# Big Data & Hadoop

Darshil Doshi[1], Charan Tandel[2],Prof. Vijaya Chavan[3]

[1]*Student, Computer Technology, Bharati Vidyapeeth Institute of Technology, Maharashtra, India*
[2]*Student, Computer Technology, Bharati Vidyapeeth Institute of Technology, Maharashtra, India*
[3]*Professor, Computer Technology, Bharati Vidyapeeth Institute of Technology, Maharashtra, India*

## ABSTRACT

*Big data is a process of collecting the large amount of data and then processing it. Big data are used by many big companies for certain surveys. Hadoop is a platform provided that is used for big data. Hadoop stores massive amount of data that have massive power and can process multiple things at a single time. The architecture of the hadoop is also explained and its respective components and its working. HDFS is Hadoop Files System generally termed as HDFS.*

**Keywords:** *Big data, physical architecture, hadoop, HDFS*

## 1. INTRODUCTION

*Big data usually includes storing data sets which cannot be stored and aregenerally used to capture, manage, and process data within a very short time. Today's business companies owe a huge part of their success to an economy that is firmly knowledge based.Data drives the modern organizations of the world and hence making sense of this data and undo the various patterns and revealing unseen connections within the big sea of data becomes sensitive and a hugely rewards to achieve it.Big data should be correct so that it lead to more positiveand satisfactory decisions resulting in greater effective, cost reduction and reduced risk. Big data is going to be used by the big companies for analysis and other purpose. Big data majorly consist 3 V's that are Volume, Velocity and Variety. Hadoop is a free programming framework that supports the processing of large data sets in a distributed system. Hadoop is an open source software framework which provides huge data storage facility. In hadoop cluster, a number of machines are connected in a network and when a request comes from the client the number of computers or machine will process the data and will provide the output in a very short period of time. Cluster is nothing but group of machines connected in a network. The use of Hadoop makes it possible to run applications on systems with thousands of computers involving thousands of terabytes. Its distributed file system helps in rapid data transfer rates among computers and allows the system to continue process uninterrupted in case of a node failure. Hadoop is providing a platform to the purpose of the Big Data.*

## 2. LITERATURE REVIEW

We can live with many of the doubts of big data for now, with the hope that its benefits will remove its harms, but we shouldn't blind ourselves to the possible irreversibility of changes-whether good or bad-to society[1]. The University of Wollongong discusses the Computer magazine special issue on "Big Data: New Opportunities and New Challenges".

Big data, because it can mine more knowledge for economic growth and technical innovation, has recently received more attention[2], and many research efforts are been done for big data processing due to its high volume, velocity, and variety challenges.

Big data is a term used for large amount of data sets having large[3] and group structure with the difficulties of storing, analyzing and visualizing the data for further processes or results. The process of research into massive amounts of data is to reveal hidden patterns and secret correlations that can be used by the big companies named as big data analytics.

Big data is state as a large amount of data which requires new technology and architectures such as Hadoop so that it becomes possible to understand the dataanalysis process that can be used[4]. Due to such large size of data it becomes very difficult for human to perform successful analysis using the existing traditional techniques that are used for analysis.

## 3. METHODOLOGY

### 3.1 3V's in Big Data

1) Volume
2) Velocity
3) Variety

1) Volume –Many companies are now collecting data from various sources, including business transactions, social media and from survey. In the past, storing that huge amount of data was very difficult – but new technologies such as hadoop have made it easy to store.

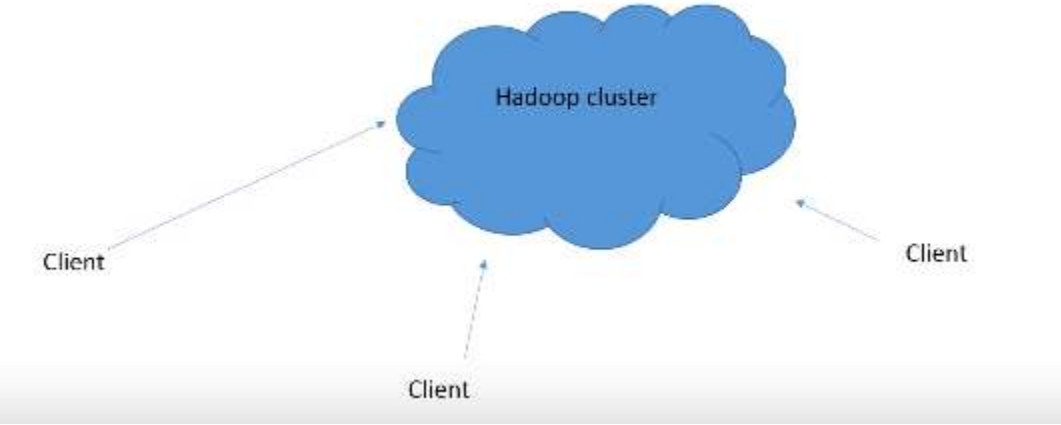
2) Velocity - Data streams is at now unstoppable speed and must be dealt with in a timely manner. RFID tags, smart metering are driving the need to deal with torrents of data in near-real time.

3) Variety - Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

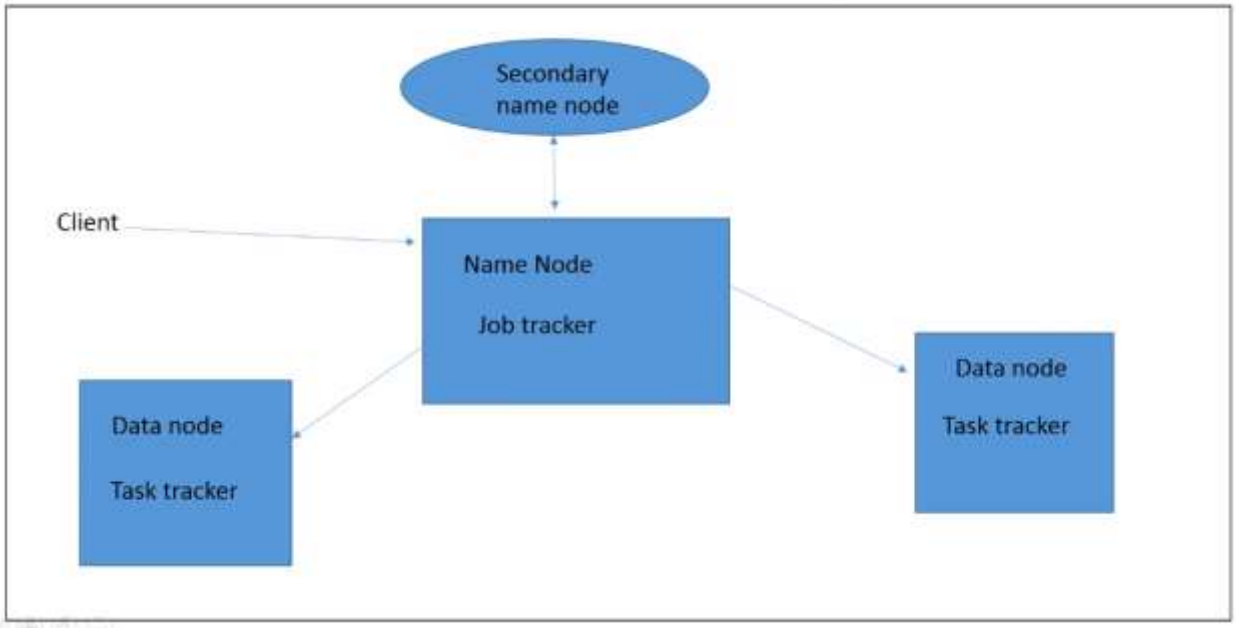
The importance of big data not only includes how much data you have, but what is your purpose. You can take data from any source and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

## 4. Hadoop

Hadoop is a way of storing huge amount of data across distributed group of servers and then running it from "distributed" analysis applications in each cluster. It's designed to be strong, in that your Big Data applications will continue to run even when individual servers — or clusters — fail. And it is also designed to be more efficient, because it doesn't require your applications to transfer huge amount of data across your network. In hadoop, there are several numbers of client and in hadoop cluster number of machines or computers are connected through a network which process the job given by the client to the hadoop. It is divided in to small parts which is explained in the physical architecture of the hadoop.



**4.1 Physical Architecture of Hadoop**




**4.1.1 Process**

Client provides its job request to hadoop. This request is accepted by the name node. Name node is master in hadoop.It also contains a job tracker which is again a part of master.This job is divided into task's and job tracker provides it to the data node. Now the data node is a slave and it possesses task tracker which actually performs the task. And job tracker continuously communicates with task tracker and if anytime it fails to reply then it assumes that the task tracker may have crashed and it assigns the job to the new task tracker.Physical architecture is a master slave combination.Because every cluster have only one master i.e. Name node. Data node is a slave.

**4.1.2 Components present in Physical Architecture of hadoop**

1) Name Node (NN)
2) Data Node (DN)
3) Job Tracker (JT)
4) Task Tracker (TT)
5) Secondary Name Node (SNN)

1) Name Node (NN)

i) The job from the client goes to the name node and it's on the name node that accepts the job given by the client.
ii) It is the master of HDFS i.e. Hadoop File System.
iii) It has job tracker which keeps track of files distributed to data nodes.
iv) Name node is the only single point to failure.
v)If the name node fails then the whole structure is crashed.

2) Data Node (DN)

i)The job is given to the data node by the job tracker. Data node will then further pass the job to the task tracker to be performed.
ii) It is the slave of HDFS.
iii) It takes client block address from name node.
iv) For replication purpose it can communicate with other name node.

v) Data node informs local changes/ updates to name node.
vi) Each node in the structure can communicates with each other.

3) Job Tracker (JT)

i)Job tracker divides the job in to number of small parts that are easy to be executed and are passed to the data nodes.
ii) It determines the files to process,
iii) Only one jobtracker per hadoop cluster is allowed.
iv) It runs on a server as a master node of cluster.

4) Task Tracker (TT)

i)Task tracker does the job given by the data node.
ii) There is a single task tracker per slave node.
iii) It may handle multiple taskparallelly.
iv) Individual tasks are assigned by job tracker to task tracker.
v) Job tracker continuously communicates with task tracker and if anytime it fails to reply then it assumes that the task tracker has crashed.

5) Secondary Name Node (SNN)

i) State monitoring is done by SNN.
ii) Every cluster has one SNN.
iii) SNN resides on its own machine.
iv) On that machine or server no other daemon (DN or TT) can work.
v) SNN takes snapshot of HDFS metadata at constant intervals.

## 5. Future Scope

Big data is a way of storing huge amount of data in a single cluster which is resolving the need to store data in small units. This is a huge revolution in the world of storing data. It is going to be used by big companies for analysis purpose and targeting the purpose what the user exactly need. Hadoop is going to be big in future as it is developing day-by-day and is allowing a platform thousands and lacs amount of data in a single cluster resolving the pervious issue of storing the data. Talking about the future, the demand for big data consultants and data scientists is going to be huge. The increasing demand for big data management needs to be satisfied with adequate supply of talent. Hence, the need for skilled data scientists is going to be huge.

## 6. Conclusion

In this, we have explained how the data is processed in hadoop. The physical architecture of hadoop is explained and the working of each component in the architecture is also explained what is the role of each components. In this way big data can be a big turning point towards storing and processing the data.

## 7. REFERENCES

[1]Katina Michael and Keith W. Miller "Big Data: New Opportunities and New Challenges [Guest editors' introduction]"
https://ieeexplore.ieee.org/abstract/document/6527259/
[2]Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu and Jun Shao "Toward efficient and privacy-preserving computing in big data era"
https://ieeexplore.ieee.org/abstract/document/6863131/
[3]SerefSagiroglu and DuyguSinanc "Big data: A review"

https://ieeexplore.ieee.org/abstract/document/6567202/
[4]AvitaKatal, Mohammad Wazid and R. H. Goudar "Big data: Issues, challenges, tools and Good practices" https://ieeexplore.ieee.org/abstract/document/6612229/
[5]Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding "Data mining with big data" https://ieeexplore.ieee.org/abstract/document/6547630/
[6]Sachchidanand Singh and Nirmala Singh "Big Data analytics" https://ieeexplore.ieee.org/document/6398180/
[7]Ibrahim AbakerTargioHashema, IbrarYaqooba, Nor BadrulAnuara, SalimahMokhtara, AbdullahGania and SameeUllahKhanb "The rise of "big data" on cloud computing: Review and open research issues" https://www.sciencedirect.com/science/article/pii/S0306437914001288?via%3Dihub
[8]https://en.wikipedia.org/wiki/Big_data
[9]BinaKotiyal, Anikit Kumar, BhaskarPany and R H Goudar"Big data: Mining of log file through hadoop" https://ieeexplore.ieee.org/document/6887797/
[10]Shankar Ganesh Manikandan and Siddarth Ravi "Big Data Analysis Using Apache Hadoop" https://ieeexplore.ieee.org/document/7021746/
[11] MrunalSogodekar, ShikhaPandey, IshaTupakari and AmitManekar "Big data analytics: hadoop and tools" https://ieeexplore.ieee.org/document/7940204/
[12]JyotiNadimath, Ekata Banerjee, AnkurPatil, Pratimakakde, SaumitraVaidya and DivyanshChaturvedi "Big data analysis using Apache Hadoop" https://ieeexplore.ieee.org/document/6642536/