

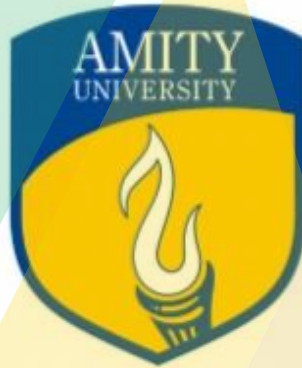
Term Paper Report

on

Big Data

Submitted to

Amity University Uttar Pradesh



In partial fulfilment of the requirements for the award of the degree
of

Bachelor of Technology
in

Computer Science and Engineering
by

AMINOTES.COM
(A2305219999)

Under the guidance of

Ms Faculty Name

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY

AMITY UNIVERSITY UTTAR PRADESH

JULY 2017

Declaration

I Aminotes student of B.Tech (3-C.S.E.-21(Y) hereby declare that the project titled “**Big Data**” which is submitted by me to Department of Computer science, **Amity School of Engineering Technology**, Amity University Uttar Pradesh, Noida, in partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

The Author attests that permission has been obtained for the use of any copy righted material appearing in the Dissertation / Project report other than brief excerpts requiring only proper acknowledgement in scholarly writing and all such use is acknowledged.

Place:

Date: _____

Name and Signature of Student

Aminotes

A2305219999

3CSE-21

(2016-2020)

CERTIFICATE

This is to certify that Mr Aminotes, student of B.Tech in Computer Science and Engineering has carried out work presented in the project of the Term paper entitle “**Big Data**” as a part of First year program of Bachelor of Technology in Computer Science and Engineering from Amity University, Uttar Pradesh, Noida under my supervision.

Ms Faculty Name

Department of Computer Science and Engineering
ASET, Noida

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I would like to thank Prof (Dr) Abhay Bansal, Head of Department-CSE, and Amity University for giving me the opportunity to undertake this project. I would like to thank my faculty guide Ms Richa Gupta who is the biggest driving force behind my successful completion of the project. She has been always there to solve any query of mine and also guide me in the right direction regarding the project. Without her help and inspiration, I would not have been able to complete the project. Also I would like to thank my batchmates who guided me, helped me and gave ideas and motivation at each step.

Aminotes

INDEX

S.NO.	CONTENTS	PAGE NO.
1.	ABSTRACT	5
2.	INTRODUCTION	6
3.	REVIEW	
	3.1.CHARACTERISTICS OF BIG DATA	8
	3.2 PROBLEMS IN BIG DATA & THEIR SOLUTIONS	10
	3.3.ARCHITECTURE	12
	3.4.BIG DATA OPTIMISATION TECHNIQUES	15
4.	CONCLUSION	20
5.	REFERENCE	21

Big Data

1. Abstract

Big data is manipulating of the general collection of Raw data collected from different resources and services. And this collected data is used in order to perform better than others. Big data is collecting the huge amount of data sets from conventional and electronic sources to differentiate the trends and patterns of a specific item. That information is used by companies to improve what they know about customers wants and needs. The goal should be to make solid decisions based on data and not just hunches. People are increasingly willing to hand over their personal data in return for product and services that make their lives easier and these all information is transferred Back to systems for an additional investigation that allows for new sorts of queries to Be requested, for example, *What will customer response be if we avail these kinds of goods?*

The Arrival of Technology like hardware and software which immediately assesses natural Human speech, and enormous amounts and sorts of significant data flowing from Detectors, cellular devices, and the internet, are helping the current data leaders locate Replies to questions such as, *How do customers feel about my merchandise?* and *What happens when we place a Wind farm here rather than there?* And identify patterns and patterns on the fly, then introduce them in a means that is Easy for individuals to understand.

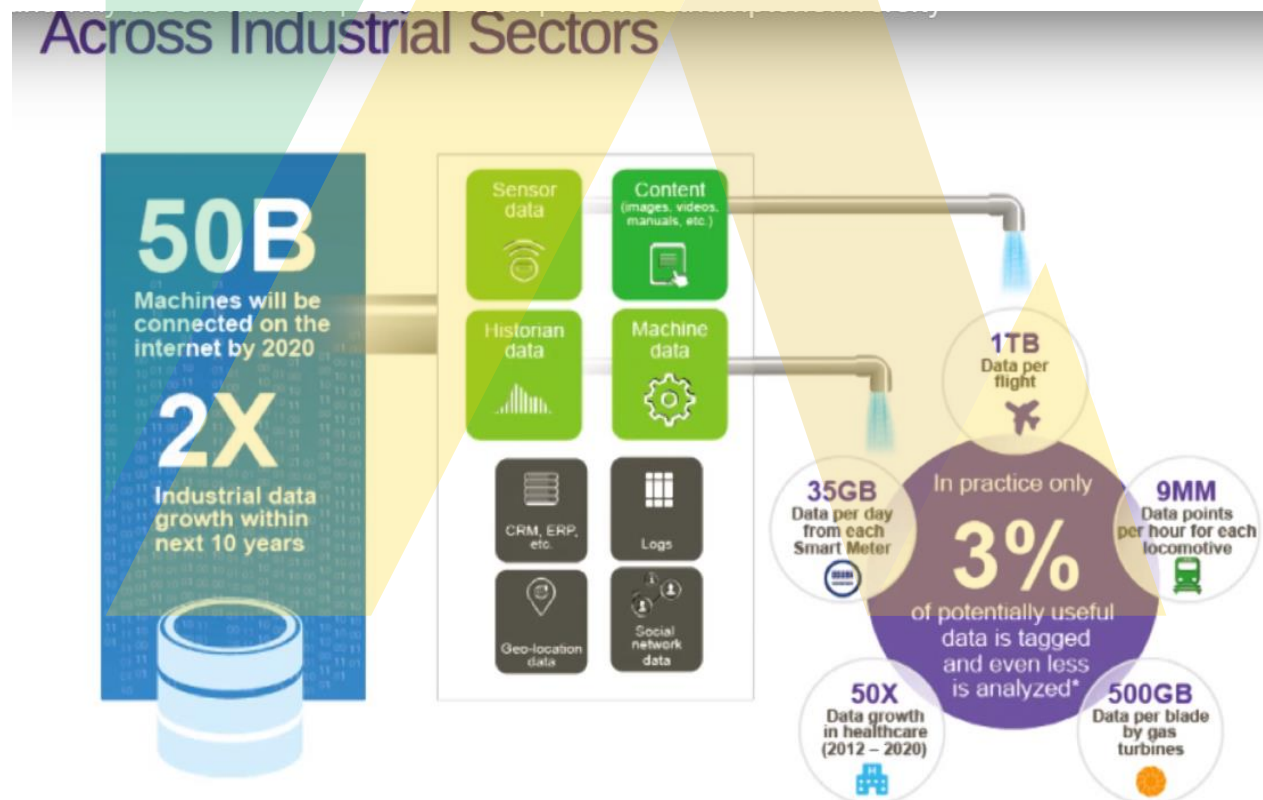
2.Introduction

Before understanding what's "Big Data" it is more necessary to know about what is "Data"?

Data is defined as "Basic values or facts" which can be taken from any Individual Person or some other Agencies.

"Big Data" more specifically "Industrial Big Data" well most people understand that term in the context of something like twitter users will send a hundred thousand tweets every minute or Google will get two million queries a minute but really industrial big data is set to really swamp that so it was pointing comparing maximum twitter usage 2013 was about 500million of tweets a day that equates to 200 Billion tweets per year and if you heard about the "Large Hadron Collider"(LHC) it generates (distribute and analyses)the 15 petabytes (15 million gigabytes) of data generated every year and that is really a "Big Data" aka "Industrial Big Data".

Big data is difficult to track down it represents the amount all the digital information which is uneasy to store, transport and analyze. It is so immense that it overwhelms the technology of the today and challenges us to create the next generation of data storage tools and techniques



In this Picture you can see those industrial sources and their impact on the Global Data occupancy these all are Necessary sectors of any economy:

Health Care

Electricity

Transportation

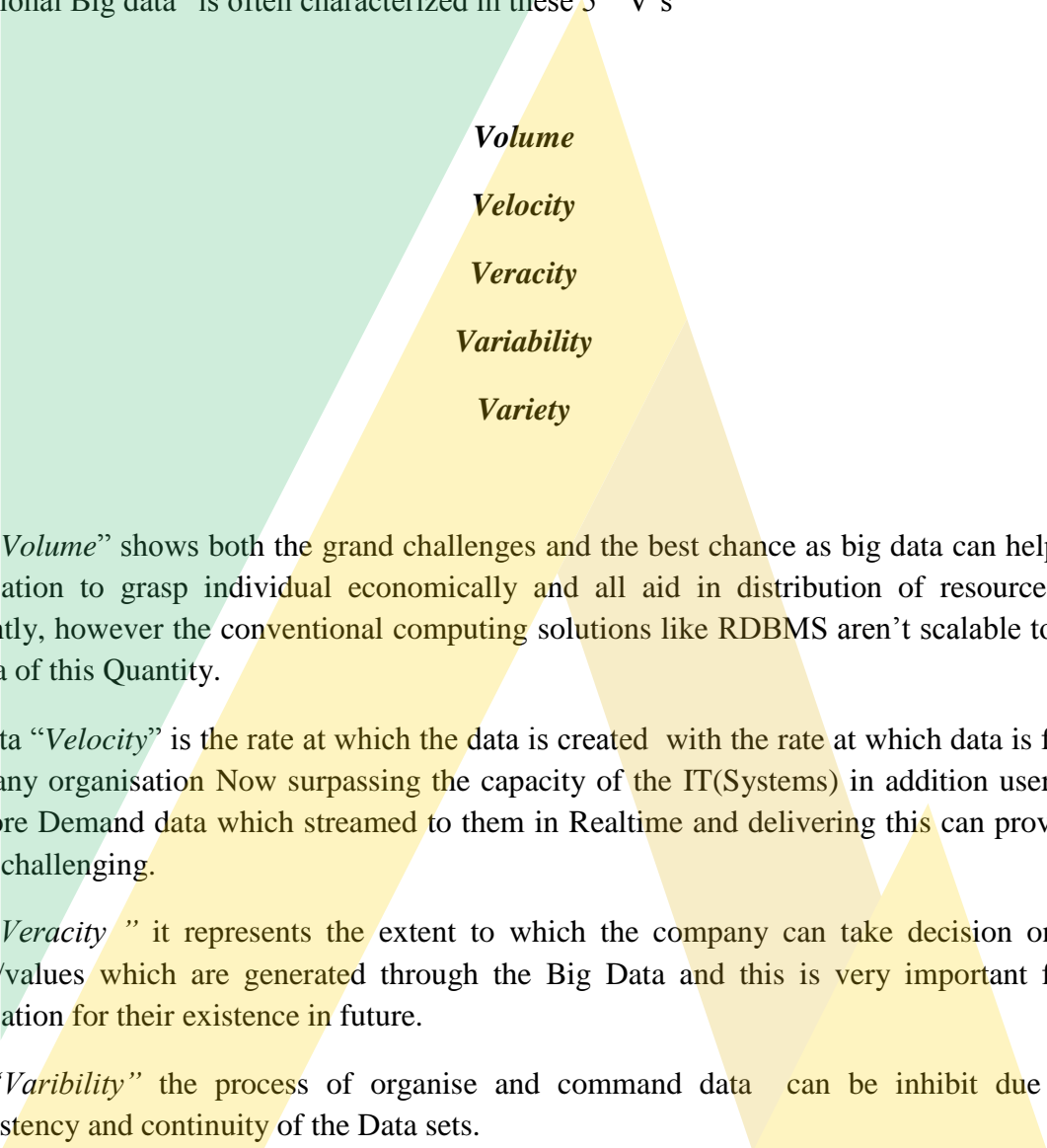


3.Review

3.1.Characteristics of Big Data

As the increasing population the generation of Big Data is also increasing so to Controlling is also hard .And a Good Big Data characteristics.

“Traditional Big data” is often characterized in these 5 “V’s”



Volume
Velocity
Veracity
Variability
Variety

Here, “*Volume*” shows both the grand challenges and the best chance as big data can help many organisation to grasp individual economically and all aid in distribution of resources more efficiently, however the conventional computing solutions like RDBMS aren’t scalable to tackle the data of this Quantity.

Big Data “*Velocity*” is the rate at which the data is created with the rate at which data is flowing into many organisation Now surpassing the capacity of the IT(Systems) in addition users more and more Demand data which streamed to them in Realtime and delivering this can prove to be quite a challenging.

Data “*Veracity* ” it represents the extent to which the company can take decision on those figures/values which are generated through the Big Data and this is very important for and organisation for their existence in future.

Data “*Varibility*” the process of organise and command data can be inhibit due to the inconsistency and continuity of the Data sets.

The “*Variety*” of data types to be processed is turning more and more diverse gone are the days when data centres only had to handle with traditional data today the whole system captured by Big data is being piled into many a corporate data silo many such big sources are also unstructured and hence not easy to categorise let alone process with traditional computing Techqunies.



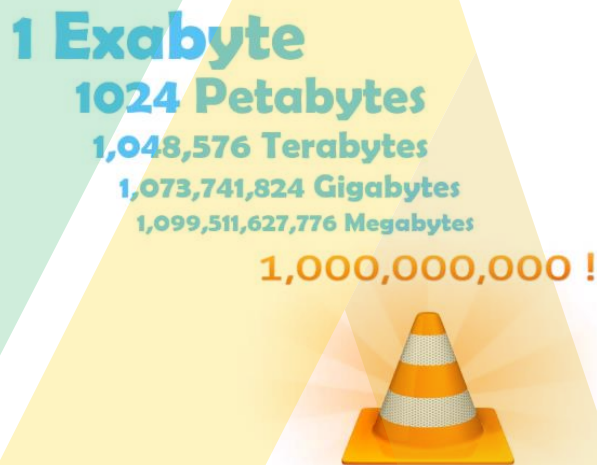
3.2 Problems in Big Data & their Solutions

- Storage and transport Issues
- Management Issues
- Processing Issues

These are some major issues in big data lets now discuss them briefly

First, comes “**Storage and transport issues**” the data is created by everything and everyone due to social media and internet of thing approximately “1 Exabyte” of data is created every day.

1 Exabyte of data equals to processing “1 Billion” movies every day



Now the problem is current storage technology allows us to store only “4 Tera Bytes” per bytes so we need “25,000 disks” to store that huge amount data.

Solution to Problem :

- Processing data in place - Also called as bring code to the “Data”. Where the data is stored at the node we implement code programs on that node and we got the result of the processing.
- Critical data transportation - In this method, we only send critical data to next stage instead of sending the complete data to next stage for processing this reduces transportation cost.
- HDFS (Hadoop distributed file system) its special type of file system for keeping the Big data. It is storage components of Hadoop.

Management Issues-

Management is a most difficult problem to address big data unlike, the collection of data through the manual method were accurate protocols are used to maintain the Data validity and the accuracy data.

The digital data collection is much more reliable. Consider social media where data is created by Everything and everyone it is very hard to verify the data accuracy and validity the data often lacks meta-data to describe them.

Solution to Management :

1. Remove Duplicates
2. Verify New Data
3. Update Data
4. Implement Consistent Data Entry

Processing Issues -

Normal data like small in size will be processed easily for decision making but for the huge amount of Data it is not easier to make the decision quickly .

Problems are:

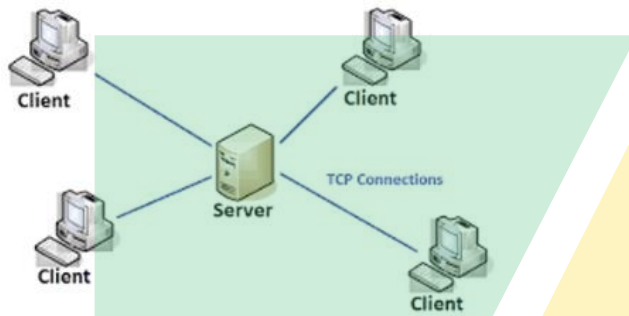
- Heterogeneity and Incompleteness
- Scale
- Timeliness
- Privacy

Big data needs parallel processing Parallel processing or we can say using multiple resources.



3.3.Architecture

Traditional Data

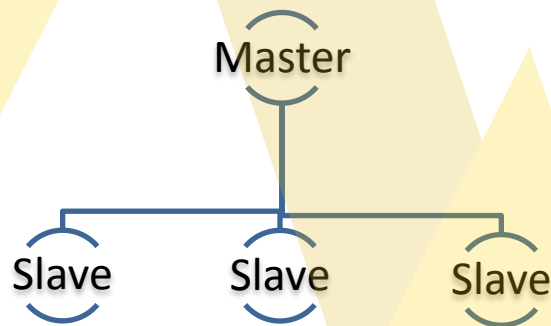


Traditional data is processed with client server architecture

In which all the data which is processed by the Client is being stored in Server (RDBMS)

i.e ATM

Big Data



Big data is processed with Master Slave Architecture

In which the all the slave will process that specific process which is assigned to that Slave and Master is the main Process.

Main Components of Big Data:

- HDFS (Hadoop distributed file system)

It's a storage format in which the Digital Data can be stored

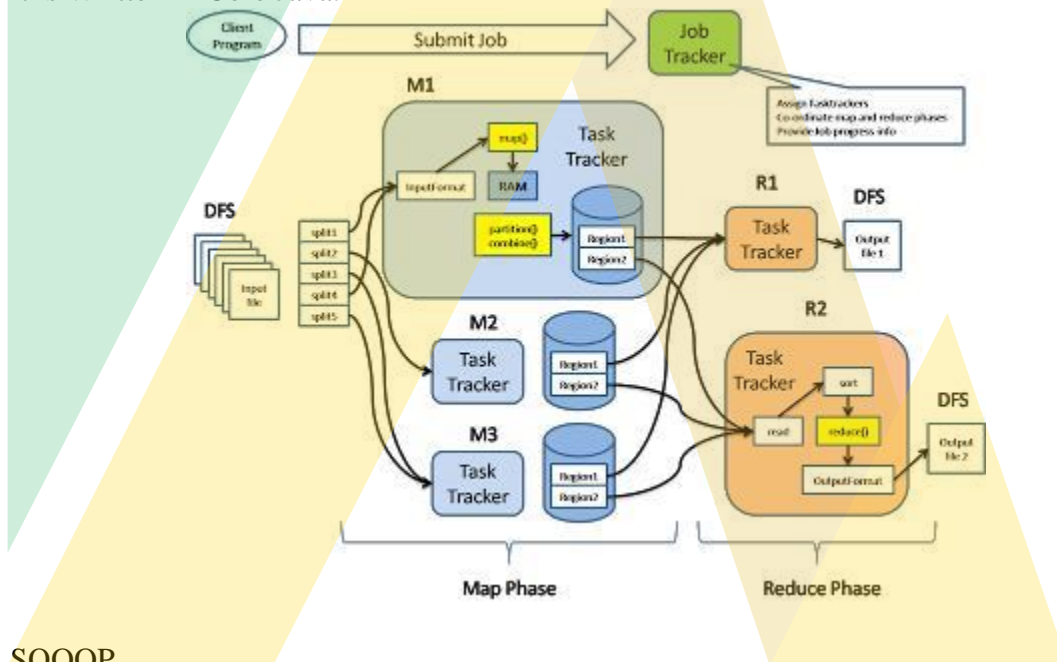
i.e

- Audio
- 3D-Videos
- And high Graphic Content
- Text file
- Log Files

- MR(Map Reduce)

It is used to write the Business Logic. The that Business logic is used to process the big data which gives the valuable role in decision making.

It is Written in Core Java.



- SQOOP

It is used to import & export the data from the Structured Query language data(like MySQL or from any) to HADOOP data vice-versa.

- HIVE(Dataware House)

Information Stockroom Programming for Questioning And Overseeing Expansive

- HBASE

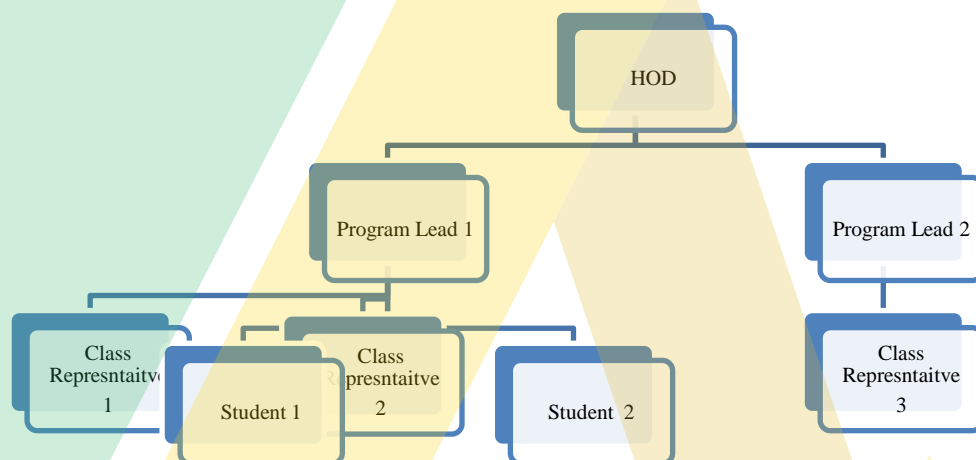
It has all the normal data (Images, MS Word) except SQL component.

- OOZIE

It is the workflow of any task. If any of the intermediate steps of the workflow is skipped the goal of the workflow can't accomplish.

i.e

If Student 1 wants some information passed to HOD
 Then He/She first goes to Class Representative 2
 Then Class Representative 2 passes that information to Program Lead 1
 Then Program Lead 1 informs that to HOD
 Since this is how a workflow works.



- FLUME

It is used for continuous streaming of data like on twitter, facebook, Instagram etc.
 The data can be of any type: Video, images, text, GIF file

- PIG

It is a predefined component used for processing. It is like MapReduce, the only difference is that in MapReduce we have to apply our own Query. But in PIG we have predefined Query which will automatically be applied to Big Data and dig out the Business logic.

3.4. Big Data Optimisation Techniques

The most famous tool that is used for Optimization of “big data” is **Hadoop** it is the product of Apache and it is open sources. Microsoft also own one more big data technology named “COSMOS” which is currently used in their “BING” search engine. Apache is also developing another Big Data Technology “SPARK” and its scripting language is SCALA.

Hadoop runs on Linux OS and it is developed under the JAVA and it is adopted by many MNC's like Facebook , Yahoo, Amazon, Netflix, Google, Twitter and many more tycoons companies.

Why are the people getting into Hadoop?

- It is a very flexible file system.
- It is very Scalable system
- Inexpensive- it's almost free can be download for Apache.

HDFS(Hadoop distributed file system)Overview

- HDFS follows the file system for storage of data/information is based on Hierarchical UNIX.
- Splitting of large files into the smaller blocks and Distribution and replication of blocks to nodes on different server
- Two main key's for service are :
 - Master Name Node

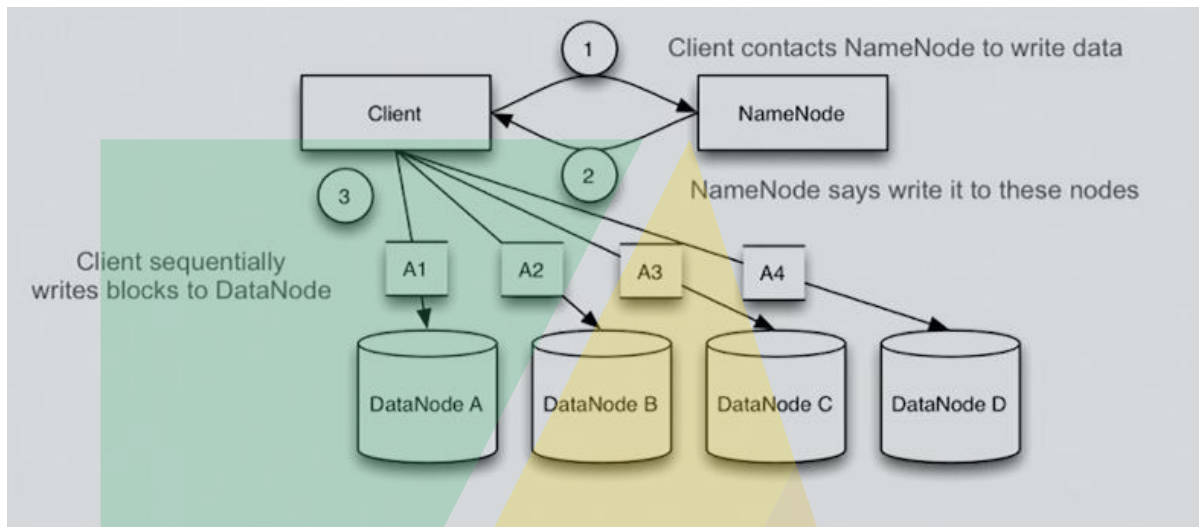
It's a single service that is basically a giant name server of knowing where their blocks are inside the system.

Single point of failure(HDFS 1.x) it stores file to block to location mappings in the namespace and all transactions are logged to disk.

- Many Data Nodes
They are actually restoring the data. There are up to 4,000 nodes present. It stores blocks on local side and send frequently heartbeats to Name Node also sends the block report to Name node and client connect to Data Node for Input and output (I/O).
- Checkpoint Node –It is different from Name Node and Data Node it performs checkpoints of the Name Node's namespace and logs. Checkpoint Node is in HDSF 2.0.

Working of HDFS

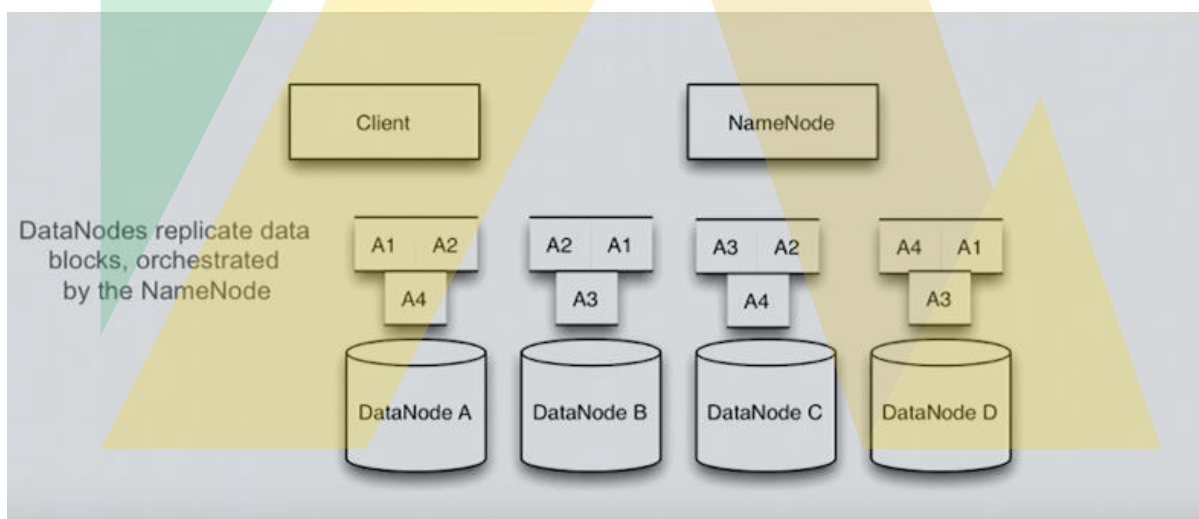
How is Writes?



First, the Client will contact the Name Node to write the data specifying the Name of the file
Then the Name Node say write all this to these nodes at which point the client server starts consecutive write blocks to Data Node

For example

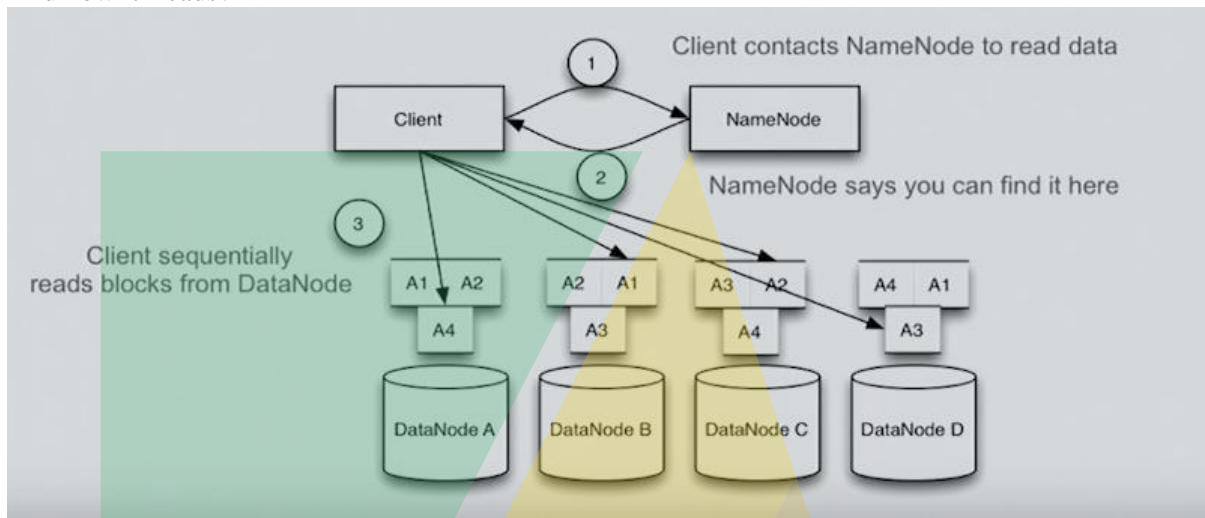
–I have a file named “A” which split into 4 block’s (256 megabytes file) split into Four (64-megabyte) block named (A1, A2 , A3, A4) from which the client server connects to the datanode and write a block out and it will connect to the very next datanode as arranging by the name it said where to put all these blocks and it will write out that very next block.



And ,when the replication is done.The datanodes kind of rear epic eight all of this data in-between Themselves so the client server is only writing the file only once and it’s actually make two more copies(as you can see A1 is present in Datanode A and as well its copy is distributed among the other Datanode B & Datanode C) in Hadoop distributed file system it’s all spread and copied and block off. And this

whole process is arranged by the Namenode. And Namenode not in real have an influence except for replicating blocks to others datanodes and serving the request from which is provided by the client.

And how it Reads?



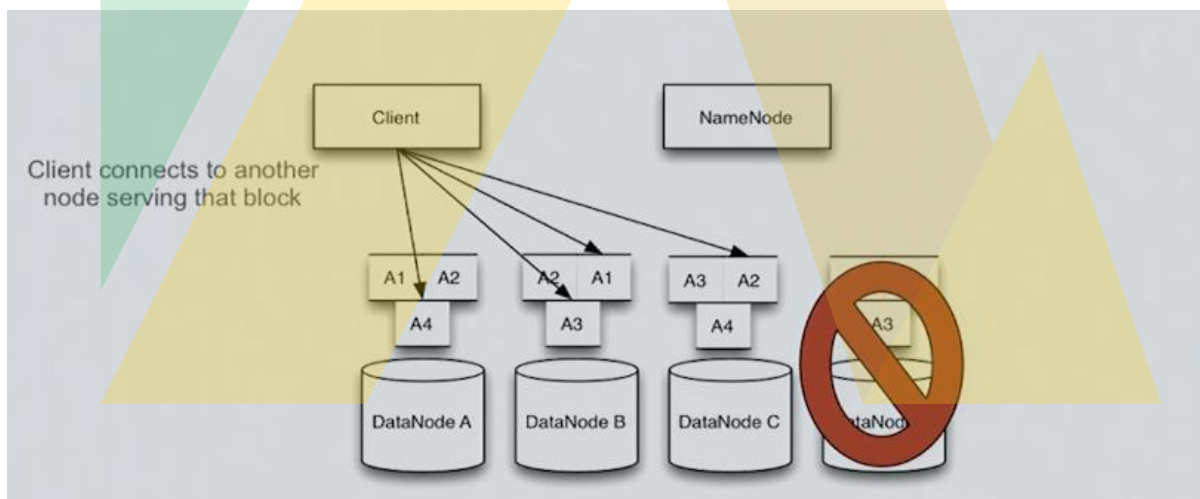
First, client start contacting the NameNode

Then NameNode responds to the client and say like you can transverse whole data from the file you would like to see or read out can be transverse on these datanodes are all the blocks consist at on which point the client can consecutively read the blocks from these data nodes.

i.e. First this read the datanode A completely then it moves to datanode B read that completely and so on until it detects that particular type of file that client wants to read.

Client are built in the way and implemented they acts as the local to the data so when it got that block report back from that Name Node.

And how does the HDFS tackle the Failure?



If in case Data Node “D” goes down or we can say failure of Data Note “D” because there is always replication of these Blocks in the file system that the client still can re-established the connection with A1, A2, A3 cause of that block somewhere in the system and the Name Node also identify that this datanode is crashed at which point it starts the re-replicating of all the blocks at that node .

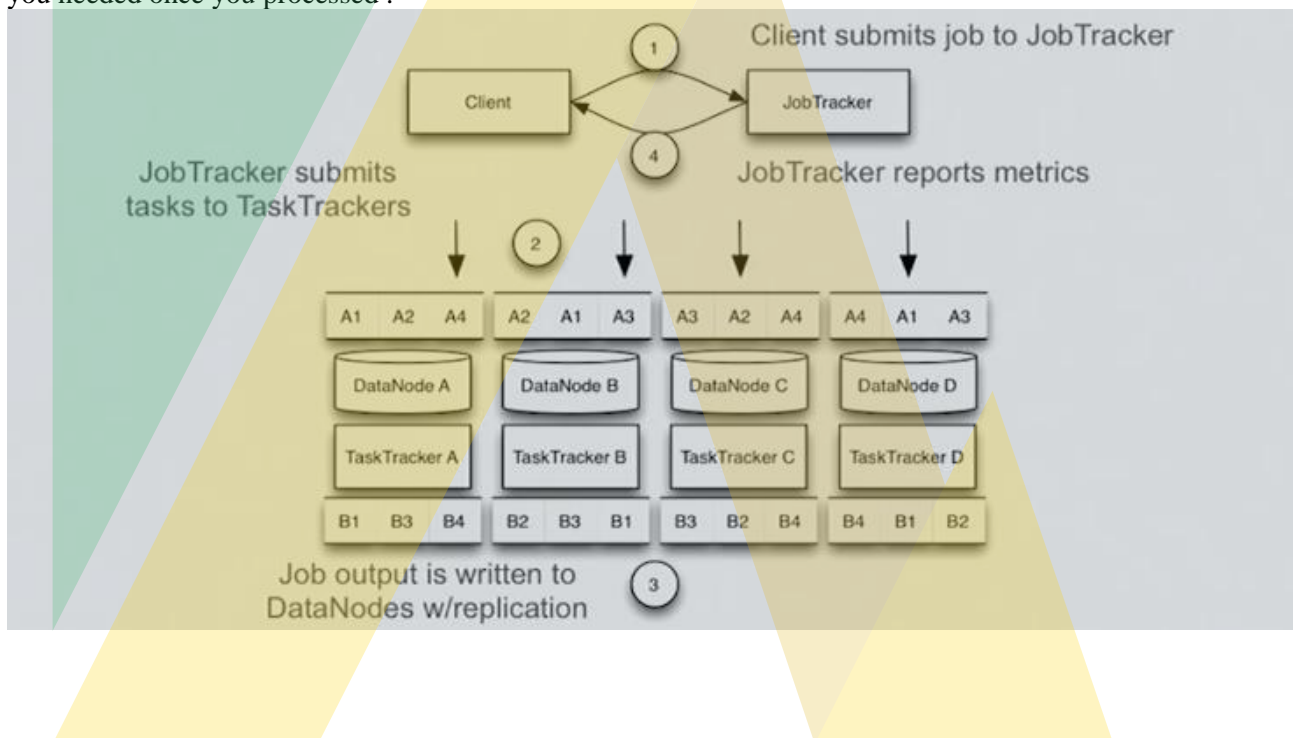
Hadoop another important techniques is “*Hadoop MapReduce*”

Hadoop MapReduce 1.x Version

- MapReduce moves/shift the code to data.
- There are generally two types of services offered in MapReduce:
 - Job Trackers- It monitors all the master service like monitoring Jobs.
 - Task Tracker-Its uses multiple services to run the task and it's the same physical machine as a Data Node.
- There are no definite limits to jobs they can contain multiple task and a task can also contain one or more attempts

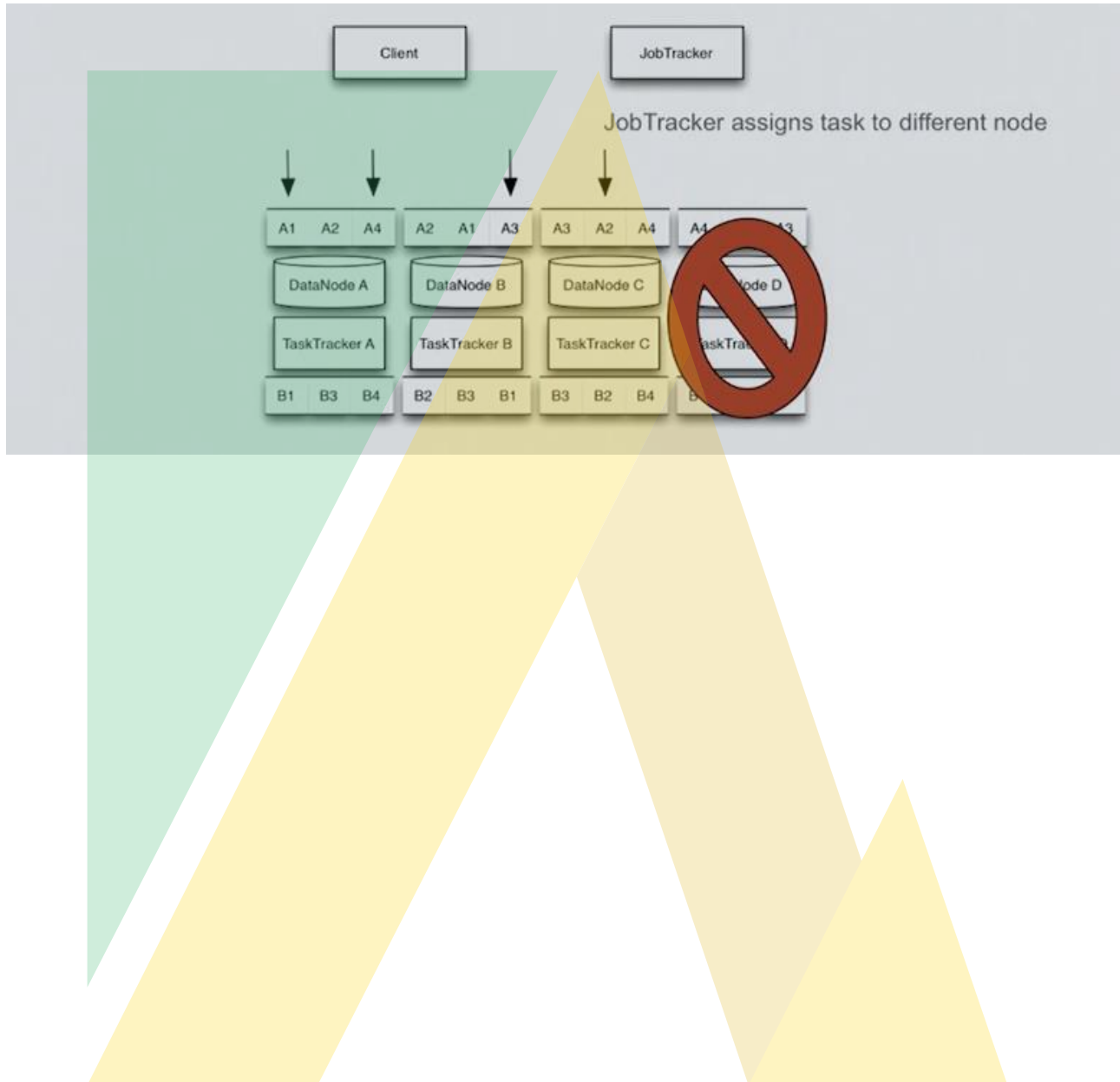
How does it works?

First the client submits job to the Job Tracker(It contains all the Files, folders as well the input path & format also) then the job tracker submits tasks to the task trackers which is running that time and all this job output is written to datanode with replication and After the work successfully it will record back data to you do not really get your desired output just like you may be accustomed to using RDBMS it'll only say my job will be completed at which time you may go fetch your information it is possible to export it into some other system you can perform anything on it when you needed once you processed .



In Case of the node fails

Job tracker gets notification of this and we will simply assign it to some other task tracker so the way the HDFS is quite fault tolerant and replication of data can be seen here in copying real processing jobs.



4. Conclusion:

It is concluded from the Term paper that “Big data” is powerful it has impact on all over the world it helps us to understand the pattern and analyze them to develop the techniques for and any Organisation. It also gives the analytics records which is very supportive for Marketing Agencies which will help them to target the correct audience. Big data also have important role in understanding the behaviour of the people it takes data from all over places like Hospitals, Fitness Centre and determine the particular behaviour of the pattern of the people this helps the firms (like Pharmaceuticals)to develop the medicines/drug more effective way and help to tackle the cure of disease. Big data is totally linking the world together making it a better place using it's all necessary technologies and drive and the server tries to serve us better. Big Data is the very emerging field which would make things different in future.

5.References

<https://www.dsayce.com/social-media/tweets-day/>
<http://home.cern/topics/large-hadron-collider>
<https://metricbuzz.com/blog/will-humans-worship-big-data-in-future/>
documents.tips/technology/hadoop-just-the-basics-for-big-data-rookies.html
<https://www.slideshare.net/Pivotal/hadoo...es-springone2gx-2013>
<https://www.slideshare.net/SpringCentral/hadoop101-spring-one2013>
qiang129.blogspot.com/
<https://sanwen.net/a/asexvqo.html>
ijmas.com/upcomingissue/05.02.2015.pdf
<https://www.coursehero.com/file/pd389n/I...udies-ISSN-NO-2348/>
<https://usegalaxy.org/u/aun1/p/galaxy101>
<https://github.com/chrisquince/DESMAN>