# Data Intensive Computing in the Clouds

N Suresh Goud


**Department of Computer Science and Engineering**

**National Institute of Technology Rourkela**

**Rourkela-769 008, Odisha, India**

# Data Intensive Computing in the Clouds

*Thesis submitted to the department of*

***Computer Science and Engineering***

*of*

***National Institute of Technology Rourkela***

*in partial fulfillment of the requirements*

*for the degree of*

## Master of Technology

*by*


## N Suresh Goud

[ Roll No. 211CS1044 ]




**Department of Computer Science and Engineering**

**National Institute of Technology Rourkela**

**Rourkela-769 008, Odisha, India**

# Acknowledgment

I would like to express my gratitude to my friends for the useful comments,remarks and engagement through the learning process of this master thesis.The flexibility of work offered me has deeply encouraged me producing the research. Furthermore I would like to thank Dr. Ashok Kumar Turuk for being a source of support and motivation for carrying out quality work.

I wish to thank all faculty members and secretarial staff of the CSE Department for their sympathetic cooperation.

I am really thankful to my all friends. My sincere thanks to everyone who has provided me with kind words, a welcome year, new ideas, useful criticism, and their invaluable time, I am truly indebted.

My thanks and apologies to those who I have inadvertently missed out.

Finally, Above all ***I praise God*** for helping me in completing my work.

***N Suresh Goud***

# Abstract

Data intensive computing deals with computational methods and architectures to analyze and discover intelligence in huge volumes of data generated in many application domains. Cloud computing is a major aim for data intensive computing, as it allows scalable processing of massive amount of data. It is a promising next-generation computing paradigm given its many advantages such as scalability, reliability, elasticity, high availability, and low cost. Here the problem is huge data and is growing exponentially. Here we are using MapReduce model to compress data. But the problem with MapReduce is fault tolerance and reliability, this problem can be overcomed by using Hadoop framework. Hadoop framework will provide high fault tolerance, high throughput and is suitable for applications with large data sets, streaming access to file system data and it can be built out of commodity hardware.

# Contents

**Certificate** ii

**Acknowledgement** ii

**Abstract** iii

**List of Figures** vi

**List of Tables** viii

**1 Introduction** 1

1.1 Introduction . . . . . 1

1.2 Cloud Computing . . . . . 2

1.3 Cloud computing service . . . . . 3

1.4 Motivation . . . . . 4

1.5 Problem Statement . . . . . 4

1.5.1 How the data is generated . . . . . 5

1.5.2 Sample events generating data . . . . . 7

1.6 Objective . . . . . 8

1.7 Organisation of Thesis . . . . . 8

**2 Related work** **10**

2.1 Basic Definitions: . . . . . 10

2.2 DATA INTENSIVE COMPUTING . . . . . 15

2.2.1 DATA PROCESSING PIPELINE . . . . . 16

**3 Proposed Method** **23**

3.1 MapReduce . . . . . 23

3.2 Hadoop . . . . . 24

3.2.1 Fault tolerance . . . . . 27

**4 Conclusion** **29**

**Bibliography** **30**

# List of Figures


1.1 Without Cloud . . . . . 3

1.2 Cloud . . . . . 3

1.3 Growth of data speed and size . . . . . 5

1.4 Top ten largest databases in (2007) . . . . . 6

1.5 Facebook (21 petabytes in 2010) . . . . . 7

2.1 Platform as a Service . . . . . 12

2.2 Private cloud . . . . . 13

2.3 Public Cloud . . . . . 14

2.4 Hybrid Cloud . . . . . 14

2.5 Without Cloud . . . . . 16

2.6 Without Cloud . . . . . 21

3.1 MapReduce . . . . . 24

3.2 Concept of Hadoop . . . . . 25

3.3 Slave in Hadoop . . . . . 25

3.4 Hadoop . . . . . 26

3.5 Fault Tollerance in Hadoop . . . . . 27

3.6 Traditional Database System . . . . . 28

# List of Tables

1.1 . . . . . 5

1.2 . . . . . 6

# Chapter 1

# Introduction

## 1.1 Introduction

Data intensive computing is the process of collecting, managing, analyzing, and understanding data at volumes and rates that push the frontier of current technologies. It deals with computational methods and architectures to analyse and discover intelligence in huge volumes of data generated in many application domains.

Cloud computing is a major aim for data intensive computing, as it allows scalable processing of massive amount of data [1]. It is a promising next-generation computing paradigm having many advantages such as scalability, reliability, elasticity, high availability, and low cost.

## 1.2 Cloud Computing

Cloud is the computer hardware that holds all digital data. It is a computing platform capable of distributing applications and data to geographically dispersed location. There is no cloud without data. You can use the cloud through internet, remote location and mobile applications.

Cloud computing is a pay for use model and you only pay what you use, when you use, like renting furnished home. The users are charged according to whatever they use in cloud services like computing, storage, and network service [2]. It is emerging information technology and delivery model that enables real time delivery of configurable computing resources. It stores the data in many places through virtualization. It can help you create more secure and efficient infrastructure with private and public. The infrastructure service whole thing is maintained by providers like softlayer amazon, Microsoft, and rockspace. It becomes very significant in the information technology and business model, it have limited storage and computing capacity. The user no needs to buy software and hardware through internet they can use. The key of cloud computing is not focusing on the hardware is just focused on your website.

Salesforce.com is the first mover of cloud computing (1999). They introduced the concept of delivering enterprise applications via simple website. Amazon is the next mover(2002). Google docs brought the cloud computing to the fore front of public consciousness (2006). Amazon is introduced Elastic Compute Cloud (2006) and recently Microsoft started Microsoft Azure (2010) in cloud computing.

Without cloud the organizations in past is

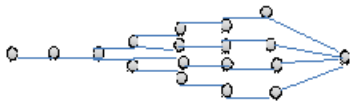

Figure 1.1: Without Cloud

When an organization started that will be developed like this, but here it will take a lot of time, money and everything. It is very difficult to maintain a organization in the past. But, presently using cloud no need of anything you can get it just through internet.


Figure 1.2: Cloud

## 1.3 Cloud computing service

- Software as a service (SaaS)

- Platform as a service (PaaS)
- Infrastructure as a service (IaaS)

## 1.4 Motivation

Data is growing and growing at an enormous rate. Each day 2.5 quintilion bytes of data is generated and available across different domains life and work. Due to the enormous size of data, it becomes difficult to store and analyze data efficiently. The traditional systems are incapable to scaling of data at this rate. Hence, we must come up with efficient and novel storage and processing solutions to handle this data. In this regard, we propose a novel based on cloud to aid in storage and processing of big data.

## 1.5 Problem Statement

The growing of data is a big problem. Big data is creating large growing files and these files are of terabytes or petabytes. The data comes from users, application, system and sensors. Data is so large that it becomes difficult to process using the traditional system. The big data challenges are Velocity, Volume, and Varity.

The business people find it difficult to analyze the unstructured data, due to which unstructured data gets wasted and is not captured, even if captured its not easy to analyze

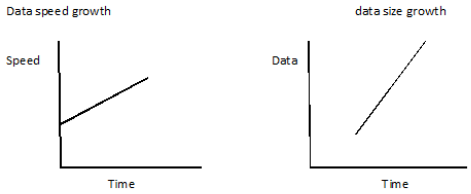


Figure 1.3: Growth of data speed and size

| Parameters | 90's | 00's | 10's |
|---|---|---|---|
| Capacity | 2.1GB | 200GB | 3000GB |
| Price | $157 | $0.5 | $0.05 |
| Speed | 16.6MBPS | 56.5MBPS | 210MBPS |
| Time Required | 126sec | 58min | 4hours |

Table 1.1:

### 1.5.1 How the data is generated

Every person, every employee, every organization and everybody in the world needs data. Their generated data is very large. In today's world the growing of data is a big challenge. Previously, the user brought the data from single processor but now everybody is using multiple servers and processers.

Data is being generated and gathered by workers. The data is stored in the form of digital. In these days Facebook is 1st place to generate more data,

| Unsorted Data | Structured Data |
|---|---|
| 90% | 10% |

Table 1.2:

everybody is maintaining accounts and sharing photos and videos. It is clearly mentioned in the below figure about data generation. The data is generated through social media (like Facebook and Google), YouTube's, text messages, videos, climate modelling, twitter tweets, blogs emails, photos and music. Every organization and business and everything in the world running based on the data only. So it is estimated that data could be 33 zettabytes by 2020.

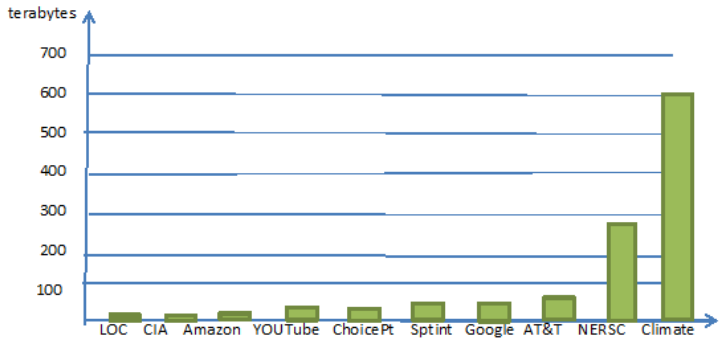


Figure 1.4: Top ten largest databases in (2007)

Every day, we create 2.5 quintillion bytes of data - so much that 90% of the data in the world today has been created in the last two years alone. This data comes

from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

Data is the most precious raw material in the 21st century. The data may be structured or unstructured, private or public.

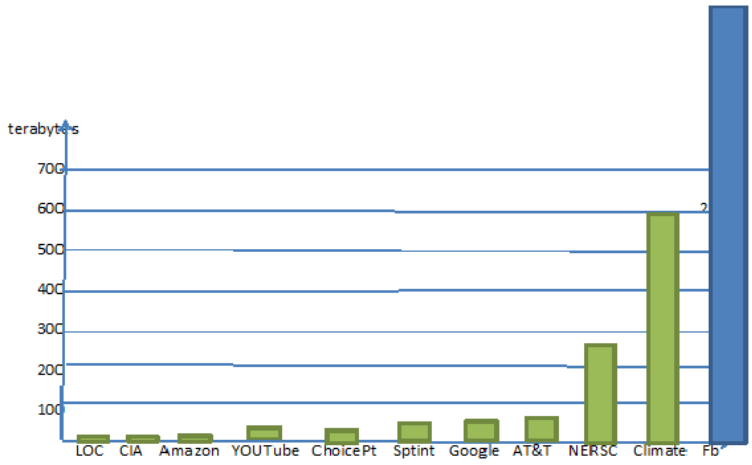


Figure 1.5: Facebook (21 petabytes in 2010)

### 1.5.2 Sample events generating data

- Facebook has more than 800 million active users and more than 75% are outside of U.S. More than 250 million photos are uploaded to Facebook every day. Facebook generates 25 Petabytes of data every day.

- The number of text messages that will be sent today is greater than the

population of the entire world.

- 29 billion emails are sent every day. That's one every .00000015 seconds
- 97,000 tweets sent every second.
- 400 tweets per minuet contain youtube link.
- Air bus generate 10 TB every 30 minutes, About 640 TB is generated in one flight.
- 5 billion camera phones are there worldwide most of them have location awareness (GPS).
- In 2009 the total data was estimated to be 1 ZB, in 2020 it is estimated to be 35ZB.

## 1.6 Objective

The objective of this thesis is to propose a novel cloud solution to handle and process big data. We intent to review the existing literature on mechanism that store and analyze big data. After establishing a foundation on the subject matter we would like to propose modifications to improve the performance metrics

## 1.7 Organisation of Thesis

**Chapter 1:Introduction** In this chapter we are going to discuss the concepts of Data Intensive Computing and Cloud Computing.

**Chapter 2:Related Work** In this chapter we are going to discuss background concepts of Data Intensive Computing and the problems that occur in Data Intensive Computing

**Chapter 3:Proposed Method** In this chapter we are going to discuss the disadvantage of Big Data and how we are going to solve the problem of Big Data by using Hadoop. **Chapter 4:Conclusion**

# Chapter 2

# Related work

## 2.1 Basic Definitions:

What is cloud?

Cloud is the computer hardware that holds all digital data. It is a computing platform capable of distributing applications and data to geographically dispersed locations. There is no cloud without data. You can use the cloud through internet, remote locations and mobile app as well.

What is Cloud Computing?

Cloud computing is a pay for use model and you only pay what you use, like renting furnished home. The users are charged according to whatever they use in cloud services like computing, storage, and network services. It is an emerging information technology and delivery model that enables real time delivery of configurable computing resources. It stores the data in many places through virtualization. It can help you create more secure and efficient infrastructure with

private and public clouds. The whole infrastructure service thing is maintained by providers like softlayer, amazon, Microsoft, and rockspace. It has become very significant in the information technology and business model, it has limited storage and computing capacity. The user need not buy software or hardware they can use through internet . The key of cloud computing is not focusing on the hardware is just focused on your website.

Salesforce.com is the first mover of cloud computing (1999). They introduced the concept of delivering enterprise applications via simple website. Amazon is the next mover(2002). Google docs brought the cloud computing to the fore front of public consciousness (2006). Amazon introduced Elastic Compute Cloud (2006) and recently Microsoft started Microsoft Azure (2010) in cloud computing.

Cloud computing services:

- Software as a service (SaaS)
- Platform as a service (PaaS)
- Infrastructure as a service (IaaS)

**Software as a service (SaaS):**

It is at the top most layer of the cloud computing stack. It is a model for delivering applications online. Here the cloud providers have created full applications running on server farms that may themselves be geographically distributed for time. Still Salesforce.com has been held up as the premier SaaS provider , Facebook, Flickr, eBay, Yahoo Stores, Amazon Marketplace, the

backup-storage provider Carbonite, and even the financial assistant Mint.com offer SaaS as well. [4]

**Platform as a service (PaaS):**

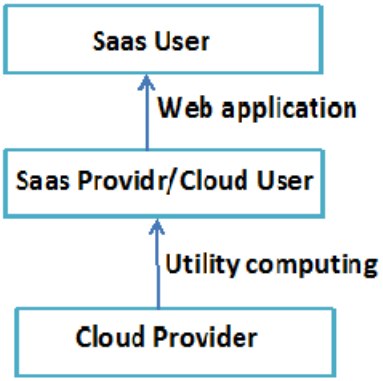


Figure 2.1: Platform as a Service

It is one step up vendors provide preconfigured computers running operating systems (OS) and applications. Amazon provides users with a variety of highly specialized offerings and also provides scalable databases, Web-accessible storage, message queues, and the "Mechanical Turk" system for organizing human computation. One of another example is the Google App Engine, it is a system for high performance computing. Microsoft is also a supplier: its Azure service provides preconfigured computers running Windows and SQL Server. All of these are elastic pay-as-you-go conception.

**Infrastructure as a service (IaaS):**

Here clients can buy processing, storage, and network services (typically using a Web-based self-service tool) and then build their own systems on top of this

infrastructure. Amazon and Rackspace is the best known IaaS. While clients have the illusion that they are renting specific servers, network switches, and hard drives, this equipment is in fact simulated by virtualization software, allowing multiple clients to be served by the same physical device.

Private cloud:

This cloud is very secure and is used by a single organization or single company. It is very expensive, and nothing will be shared for third party

Public cloud

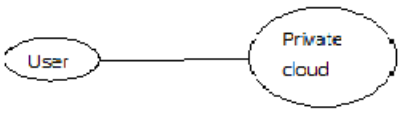


Figure 2.2: Private cloud

This cloud can be used any one, the cloud provider makes available everything public, example of this cloud is Facebook, Google and etc. here a group of people can use this cloud, but you don't know who are using. Hybrid cloud

Hybrid cloud is a combination of public cloud and private cloud. If user

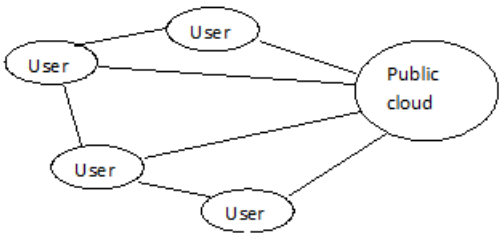


Figure 2.3: Public Cloud

wants to share data he can share if not he can stop. So this cloud is very popular because of this reason. Cloud computing is a major aim for data intensive

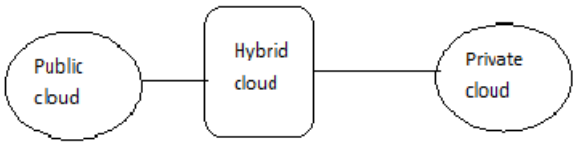


Figure 2.4: Hybrid Cloud

computing, as it allows scalable processing of massive amount of data. It is a promising next-generation computing paradigm given its many advantages such

as scalability, reliability, elasticity, high avail-ability, and low cost [5].

## 2.2 DATA INTENSIVE COMPUTING

Data intensive computing is managing, analysing, and understanding data at volumes and rates that push the frontiers of current technologies. Data intensive computing facilitates human understanding of complex problems. Data intensive applications provide timely and meaningful analytical results in response to exponentially increasing data complexity and associated analysis requirement of new classes of software, hardware, and algorithms.

Bayesian compression and aggregation method [6] is successfully applied in many different fields such as computer science, genetics, business, economics, imaging, epidemiology and political science. The compression and aggregation schema is compresses each partition into a synopsis and aggregates the synopsis into an overall estimate without accessing the raw data. This schema can find applications in OLAP for data cubes, cloud computing and stream data mining. It saves computation time when it processes each partition only once and it allows parallel processing.

The proposed C and A scheme is well suited for performing Bayesian analysis on bulky datasets in the cloud. For example, the compression and aggregation phases in a C and A scheme match the mapping and reducing stages, respectively, of the well-known Map Reduce algorithmic framework for cloud computing.

Data-intensive computing [4] is managing, analysing, storing, acquiring and understanding the data at volumes and rates that push the frontiers of current technologies. Data-intensive computing facilitates human understanding of difficult problems. Data-intensive applications provide timely and meaningful analytical results in response to exponentially growing data complexity and associated analysis requirements through the development of new classes of, software, hardware and algorithms.

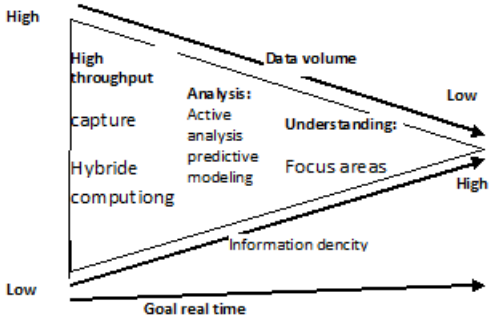


Figure 2.5: Without Cloud

### 2.2.1 DATA PROCESSING PIPELINE

As Figure shows, processing pipelines start with large data volumes with low information content. The subsequent processing steps in the pipeline reduce this data to create relatively small datasets rich in information and suitable for visualization or human understanding.

The researchers capture and stored the raw data that originates from a

scientific instrument or a simulation. The first stage of processing typically applies techniques to reduce the data's size by removing noise or by indexing, summarizing, or marking it up so that downstream analytics can manipulate it more efficiently. Second stage: These algorithms create information or knowledge that humans or further computational processes can digest.

Map Reduce distributes data and processing across clusters of commodity computers and processes the data in parallel locally at each node. In this way, massively parallel processing can achieve using clusters that comprise thousands of nodes. In addition, the supporting runtime environment provides transparent fault tolerance by automatically duplicating data across nodes and detecting and restarting computations that fail on a particular node. The first challenge to data-intensive computation is the incoming data, obtained from multiple sources, types, scales, and locations with varying degrees of quality and reliability. [4]

The basic problem is basically huge amount of data is stored in distributed data centres. When one job needs several datasets located in different data centres, the moving of data become a challenge task. So here Workflow technologies [5] are facilitated to automatize these scientific applications. Scientific workflows are commonly very difficult. They usually have a large number of jobs and required a long time for execution. Presently, popular scientific workflows are display in grid systems because they have very high performance and big storage. However, building a grid system is very expensive and it is not available for scientists all over the world to use.

The emergence of cloud computing technologies gives a novel way to develop

scientific workflow in the systems. Cloud computing is proposed as the next generation of IT platforms that can be deliver computing as a kind of utility. Foster et al. done a comprehensive comparison of grid computing and cloud computing. Some of the features of cloud computing also meet the requirements of scientific workflow systems.

Firstly, cloud computing systems can provide high performance and huge storage required for scientific applications in the same way as the grid systems, but with a lower infrastructure construction cost among many other features, because cloud computing systems are composed of data centres which can be clusters of commodity hardware. Secondly, cloud computing systems offer a new paradigm that scientists from all over the world can integrate and conduct their research together. Cloud computing systems are based on completely Internet, and so are the scientific workflow systems deployed on the cloud. Dispersed computing facilities (like clusters) at different institutions can be viewed as data centres in the cloud computing platform. Scientists can upload their data and launch their applications on scientific cloud workflow systems from anywhere in the world via the Internet. As all the data are managed on the cloud, it is easy to share and distribute the data among scientists.

When one job needs to process data from different data centres, moving data becomes a challenging task [7]. Some application data are too large to be distributing efficiently, some may have exact locations that are not feasible to be moved and some have to be located at fixed data centres for processing, but these are only one view of this challenge. In this paper [5] they are using

k-means clustering. Scientific workflows can be very complex, one task require many database for execution and one database may need many tasks, here the database and tasks are dependents to each other.

Managing and processing exponentially increasing data volumes often are arriving in time-sensitive streams from arrays of sensors and instruments or as the outputs from the simulations; and significantly reducing data analysis cycles so that researchers can make timely decisions. Data intensive computing [8] is mainly concerned with creating scalable solutions for capturing, analysing, managing and understanding multi-terabyte and petabyte data volumes.

Simply capturing the data at the rates it is emitted from a complex simulation, high resolution instrument or high-speed network is alone a challenging problem. To address this, the first stage of processing typically applies techniques to reduce the data in size, or process it so that it can be more efficiently manipulated by downstream analytics [9].

The challenge of data analysis in a scenario where data is stored across a local cluster and cloud resources. Here they are describing a software framework to enable data-intensive computing with cloud bursting that is: using a combination of compute resources from a local cluster and a cloud environment to perform Map-Reduce type processing on a data set which that is geographically distributed. Our evaluation is with three different applications shows that data-intensive computing with cloud bursting is feasible and scalable too. Particularly, as comparing with the situation where the data set is stored at one location and processed using resources at that end [10].

Many applications are already taking hours to days of computation time to process the data to the day, and it will take much longer over time as the data volume grows. Such big processing delays put Enter-prises at competitive disadvantages since they cannot react fast enough.

Cloud could provide a much more reliable and secure infrastructure than most Enterprises could afford. For example like Amazon has multiple data centers and each data center also has multiple backup power generators. Amazon also has strong security measures to guard against not only physical security breach but also hacker attacks [11]

Due to the explosive increase in the amount of information in computer systems, we need a type of system which it can process large amounts of data efficiently. Cloud computing system is an efficent means to achieve this capacity and has spread throughout the world.

Hybrid cloud environments, users can access public clouds and private clouds [12]; private clouds are secure clouds built using the secure resources of the user company, and public clouds can provide scalable resources if the user pays metered rates. Combining these clouds can address shortcomings related to safety and scalability. For data-intensive jobs, hybrid clouds are appropriate. For increasing amounts of data, hybrid clouds can provide secure and scalable processing.

Performance and cost must be balanced. We will increase speed, but the metered cost will also be greater. In contrast, if these jobs are processed using private cloud resources al-most exclusively, users will not have to pay metered rates, but the job execution time will be longer. Thus, we need a system that

can determine optimal job placement based on cost limitations and necessary performance to ensure efficient processing in hybrid cloud environments.

Public clouds can be used through the Internet, and users can use cloud services scalable if they pay metered rates to the cloud provider. [13] Getting the right cloud provider that minimizes the amount of transferred data and maximizes the global throughput is a major issue.

Every site has different computational capabilities and data storage, and the input datasets are replicated among them with the assumption that most data-intensive applications have large amount of input datasets [14].

Sending the job nearer to its input data might not be proper to meet the application deadlines because of the huge output data produced by the application, which might be transferred back to another cloud site.

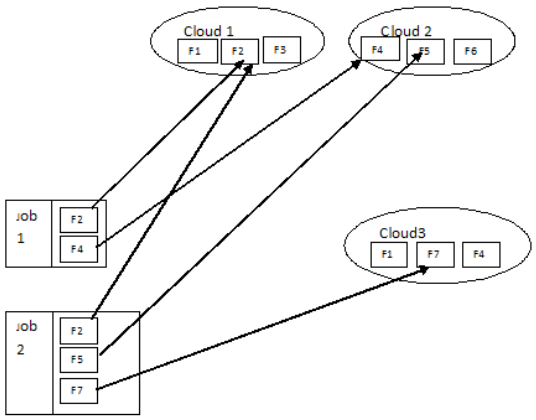


Figure 2.6: Without Cloud

The emergence of the cloud's pay-as-you-go model, users are increasingly storing portions of their data remotely and allocating compute nodes on-demand

to meet deadlines.The elasticity of clouds for different services, including a transactional data store , data-intensive web services, a cache that accelerates data-intensive applications [15], and for execution of a bag of tasks.

The ability to control the distribution and placement of the computation units (workers) is critical in order to implement the ability to move computational work to the data. This is required to place data network transfer low and in the case of commercial Clouds the monetary cost of computing the solution is low. The heterogeneity and evolution of large-scale distributed systems, the fundamentally distributed nature of data and its exponential increase collection, storing, processing of data, it can be argued that there is a greater premium than ever before on abstractions at multiple levels. The emergence of Clouds as important distributed computing infrastructure, we need abstractions which can support existing and emerging programming models for Clouds. Inevitably, the unified concept of a Cloud is evolving into different flavours and implementations on the ground. [16]

# Chapter 3

# Proposed Method

## 3.1 MapReduce

MapReduce is a programming model which Google has used successfully for processing its big-data [17]. A map function extracts some intelligence from raw data. A reduce function aggregates according to the data outputed by the map. Users specify the computation in terms of a map and a reduce function, Underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, and Underlying system also handles machine failures, efficient Communications, and performance issues [18]

There are 4 steps in Map Reduce

- Splitting of data
- Passing the output of map functions to reduce functions.
- Sorting the inputs to the reduce function based on the intermediate keys.

- Quality of services.

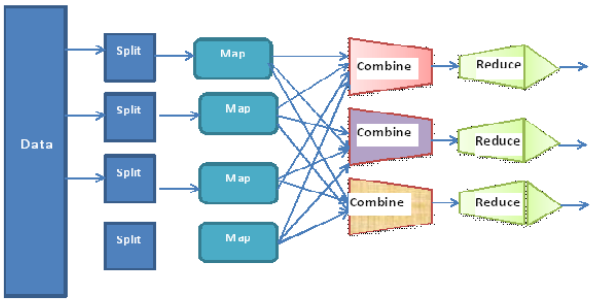


Figure 3.1: MapReduce

At Google, MapReduce operation are run on a special file system called Google File System (GFS) [19] that is highly optimized for this purpose. GFS is not open source. Doug Cutting and Yahoo! reverse engineered the GFS and called it Hadoop Distributed File System (HDFS).The software framework that supports HDFS, MapReduce and other related entities is called the project Hadoop or simply Hadoop. This is open source and distributed by Apache.

## 3.2 Hadoop

The objective of this tool is to support running application on big data with the help of map and reduce function.It is open source software that uses cluster of commodity machines to process large amount of data.

Here the big data [20] can store powerful computation but whenever the data is

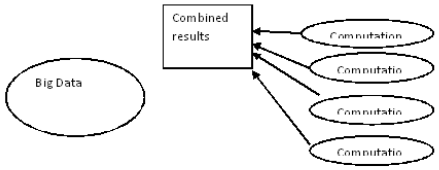


Figure 3.2: Concept of Hadoop

lower then it uses traditional systems. If the data size increases then the traditional system fails.

So here we are using Hadoop concept to solve this problem. Hadoop is the object of the tool, it is support running application on big data.

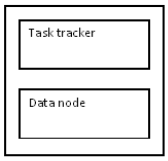


Figure 3.3: Slave in Hadoop

Hadoop is a set of tools like map reduce [21], file system hbase and etc. Hadoop is a Linux based set of tools, so we have Linux on this entire low cost computer. This entire computer will have two components of hadoop.

- Task tracker
- Data nodes

Task tracker is to process the small pieces of task that have been given to a particular node.Data node manages the pieces of data that have been given to a particular node.

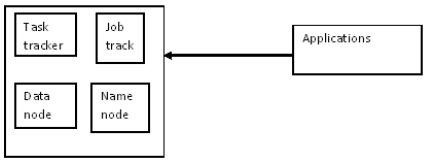


Figure 3.4: Hadoop

Hadoop is a batch processing set of tools. The above components, task tracker and data node belongs to slaves.

The goal of job tracker is to break bigger tasks into smaller pieces, to send each small pieces of computation to task tracker. The task tracker after finishing its work sends back the results to applications.

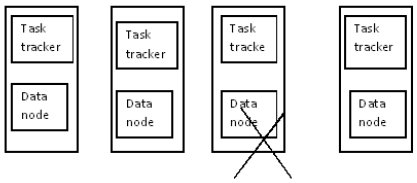


Figure 3.5: Fault Tollerance in Hadoop

The "name node" running on master is responsible to keep an index of which data is responsible and data node when application contacts the name node. It tells the application in which computer its data is stored.

### 3.2.1 Fault tolerance

Job tracker will detect failures. When one node fails then the process is given to the neighbour node. So here the user does not lose data. The data will be safe. Hadoop keeps two masters, main master and backup master. If main master fails the backup master will do all processing in the computer. It is highly scalable, this system consists of 1 computer or 100 computers and the scalable cost is linear.

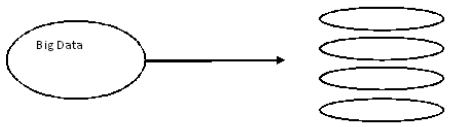


Figure 3.6: Traditional Database System

# Chapter 4

# Conclusion

Previously there were many techniques applied on the big data. But none of them are fully successful. The techniques were XML Schema, scientific workflows techniques and pipeline techniques. XML schema is applicable only on small database, and scientific workflows are taking long time to execute the applications. But here we have applied Hadoop model, this model overcomes all the drawbacks of previous technologies and it will provide many features like high throughput, highly fault tolerance etc.

# Bibliography


[1] Jawwad Shamsi, Muhammad Ali Khojaye, and Mohammad Ali Qasmi. Data-intensive cloud computing: Requirements, expectations, challenges, and solutions. *Journal of Grid Computing*, pages 1–30, 2013.

[2] Junaid Arshad, Paul Townend, and Jie Xu. A novel intrusion severity analysis approach for clouds. *Future Generation Computer Systems*, 2011.

[3] Anton Beloglazov, Jemal Abawajy, and Rajkumar Buyya. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Computer Systems*, 28(5):755–768, 2012.

[4] Richard T Kouzes, Gordon A Anderson, Stephen T Elbert, Ian Gorton, and Deborah K Gracio. The changing paradigm of data-intensive computing. *Computer*, 42(1):26–34, 2009.

[5] Xin Liu and A. Datta. Towards intelligent data placement for scientific workflows in collaborative cloud environment. In *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on*, pages 1052–1061, 2011.

[6] Ruibin Xi, Nan Lin, Yixin Chen, and Youngjin Kim. Compression and aggregation of bayesian estimates for data intensive computing. *Knowledge and information systems*, 33(1):191–212, 2012.

[7] Dong Yuan, Yun Yang, Xiao Liu, and Jinjun Chen. A data placement strategy in scientific cloud workflows. *Future Generation Computer Systems*, 26(8):1200–1214, 2010.

[8] I. Gorton, Paul Greenfield, A. Szalay, and R. Williams. Data-intensive computing in the 21st century. *Computer*, 41(4):30–32, 2008.

[9] Ian Gorton. Software architecture challenges for data intensive computing. In *Software Architecture, 2008. WICSA 2008. Seventh Working IEEE/IFIP Conference on*, pages 4–6. IEEE, 2008.

[10] Tekin Bicer, David Chiu, and Gagan Agrawal. A framework for data-intensive computing with cloud bursting. In *Cluster Computing (CLUSTER), 2011 IEEE International Conference on*, pages 169–177. IEEE, 2011.

[11] Huan Liu and D. Orban. Gridbatch: Cloud computing for large-scale data-intensive batch applications. In *Cluster Computing and the Grid, 2008. CCGRID '08. 8th IEEE International Symposium on*, pages 295–305, 2008.

[12] Jawwad Shamsi, Muhammad Ali Khojaye, and Mohammad Ali Qasmi. Data-intensive cloud computing: Requirements, expectations, challenges, and solutions. *Journal of Grid Computing*, pages 1–30, 2013.

[13] Yumiko Kasae and Masato Oguchi. Proposal for an optimal job allocation method for data-intensive applications based on multiple costs balancing in a hybrid cloud environment. In *Proc. the 7th ACM International Conference on Ubiquitous Information Management and Communication (ICUIMC2013)*, pages 9–3, 2013.

[14] N.A. Mehdi, B. Holmes, A. Mamat, and S.K. Subramaniam. Sharing-aware intercloud scheduler for data-intensive jobs. In *Cloud Computing Technologies, Applications and Management (ICCCTAM), 2012 International Conference on*, pages 22–26, 2012.

[15] Tekin Bicer, David Chiu, and Gagan Agrawal. Time and cost sensitive data-intensive computing on hybrid clouds. In *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*, pages 636–643. IEEE Computer Society, 2012.

[16] Chris Miceli, Michael Miceli, Shantenu Jha, Hartmut Kaiser, and Andre Merzky. Programming abstractions for data intensive computing on clouds and grids. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, pages 478–483. IEEE Computer Society, 2009.

[17] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[18] Hanli Wang, Yun Shen, Lei Wang, Kuangtian Zhufeng, Wei Wang, and Cheng Cheng. Large-scale multimedia data mining using mapreduce framework. In *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*, pages 287–292. IEEE, 2012.

[19] Hanli Wang, Yun Shen, Lei Wang, Kuangtian Zhufeng, Wei Wang, and Cheng Cheng. Large-scale multimedia data mining using mapreduce framework. In *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*, pages 287–292. IEEE, 2012.

[20] Andrew McAfee, Erik Brynjolfsson, et al. Big data: the management revolution. *Harvard business review*, 1, 2012.

[21] Julio Anjos, Wagner Kolber, Claudio R Geyer, and Luciana B Arantes. Addressing data-intensive computing problems with the use of mapreduce on heterogeneous

environments as desktop grid on slow links. In *Computer Systems (WSCAD-SSC), 2012 13th Symposium on*, pages 148–155. IEEE, 2012.