IMPLEMENTING BIG DATA AND ANALYTICS SOLUTIONS IN CLOUD COMPUTING ENVIRONMENT

Dissertation proposal submitted in partial fulfillment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

BALRAJ

Registration number

11617421

Supervisor

Mrs. ROSY (UID-19397)

Assistant Professor

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

December (2017)

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0344

Supervisor Name : Rosy **UID :** 19397 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Balraj	11617421	2016	K1637	9815112936

SPECIALIZATION AREA : System Architecture and Design **Supervisor Signature:** _____

PROPOSED TOPIC : Implementing Big data and analytics solutions in cloud computing environment

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	6.14
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.57
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.	6.43
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.14
5	Social Applicability: Project work intends to solve a practical problem.	6.43
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.71

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): Yes
PAC Member 2 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 3 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): NO
PAC Member 5 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 6 Name: Harleen Kaur	UID: 14508	Recommended (Y/N): Yes
PAC Member 7 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 8 Name: Tejinder Thind	UID: 15312	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Implementing Big data and analytics solutions in cloud computing environment**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

11/25/2017 10:45:14 AM

Abstract

The big data is the type of data which is non-relational and large in quantity. The big data is generally uploaded on the cloud servers. The term “big data” is relatively new in IT and business. The Big data is a term used where the large volume of data is difficult to process, store and analyze by using traditional existing database technologies. As the nature of big data is indistinct so, there is need to involves considerable processes to identify and translate the data into new insights. There are number of definitions of big data some researchers also define big data as a large volume of scientific data for visualization. To manage the data of cloud servers, the technique of HDFS is required, which divides data into small chunks. When the data is divided into small chunks, sometimes, data gets un-managed due to overloading. This reduces the network efficiency in terms of various parameters. The module of load balancing will be introduced which will balance network load and increase its efficiency in terms of various parameters.

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude to my supervisor, **Mrs. Rosy**, Assistant Professor, Computer Science and Engineering Department, **Lovely Professional University**, **Phagwara**, for her invaluable help and guidance during the course of thesis. I am highly indebted to her for constantly encouraging me by giving her critics on my work. I am grateful to her for giving me the support and confidence that helped me a lot in carrying out the research work in the present form. And for me, it's an honor to work under her. I also take the opportunity to thank **Mr. Dalwinder Singh, HOD**, Computer Science and Engineering Department, **Lovely Professional University, Phagwara,** for providing us with the adequate infrastructure in carrying the research work. I would also like to thank my parents and friends for their inspiration and ever encouraging moral support, which went a long way in successful partial completion of my thesis. Above all, I would like to thank the almighty God for His blessings and for driving me with faith, hope and courage in the thinnest of the times.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation proposal entitled **"IMPLEMENTING BIG DATA AND ANALYTICS SOLUTIONS IN CLOUD COMPUTING ENVIRONMENT"** in partial 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 **Mrs. Rosy** . 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 the dissertation proposal represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation proposal work.

Signature of the Candidate

Balraj

Registration No…11617421

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation proposal entitled **"IMPLEMENTING BIG DATA AND ANALYTICS SOLUTIONS IN CLOUD COMPUTING ENVIRONMENT"** .Submitted by Balraj 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

Name: …….....

Date: …..……

Counter Signed by:

(1) Concerned HoD:

HoD's Signature: …….

HoD Name: ……..

Date: …….

(2) Neutral Examiners:

External Examiner

Signature: …….

Name: ……..…….

Affiliation: ……

Date: ……..

Internal Examiner

Signature: ……..

Name: ……

Date: ……..

TABLE OF CONTENTS

CONTENT **PAGE NO.**

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction to Big data

The term “big data” is relatively new in IT and business. The Big data is a term used where the large amount of records is hard to process, hoard and evaluate by using traditional existing record technology. As the scenery of large information is indistinguishable so, there is need to involves considerable processes to identify and translate the data into new insights. There are number of definitions of big data some researchers also define big data as a huge size of technical data for interpretation. Other researchers identify big data as “the quantity of records just away from expertise ability to hoard, handle, and development proficiently [1].”

The major challenges of statistical science are big data. A lot of recent reference starts to think about the frequent penalty of this new context from the algorithmic viewpoint and for the theoretical implications of this recent framework. Massive data always involve in big data. Data streams and data heterogeneity often included by them [2].

1.2 Applications of Big data

Here are some examples of Big Data applications:

- **Smart Grid case:** it is crucial to monitor and manage the smart grids operations and the national electronic power consumption in real time. To achieve the objective above mentioned there is need to make multiple connections among smart meters, sensors, control centers and other infrastructures. To detect the abnormal behaviors of the connected devices and to identify at-risk transformers Big Data analytics need to use. With the help of Big data Grid Utilities can choose the most excellent action or achievement. The concurrent study of the generated huge information allow to model incident scenarios.

E-health: To personalize health users are already using services connected health platforms. Large data is generated from different kinds of sources (e.g., laboratory and medical data,

patients symptoms uploaded from remote sensors, hospitals operation, and medicine data). There are number of beneficial applications for using advanced analysis of medical data sets. It enables to personalize

- Health services (e.g., doctors can monitor online patients symptoms in order to adjust prescription); to adapt public health plans according to population symptoms, disease evolution and other parameters. To decrease the cost expenditure and optimize the operations of hospital the big data has been used.
- **Internet of Things:** Large amount of Data applications are the core market of IoT. Because of the soaring range of objects, the applications of IoT are continuously growing. These days, there are various Large Data applications sustaining for logistic enterprises. With the help of sensors, wireless adapters and GPS it becomes possible to detect the position of vehicles. Thus, such data driven applications enable companies not only to oversee and supervise employees but also to best rescue route.
- **Public utilities:** In complex water supply network the sensors have been placed in pipelines to monitor the flow of water. The real-time monitoring system is implanting in the Bangalore Water Supply and Sewage Board published in the press. This system is build to detect leakages, illegal connections and remotely control valves to ensure impartial supply of water to different areas of the town. It helps to reduce the need for regulator operators and to timely identifying and fixing irrigate pipes that are leaking.
- **Shipping and arrangement:** The RFID (Radiofrequency Identifi- cation) and GPS have been used by many public road transport companies to track buses and explore interesting data to improve their services. We are able to choose best bus routes and the regularity of trips by collecting data about the many passengers using the buses in different routes. Various valid-point in time systems has been implement not only to provide
- Passengers with recommendations but also to offer important information on when to guess the next bus which will take him to the most wanted target. By predicting the demand about public or private networks the by using mining of Big Data helps in improving the travelling business. Making predictions from such data is a complicated issue because it depends on some factors such as weekend, festival, nighttime train, opening or midway place. By using machine learning algorithm, it is possible to supply

and relate advanced analytics on precedent and new big data collection. In fact advanced analytics can ensure high accuracy of results regarding many issues.

- **Political services and government monitoring:** Many governments such as United States and India are removal statistics to monitor opinionated trends and analyze people sentiments. Present many applications that combine many statistics source: public net connections, individual interview, and elector composition. Such systems enable also to detect local issue in addition to nationalized issue. Furthermore, government is capable of use Big Data systems to optimize the use of valuable resources and amenity. For instance, sensors can be placed in the pipelines of irrigate supply chains to monitor irrigate flow in large networks. Thus it is possible for many countries to rely on real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure impartial bring in of water to different areas of the town [3].

1.3 Issues in Big data

There are numerous issues arising within this technology some of which are liste1d below:

- **Storage space and transfer issue:** Every moment while latest storage space means has been made-up the amount of records has exploded. Now there is no storage area is left due to the increase in public medium. Moreover, everything and everyone produced the data (e.g., devices, etc) by professionals such as scientist, journalists, writers, etc. present disk technology limits are about 4 terabytes per disk. So, 1 Exabyte would require 25,000 disks. Even if an Exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Current communication networks has been overwhelm due to access to that data. Suppose that a 1 gigabyte per second network has a useful sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes [4]. Thus, transferring an Exabyte would take about 2800 hours, if we suppose that a sustained transfer could be maintained. It would take longer to put out the data from a collection or storage point to a processing point than it would to actually process it.
- **Privacy**: In big data privacy and security is main concern. The Big data security model gets disabled in case of result of difficult applications. However, in its deficiency, data

can constantly compromise easily. Privacy is the benefit to have some control in excess of how the individual information is collected and used. In case of information privacy it define the capacity of an individual or group to stop in sequence about them from becoming identified to people other than the people whom that information is need to send. The recognition of individual information through communication over the Internet is one of the serious issues of user privacy issue [5].

1.4 Introduction to Cloud

Cloud computing is a highly scalable and gainful communications for operation number of applications such as HPC, project and Web applications. However, there is one big critical issue in cloud computing which have been emerging due to its growing demand which have drastically increased the consumption of energy in data centers. The issue of high consumption not only increase the operation value which reduces the profit of cloud source but it also affect the environment as the high consumption of energy lead of high emission of carbon. Hence, energy-efficient solutions are essential to reduce the impact of Cloud computing on the environment. The objective of making cloud which is environment friendly can be achieve by the use of green cloud computing.

Over other existing computing techniques the cloud computing is advantageous and it greatly pick up the accessibility of IT resources. So with the use of cloud computing users are able to use the infrastructure of IT and pay for that only which will accumulate the cost to get the physical resources that may be empty when it is not in use. It is a analogy to explain web as a space where computing has been pre installed and exist as a service; data, operating systems, applications, storage and processing power exist on the web ready to be shared [7].

Cloud computing consists of mainly five essential characteristics of cloud and the absence of any one of those characteristics can make a cloud computing environment inappropriate to use [6].

- **On demand capabilities:** The cloud computing services are made sure to be secured by the cloud computing environment which is also known as a software vendor in business terms. The services can be accessed as well as changes by the online control panel by the user. There is no need of interacting with the server and can be done directly by the provider [6].

- **Broad network access:** Today, the digital devices can access the broad networks anywhere and can connect through a simple network access point. This property is very useful in business fields as the employees can connect and communicate with each other even after the office hours.
- **Resources Pooling:** The employee can share the information or services at the same time from any location with the help of cloud computing. This is done within business management software hosted at the cloud.
- **Rapid elasticity:** The users can be added or removed as per your need by the cloud computing. The flexibility and scalability is given up to an extent.
- **Measured service:** The services can be accessed. The services however, are paying. The source as well as the user side is monitored and the transparency of the network is thus improved [7].

1.5 How Big Data will be used with Cloud

The recent advances in technologies and architecture enabled a new data analyzer concept i.e., Big Data. The use of large information is prohibited in little and average sized businesses due to its huge assurance of hardware and processing resources cost. The concept of Cloud in Big Data has enabled a use of Big Data in little and average sized businesses. MapReduce is used as a programming paradigm for processing Big Data that required parallel processing and networked attached storage. Provided by outside entity, a Cloud computing is an on-demand network access to computing resources and computing done by using MapReduce is not possible for small and medium sized business.

The three main reasons behind use of cloud computing in Big Data technology implementation in little and average sized businesses are given below:

- Hardware cost reduction
- Processing cost reduction
- Capability to analysis the value of Big Data

The security and loss of control are main issue in cloud computing.

CHAPTER 2
SCOPE OF STUDY

The big data is the type of data which store any type of information and remove the constraints of type data and relational data bases. The data which is stored in the database is highly sensitive due to which various attacks need to breach security of the database. The big data has very dynamic nature due to its non relational nature. The cloud computing has the architecture in which virtual servers are involved to store the data. The data which is stored on the virtual servers is the big data. The Map reduce is the technique will is applied to analyze, perform operations on the big data. In the Map reduce, the HDFC file system is used which is hierarchical distributed file system. In this file system, the data is divided into small chunks and each chunk is treated individually. The chunk is assigned to each virtual machine on which different process is executed. In this research proposal, the HDFC file system will be used to analyze cloud data, the data is divided in such a way that load in the network can be balanced. The proposed improvement leads to increase efficiency in terms of various parameters.

CHAPTER 3

OBJECTIVE OF THE STUDY

Following are the objectives of this research work:-

1. To study and analyze various data analysis algorithms for big cloud data

2. To propose improvement in HDFC file to balance network load for efficient execution of the system

3. The proposed improvement will be based on the load balancer module of HDFC file system

4. Implement proposed algorithm and compare with existing in terms of various parameters

CHAPTER 4
REVIEW OF LITERATURE

Peter Brezany, et.al (2017), presented that the evolution of cloud computing towards the Dew computing will help in providing various advancements in the scientific computational productivity with the usage of automation. There are various advancements being made recently for maximizing the productivity of the applications related to big data scenarios. Various measures have been proposed to generate automated data science platforms. However, most of the platforms generated fall into the category of business and engineering application areas. The automatic data analysis which is generated by Cloud-Dew computing is presented in this paper. The two application domains namely breath das analysis and brain damage restoration are focused upon in this paper. A novel Dew-enabled balance disorder rehabilitation approach was utilized for presenting various guidelines in order to provide improvements in these techniques. On the basis of various experiments conducted and comparisons made it was seen that various enhancements in terms of accuracy were achieved with the application of this proposed approach.[9]

Mohammadhossein Ghahramani, et.al (2017) have recommended, that there is a huge growth in the percentage of processed heterogeneous data with the increase in amount of data being generated each day. The popularity of mobile phones for example is growing on huge rate due to the presence of sensors and the cost effectiveness they include. The collection of contextual data which can further be utilized in engineering and business domains is done in a very easy manner due to such technologies. Amongst the various challenges the researchers are facing related to this technology, the amount of data being generated and the need to analyze this information closer to real-time are gaining attention these days. From the academia, industry and government applications these days, the demand of big data has been arising lately. New technologies are to be presented such that the huge amount of data can not only be processed and analyzed but also ingested quickly at an easy location. A dynamic data analysis framework is proposed in this paper which explores and analyzes the mobile phone data being generated. An interactive exploratory spatial data analysis algorithm is presented in this paper once the data of cell phone

communication records is processed. A neighborhood function is defined here using the nearest neighbor function. The frequency of calls at each cell tower is also analyzed. On the basis of various calculations made, the huge amount of data is processed and stored in efficient manner within less time duration [10].

Giuseppe Agapito, et.al (2017), have concluded that Both, Parallel Bioinformatics Algorithms and Cloud-based Healthcare and Biomedicine Services and Systems are reviewed in this paper which is a part of the parallel computing and cloud computing in life sciences. These methods have been utilized within the parallel preprocessing and statistical and data mining analysis of omics data as well as large scale applications respectively. There are various issues that arise when such platforms are utilized in order to store and analyze the health data which are also presented in this paper. The major focus here is made on preserving the security and privacy of the records of patients. Further, the study is proposed related to the parallel and distributed modeling and simulation within the fields of medicine and biology. High performance methods were reported to model, simulate and design these cases present in biological and clinical applications. In order to verify the speed-up of the proposed mechanism, the algorithms and applications were tested and validated with the datasets of real clinics. The performances of these techniques were measured within the parallel computation environments which showed that the proposed technique provided better results [11].

Albino Altomare, et.al (2017) presented, that the minimization of power consumption of cloud data centers is a major concern within the consolidation of virtual machines. Thus, various studies have been presented within this area. Along with the satisfaction of Service Level Agreement made by the users, it is the objective of consolidation to allocate the virtual machines on minimum number of physical server possible. On the basis of forecast of the virtual machine resource that is required, the effectiveness of the consolidation strategy can change. In order to develop intelligent consolidation policies, the data-driven predictive models are exploited. The various consolidation techniques of virtual machines in cloud systems that are driven by the predictive data mining models are compared in this paper. In order to allocate the requirements present on the present servers, the migrations of future computational requirements of virtual machines are made. There is huge improvement seen within the results in terms of energy saving and most efficient consolidation techniques as per the simulation results achieved [12].

Jiangfan Peng, et.al (2016) has aimed, that there is a need to attain the space information from the tunnel simulation scenario, in which the data is to be analyzed and compared to provide reliability and quality. Two technology solutions are designed in this paper as per the engineering application of tunnel measurement which utilizes the 3-d laser scanner individually with separate distance measurement principle. In order to detect the target recognition and measure the accuracy, numerous styles identify the cloud of target in both types of 3-d laser scanner. The factors that are related to the design of the technical solution are tested before the implementation process. The accuracy and log size of the two station 3-D laser scanner is tested along with the impact of incident angle on accuracy as well as technology of the systems. As per the results achieved after conducting experiments, it is seen that the error is very less and the precision and reliability have enhanced with the application of proposed technique [13].

Zakia Asad, et.al, (2015), have concluded that pressure is increased on the data centers network because of movement of massive volume of data in cloud. So, in this paper the authors have used the mixing technique, spate coding along by software defined network control to propose a new scheme to dynamically reduce the volume of communication. For this purpose they have introduced a novel spate coding algorithm which helps in achieving the real world use cases for networks of data centers. Further through performing a proof of idea implementation on the proposed system they continuously bridge the gap between theory and practice. The experiment results of proposed coding based scheme is compared with vanilla Hadoop implement tation, Combiner-N-Code and an in network combiner and shows the better performance. The results are compared in terms of communication volume which is up to 62% better than existing schemes, in terms of good put it is improved by 76%, disk utilization is by 38% and in terms of number of bits that can be transmitted per joule of energy is up to 200%. The results show that the proposed scheme is advantageous from the presented techniques in terms of different parameters [14].

Zakia Asad, et.al, (2015), have proposed a network coding technique CodHoop employ by system for the same purpose mentioned in previous paper. In this paper, authors have used a network middle box service and specifically index coding for controlling the dynamically reduction in communication volume. Further they have presented the motivating use case for this

class of applications and used Hadoop as a representative. The results of the proposed scheme are compared with the Hadoop in terms of number of parameters. A result shows that the proposed scheme is in average 31% better than use case translates depending vanilla Hadoop implementation. This shows that there is 31% less utilization of equipment energy in Hadoop scheme and in proposed scheme 31% jobs can run simultaneously or can say job completion time is reduced by 31%. The coding based scheme used in this requires a XORing of packets whose operations are computationally very fast. In this case the given memory has larger bandwidth by which authors are capable to process closer to link rate. Still in the worst case this coder has 809 Mbps of throughput on a 1 Gbps link [15].

Xuelian Lin, et.al, (2012), have recommended that for job analysis and optimization of MapReduce the accurate performance model is required. The numbers of steps are need to be perform in case MapReduce that make it a challenging task. In MapReduce the number of steps is directly proportional to complexity, with increase in number of steps complexity increased at steady rate. In this paper to measure the MapReduce task complexity, authors have used a new concept which helps in analyzing the detail composition. The concept of SP, CEF and RCC is also defined in this paper to precisely measure the cost of Map or Reduce function. To calculate the cost of each item they have decomposed the main cost objects and make a new cost model based on vector, equation. The result of model is verified on a several clusters of Hadoop and it shows the effectiveness of proposed model. In this proposed model, authors have not considered the combine operation and serialization cost. By improving the proposed scheme the results can be improve in terms of serialization. In case of resource contentions in the cluster, proposed model will not be able to accurately predict the execution time of task [16].

Chang Liu, et.al, (2013), have recommended that for information serious computation in application of big data, a low cost and high efficiency can be achieved by an environment of cloud computing. The cloud computing is cost effective and very flexible but it will restrict the control of user on their own data which results in data security problem. The authors have planned a Cloud Background Hierarchical Key Exchange (CBHKE) novel hierarchical scheme. This key exchange scheme will help in achieving the safe and proficient scheduling for cloud computing environment. They have designed a level by level iterative key exchange strategy to get a more capable Authentication Key Exchange even without compromising the data security.

The experimental and abstract outcome of future scheme Cloud Background Hierarchical Key Exchange and Internet Key Exchange (IKE) is superior in terms of efficiency. The proposed scheme CBHKE key exchange scheme help in improving the efficiency but at the same time they become slow in case of large datasets [17].

Vidushi Vashishth, et.al, (2017), have recommended that with the development in cloud and Internet of Thing (IoT) integration will continuously generate a stream of sensor data. Because of the above mentioned reason number of researchers has started working on the integration of Cloud with Big Data. In this paper, the authors have proposed a predictive scheme for task scheduling on the cloud in case of high velocity (Big Data). The purpose of this is to reduce the overhead incurred when Big Data is processed on the Cloud. The results of proposed algorithm are compared with existing traditional algorithms and it shows that the proposed scheme is 10 times faster than traditional algorithm. Due to volume of Big Data, fast processing is required which is achieved by using a proposed scheme. In case of small datasets acceptable accuracy can be achieved by appropriate choice of classifiers. By using the classifiers for allocation tasks, researchers are able to achieve the load balancing allocation [18].

E. Goldin, et.al, (2017), have aimed to make a novel infrastructure of cloud computing for the Big Data analytics. In real time closed control loops such as process control industry a new model is developed and deployed which is based on the analysis of historical sensor data, machine learning based optimization model. Current innovations in the field of Process Analyzer Techniques (PAT), big data and wireless technologies have created a new environment. In this environment with the help of different techniques almost all stages of the industrial process can be recorded and utilized for safety, real time optimization. The sensors of Big Data continuously record a data which require a huge investment in hardware and software. In this paper, authors have presented pilot cloud based architecture for data driven modeling applications. In process control field pilot based architecture will help in getting the optimal control configuration. As it was presented, these developments have been carried in close relationship with the process industry. They also overlay a way for a generalized application of the cloud based approaches, towards the future of Industry 4.0 [19].

Chu-Hsing Lin, et.al (2017) proposed in this paper a study in which the data mining was performed on the land price data of past ten years of Taichung City. The clustering algorithms

were utilized here in order to perform data extraction techniques. The Hadoop HDFS and MapReduce methods were combined along with R language and the results achieved were seen on Google Maps. The K-means and Fuzzy C-means clustering algorithms were executed in Hadoop cloud and a stand-alone PC in order to analyze their respective performances. As per the achieved results it was seen that in a cloud that included 9 compute nodes, an acceleration of around 3.5 times was achieved. Thus, it was concluded that in order to solve insufficient memory related problems within the big data applications, the Hadoop cloud with R provided better results. With the execution of proposed method the computation time was minimized to great extent. Also, with the utilization of adequate number of computing nodes, the issue related to insufficient memory was resolved [20].

Ahmed S. Kaseb, et.al (2017) presented that there are innumerable applications that use network cameras in order to visualize real time data within different environments. However, there is a need of adequate number of resources in order to analyze such huge amount of data which is being generated regularly. There are numerous resources available these days on clouds which are charged as per the usage. Many issues arise while managing all such cloud resources along with ensuring that the performance requirements are fulfilled. A cloud resource manager is proposed in this paper that helps in solving all such problems. The resources that are required to analyze the data stream of each camera are predicted by this manager. A heuristic algorithm is utilized in order to formulate the resource allocation issue. The allocated resources are monitored with the help of proposed manager. Any requirement of a new resource within the application is fulfilled and in case there are any unused resources present, they are removed from the application which results in minimizing the overall cost of the system. Experiments were conducted in this paper to analyze the performance of proposed system. As per the results achieved it was seen that the proposed method minimized the overall cost of the system by up to 60% [21].

Ming-Shen Jian, et.al (2017) presented that there are numerous sites available on Internet these days which help in selecting an appropriate location for spending your vacations as per your comforts. However, with so many choices available, it becomes difficult to select an appropriate destination. A solution to this problem was proposed in this paper which combined cloud computing with big data. The information present on the web was collected and sorted in a

proper manner. Further, the results achieved were ranked on the basis of the analysis made such that they could be read and understood very easily. The emotions of sentences present on web helped in generating the data which could be given as input to Hadoop in order to perform distributed computing. In order to sort all the data, the K-means algorithm was used here which also helped in updating the database on daily basis. The best choice for the user was selected with the help of intelligent learning mechanism. With the help of this generated system, the travelers were easily able to select perfect destination as per their requirements and choices [22].

Rezvan Pakdel, et.al (2016) proposed in this paper a mechanism which can handle all types of unstructured data present within the medical applications. Here, both image and textual data needs to be handled in a proper manner. The proposed method needs to be designed in such a manner that it is very general and highly efficient so that the all different types of data can be analyzed easily. As the solution provided here is cloud-based, there is a dynamic improvement in the efficiency which relies on the real-time performance of the computing nodes. Various experiments are conduct to analyze the performance of proposed system. As per the results achieved, it is seen that a scalable solution is provided through this framework. The analysis performance can be enhanced to greater extents in case when there are larger datasets are available. With the help of proposed framework, all types of unstructured data can be processed easily [23].

CHAPTER 5

RESEARCH METHODOLOGY

The big data is the type of data which is large in quantity and do not have any relation with each other. The data which is uploaded on the cloud servers is the big data. The HDFC is the system which is applied to analyze the big data. In the HDFC file system input data is divided into small chucks. Each chuck is assigned to each virtual machine for the execution. The load balancer is the technique will can be applied with the HDFC file system to balancer the network load on each virtual machine. The load balancer will analyze each chuck, will is further assigned to virtual machine. When the virtual machine has large length chuck, then load of that machine will be further divided to balance network load. The proposed improvement leads to increase network performance in terms of resource consumption

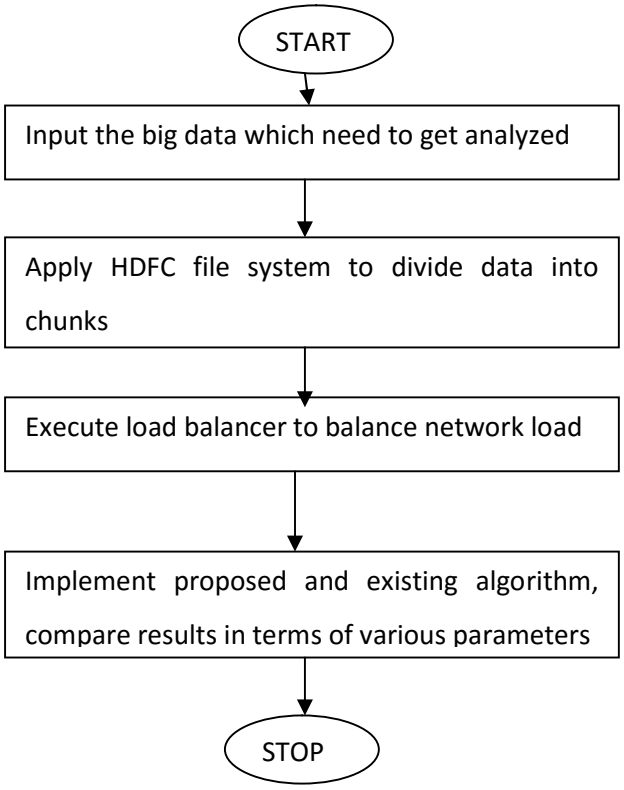

Fig 1: Proposed Flowchart

CHAPTER 6
EXPECTED OUTCOMES

Following are the various expected outcomes of this research:

1. This research is based on analyzing the cloud data. The data when divided into chunks due to non-management of the data, the overloading occurs which reduces the network efficiency. With this improvement, the network efficiency also gets enhanced with respect to various parameters.

2. When the data is not managed in an accurate manner, then the execution time is increased at steady rate. With this improvement, the execution times get reduced.

CHAPTER 7

SUMMARY AND CONCLUSION

In this research work, it has been concluded that the data which is uploaded on the cloud servers is non-relational. The term “big data” is relatively new in IT and business. The Big data is a term used where the large data is hard to process, hoard and examine by using traditional existing database technologies. As the character of big data is indistinguishable so, there is need to involves significant processes to recognize and interpret the data into new insight. There are number of definitions of big data some researchers also define big data as a large amount of scientific data for interpretation. To handle the data which is not managed, the technique of HDFS came into existence. In this technique, the data will be divided into equal number of chunks. When the chunks are not managed properly, network gets overloaded, which increases the execution time and reduces its efficiency. The technique of load balancing will be proposed which will manage the network load and maintain the network efficiency.

CHAPTER 8
REFERENCES

[1] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan, "The rise of "big data" on cloud computing: Review and open research issues", Elsevier Information Systems, vol.47, pp.98–115, 2015.

[2] Nada Elgendy and Ahmed Elragal, "Big Data Analytics: A Literature Review Paper", Springer International Publishing Switzerland 2014, vol.21, pp. 214–227, 2014.

[3] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih, "Big Data technologies: A survey", Journal of King Saud University – Computer and Information Sciences, vol.27,pp.1-18, 2017.

[4] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", IEEE 2013 46th Hawaii International Conference on System Sciences, vol.13, pp.995-1004, 2013.

[5] Priyank Jain, Manasi Gyanchandani and Nilay Khare, "Big data privacy: a technological perspective and review", Springer Jain et al. J Big Data, vol.25, pp.1-25, 2016.

[6] George Suciu, Cristina Butca, Victor Suciu, Alin Geaba, Alexandru Stancu, Stefan Arseni, "Basic Internet Foundation and Cloud Computing", IEEE 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, vol. 56, pp. 278-284, 2015.

[7] Bharath Balasubramanian, Mung Chiang, and Flavio Bonomi, "Introduction", IEEE, vol. 41, pp. 304-313, 2015.

[8] Bernice M. Purcell, "Big data using cloud computing", Journal of Technology Research, vol. 2, pp. 1-18, 2013.

[9] Peter Brezany, Thomas Ludeschery and Thomas Feilhauer, "Cloud-Dew Computing Support for Automatic Data Analysis in Life Sciences", MIPRO, vol. 27, pp. 401-408, 2017

[10] Mohammadhossein Ghahramani, MengChu Zhou, and Chi Tin Hon, “Analysis of Mobile Phone Data under a Cloud Computing Framework”, IEEE, vol. 7, pp. 1021-1027, 2017.

[11] Giuseppe Agapito, Barbara Calabrese, Pietro H. Guzzi, Gionata Fragomeni, “Parallel and Cloud-based Analysis of Omics Data: Modelling and Simulation in Medicine”, 25th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, vol. 13, pp. 181-189, 2017.

[12] Albino Altomare, Eugenio Cesario, “A Comparative Analysis of Data-Driven Consolidation Policies for Energy-Efficient Clouds”, 25th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, vol. 13, pp. 207-215, 2017.

[13] Jiangfan Peng, Xingwang Shen, Ming Guo, “Research on Processing and Analysing of Point Cloud Data of a variety of Lidar”, Fourth International Workshop on Earth Observation and Remote Sensing Applications, vol. 20, pp. 232-239, 2016.

[14] Zakia Asad, Mohammad Asad Rehman Chaudhry, David Malone, “Greener Data Exchange in the Cloud: A Coding Based Optimization for Big Data Processing”, IEEE Journal on Selected Areas in Communications, vol. 5, pp.1-18, 2015.

[15] Zakia Asad, M. Asad Rehman Chaudhry, D. Malone, “Codhoop: A system for optimizing big data processing”, in IEEE International Systems Conference (SysCon), 2015, pp. 295–300.

[16] Xuelian Lin, Zide Meng, Chuan Xu, Meng Wang, “A practical performance model for hadoop mapreduce”, in IEEE CLUSTER Workshops, vol.4 pp. 231– 239, 2012.

[17] Chang Liu, Xuyun Zhang, Chengfei Liu, Yun Yang, Rajiv Ranjan, Dimitrios Georgakopoulos, Jinjun Chen, “An Iterative Hierarchical Key Exchange Scheme for Secure Scheduling of Big Data Applications in Cloud Computing”, 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, vol. 4, pp. 9-15, 2013.

[18] Vidushi Vashishth, Anshuman Chhabra, Apoorvi Sood, “A predictive approach to task scheduling for Big Data in Cloud environments using classification algorithms”, IEEE 2017 7th

International Conference on Cloud Computing, Data Science & Engineering – Confluence, vol.7, pp. 1888-192, 2017.

[19] E. Goldin, D. Feldman, G. Georgoulas, M. Castano, G. Nikolakopoulos, “Cloud Computing for Big Data Analytics in the Process Control Industry”, 2017 25th Mediterranean Conference on Control and Automation (MED), vol. 5, pp.1373- 1378, 2017

[20] Chu-Hsing Lin, Jung-Chun Liu, Tsung-Chi Peng, "Performance Evaluation of Cluster Algorithms for Big Data Analysis on Cloud", 2017, IEEE

[21] Ahmed S. Kaseb, Anup Mohan, Youngsol Koh, Yung-Hsiang Lu, "Cloud Resource Management for Analyzing Big Real-Time Visual Data from Network Cameras", 2017, IEEE

[22] Ming-Shen Jian, Yi-Chi Fang, Yu-Kai Wang, Chih Cheng, “Big Data Analysis in Hotel Customer Response and Evaluation based on Cloud”, 2017, ICACT

[23] Rezvan Pakdel, John Herbert, “Scalable Cloud-based Analysis Framework for Medical Big-data”, 2016 IEEE 40th Annual Computer Software and Applications Conference

CHAPTER 9

APPENDIX

IOT Internet of Things

RFID Radio Frequency Identification

HPC High Performance Computing

IAAS Infrastructure as Services

PAAS Platform as Services

SAAS Software as Services

HDFC Hadoop Distributed File System

GBPS Gigabits per Seconds

CBHKE Cloud Background Hierarchal Key Exchange

AKE Authentication Key Exchange

IKE Internet Key Exchange

PAT Process Analyzer Techniques