changing system parameters and results are discussed in section IV. The conclusion of the whole exercise is given in section V.

II. HADOOP FRAMEWORK AND EXPERIMENTAL SET-UP

As discussed in section-I, Hadoop is divided into two parts one is HDFS & other one is MapReduce process. HDFS stores the BigData on the system &MapReduce process the stored data with the help of mappers and reducers [11].
The overall controlling of the processing of the data in the Hadoop framework is done by jobtrackers and tasktrackers. The role of jobtracker is to communicate between HDFS and MapReduce process (Fetching the data from HDFS for processing in MapReduce) & controlling the functioning of tasktrackers. The tasktrackers are primarily involved in the running of the mappers and reducers. Since the jobtrackers and tasktrackers are mitochondria of Hadoop framework, the experiment is focused on tasktrackers. The developed algorithm is based on Map Reduce process, therefore the role of tasktrackers is prominent and the experiment is focused on its optimization.

The main objective of research discusses the applicability of Hadoop framework in the social networking sites; the processing of BigData is involved in such platforms. Various experiments were set on single & multiple systems with different configurations in term of RAM; to understand the usefulness of Hadoop framework. The experiment was set in two steps;

- Firstly a graph search algorithm is developed for keyword search operation in social network; fundamentally the algorithm is based on MapReduce search process.
- The Hadoop is an open source framework, which provides the support to run the MapReduce programs for BigData, hence in the next part the algorithm was processed &optimized on the Hadoop framework.

III. SYSTEM PARAMETER WHICH AFFECT THE PERFORMANCE OF THE SINGLE AND MULTIPLE SYSTEM

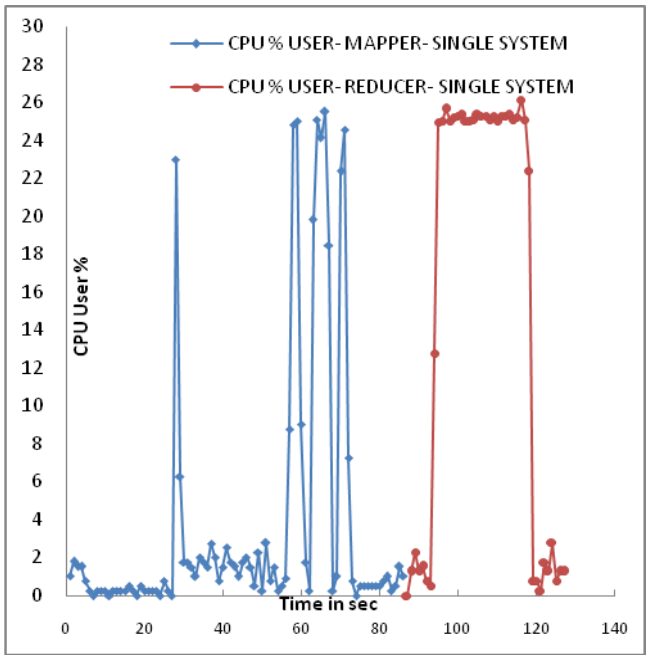
The performance of the system is assessed by running the algorithm on single system. In the results it was found that

- ✓ The RAM value used in single system &Hadoop is almost same for the BigData, but the time complexity is reduced when Hadoop framework used [5] in multiple systems as compared to single system.
- ✓ Hadoop framework will also require more resource in the form of RAM to perform the BigData of the algorithm on single system.

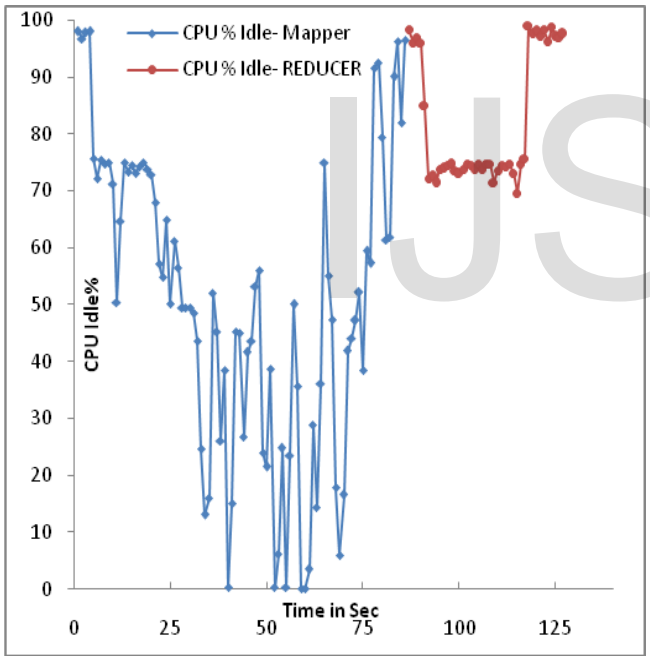
In the second experiment the Hadoop framework is used on multiple systems to process the same algorithm. The results show that BigData analysis can be efficiently done by using multiple systems; this reflects that the time complexity in processing BigData can be optimized by using multiple systems (Hadoop combine all RAM and Processor of the all systems). The results of experiments are studied by executing graph search algorithm on single system without using Hadoop framework for 1GB data. Parameters like CPU usage, Input-Output (IO) and RAM are studied while executing the algorithm. The algorithm is programmed in two parts where one part of code is for Mapping of the data and second one is for Reducer. Monitoring of the system is done for mappers and reducers.

Graph-2.1&2.2 show the CPU used in running of the program in user &Idle percentage respectively, against the time elapsed in seconds for executing the task.

International Journal Of Scientific & Engineering Research, Volume 7, Issue 10, October-2016
ISSN 2229-5518

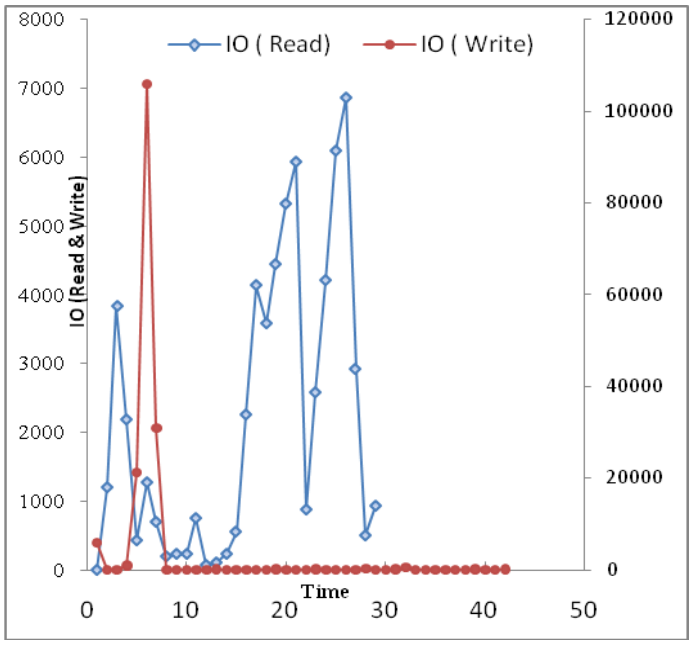


Graph 2.1: CPU User Percentage at the Time of Program execution



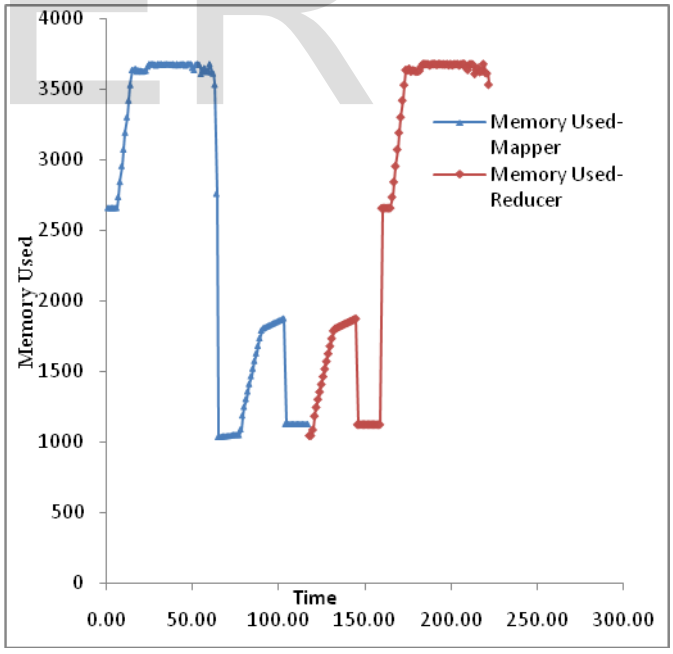
Graph 2.2: CPU idle Percentage at the Time of Program execution

Graph-2.3 studies the IO system use by the Search program. It also shows the read and writes time in seconds against the system performance.



Graph 2.3: IO (Read & Write) at the Time of Program execution

Graphs 2.1to 2.3 shows that there is little affection CPU & IO due to running of the algorithm. It could be due to smaller sample size of the BigData, but Graph 2.4 shows that size of RAM is a variable which may affect the performance of the system in a greater way while doing the BigData analysis using Hadoop.



Graph 2.4: Memory used at the Time of Program execution

Graph 2.4 depicts the monitoring of the RAM as a main parameter. All future experimental results are done by varying the capacity of RAM and studying the performance of the BigData analysis with the variation of RAM.

International Journal Of Scientific & Engineering Research, Volume 7, Issue 10, October-2016
ISSN 2229-5518

The sample of the analytical data is performed on 4GB RAM. RAM used by the program is presented against the time in seconds.

The sample of BigData taken in the algorithm varied from 500MB to 10GB in the experiment.

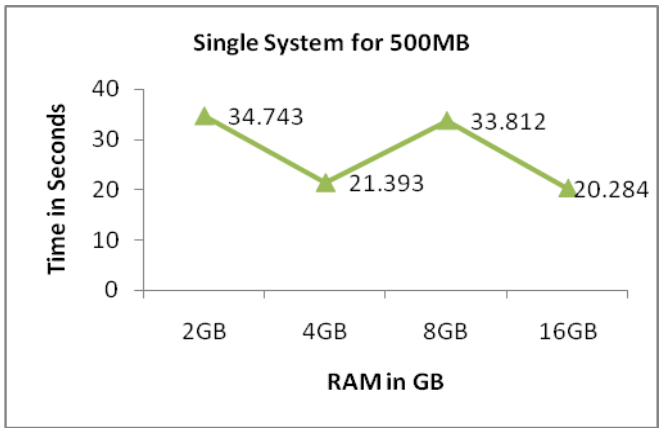
The Hadoop framework has limited usability in the single systems as compared to multiple systems. The algorithm is run on single system & multiple systems to process the BigData from 0.5 GB TO 32GB. The RAM size variation in single and multiple systems is done to swiftly process the BigData.

The results are discussed in section IV, it clearly shows single system is not able to reduce the time complexity by increasing the RAM size, but the multiple system can reduce the time complexity to certain extent by using cluster of systems and increasing the size of RAM.

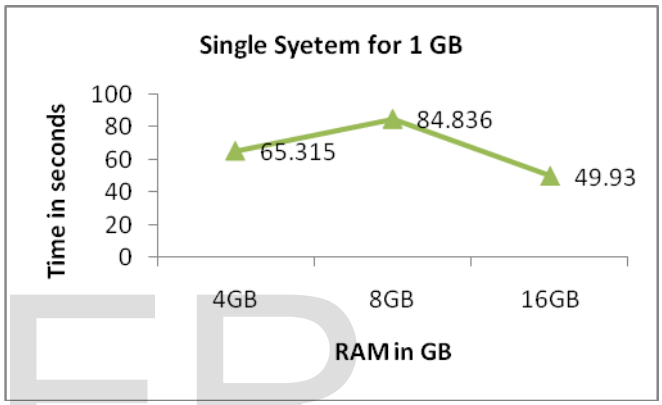
IV. Experimental Results Of Single & Cluster System In Hadoop Framework

The algorithm is designed to search and analyze the data where Mapper work for filtering, analyzing and searching the data and Reducer program works to calculate the search. The algorithm consumes RAM when it runs; the same is shown in above result (Graph 2.4). The paper compares the capacity and performance of the single system with the multiple cluster system of Hadoop, by running the algorithm. We are trying to conclude through experimental results that the performance of the system depends on RAM where all other processing parameters like CPU, IO, etc are available in threshold quantity.

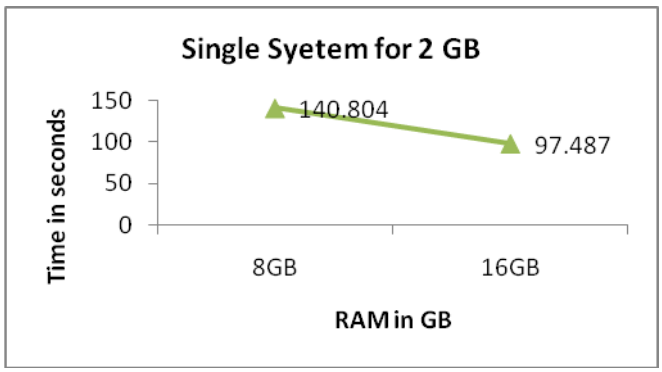
The algorithm is processed in a single system by varying the capacity of RAM from 2GB to16 GB. The BigData from 500 MB to 2GB is processed and performance of the system is monitored in terms of time taken in seconds to perform the BigData processing. The performance monitoring of the systems with various RAM capacities are studied to understand the reduction in time complexity with increase of RAM and increase in size of BigData. Graphs 4.1, 4.2 & 4.3 depict the system performance for 0.5, 1 & 2GB BigData respectively, along with the different RAM capacities. The graphs 4.1, 4.2 & 4.3 clearly show that application of Hadoop framework in single system does not reduce the time complexity by increasing the RAM.



Graph 4.1: Single system Graph- 500 MB data



Graph 4.2: Single system Graph- 1GB data



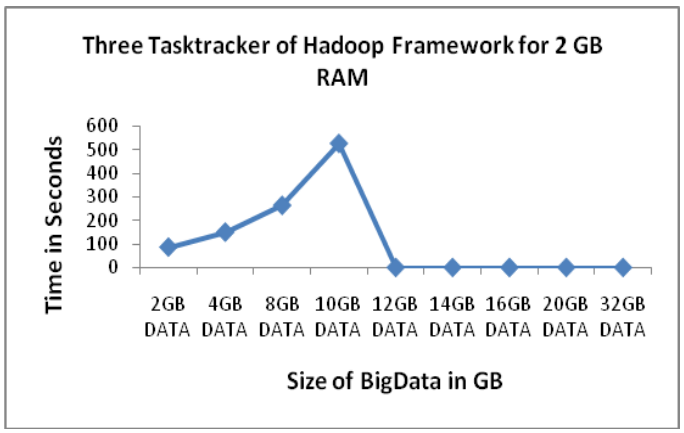
Graphs 4.3: Single system Graph- 2GB data

In next set of experiments algorithm is processed in multiple systems by varying the capacity of RAM. The performance of the system in terms of processing time of BigData is monitored against the capacity of RAM in different multiple systems. Multiple systems of Three, Four, Ten & Fourteen task trackers of Hadoop are used in1 the experiments.

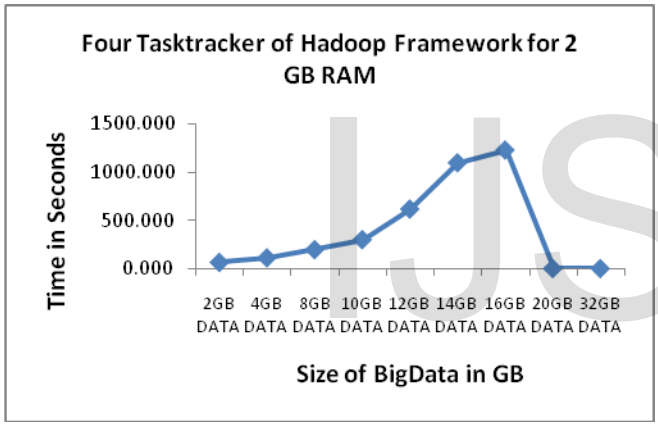
In the first experiment the three task trackers of Hadoop framework with 2GB RAM is used to process data from 2GB to 32GB. The graph 4.4 shows that BigData

International Journal Of Scientific & Engineering Research, Volume 7, Issue 10, October-2016
ISSN 2229-5518

processing using Hadoop framework improves the time complexity. The three tasktracker system with 2GB RAM can process upto 10GB BigData, after 10 GB the system is not able to process the algorithm.



Graph 4.4: Three Tracker Multiple systems Graph- 2GB RAM &BigData 2-32GB

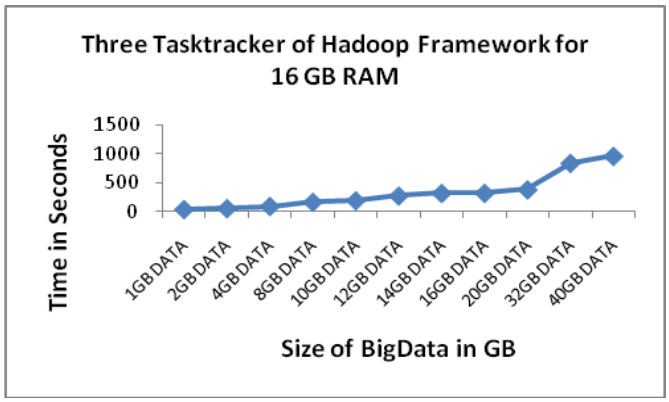


Graph 4.5: Four Tracker Multiple systems Graph- 2GB RAM &BigData 2-32GB

In the second experiment the four task trackers is used with the same configuration of RAM & BigData Size from 2GB to 32GB [7]. 11The graph 4.5 shows that BigData processing using Hadoop framework improves the time complexity. The four tasktracker system with 2GB RAM can process upto 16GB BigData, & after 16 GB the system is not able to process the algorithm.

In Third and Fourth experiments 10 tasktracker & 14 tasktracker have used with 2GB RAM. Both the experiments fetched same results in terms of processing time of BigData from 2GB to 32GB. This shows that after certain increment in the RAM capacity in form of multiple systems the processing time of BigData cannot be optimized.

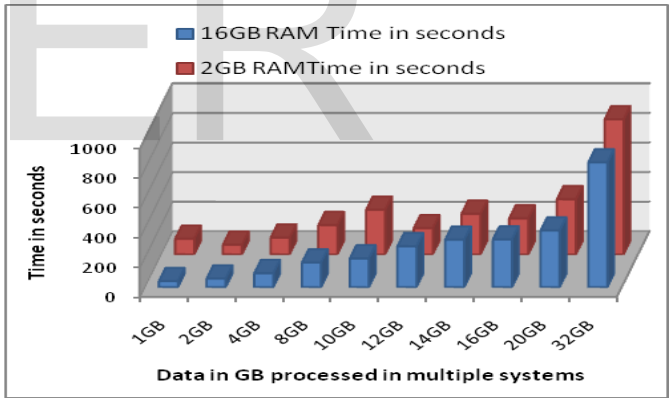
In Fifth experiment the capacity of RAM has increased to 16GB and used three tracker systems to process the data from 1GB to 32 GB. The graph 4.6 shows that by increasing the RAM the algorithm is able to run on the system and process the BigData, but the time complexity is not improved as compared to previous systems of 2GB RAM.



Graph 4.6: Four Tracker Multiple systems Graph- 16GB RAM &BigData 1-32GB

Graph 4.7 shows the Comparison of Fourteen tasktracker & three tasktracker [6] multiple systems with 2GB & 16GB RAM respectively& processing of BigData from 1 to 32GB. The bar chart shows that processing time of BigData in 2GB and 16GB is almost similar.

.



Graphs 4.7: Comparison Three Tracker Multiple systems-16GB RAM Graph-Fourteen tasktracker with 2GB RAM &BigData 1-32GB

Graph 4.7 data is presented in tabular format in Table 4.1. This BigData is processed in 16GB and 2GB RAM multiple systems of different configurations [7]. The time taken to process the data is lesser in 16GB RAM system as compared to 2GB RAM system, but as the size of BigData increases this processing time in both the systems is almost same or there is insignificant difference between the processing times of BigData.

International Journal Of Scientific & Engineering Research, Volume 7, Issue 10, October-2016
ISSN 2229-5518

TABLE-4.1 COMPARISON THREE TRACKER MULTIPLE SYSTEMS-16GB RAM GRAPH-FOURTEEN TASKTRACKER WITH 2GB RAM &BIGDATA 1-32GB

Data Size	16GB RAM Time in seconds	2GB RAM Time in seconds
1GB	39.99	104.329
2GB	56.893	63.855
4GB	93.682	111.855
8GB	166.198	196.703
10GB	193.359	300.138
12GB	275.603	175.846
14GB	320.442	272.751
16GB	321.674	241.565
20GB	382.278	371.123
32GB	843.437	912.095

V. CONCLUSION

The results of running the algorithm on the single system with different RAM capacities show that there is limited applicability of Hadoop framework on single system. The results on single system prompted us to run the algorithm on multiple systems with different RAM capacities to process the BigData in the Hadoop Framework.

The experimental results show that the Hadoop framework in BigData processing can reduce the time complexity and improve the speed of processing of the BigData by using the multiple systems.

It can also be concluded from the results of the running the algorithm on multiple system that with the increase in RAM capacity the BigData can be processed which was not possible in single system.

The time of processing of BigData in multiple system can be optimized upto a level, by increasing the capacity of RAM in different configurations of the multiple systems, after achieving the optimum level, the RAM capacity will also not reduce the time complexity of BigData processing.
In Hadoop framework fourteen & three tasktrackers systems with 2GB & 16GB respectively are also compared to understand the affect of RAM size on processing of the BigData. Here we can conclude that the time complexity is reduced to very less extent by increasing the RAM size in multiple systems.
Hence for future experiments it can suggested to redesign the algorithm to achieve the optimum time for processing the data in Hadoop Framework

REFERENCES

[1] Firat Tekiner1, John A. Keane, (2013) "BigData Framework", International Conference on Systems, Man, and Cybernetics, IEEE.

[2] Jean Yan, (2013) "BigData, Bigger Opportunities". Available:http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf

[3] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, (2014), "A Review Paper on BigData and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October.

[4] AvitaKatal, Mohammad Wazidand R H Goudar, "BigData: Issues, Challenges, Tools and Good Practices", Department of CSE, Graphic Era University, Dehradun, India, avita207@gmail.comwazidkec2005@gmail.com, rhgoudar@gmail.com

[5] J. Xie, S. Yin; X.J. Ruan; Z.Y. Ding, Y. Tian, M. J. Manzanares and X. Qin, "Improving MapReduce performance through data placement in heterogeneous Hadoop clusters," Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on , vol., no., pp.1,9, 19-23 April 2010

[6] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis,."Evaluating MapReduce for multi-core and multiprocessor systems." In Pro-ceedings of HPCA. IEEE Computer Society, 2007.

[7] C. Doulkeridis and K. Norvag, "A survey of large scale analytical query processing in MapReduce", VLDB J., 23(3)(2014), pp. 355-380.

[8] Lewis W.,Chu T., Salehi-Abari W., (2010), "Media Monitoring Using Social Networks", Social Computing (Social Com), 2010 IEEE Second International Conference on, vol., no., pp.661-668, 20-22 Aug. 2010 doi: 10.1109/SocialCom.

[9] Bo Li, "Survey of Recent Research Progress and Issues in BigData", boli@seas.wustl.edu (A paper written under the guidance of Prof. Raj Jain)

[10] NESSI – BigData White Paper; NESSI_WhitePaper_BigData.pdf

[11] Tom White, "Hadoop: The Definitive Guide", publication is Oreilly and Yahoo press, 2009.