Thesis Report on

"The Rise of Big data Clustering on Cloud Computing: More Flexible or Not"

A Thesis Report Submitted By

Mahbuba Afrin Chowdhury

ID:16103029

Abdullah Al Mahmud Khosru

ID:16103010

In Partial Fulfillment of the Requirements for the Award of

Bachelor of Computer Science and Engineering

Department of

Computer Science and Engineering

College of Engineering and Technology

IUBAT– International University of Business Agriculture and Technology

Summer 2019

“The Rise of Big data Clustering on Cloud Computing: More Flexible or Not”

Mahbuba Afrin Chowdhury (ID#16103029)
Abdullah Al Mahmud Khosru (ID#16103010)

A Thesis report submitted in partial fulfillment of the requirements for the degree of

Bachelor of Computer Science and Engineering (BCSE)

The Thesis has been examined and approved,

_____

Prof Dr. Md Abdul Haque
Chair and Professor
Dept. of Computer Science and Engineering
IUBAT – International University of Business Agriculture and Technology

_____

Prof Dr. Utpal Kanti Das
Coordinator and Associate Professor
Dept. of Computer Science and Engineering
IUBAT – International University of Business Agriculture and Technology

_____

Mohammad Sajid Shahriar
Lecturer
Dept. of Computer Science and Engineering
IUBAT – International University of Business Agriculture and Technology

Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Summer 2019

LETTER OF TRANSMITTAL

To

The Chairman

College of Engineering and Technology (CEAT)

IUBAT– International University of Business Agriculture and Technology

4 Embankment Drive Road, Sector 10, Uttara Model Town

Dhaka 1230, Bangladesh

Dear Sir,

With due respect, this is our pleasure to present our report entitled. The rise of big data clustering on cloud computing: More flexible or not. We prepared this report as partial fulfillment of the thesis. We have tried our level best to prepare this thesis report to the required standard.

It was certainly a great opportunity for us to work on this paper to actualize our theoretical knowledge in the practical arena. Now we are looking forward for your kind appraisal regarding this thesis report.

We shall remain deeply grateful to you if you kindly go through this report and evaluate our performance. We hope that you would find the report competent augmented.

Yours sincerely,

_____

Mahbuba Afrin Chowdhury
ID # 16103029

_____

Abdullah Al Mahmud Khosru

ID # 16103010

THESIS TITLE

“The Rise of Big Data Clustering on Cloud Computing: More Flexible or Not”

STUDENT NAME:

STUDENT ID:

Mahbuba Afrin

ID:16103029

Abdullah Al Mahmud Khosru

ID:16103010

DEPARTMENTOF COMPUTER SCIENCE AND ENGINEERING

COLLEGE OF ENGINEERING AND TECHNOLOGY

IUBAT—INTERNATIONAL UNIVERSITY OF BUSINESS AGRICULTURE AND TECHNOLOGY

Summer, 2019

THESIS TITLE

"The Rise of Big Data Clustering on Cloud Computing: More Flexible or Not"

By

Mahbuba Afrin

ID:16103029

Abdullah Al Mahmud Khosru

ID:16103010

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE FOR BACHELOR OF SCIENCE IN COMPUTER ENGINEERING (BCSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERINGCOLLEGE OF ENGINEERING AND TECHNOLOGY

IUBAT—INTERNATIONAL UNIVERSITY OF BUSINESS AGRICALTURE AND TECHNOLOGY

Summer, 2019

DECLARATION

This thesis report has been prepared after seven months of research on cloud "The Rise of Big Data Clustering on Cloud Computing: More Flexible or Not". The thesis is solely for academic requirement of the course CSC 488 and has not been submitted in part or full elsewhere for any other degree, reward or for any other purpose. I do solemnly and sincerely declare that all and every right's in the copyright of this thesis belong to IUBAT-International University of Business Agriculture and Technology. Any reproduction or use in any form or by any means whatever is prohibited without the written consent of IUBAT.

……..

Mahbuba Afrin
Student ID: 16103029

……..

Abdullah Al Mahmud Khosru
Student ID: 16103010

ACKNOWLEDGEMENT

We would like to express my deepest appreciation to all those who provided us the possibility to complete this report. A special gratitude we want to give to my honorable faculty Mohammad Sajid Shahriar sir, our supervisor whose contribution in stimulating suggestions and encouragement, helped us to coordinate our thesis, especially in writing this report.

Furthermore, we would also like to acknowledge with much appreciation the crucial role of the staff of internet related or all the sources like articles, research papers, blogs and other sources who gave the permission to use all required equipment and the necessary materials to complete the task. The special thanks goes to our friends also who helped and gave idea of parts and gave suggestion about the task. Last but not least, my parents are also a great inspiration for me. So, with due regards I also express my gratitude's to them. I have to appreciate the guidance given by other faculty and supervisor as well, Thanks all.

…….

Mahbuba Afrin
ID:16103029

……..

Abdullah Al Mahmud Khosru
ID: 16103010

THESIS TITLE

“The Rise of Big Data Clustering on Cloud Computing: More Flexible or Not”

Candidates

…….
Mahbuba Afrin
ID: 16103029

…….
Abdullah Al Mahmud Khosru
ID:16103010

Supervisor

…….
Mohammad Sajid Shahriar
Department of Computer Science and Engineering

ABSTRACT

Today, we're surrounded by data like oxygen. The exponential growth of data first presented challenges to cutting-edge businesses such as Google, Yahoo, Amazon, Microsoft, Facebook, Twitter etc. Data volumes to be processed by cloud applications are growing much faster than computing power. This growth demands new strategies for processing and analyzing information. MapReduce has become a powerful Computation Model addresses to these problems. It became more popular amongst all the Big Data tools as it is open source with flexible scalability, less total cost of ownership & allows data stores of any form without the need to have data types or schemas defined. MapReduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes. In this report, we explained MapReduce programming model, its various applications and implementations in Cloud Environments, how it works, comparing with other research papers, what are the extra benefits doing clustering on cloud environment, which flexibility and facilities we got about services of our data. And it's our pleasure to see the result of our research that was retrieving data from cloud is very fast and easy for analyzing and growing our business. At the same time our data is more secure than the normal server. Whenever we cluster our massive amounts of data on cloud, it took the partition on parallel way. That's why for processing and retrieving data it take less time and it is more flexible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

SYMBOL	ABBREVIATION
BD	Big Data
KM	K Means
M-Reduce	Map Reduce
D-Clustering	Data clustering
SaaS	Software as a Service
PaaS	Platform as a Service
IaaS	Infrastructure as a Service

CHAPTER 1: INTRODUCTION

1.1 Introduction

The continuous increase of computational power has produced an overwhelming flow of data. Big data is not only becoming more available but also more understandable to computers. For example, modern high-energy physics experiments, such as DZero, typically generate more than one Terabyte of data per day. The famous social network Website, Facebook, serves 570 billion-page views per month, stores 3 billion new photos every month, and manages 25 billion pieces of content.[2]

Google's search and ad business, Facebook, Flickr, YouTube, and LinkedIn use a bundle of artificial-intelligence tricks, require parsing vast quantities of data and making decisions instantaneously. Multimedia data mining platforms make it easy for everybody to achieve these goals with the minimum amount of effort in terms of software, CPU and network. On March 29, 2012, American government announced the "Big Data Research and Development Initiative", and big data becomes the national policy for the first time. All these examples showed that daunting big data challenges and significant resources were allocated to support these data intensive operations which lead to high storage and data processing costs. [3]

"Big Data" refers to enormous amounts of unstructured data produced by high-performance applications falling in a wide and heterogeneous family of application scenarios: from scientific computing applications to social networks, from e-government applications to medical information systems, and so forth. Data stored in the underlying layer of all these application scenarios have some specific characteristics in common, among which we recall: (i) large-scale data, which refers to the size and the distribution of data repositories; (ii) scalability issues, which refers to the capabilities of applications running on large-scale, enormous data repositories (i.e., big data, for short) to scale over growing-in-size inputs rapidly; (iii) supporting advanced Extraction-Transformation-Loading (ETL) processes from low-level, raw data to somewhat structured information; (iv) designing and developing easy and interpretable analytics over big data repositories in order to derive intelligence and extract useful knowledge from them.

On other hand, Cloud computing is a successful computational paradigm for managing and processing big data repositories, mainly because of its innovative metaphors known

under the terms “Database as a Service” (DaaS) [10] and “Infrastructure as a Service” (IaaS). DaaS defines a set of tools that provide final users with seamless mechanisms for creating, storing, accessing and managing their proper databases on remote (data) servers. Due to the naïve features of big data, DaaS is the most appropriate computational data framework to implement big data repositories [2]. MapReduce [8] is a relevant realization of the DaaS initiative. IaaS is a provision model according to which organizations outsource infrastructures (i.e., hardware, software, network) used to support ICT operations. The IaaS provider is responsible for housing, running and maintaining these services, by ensuring important capabilities like elasticity, pay-per-use, transfer of risk and low time to market. Due to specific application requirements of applications running over big data repositories, IaaS is the most appropriate computational service framework to implement big data applications.

The current technologies such as grid and cloud computing have all intended to access large amounts of computing power by aggregating resources and offering a single system view. Among these technologies, cloud computing is becoming a powerful architecture to perform large-scale and complex computing and has revolutionized the way that computing infrastructure is abstracted and used. In addition, an important aim of these technologies is to deliver computing as a solution for tackling big data, such as large scale, multi-media and high dimensional data sets.

The data collected is of very large amount and there is difficulty in collecting and assessing big data. Clustering algorithms means to put similar kind of data together. As if whenever we want to retrieve our data from cloud, it’s easy to get our data and analyze the data. Clustering is always taken less time for retrieving data. [4]

Necessity of big data clustering is very high. Because of rising massive amount of data Although there is a different formatted data like structured and unstructured and we can cluster it on cloud by using different algorithm like K-mean algorithm, map-Reduce algorithm etc. Whenever we will do the clustering on cloud that time it will consume our time and it work so faster than the normal server. So, it’s really important and helpful to manage our massive amount of data within shortest time. And it also reduces our employees, spaces etc. and we will achieve high profit from our business.

Previously many teams did this research on normal server, but we are using cloud as our server. Because the problem we are facing for this large amount of data and the possible solution of tackling our massive amount of data can be the clustering on cloud

computing. And specifically, we want to use Map-reduce algorithm. This map reduce algorithm split our data into parallel and then reduce the different node and merge it and we will get HDFS data.

1.2 Definition and characteristics of big data:

Big data is a term utilized to refer to the increase in the volume of data that are difficult to store, process, and analyze through traditional database technologies. The nature of big data is indistinct and involves considerable processes to identify and translate the data into new insights. The term “big data” is relatively new in IT and business.

However, several researchers and practitioners have utilized the term in previous literature. For instance,[6] referred to big data as a large volume of scientific data for visualization. Several definitions of big data currently exist. For instance,[7] defined big data as “the amount of data just beyond technology's capability to store, manage, and process efficiently.” Meanwhile,[8] and [9]defined big data as characterized by three Vs: volume, variety, and velocity. The terms volume, variety, and velocity were originally introduced by Gartner to describe the elements of big data challenges. IDC also defined big data technologies as “a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis.” [10] specified that big data is not only characterized by the three Vs mentioned above but may also extend to four Vs, namely, volume, variety, velocity, and value (Fig. 1, Fig. 2). This 4V definition is widely recognized because it highlights the meaning and necessity of big data.

The following definition is proposed based on the abovementioned definitions and our observation and analysis of the essence of big data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

(i) Volume: It refers to the amount of all types of data generated from different sources and continue to expand. The benefit of gathering large amounts of data includes the creation of hidden information and patterns through data analysis. Laurel et al. [11] provided a unique collection of longitudinal data from smart mobile devices and made this collection available to the research community. The aforesaid initiative is called mobile data challenge motivated by Nokia [11]. Collecting longitudinal data requires considerable effort and underlying investments. Nevertheless, such mobile data

challenge produced an interesting result like that in the examination of the predictability of human behavior patterns or means to share data based on human mobility and visualization techniques for complex data.

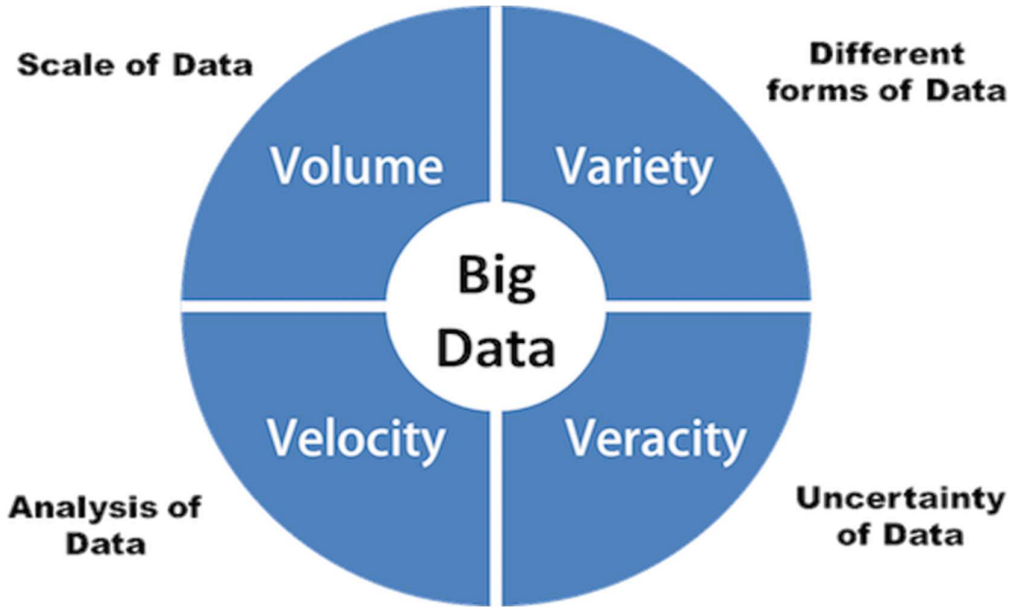

Fig 1.1: Four Vs of big data

(ii) Variety: It refers to the different types of data collected via sensors, smartphones, or social networks. Such data types include video, image, text, audio, and data logs, in either structured or unstructured format. Most of the data generated from mobile applications are in unstructured format. For example, text messages, online games, blogs, and social media generate different types of unstructured data through mobile devices and sensors. Internet users also generate an extremely diverse set of structured and unstructured data.[12]

(iii) Velocity: It refers to the speed of data transfer. The contents of data constantly change because of the absorption of complementary data collections, introduction of previously archived data or legacy collections, and streamed data arriving from multiple sources.[9]

(iv) Value: It is the most important aspect of big data; it refers to the process of discovering huge hidden values from large datasets with various types and rapid generation.[13]

1.3 Classification of big data

Big data are classified into different categories to better understand their characteristics. Fig. 2 shows the numerous categories of big data. The classification is important because of large-scale data in the cloud. The classification is based on five aspects:

(i) data sources,
(ii) content format,
(iii) data stores,
(iv) data staging and
(v) data processing.

Each of these categories has its own characteristics and complexities as described in Table 2. Data sources include internet data, sensing and all stores of transnational information, ranges from unstructured to highly

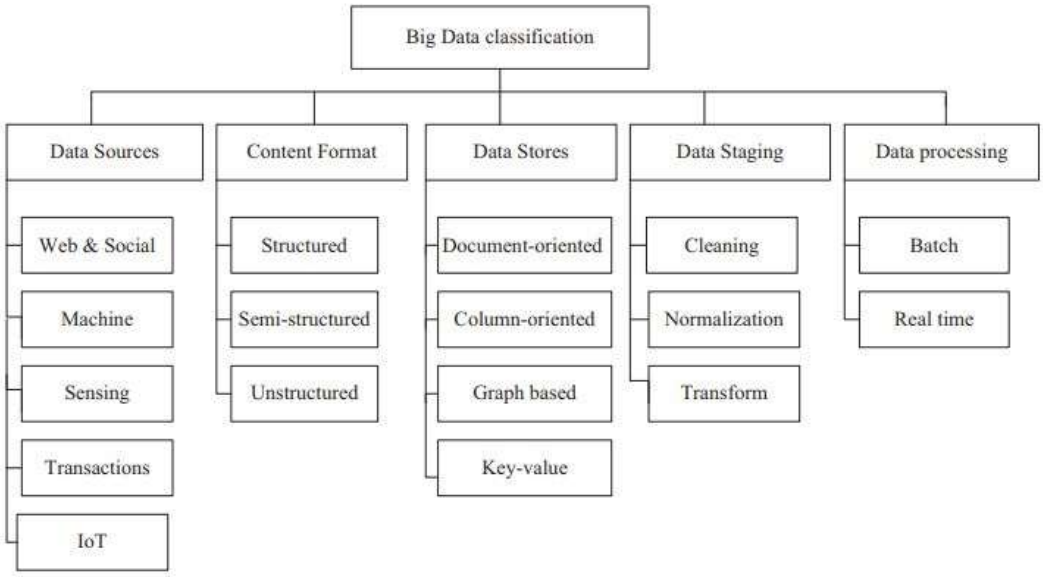

Fig 1.2: Big data classification.

structured are stored in various formats. Most popular is the relational database that come in a large number of varieties. [14] As the result of the wide variety of data sources, the captured data differ in size with respect to redundancy, consistency and noise, etc.

1.4Open problems and actual research trends of big data analytics

There are a number of open problems and actual research trends related to big data analytics. In the following, we provide an overview on some of the most significant of them.

(a) Data Source Heterogeneity and Incongruence. Very often, data sources storing data of interest for the target analytics processes (e.g., legacy systems, Web, scientific data repositories, sensor and stream databases, social networks, and so forth) are strongly heterogeneous and incongruent. This aspect not only conveys in typical integration problems, mainly coming from active literature on data and schema integration issues, but it also has deep consequences on the kind of analytics to be designed.

(b) Filtering-Out Uncorrelated Data. Due to the enormous size of big data repositories, dealing with large amount of data that are uncorrelated to the kind of analytics to be designed occurs very frequently. As a consequence, filtering-out uncorrelated data plays a critical role in the context of analytics over big data, as this heavily affects the quality of final analytics to be designed.

(c) Strongly Unstructured Nature of Data Sources. In order to design meaningful analytics, it is mandatory that input big data sources are transformed in a suitable, structured format, and finally stored in the HDFS. This poses several issues that recall classical ETL processes of Data Warehousing systems, but with the additional challenges that data alimenting big data repositories are strongly unstructured (e.g., social network data, biological experiment result data, and so forth) in contrast with less problematic unstructured data that are popular in traditional BI tools (e.g., XML data, RDF data, and so forth). Again, here transformations from unstructured to structured format should be performed on the basis of the analytics to be designed, according to a sort of goal-oriented methodology.

(d) High Scalability. High scalability of big data analytics is one of the primerfeatures to be ensured for a MAD-inspired big data analytics system. To this end, exploiting the cloud computing computational framework seems to be the most promising way to this end [2]. The usage of the IaaS-inspired cloud computing computational framework is meant with the aim of achieving some important characteristics of highly-scalable big data analytics systems, among which we recall: (i) "true" scalability, i.e. the effective scalability that a powerful computational infrastructure like clouds is capable of ensuring; (ii) elasticity, i.e. the property of rapidly adapting to massive updates and fast evolutions of big data repositories; (iii) fault-tolerance, i.e. the capability of being robust to faults that can affect the underlying distributed data/computational architecture; (iv) self-manageability, i.e. the property of automatically adapting the framework configuration (e.g., actual load balancing) to rapid changes of the surrounding data/computational environment; (v) execution on commodity machines, i.e. the capability of scale-out on thousands and thousands of commodity machines when data/computational peaks occur.

(e) Combining the Benefits of RDBMS and NoSQL Database Systems. One of the more relevant features to be achieved by big data analytics systems is represented by flexibility, which refers to the property of covering a large collection of analytics scenarios over the same big data partition. In order to obtain this critical feature, it is necessary to combine the benefits of traditional RDBMS database systems and those of next-generation NoSQL database systems, which propose representing and managing data via horizontal data partitions by renouncing to fixed table schemas and, consequentially, resource-expensive join operations [4].

(f) Query Optimization Issues in HiveQL. Several open issues arise with respect to query optimization aspects of HiveQL. Among 102 these, noticeable ones are the following: (i) moving towards more expressive, complex aggregations, e.g. OLAP-like rather than SQLite, hence enforcing the User Defined Function (UDF) and the User Defined Aggregate Function (UDAF) [5] paradigms; (ii) covering advanced SQL statements such as nested queries and order-by predicates; (iii) incorporating data compression paradigms in order to achieve higher performance; (iv) devising novel costbased optimizations, e.g. based on table or column statistics; (v) integration with third-part BI tools.

1.5 Clustering:

Clusteris a group of objects that belongs to the same class. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory datamining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

➢A cluster of data objects can be treated as one group.

➢While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.

➢The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

1.6 Why Do We Need Cluster Analysis:

Cluster analysis can be used to identify homogeneous groups of potential customers/buyers. Based on the previous purchase history of the product. Cluster analysis basically involves formulating one problem, selecting its approach and selecting a clustering algorithm. After deciding the number of clusters or groups, you can look at the final data which is beneficial for marketing or business purposes.

1.7 Algorithms in Cluster Analysis:

Cluster analysis uses many different types of clustering techniques/algorithms to figure out the final outcome. It is different than data processing but can be considered as a step depending on its use. Cluster techniques/algorithms can be classified based on their separating cluster model. In this overview, we are going to look at some of the prominent examples of cluster analysis.[15]

- Connectivity-based clustering aka hierarchical clustering: Hierarchical clustering is based on the concept of objects being related to their nearby entities then those are far away. The algorithm connects objects to form clusters purely based on distance relationships. A cluster can be defined based by a maximum distance. By this, different clusters will be formed based on different distances. This will produce a series of hierarchical clusters based on data distance which we can represent through a dendrogram.
- Centroid-based clustering: In centroid-based clustering, data groups are represented by a central vector point, which may or may not be a part of a data group. When these numbers of clusters are fixed to k, it gives a formal definition of an optimization problem. After finding the cluster centers, you can assign the objects to the nearest data group and so on. This will create a centroid-based clustering graph.
- **Distribution-based clustering:** This cluster distributing model is based on distributing data results. The theoretical foundation of these methods is surely excellent but lacks practicality as it suffers from overfitting problem. In order to get a hard cluster of data, then we need to assign to the Gaussian distribution method. However, for soft data clusters, it is not necessary.
- **Density-based clustering:** In density-based clustering, clusters are defined in such as way that data with high density are taken into consideration. The most popular density-based data clustering method is known as DBSCAN where it deals with density reachability. Here we make data groups based on connected density objects which can form different shapes. However, this data clustering method requires a density drop in order to detect variation. So this method doesn't work when applied to same density data groups.

1.8 Applications of Cluster Analysis:

Cluster analysis is a set of techniques or methods which were used to classify objects, cases, figures into relative groups. These related groups are further classified as clusters. Cluster analysis is also known by the name of numerical taxonomy or classification analysis. In cluster analysis, there is no information directly related to groups or clusters. It is used for various purposes in marketing products to specific targets.

➢Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.

➢Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

➢In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.

➢Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.

➢Clustering also helps in classifying documents on the web for information discovery.

➢Clustering is also used in outlier detection applications such as detection of credit card fraud.

➢As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

1.9 Cloud computing

Simply put, cloud computing is the delivery of computing services—including servers, storage, databases, networking, software, analytics, and intelligence—over the Internet ("the cloud") to offer faster innovation, flexible resources, and economies of scale. You typically pay only for cloud services you use, helping you lower your operating costs, run your infrastructure more efficiently, and scale as your business needs change.[16]

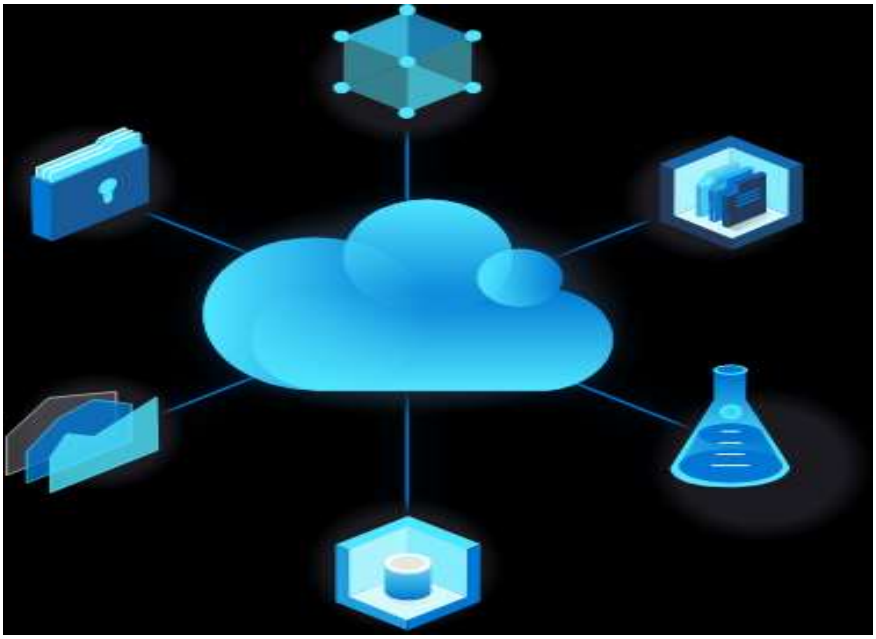

Fig 1.3: cloud computing

1.9.1 Top benefits of cloud computing

Cloud computing is a big shift from the traditional way businesses think about IT resources. Here are seven common reasons organizations are turning to cloud computing services:

Cost

Cloud computing eliminates the capital expense of buying hardware and software and setting up and running on-site datacenters—the racks of servers, the round-the-clock electricity for power and cooling, and the IT experts for managing the infrastructure. It adds up fast.

Speed

Most cloud computing services are provided self service and on demand, so even vast amounts of computing resources can be provisioned in minutes, typically with just a few mouse clicks, giving businesses a lot of flexibility and taking the pressure off capacity planning.

Global scale

The benefits of cloud computing services include the ability to scale elastically. In cloud speak, that means delivering the right amount of IT resources—for example, more or less computing power, storage, bandwidth—right when they're needed, and from the right geographic location.

Productivity

On-site datacenters typically require a lot of "racking and stacking"—hardware setup, software patching, and other time-consuming IT management chores. Cloud computing removes the need for many of these tasks, so IT teams can spend time on achieving more important business goals.

Performance

The biggest cloud computing services run on a worldwide network of secure datacenters, which are regularly upgraded to the latest generation of fast and efficient computing hardware. This offers several benefits over a single corporate datacenter, including reduced network latency for applications and greater economies of scale.

Reliability

Cloud computing makes data backup, disaster recovery, and business continuity easier and less expensive because data can be mirrored at multiple redundant sites on the cloud provider's network.

Security

Many cloud providers offer a broad set of policies, technologies, and controls that strengthen your security posture overall, helping protect your data, apps, and infrastructure from potential threats.

1.9.2 Types of cloud computing

Not all clouds are the same and not one type of cloud computing is right for everyone. Several different models, types, and services have evolved to help offer the right solution for your needs.

First, you need to determine the type of cloud deployment, or cloud computing architecture, that your cloud services will be implemented on. There are three different ways to deploy cloud services: on a public cloud, private cloud, or hybrid cloud.

Public cloud

Public clouds are owned and operated by a third-party cloud service providers, which deliver their computing resources, like servers and storage, over the Internet. Microsoft Azure is an example of a public cloud. With a public cloud, all hardware, software, and other supporting infrastructure is owned and managed by the cloud provider. You access these services and manage your account using a web browser.

Private cloud

A private cloud refers to cloud computing resources used exclusively by a single business or organization. A private cloud can be physically located on the company's on-site datacenter. Some companies also pay third-party service providers to host their private cloud. A private cloud is one in which the services and infrastructure are maintained on a private network.

Hybrid cloud

Hybrid clouds combine public and private clouds, bound together by technology that allows data and applications to be shared between them. By allowing data and applications to move between private and public clouds, a hybrid cloud gives your business greater flexibility, more deployment options, and helps optimize your existing infrastructure, security, and compliance.

1.9.3 Types of cloud services: IaaS, PaaS, serverless, and SaaS

Most cloud computing services fall into four broad categories: infrastructure as a service (IaaS), platform as a service (PaaS), serverless, and software as a service (SaaS). These are sometimes called the cloud computing "stack" because they build on top of one another. Knowing what they are and how they're different makes it easier to accomplish your business goals.

Infrastructure as a service (IaaS)

The most basic category of cloud computing services. With IaaS, you rent IT infrastructure—servers and virtual machines (VMs), storage, networks, operating systems—from a cloud provider on a pay-as-you-go basis

Platform as a service (PaaS)

Platform as a service refers to cloud computing services that supply an on-demand environment for developing, testing, delivering, and managing software applications. PaaS is designed to make it easier for developers to quickly create web or mobile apps, without worrying about setting up or managing the underlying infrastructure of servers, storage, network, and databases needed for development.

Serverless computing

Overlapping with PaaS, serverless computing focuses on building app functionality without spending time continually managing the servers and infrastructure required to do so. The cloud provider handles the setup, capacity planning, and server management for you. Serverless architectures are highly scalable and event-driven, only using resources when a specific function or trigger occurs.

Software as a service (SaaS)

Software as a service is a method for delivering software applications over the Internet, on demand and typically on a subscription basis. With SaaS, cloud providers host and manage the software application and underlying infrastructure, and handle any maintenance, like software upgrades and security patching. Users connect to the application over the Internet, usually with a web browser on their phone, tablet, or PC.

1.9.4 Uses of cloud computing

You're probably using cloud computing right now, even if you don't realize it. If you use an online service to send email, edit documents, watch movies or TV, listen to music, play games, or store pictures and other files, it's likely that cloud computing is

making it all possible behind the scenes. The first cloud computing services are barely a decade old, but already a variety of organizations—from tiny startups to global corporations, government agencies to non-profits—are embracing the technology for all sorts of reasons.

Here are a few examples of what's possible today with cloud services from a cloud provider:

- **Create cloud-native applications**

Quickly build, deploy, and scale applications—web, mobile, and API. Take advantage of cloud-native technologies and approaches, such as containers, Kubernetes, microservices architecture, API-driven communication, and DevOps.

- **Test and build applications**

Reduce application development cost and time by using cloud infrastructures that can easily be scaled up or down.

- **Store, back up, and recover data**

Protect your data more cost-efficiently—and at massive scale—by transferring your data over the Internet to an offsite cloud storage system that's accessible from any location and any device.

- **Analyze data**

Unify your data across teams, divisions, and locations in the cloud. Then use cloud services, such as machine learning and artificial intelligence, to uncover insights for more informed decisions.

- **Stream audio and video**

Connect with your audience anywhere, anytime, on any device with high-definition video and audio with global distribution.

- **Embed intelligence**

Use intelligent models to help engage customers and provide valuable insights from the data captured.

- **Deliver software on demand**

Also known as software as a service (SaaS), on-demand software lets you offer the latest software versions and updates around to customers—anytime they need, anywhere they are.[16]

So from the above services and topics,we can say that,

- Cloud computing is a fast-growing technology that has established itself in the next generation of IT industry and business. Cloud computing promises reliable software, hardware, and IaaS delivered over the Internet and remote data centers [04].
- Cloud services have become a powerful architecture to perform complex large-scale computing tasks and span a range of IT functions from storage and computation to database and application services. The need to store, process, and analyze large amounts of datasets has driven many organizations and individuals to adopt cloud computing [05].
- A large number of scientific applications for extensive experiments are currently deployed in the cloud and may continue to increase because of the lack of available computing facilities in local servers, reduced capital costs, and increasing volume of data produced and consumed by the experiments [10].
- In addition, cloud service providers have begun to integrate frameworks for parallel data processing in their services to help users access cloud resources and deploy their programs [09]. Cloud computing "is a model for allowing ubiquitous, convenient, and on-demand network access to a number of configured computing resources (e.g., networks, server, storage, application, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [11].
- Cloud computing has a number of favorable aspects to address the rapid growth of economies and technological barriers. Cloud computing provides total cost of ownership and allows organizations to focus on the core business without worrying about issues, such as infrastructure, flexibility, and availability of resources. Moreover, combining the cloud computing utility model and a rich

set of computations, infrastructures, and storage cloud services offers a highly attractive environment where scientists can perform their experiments. Cloud service models typically consist of PaaS, SaaS, and IaaS.

- **PaaS,** such as Google's Apps Engine, Salesforce.com, Force platform, and Microsoft Azure, refers to different resources operating on a cloud to provide platform computing for end users.
- **SaaS,** such as Google Docs, Gmail, Salesforce.com, and Online Payroll, refers to applications operating on a remote cloud infrastructure offered by the cloud provider as services that can be accessed through the Internet [07]
- **IaaS,** such as Flexi scale and Amazon's EC2, refers to hardware equipment operating on a cloud provided by service providers and used by end users upon demand.

As IaaS is a provision model according to which organizations outsource infrastructures (i.e., hardware, software, network) used to support ICT operations. The IaaS provider is responsible for housing, running and maintaining these services, by ensuring important capabilities like elasticity, pay-per-use, transfer of risk and low time to market. Due to specific application requirements of applications running over big data repositories, IaaS is the most appropriate computational service framework to implement big data applications.

The increasing popularity of wireless networks and mobile devices has taken cloud computing to new heights because of the limited processing capability, storage capacity, and battery lifetime of each device [04]. This condition has led to the emergence of a mobile cloud computing paradigm. Mobile cloud facilities allow users to outsource tasks to external service providers. For example, data can be processed and stored outside of a mobile device [14]. Mobile cloud applications, such as Gmail, iCloud, and Dropbox, have become prevalent recently. Juniper research predicts that cloud-based mobile applications will increase to approximately 9.5$ billion by 2014 . Such applications improve mobile cloud performance and user experience. However, the limitations associated with wireless networks and the intrinsic nature of mobile devices have imposed computational and data storage restrictions

1.10Problem statement

Cloud computing is a powerful technology to perform massive- scale and complex computing. It eliminates the need to maintain expensive computing hardware, space and software. Massive growth in the scale of data or big data generated through cloud computing has been observed. Addressing big data is a challenging and time demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The continuous increase in the volume and detail of big data captured by organization, such as the rise of social media, Internet of thing, and multimedia has produced an over flow of data in either structured or unstructured format. Data creation is occurring at record rate. The major challenge is that data growth rate exceeds their ability to design appropriate cloud computing platforms for data analysis, update and those results as time consuming, server down etc. So, our plan was to research on how these things are affecting the flexibility of big data and by using which algorithm we can reduce these outcomings, as it is one of the most burning issue world widely. In addition, it was convenient to use MapReduce algorithm to do this research.

1.11Objective

1. To find out the big data clustering method that will create good impact on cloud computing as if rise of big data don't occur traffic on cloud when we are storing data or accessing it.
2. To show the problem that can cause by creating massive amount of data.
3. To find out the flexibility of using this method for clustering.
4. To show the way to secure our data more and make it easy to analyze for growing our business.

1.12Organization of the Report

This thesis organized with five chapters. First chapter is about research background and objective of this study. In chapter 2, elaborated literature review is presented about effect of different factors on fuel stability, material degradation and their remedies. Chapter 3 contains materials and methods for this study with detail experimental setup.

Chapter 4 includes experimental results and discussions of obtained results respectively. Conclusion of this research work is drawn in chapter 5 along with constructive recommendation and potential possibilities for further study.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In literature review part we will explain some research paper regarding to our topics. So at first focus on our research question-

2.2 Research Question:

➢ Why clustering is important?

➢What is the benefit of clustering on cloud?

➢How we can analyze our rising big data within short time?

➢How easily we can process our massive amount of data on cloud?

➢How we can manage our large amount of datasets?

Although we have gone many research papers but here we are including three papers that are very much close to our topics.

2.3 Big Data and clustering algorithm-

In this paper, researcher has talk about Data mining, which is the method which is useful for extracting useful information and data is extorted, but the classical data mining approaches cannot be directly used for bigdata due to their absolute complexity. The data that is been formed by numerous scientific applications and incorporated environment has grown rapidly not only in size but also in variety in recent era. The data collected is of very large amount and there is difficulty in collecting and assessing big data. Clusteringalgorithms have developed as a powerful metal earning tool which can precisely analyze the volume of data produced by modern applications. The main goal of clustering is to categorize data into clusters such that objects are grouped in the same cluster when they are “similar” according to similarities, traits and behavior. The most commonly used algorithm in clustering are partitioning, hierarchical, grid based, density based, and model-based algorithms. A review of clustering and its different techniques in data mining is done considering the criteria for big data. Where most commonly used and effective algorithms like K-Means, FCM, BIRCH,CLIQUE algorithms are studied and compared on big data perspective.

We know clustering algorithm have developed as a powerful Meta learning tools, and it analyze the volume of data. That means how big it is? Their main goal of clustering is to categorize data into clusters such that objects are grouped in the same cluster when they are “similar” according to similarities, traits and behavior. And for this clustering they actually compared different algorithm like k-mean, FCM, clique on big data perspective. [April 06-07, 2016]

2.4 Big Data clustering: A review-

Clustering is an essential data mining and tool for analyzing big data. There are difficulties for applying clustering techniques to big data duo to new challenges that are raised with big data. As Big Data is referring to terabytes and petabytes of data and clustering algorithms are come with high computational costs, the question is how to cope with this problem and how to deploy clustering techniques to big data and get the results in a reasonable time. This study is aimed to review the trend and progress of clustering algorithms to cope with big data challenges from very first proposed algorithms until today’s novel solutions. The algorithms and the targeted challenges for producing improved clustering algorithms are introduced and analyzed, and afterward the possible future path for more advanced algorithms is illuminated based on today’s available technologies and frameworks.

So, In this report they discussed improvement trend of data clustering algorithm. Although parallel clustering is potential and useful but the clustering of implementing algorithm is challenging. And they proved map reduce algorithm is satisfied for implementing clustering. And their result shown, M-reduce based algorithm offer impressive scalability and speed in comparison to serial counter parts while they are maintaining same quality

2.5 Optimized big data K-means clustering using MapReduce-

Here they have assumed and analyzed thatClustering analysis is one of the most commonly used data processing algorithms. Over half a century, K-means remains the most popular clustering algorithm because of its simplicity.Recently, as data volume continues to rise, some researchers turn to MapReduce to get high performance. However, MapReduce is unsuitable iterated algorithms owing to repeated times of restarting jobs, big data reading and shuffling.

In this paper, they address the problems of processing large-scale data using K-means clustering algorithm and propose a novel processing model in MapReduce to eliminate the iteration dependence and obtain high performance. They analyze and implement their idea. Extensive experiments on our cluster demonstrate that their proposed methods are efficient, robust and scalable.

Summary

These three research papers are very much close to our research topics. The difference is they did they are research on normal server, but we did our research on cloud. And we are using same algorithm as well that is map reducing algorithm. And we shown clustering massive amount of data on cloud computing is more flexible than the normal server. From 2nd paper they actually proved that, map reducing algorithm offer impressive scalability for clustering big data. As a result we did it on the cloud and we got more flexible result. For this we can manage our large amount of data set within short period of time. It's helpful for any company. Because they can retrieve and analyze their data within short time and the most important is data processing occur without any traffic.

CHAPTER 3:METHODOLOGY

3.1 Introduction

We used two processes. First one is map reducing algorithm on cloudera manager and second one is hierarchical Clustering by weka tools

We know clouding computing providing SaaS that means software as a service. So from cloud account we opened and installed the cloudera manager tools and another one is Weka tools. Now are explain the map reducing algorithm. How actually it works?

At first we have taken a cloud account then on that we made some node by map reduce algorithm. Then we made cluster. That was the first step of our work.

Map reduce is the combination of two word that is called map and reduce. The map is the first phase of the processing, where we specify all the complex logic and reduce is the second phase of processing where we specify light weight processing for example aggregation or summation .Suppose we have massive amount of data for example 10 terabyte data. So at first it will split in different nodes based on how developers specify the logic according to customer requirements. In this manner mapper split all the data in parallel way and intermediate storage device take the all data individually from mapper. And it copied and reduces the different node. Once all the mapper is finished and their output is shuffled on reduce nodes and then this intermediate output is merged and sorted. And which is go for reduce phase. And the reduce phase specify his own customer business logic as per the requirement. And output written HDFS.

Hadoop is an open-source Apache Software Foundation project written in Java that enables the distributed processing of large datasets across clusters of commodity. Hadoop has two primary components, namely, HDFS and MapReduce programming framework. The most significant feature of Hadoop is that HDFS and MapReduce are closely related to each other; each is co-deployed such that a single cluster is produced. Therefore, the storage system is not physically separated from the processing system.

3.2 Materials and Method

At firstwe used MapReduce method.The whole process goes through four phases of execution namely, splitting, mapping, shuffling, and reducing.

Consider you have following input data for your Map Reduce Program-

```
Welcome to Hadoop Class
Hadoop is good
Hadoop is bad
```

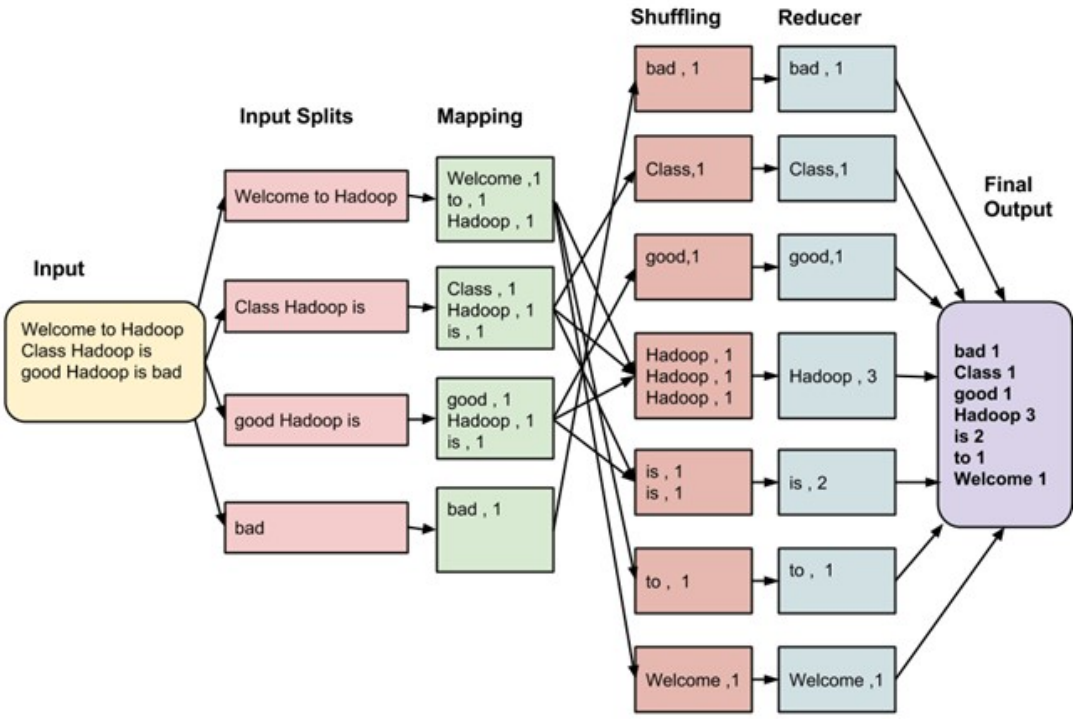

Fig 3. 1 MapReduce Architecture

Table3. 1The final output of the MapReduce algorithm

DATA	Times
Bad	1
Class	1
Good	1
Hadoop	3
Is	2
To	1
Welcome	1

The data goes through the following phases

➢**Input Splits:**

An input to a MapReduce job is divided into fixed-size pieces called **input splits** Input split is a chunk of the input that is consumed by a single map

➢**Mapping**

This is the very first phase in the execution of map-reduce program. In this phase data in each split is passed to a mapping function to produce output values. In our example, a job of mapping phase is to count a number of occurrences of each word from input splits (more details about input-split is given below) and prepare a list in the form of word, frequency

➢**Shuffling**

This phase consumes the output of Mapping phase. Its task is to consolidate the relevant records from Mapping phase output. In our example, the same words are clubed together along with their respective frequency.

➢**Reducing**

In this phase, output values from the Shuffling phase are aggregated. This phase combines values from Shuffling phase and returns a single output value. In short, this phase summarizes the complete dataset. In our example, this phase aggregates the values from Shuffling phase i.e., calculates total occurrences of each word.

3.2.1 MapReduce Architecture explained in detail

➢One map task is created for each split which then executes map function for each record in the split.

➢It is always beneficial to have multiple splits because the time taken to process a split is small as compared to the time taken for processing of the whole input. When the splits are smaller, the processing is better to load balanced since we are processing the splits in parallel.

➢However, it is also not desirable to have splits too small in size. When splits are too small, the overload of managing the splits and map task creation begins to dominate the total job execution time.

➢For most jobs, it is better to make a split size equal to the size of an HDFS block (which is 64 MB, by default).

➢Execution of map tasks results into writing output to a local disk on the respective node and not to HDFS.

➢Reason for choosing local disk over HDFS is, to avoid replication which takes place in case of HDFS store operation.

➢Map output is intermediate output which is processed by reduce tasks to produce the final output.

➢Once the job is complete , the map output can be thrown away. So, storing it in HDFS with replication becomes overkill.

➢In the event of node failure, before the map output is consumed by the reduce task, Hadoop reruns the map task on another node and re-creates the map output.

➢Reduce task doesn't work on the concept of data locality. An output of every map task is fed to the reduce task. Map output is transferred to the machine where reduce task is running.

➢On this machine, the output is merged and then passed to the user-defined reduce function.

➢Unlike the map output, reduce output is stored in HDFS (the first replica is stored on the local node and other replicas are stored on off-rack nodes). So, writing the reduce output

3.2.2How MapReduce Organizes Work?

Hadoop divides the job into tasks. There are two types of tasks:

1. Map tasks (Splits & Mapping)
2. Reduce tasks (Shuffling, Reducing).

As mentioned above .The complete execution process (execution of Map and Reduce tasks, both) is controlled by two types of entities called a

1. Jobtracker: Acts like a master (responsible for complete execution of submitted job)
2. Multiple Task Trackers: Acts like slaves, each of them performing the job

For every job submitted for execution in the system, there is one Jobtracker that resides on Namenode and there are multiple tasktrackers which reside on Datanode**.**

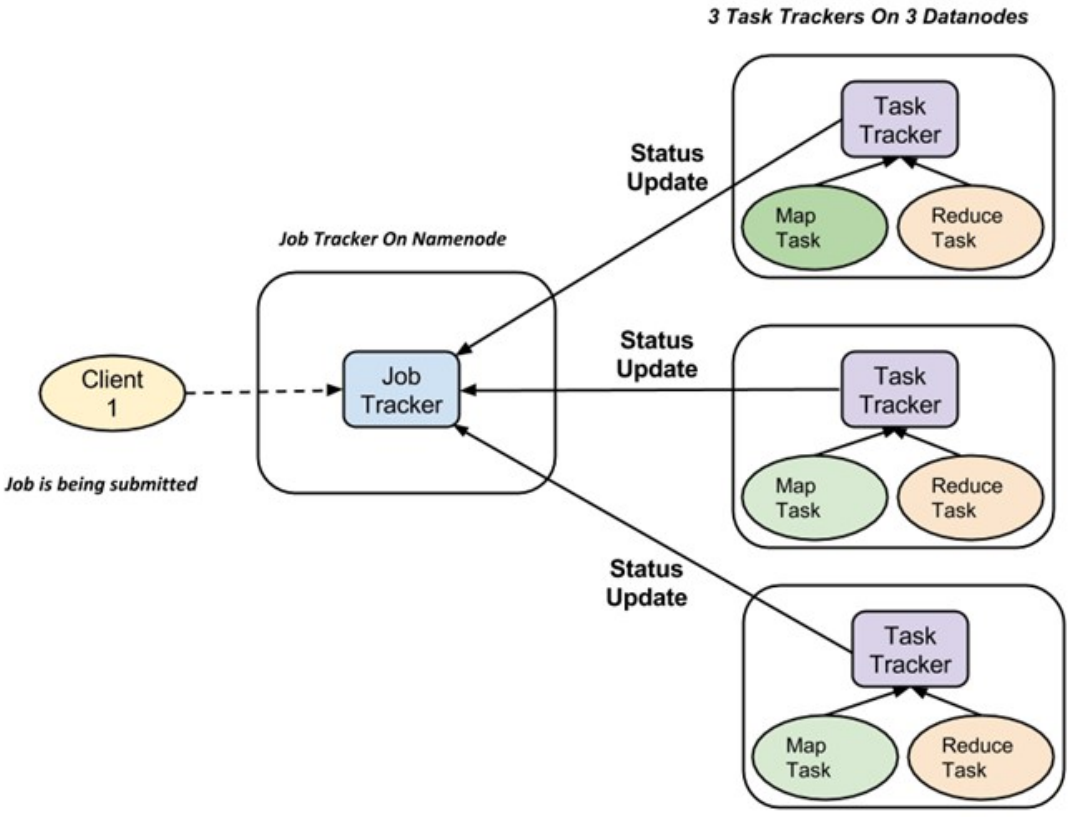

Fig 3.2. MapReduce Organizes Work

- A job is divided into multiple tasks which are then run onto multiple data nodes in a cluster.
- It is the responsibility of job tracker to coordinate the activity by scheduling tasks to run on different data nodes.
- Execution of individual task is then to look after by task tracker, which resides on every data node executing part of the job.
- Task tracker's responsibility is to send the progress report to the job tracker.
- In addition, task tracker periodically sends 'heartbeat' signal to the Jobtrackerso as to notify him of the current state of the system.
- Thus job tracker keeps track of the overall progress of each job. In the event of task failure, the job tracker can reschedule it on a different task tracker

3.3 hierarchical Clustering

Hierarchical clustering, also known *as* hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

If you want to do your own hierarchical cluster analysis, use the template below - just add your data!

3.3.1 Required data

Hierarchical clustering can be performed with either a distance matrix *or* raw data. When raw data is provided, the software will automatically compute a distance matrix in the background. The distance matrix below shows the distance between six objects.

	A	B	C	D	E
B	16				
C	47	37			
D	72	57	40		
E	77	65	30	31	
F	79	66	35	23	10

Fig: 3.3 Distance matrix

3.3.2 How hierarchical clustering works

Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps:

(1) Identify the two clusters that are closest together.

(2) merge the two most similar clusters. This continues until all the clusters are merged together. This is illustrated in the diagrams below.

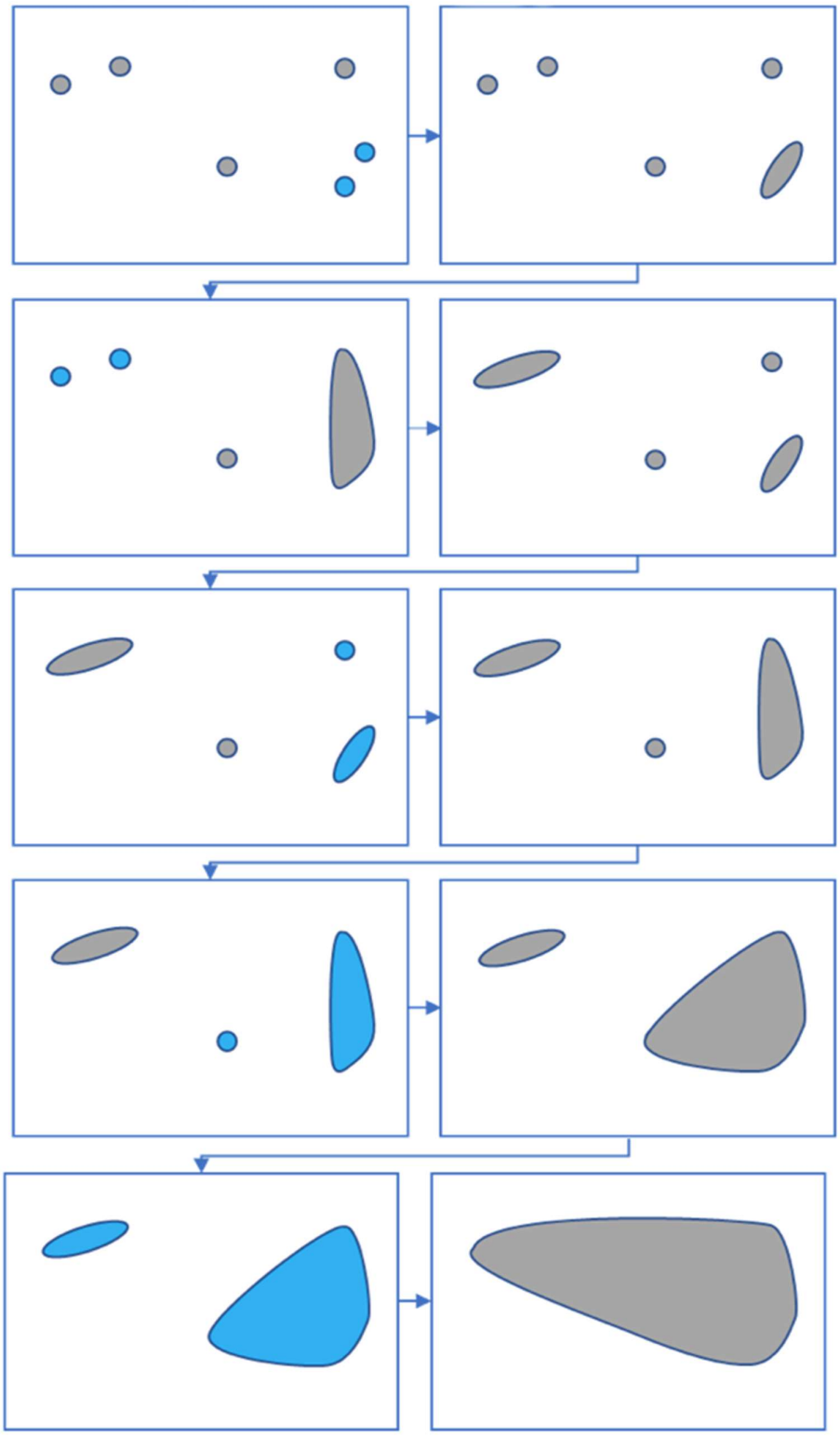

Fig: 3.4hierarchical clustering works

The main output of Hierarchical Clustering is a dendrogram, which shows the hierarchical relationship between the clusters:

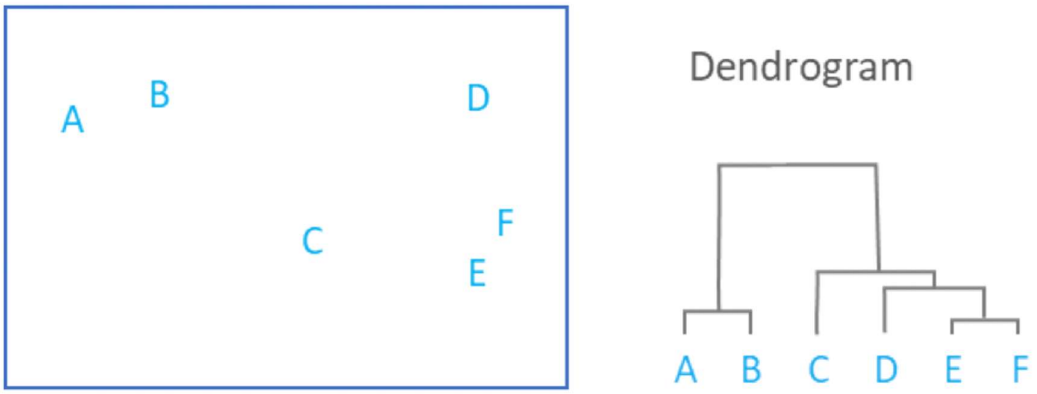

Fig: 3.5dendrogram

3.3.3 Measures of distance (similarity)

In the example above, the distance between two clusters has been computed based on length of the straight line drawn from one cluster to another. This is commonly referred to as the Euclidean distance. Many other distance metrics have been developed.

The choice of distance metric should be made based on theoretical concerns from the domain of study. That is, a distance metric needs to define similarity in a way that is sensible for the field of study. For example, if clustering crime sites in a city, city block distance may be appropriate (or, better yet, the time taken to travel between each location). Where there is no theoretical justification for an alternative, the Euclidean should generally be preferred, as it is usually the appropriate measure of distance in the physical world.

3.3.4 Linkage Criteria

After selecting a distance metric, it is necessary to determine from where distance is computed. For example, it can be computed between the two most similar parts of a cluster (single-linkage), the two least similar bits of a cluster *(*complete-linkage*),* the center of the clusters *(*mean *or* average-linkage*)*, or some other criterion. Many linkage criteria have been developed.As with distance metrics, the choice of linkage criteria should be made based on theoretical considerations from the domain of application. A

key theoretical issue is what causes variation. For example, in archeology, we expect variation to occur through innovation and natural resources, so working out if two groups of artifacts are similar may make sense based on identifying the most similar members of the cluster.

Where there are no clear theoretical justifications for choice of linkage criteria, *Ward's method* is the sensible default. This method works out which observations to group based on reducing the sum of squared distances of each observation from the average observation in a cluster. This is often appropriate as this concept of distance matches the standard assumptions of how to compute differences between groups in statistics

CHAPTER 4: RESULT AND DISCUSSION

4.1 Data Collection and Analysis:

For collecting data we used different types of camera because we need different format data. You know our rising massive amount of data is structured and unstructured. Then we took extra simple data like image, GIF, etc.Then we merged it and analyzed our data with our demo software. After completing all steps we took our data on cloud account.

For analyzing data we took different software for different result. Then we compared the appropriate result with cloud. Because whenever we got different result it's easy to find out exact result and easy to clustering on cloud as well.

Various categories data like Social media is the source of information generated via URL to share or exchange information and ideas in virtual communities and networks, such as collaborative projects, blogs and microblogs, Facebook, and Twitter. Machine data are information automatically generated from a hardware or software, such as computers, medical devices, or other machines, without human intervention.

Structured data are often managed SQL, a programming language created for managing and querying data in RDBMS. Structured data are easy to input, query, store, and analyze. Examples of structured data include numbers, words, and dates. Semi-structured Semi-structured data are data that do not follow a conventional database system. Semi-structured data may be in the form of structured data that are not organized in relational database models, such as tables. Capturing semi-structured data for analysis is different from capturing a fixed file format. Therefore, capturing semi-structured data requires the use of complex rules that dynamically decide the next process after capturing the data. Unstructured Unstructured data, such as text messages, location information, videos, and social media data, are data that do not follow a specified format. Considering that the size of this type of data continues to increase through the use of smartphones, the need to analyze and understand such data has become a challenge.

Table4. 1Various categories of big data

Classification	Description
Social media	Social media is the source of information generated via URL to share or exchange information and ideas in virtual communities and networks, such as collaborative projects, blogs and microblogs, Facebook, and Twitter
Machine-generated data	Machine data are information automatically generated from a hardware or software, such as computers, medical devices, or other machines, without human intervention
Sensing	Several sensing devices exist to measure physical quantities and change them into signals
Transactions	Transaction data, such as financial and work data, comprise an event that involves a time dimension to describe the data
IoT	IoT represents a set of objects that are uniquely identifiable as a part of the Internet. These objects include smartphones, digital cameras, and tablets. When these devices connect with one another over the Internet, they enable more smart processes and services that support basic, economic, environmental, and health needs.

We compared with previous paper as well. Because we are taking cloud for this research but they did normal server for their research. Al though the algorithm is same but different resources. They got positive result by map reduce algorithm and it's really effective but they have some limitation like it's not so faster, sometimes server is down etc. But in cloud it will be faster and no way to server down because we will do clustering.

Table4. 2 Data staging

Classification	Description
Cleaning	Cleaning is the process of identifying incomplete and unreasonable data.

Transform	Transform is the process of transforming data into a form suitable for analysis.
Normalization	Normalization is the method of structuring database schema to minimize redundancy

Table4. 3 Data processing

Classification	Description
Batch	MapReduce-based systems have been adopted by many organizations in the past few years for long-running batch jobs. Such system allows for the scaling of applications across large clusters of machines comprising thousands of nodes
Real time	One of the most famous and powerful real time process-based big data tools is simple scalable streaming system. S4 is a distributed computing platform that allows programmers to conveniently develop applications for processing continuous unbounded streams of data. S4 is a scalable, partially fault tolerant, general purpose, and pluggable platform

4.2 Tools:

We used two tools for analyzing data.

- WEKA
- CLOUDERA MANAGER

4.2.1 WEKA:

In weka software we used Iris dataset and for clustering we used Hierarchical Cluster process. Now we are adding all screenshots of our data on weka tools

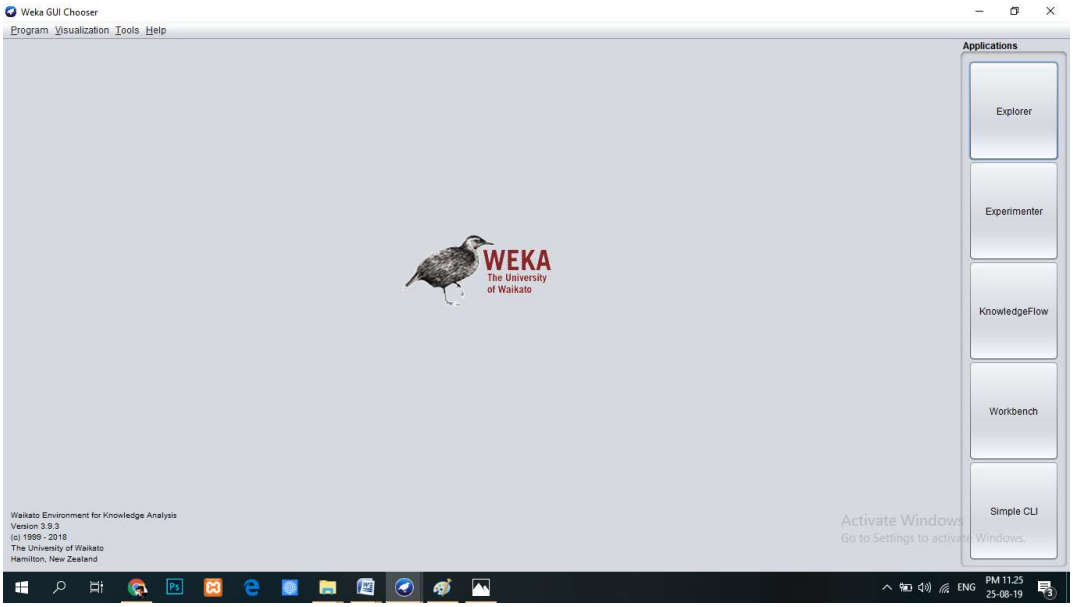

Fig: 4.1 Weka software

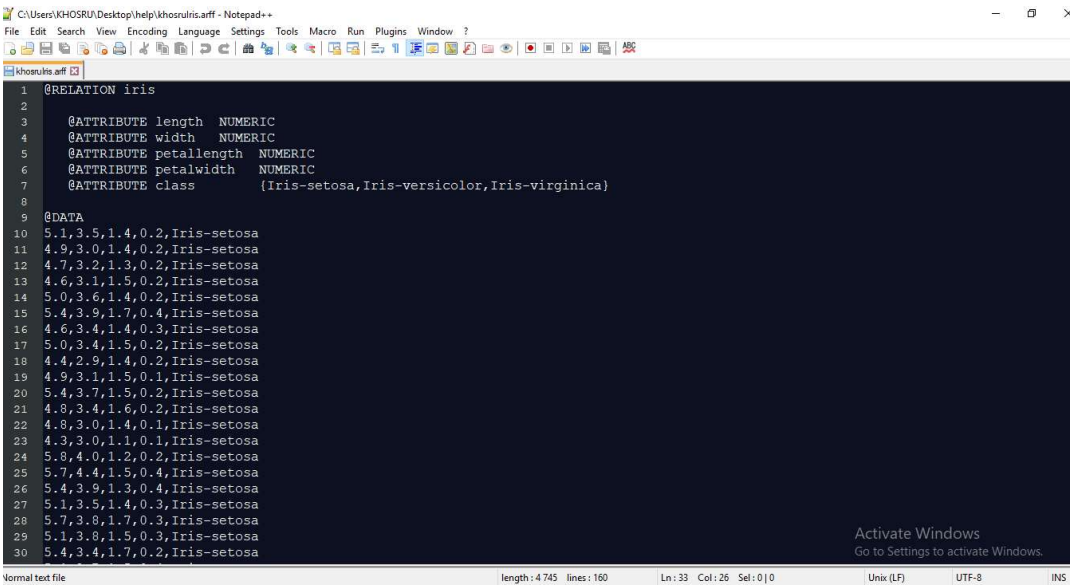

Fig:4.2 Iris data set

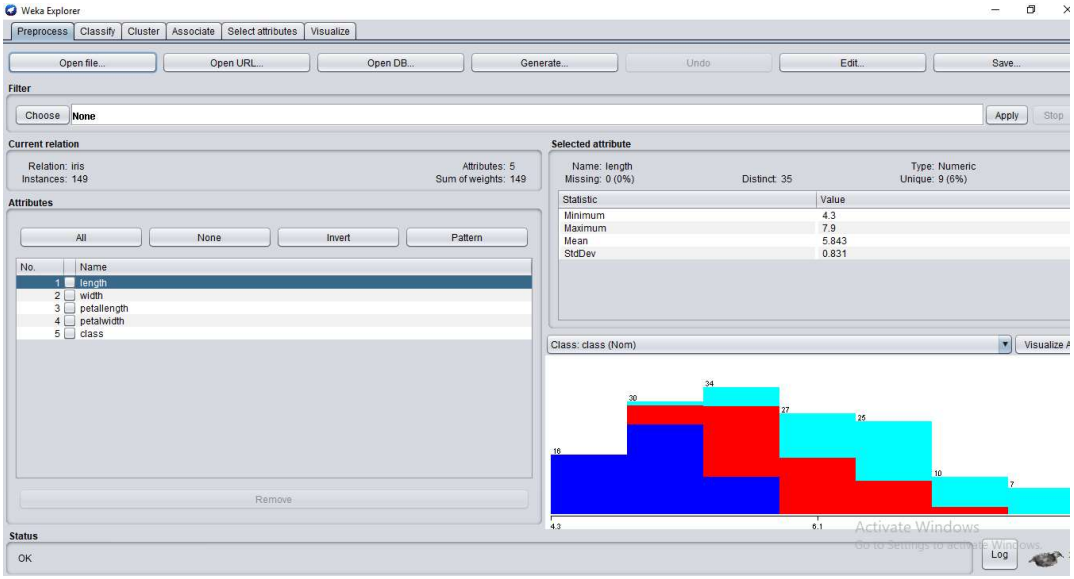

Fig: 4.3 Length class

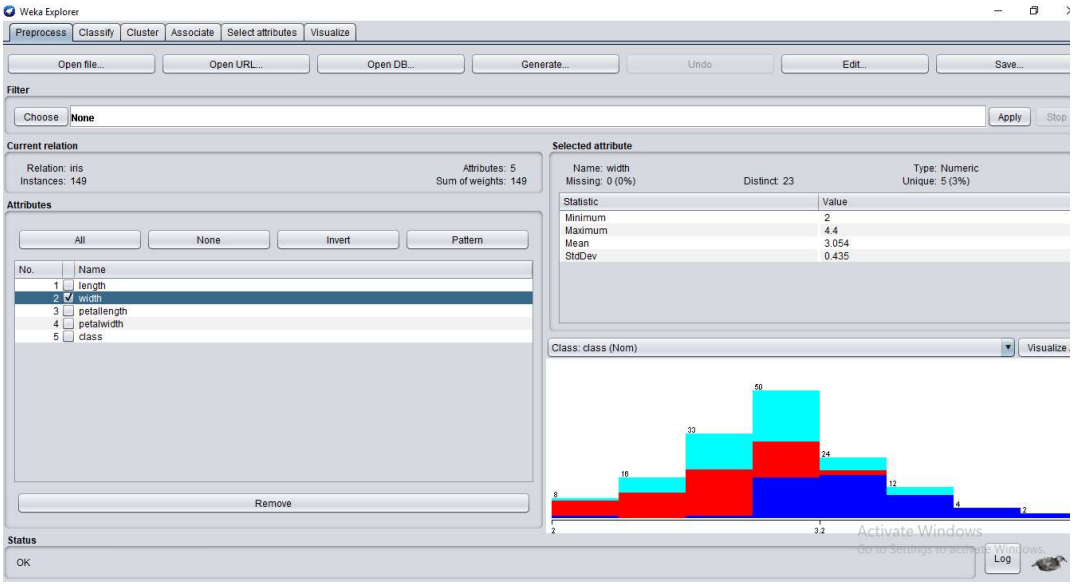

Fig: 4.4 Width class

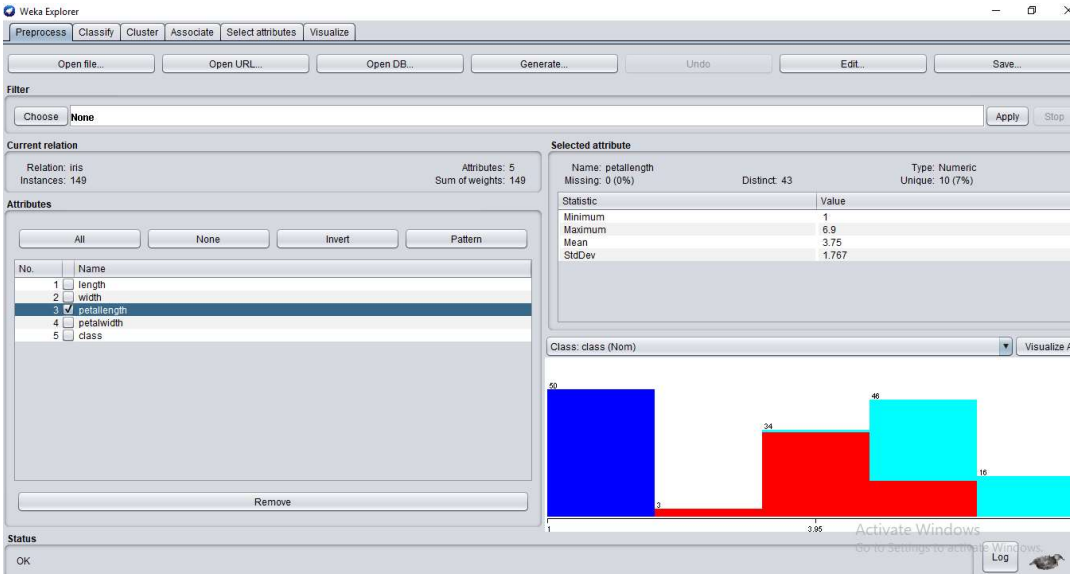

Fig: 4.5Petalength class

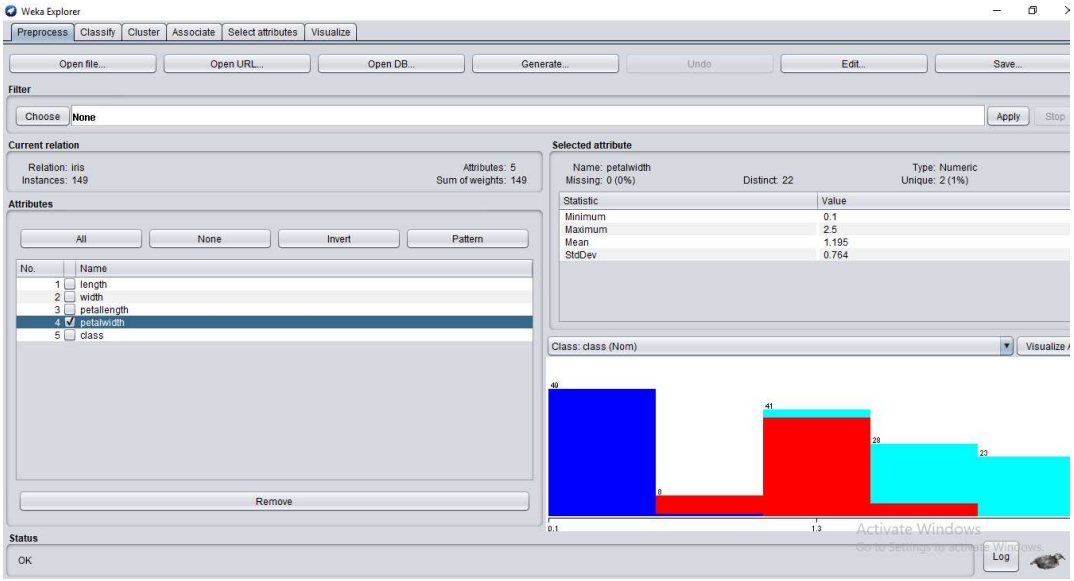

Fig: 4.6PetaWidth class

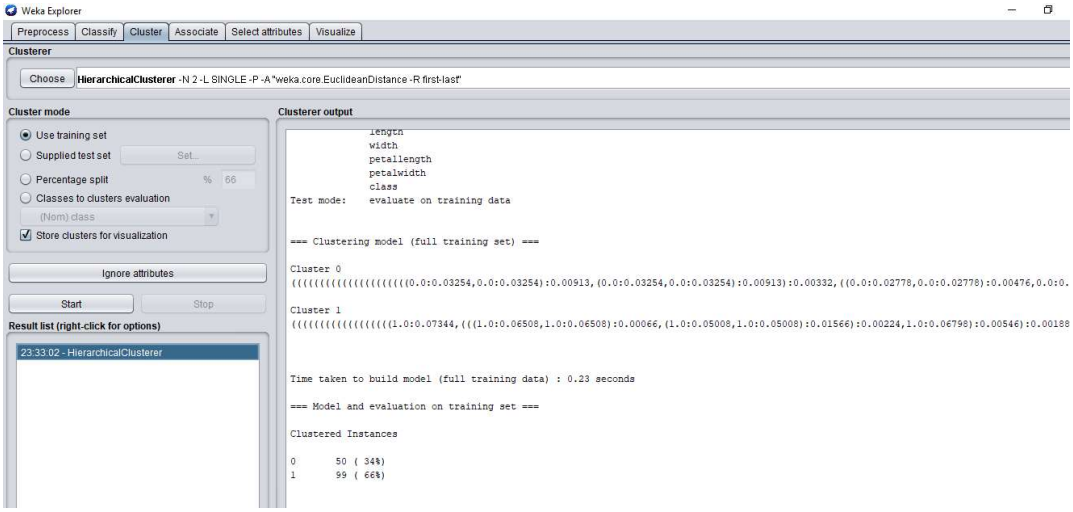

Fig: 4.7Hierarchical Cluster

4.2.2 CLOUDERA MANAGER

For map reducing algorithm we used cloudera manager and put job history web UI information for implementing map reducing clustering. Now we are adding all screenshots of our data on cloudera manager tools-

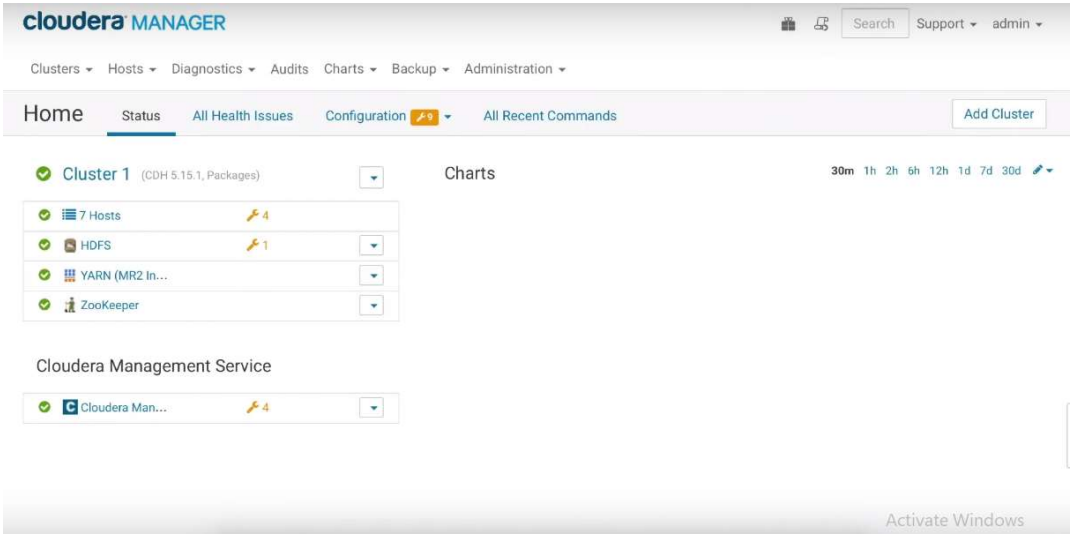

Fig: 4.8 Dashboard of cloudera manager

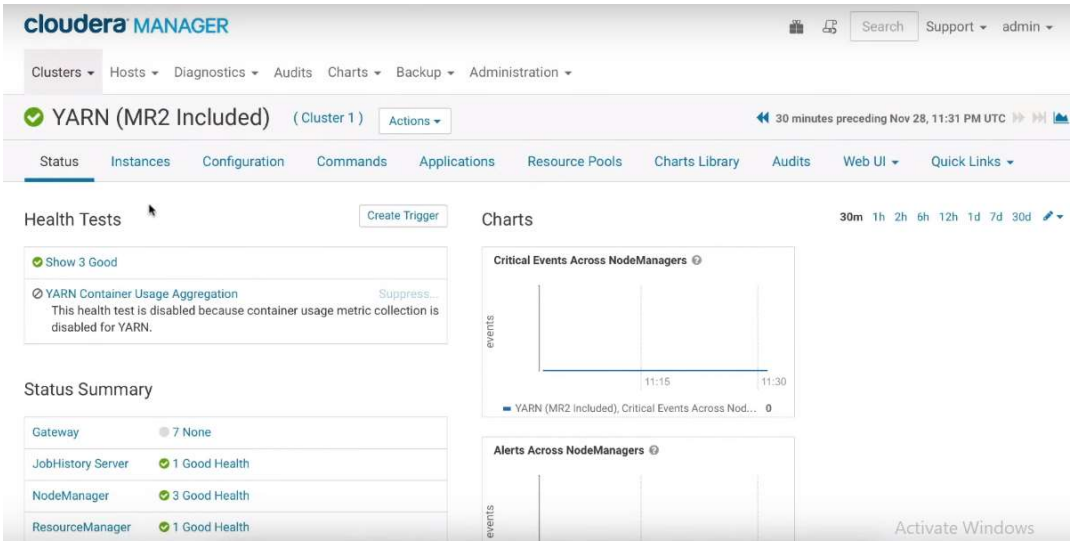

Fig: 4.9 YARN page of cloudera manager

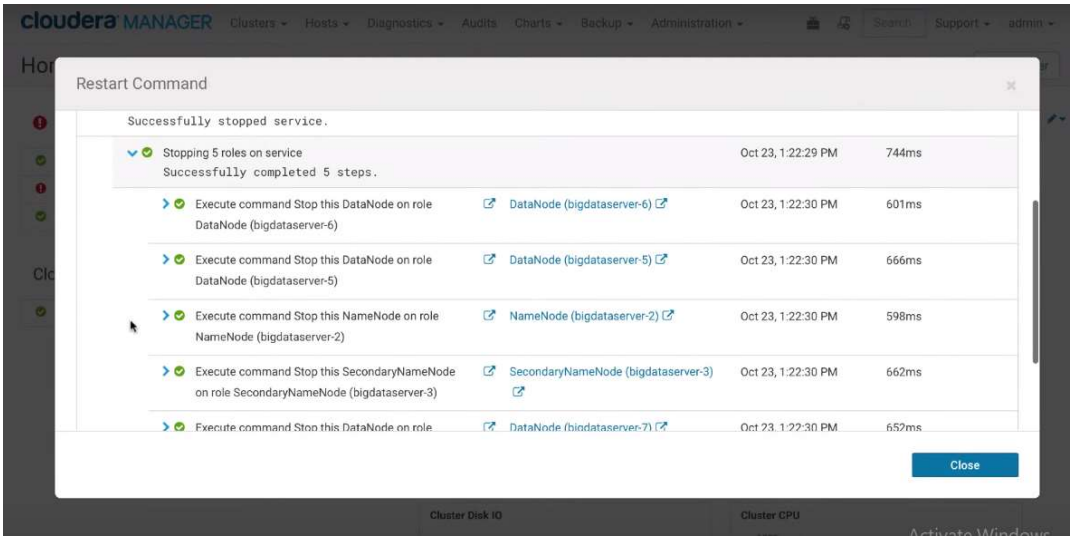

Fig: 4.10 Uploaded file

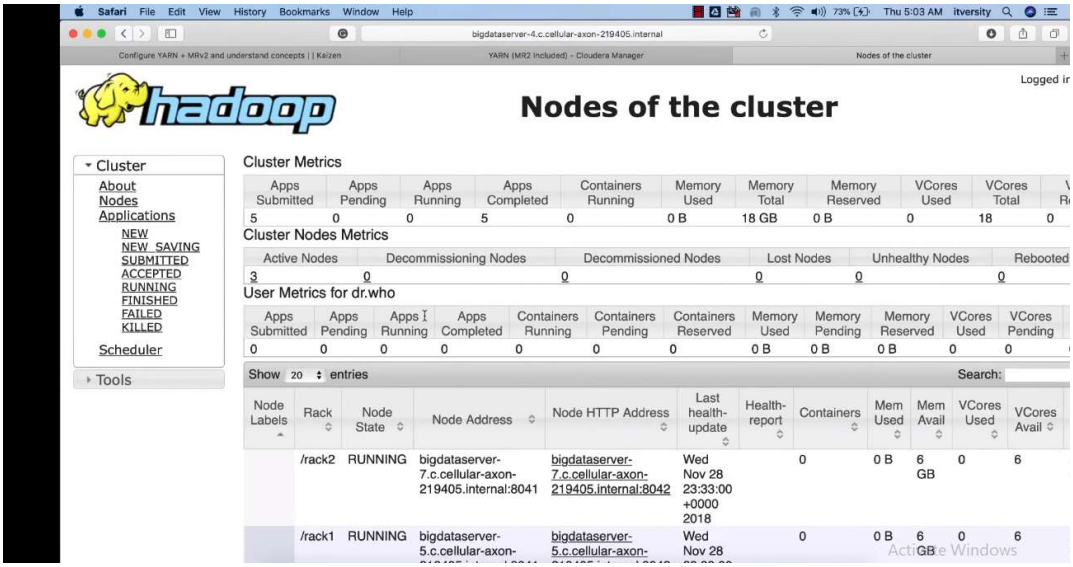

Fig: 4.8 Dashboard of cloudera manager

Fig: 4.9 YARN page of cloudera manager

Fig: 4.10 Uploaded file Fig: 4.11 All Application

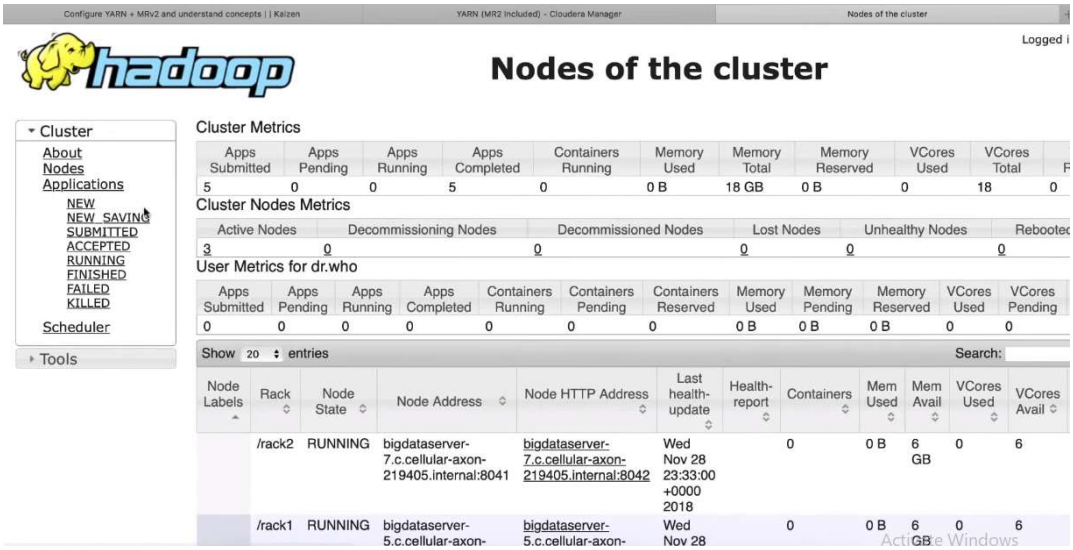

Fig: 4.12 Nodes of the cluster

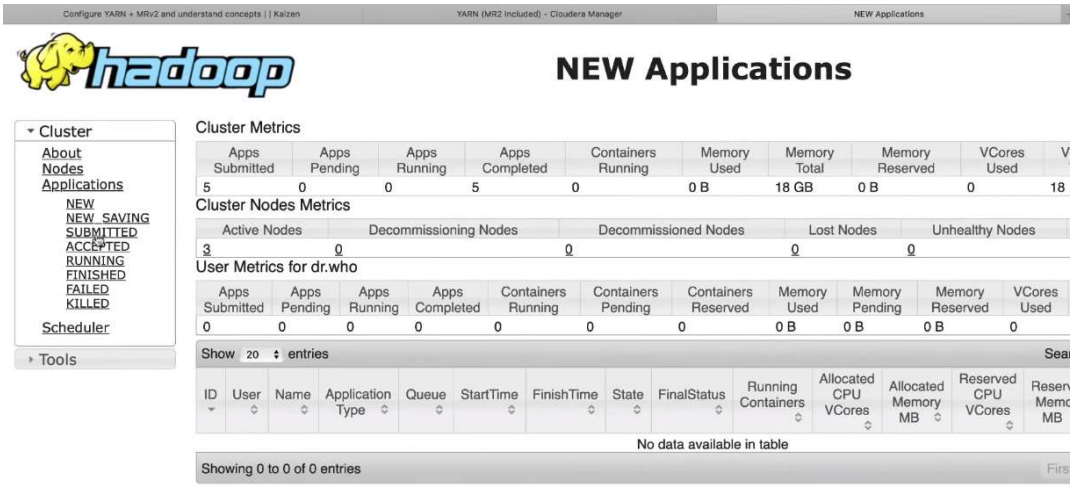

Fig: 4.13 New Application

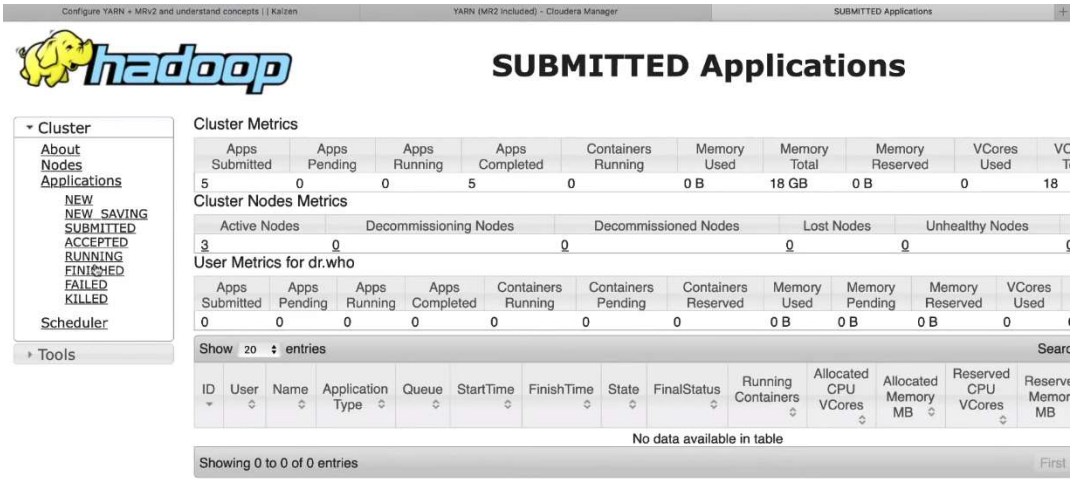

Fig: 4.14 Submitted Application

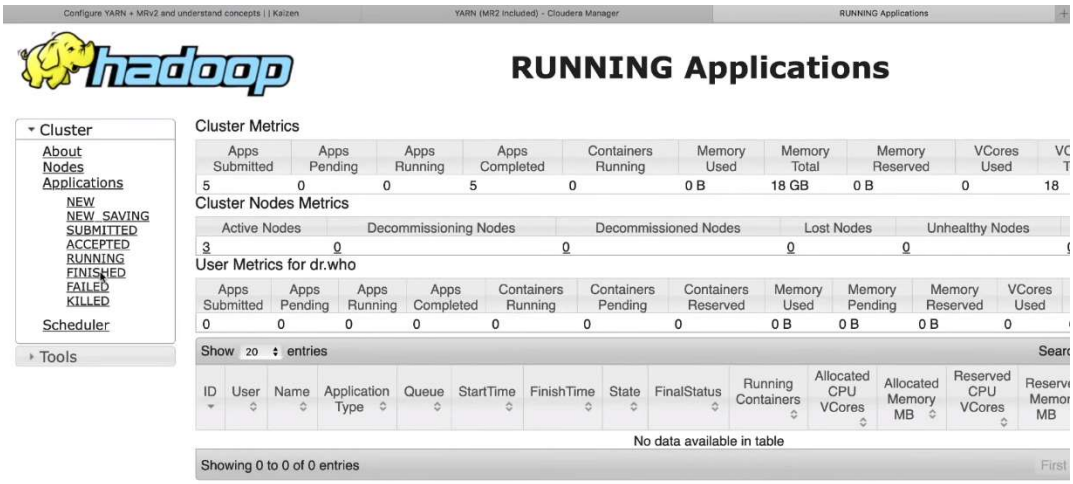

Fig: 4.15 Running Application

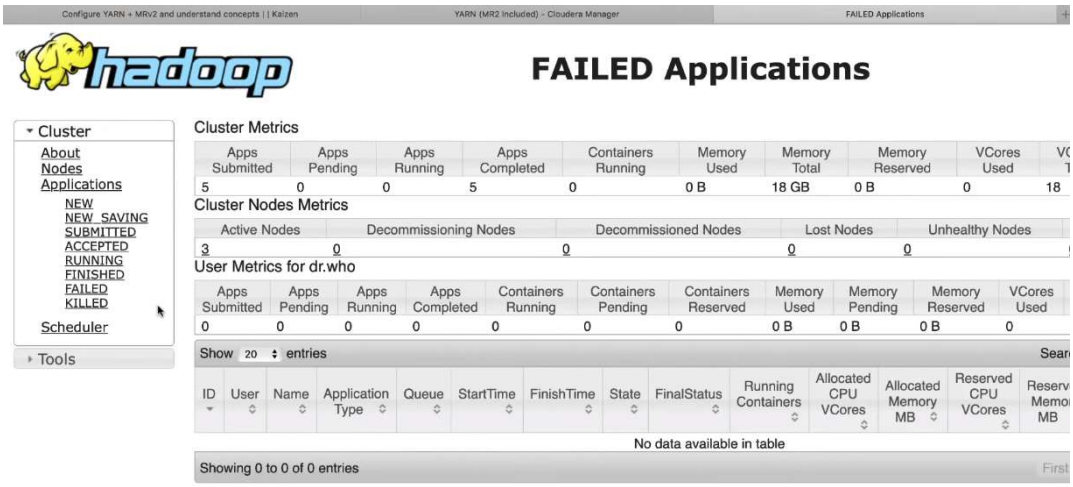

Fig: 4.16 Failed Application

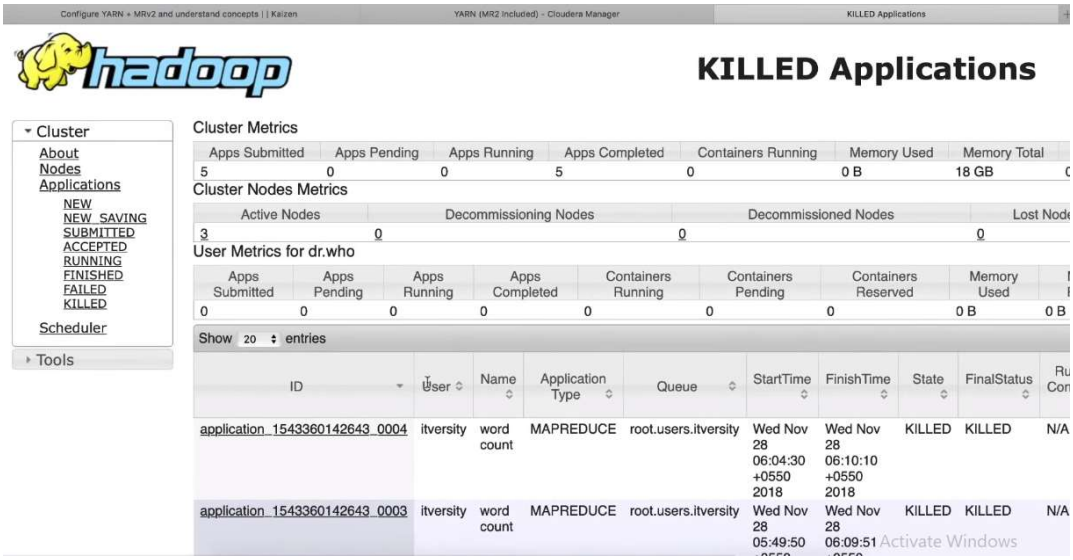

Fig: 4.17 Killed Application

4.3 Calculation

1.Sequential application of methodology which we used in our report isUsing Map reducing algorithm. We completed the data clustering on cloud as a result we got the fastest retrieving data from cloud rather than normal server. For retrieving data Map reducing algorithm are the dominating factor and here also work HDFS from hadoop. Our massive amounts of data we can easily retrieve it and process it.

2. We got the flexible result by using map reducing algorithm. Because our comparison was when we will use cloud platform for clustering our massive amount of data weather it is flexible or not. Flexibility proved that it is easier processing to retrieve and store our massive amount of data.

3.At the same time we used another clustering process that is hierarchical Clustering by using weka. It also provide us flexibility to retrieving, storing and processing data which is more flexible rather than the normal server. Hats why beside another cloud server, to show and make clear understanding, we did also use weka.

4. When we comparemap reducing algorithm with hierarchical Clustering than we got both are easy and flexible but map reducing is easier and flexible rather than the hierarchical Clustering. So we will prefer map reducing than hierarchical clustering as a suggestion.

5.After comparing with other research,we find in our research-

More flexible and easy to access data and also more secure data on cloud. No chance to loss our data and for future analysis it gives us the new method which is very fast and less time consumption. We can use both process as well. It helps the any company to get high profit because it decreases the employee and increase the fast process and analysis. Clustering data refer to high scalability. The main benefit of Clustered Cloud servers is increase redundancy and high availability through automatic failover. This is the process of transferring the virtual machines (Clustered Cloud Servers) on top of one physical node to another when the first fails, making it possible to keep critical systems available with minimal interruption. This is a valuable feature of the Clustered Cloud Servers.

These are the benefits or flexibility we got in our research which are different from others,although our method is same but as resources are different,cause we did the research after putting our data or file on cloud environment,doing different clustering and nodes,that results these flexibility.

4.4 Result Analysis

As we did quantitative research,so in our research we didn't need participants,we took data from diffetent resources. And these are the results we got in our research.

By using Map reducing algorithm we completed the data clustering on cloud as a result we got the fastest retrieving data from cloud rather than normal server. For retrieving data Map reducing algorithm are the dominating factor and here also work HDFS from hardtop. Our storing 1 terabyte data we can easily retrieve it.

- **Flexibility:**When we did cluster our dataset on cloud it gave us more flexibility in each field like data processing, data storing, data retrieving and data analysis as well. And it's so flexible and faster
- **Faster data access:** Faster data accessing is the main goal of this research. And we can retrieve our data within few second.
- **Faster processing:** Faster data accessing as well as faster data processing are possible by this research. People don't want to longer the time. So, our system minimizes the time and it faster data processing capable.

- **Scales to a very large size market:** Huge number of data it can store,analyze, cluster and retrieve. So, it's easy to do any market research.
- **Greater scalability:** As your user base grows and report complexity increases, your resources can grow in an extension.
- **Resource availability:**Increased resource availability: If one Intelligence Server in a cluster fails, the other Intelligence Servers in the cluster can pick up the workload. This prevents the loss of valuable time and information if a server fails.
- **Strategic resource usage:**You can distribute projects across nodes in whatever configuration you prefer. This reduces overhead because not all machines need to be running all projects and allows you to use your resources flexibly.

- **Increased performance:** Multiple machines provide greater processing power.

- **Simplified management:**Clustering simplifies the management of large or rapidly growing systems.

- **Load redistribution:** When a node fails, the work for which it is responsible is directed to another node or set of nodes.

- **Request recovery:** When a node fails, the system attempts to reconnect MicroStrategy Web users with queued or processing requests to another node. Users must log in again to be authenticated on the new node. The user is prompted to resubmit job requests.

- **Load balancing:** load balancing is a strategy aimed at achieving even distribution of user sessions across Intelligence Servers, so that no single machine is overwhelmed. This strategy is especially valuable when it is difficult to predict the number of requests a server will receive. Micro Strategy achieves four-tier load balancing by incorporating load balancers into the Micro Strategy Web and Web products.
- **Work Fencing:** User fences and workload fences allow you to reserve nodes of a cluster for either users or a project subscription.

4.5 Discussion

As from the result section, after implementing Map Reducing method, we got results from that which satisfy our hypothesis part. It clearly proves and show appropriate result of our assumption that as the rise of big data clustering on cloud can create lots of traffic, nonflexible things but using a appropriate clustering method can overcome the nonflexibility, create a good result.

And as we know, by using Map reducing algorithm we completed the data clustering on cloud as a result we got the fastest retrieving data from cloud rather than normal server. For retrieving data Map reducing algorithm are the dominating factor and here also work HDFS from Hadoop. Our storing 1 terabyte data we can easily retrieve it.

So, in result partwe got,

Faster data access: Faster data accessing is the main goal of this research. And we can retrieve our data within few second.

Faster processing: Faster data accessing as well as faster data processing are possible by this research. People don't want to longer the time. So our system minimize the time and it faster data processing capable. Scales to a very large size: Huge number of data it can store , analyze ,cluster and retrieve .So it's easy to do any market research. Thus, these are the importance of new results what we got after our studies about clustering and all.

And beside this, these points were the interpretation of our research, which has been described clearly above upper part

CHAPTER 5: CONCLUSION

5.1 Conclusion

At last we can say cloud is a platform as well as they provides many services. And our rising massive amount of data we can store in it by different technique. And clustering data is one of them. I showed you the time table and every step. By our research may be you don't need to use normal server, you don't need to extra employee, space, managing fee etc. Just you will put your data on cloud and it will be clustered also you can analyze your dataset for growing your business.

As our objective was to find out the big data clustering method that will create good impact on cloud computing as if rise of big data don't occur traffic on cloud when we are storing data or accessing it ,this study conclude this objective.

The size of data at present is huge and continues to increase every day. The variety of data being generated is also expanding. The velocity of data generation and growth is increasing because of the proliferation of mobile devices and other device sensors connected to the Internet. These data provide opportunities that allow businesses across all industries to gain real-time business insights. The use of cloud services to store, process, and analyze data has been available for some time; it has changed the context of information technology and has turned the promises of the on-demand service model into reality. In this study, we presented some review on the rise of big data in cloud computing.

Beside there are some limitations of our research also, that is very obvious or normal. As we are implementing Map Reducing method here to cluster data and our hypothesis becomes positive but may be or it is certain that some new types of algorithm will be invented to do clustering in more better way also.

As it's a research ,some more good effects of using this may be coming, that means for other context but on cloud plateform,may get new result, more flexible which may not available now.

5.2 Recommendations:

We explained all thing whatever we did. We used map reducing algorithm and it works and gave us positive result. Now we can try by other resources or another algorithm to compare with map reducing algorithm. For doing this research more effectively our suggestion is to take 5 months for collecting and analyzing data. Because you know our rising massive amount of data are different format. So, for clustering we have identified the each and every format because we are doing multi node cluster. Although we are using map reducing algorithm for identify the different format of data and cluster them by similarities.

And may be this field grows more in future, with more big data and fields. So for then ,it will need some other ,more flexible algorithm to face those, as all these thing is changing with the phase of time.

REFERENCES

[2] Douglas and Laney, “The importance of ‘big data’: A definition,” 2008.

[3] D. Kossmann, T. Kraska, and S. Loesing, “An evaluation of alternative architectures for transaction processing in the cloud,” in Proceedings of the 2010 international conference on Management of data. ACM, 2010, pp. 579–590.

[4] S. Ghemawat, H. Gobioff, and S. Leung, “The google file system,” in ACM SIGOPS Operating Systems Review, vol. 37, no. 5. ACM, 2003, pp. 29–43.

[6] M. Cox, D. Ellsworth, Managing Big Data For Scientific Visualization, ACM Siggraph, MRJ/NASA Ames Research Center, 1997.

[7] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H. Byers, Big data: The next frontier for innovation, competition, and productivity, (2011)

[8] P. Zikopoulos, K. Parasuraman, T. Deutsch, J. Giles, D. Corrigan, Harness the Power of Big Data The IBM Big Data Platform, McGraw Hill Professional, 2012.

[9] J.J. Berman, Introduction, in: Principles of Big Data, Morgan Kaufmann, Boston, 2013, xix–xxvi (pp).

[10] J. Gantz, D. Reinsel, Extracting value from chaos, IDC iView (2011) 1–12.

[11] J.K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, The mobile data challenge: Big data for mobile computing research, Workshop on the Nokia Mobile Data Challenge, in: Proceedings of the Conjunction with the 10th International Conference on Pervasive Computing, 2012, pp. 1–8.

[12] D.E. O’Leary, Artificial intelligence and big data, IEEE Intell. Syst. 28 (2013) 96–99.

[13] M. Chen, S. Mao, Y. Liu, Big data: a survey, Mob. Netw. Appl. 19 (2) (2014) 1–39.

[14] J. Hurwitz, A. Nugent, F. Halper, M. Kaufman, Big data for dummies, For Dummies (2013).

[15]https://planningtank.com/computer-applications/cluster-analysis-admin Category - Computer Applications Published On -23/06/2018 Last Updated On -14/07/2018

[16]https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/

[17] Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D.J.. Rasin, A., and Silberschatz, A. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. PVLDB 2(1), 2009.

[18] Agrawal, D., Das, D., and El Abbadi, A. Big Data and Cloud Computing: Current State and Future Opportunities. Proc. of EDBT, 2011.

[19] Apache Hadoop. http://wiki.apache.org/hadoop

[20] Cattell, R. Scalable SQL and NoSQL Data Stores. SIGMOD Record 39(4), 2010.

[21] Chen, Q., Hsu, M., and Liu, R. Extend UDF Technology for Integrated Analytics. Proc. of DaWaK, 2009.

Appendices

Budget: Here is the expenditure calculation:

Table:7.1-Budget

Number	Expenditure Type	US ($) Dollar
01	Cloud account	$4000
02	Software(analyzing data)	$3000
03	Transport	$2000
04	Equipment	$3000
05	Others	$2000
	Total Amount	**$14000**

Time Table:7.2-Our Research time duration following table:

Activities	Time Duration (Days)
Data Collection	30
Data Analysis	30
Data Organizing	30
Create Cloud account and make node	25
Experiment (Putting data on cloud)	35
Report Research Paper Writing	30
	Total = 6 Month