iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** IV **Month of publication:** April 2017

DOI: http://doi.org/10.22214/ijraset.2017.4020

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Survey Based on Hadoop

Chaitali Mohite[1], Nisha Ambarte[2], Rakshanda Wadichor[3]
[1, 2, 3]IT (SEM-6th), SRMCEW, RTMNU

Abstract: ***This document gives the information about the HADOOP- big data. Hadoop is the popular open source for storing the big data at a time. It is a powerful tool designed for deep analysis and transformation of very large database. Hadoop has its own file system. Hadoop applications are most widely accessible. In this document we discuss applications of Hadoop, the challenges of big data, the technology of Hadoop, comparison of Hadoop architecture and its advantages and disadvantages*** [1]***.***
Keywords: Big Data, Hadoop Framework, HDFS, MapReduce, Hadoop component

I. INTRODUCTION

In a Hadoop data is distributed into nodes of cluster which is begin loaded in. The large amount of data files partitioning into chunks which are managed by different nodes in the cluster known as Hadoop Distributed File System (HDFS). Each chunk is repeatedly across many machines; so that a single machine fails then it does not result in any data begin unavailable. An activate monitoring system then duplicate data in response to system failures which can result in partial storage. Even the file chunks are as it is and distributed across discrete machines, they form a single namespace, so their accessories are universally available [1].

A. The Challenges of Big Data

1) *Volume:* Volume refers to amount of data and represent the size of the data how the data is large. The size of the data is described in terabytes and petabytes [2].
2) *Variety:* Variety makes the data too large. The files come in various formats and of any type. It may be structured or unstructured such as text, audio, videos, log files and more [2].
3) *Velocity:* Velocity refers to the speed of data processing and the data comes at high speed. Big data is time sensitive [2].
4) *Value*: The capacity value of big data is large. It is main source for big data because it is important for businesses, IT infrastructure system to store large amount of values in database [2].
5) *Veracity*: Veracity refers to noise, partiality and abnormality. When we deals with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be messy data [2].

B. Hadoop Technology or Hadoop Components

Hadoop is a framework used for storage and processing of data [3]. It is an open source framework, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment [4]. The framework includes these major components

1) *Hadoop Common:* It is the collection of common utilities and libraries that support the other Hadoop modules [9]. Hadoop common is also known as Hadoop core. It is an essential part of Hadoop Framework [4].
2) *Hadoop Distributed File System:* A file system stored the data in efficient manner which can be used easily. A distributed file system that provides high flow capacity access to application [3]. HDFS components create many duplicates of the data block to be distributed beyond different clusters for reliable and quick data access. HDFS includes three components i.e., Secondary Name Node, Name Node and Data Node as shown in fig. HDFS operates on a master-slave architecture model. Where Name Node acts as a master node and Data Node act as a slave node [5]. Master node is used for keeping a track of the storage clusters and slave node is used for collection to the various systems with a Hadoop cluster. HDFS used for storing all the structured data and unstructured data hence, it permits processing of the stored data [5].

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

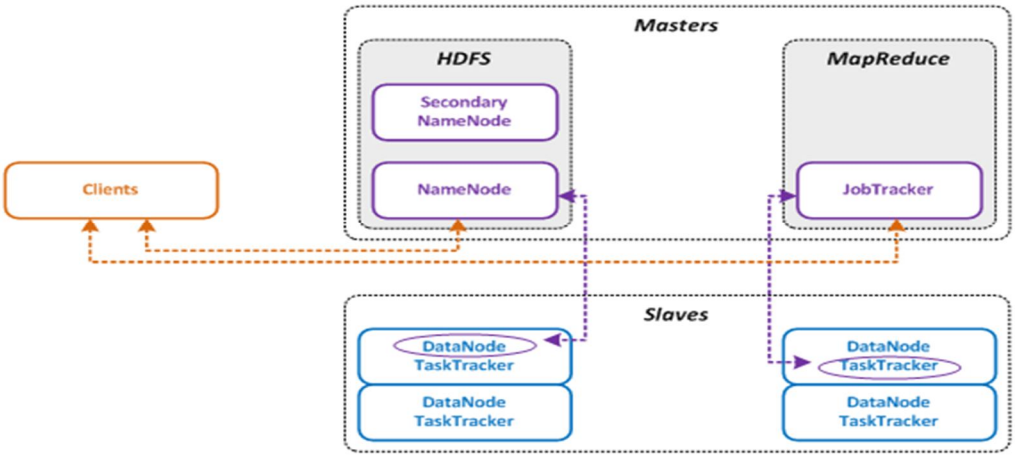


Fig 1: Master/Slave Architecture of Hadoop

3) *Hadoop MapReduce:* It is a programming module and is used for large scale data processing. Hadoop MapReduce is a software framework for easily writing applications which process large amounts of data in-parallel on large clusters [3]. The model is commonly used in functional programming. MapReduce is a java based system created by Google. The exact data stored in HDFS gets processed efficiently. MapReduce is that the "Map" job sends a query for processing to various nodes in a Hadoop cluster and the "Reduce" job collects all the results to output into a single value. Map Task takes input data and splits into independent chunks and output of this task will be the input for Reduce Task [5]. Reduce task combines Mapped data tuples into smaller set of tuples. While, both input and output of tasks are stored in a file system. MapReduce takes care of scheduling jobs, monitoring jobs and re-executes the failed task [5]. The dedication tasks of the MapReduce component are hold by two daemons- Job Tracker and Task Tracker as shown in the image below

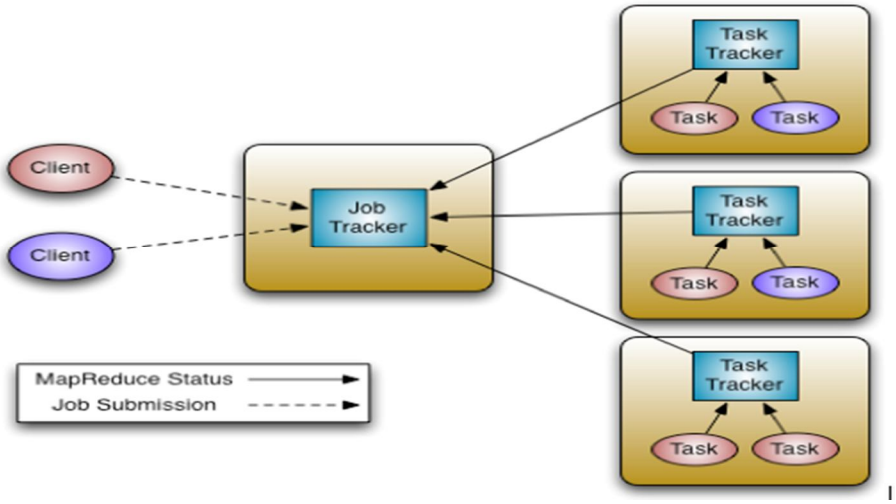


Fig 2: Hadoop MapReduce

4) *Hadoop YARN:* YARN is stands for Yet Another Resource Negotiator. It is a resource management platform, a framework for job scheduling and cluster resource management [3]. Hadoop framework as users can run various Hadoop applications without having to bother about increasing workloads. Resource manager act as a master and node manager act as a slaves. The following figure shows the components of Hadoop YARN-

a) *Resource Manager (RM):* It manages central agent and allocates cluster resources [5].

b) *Node Manager (NM):* It manages per-node agent and enforces node resource allocations [5].

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

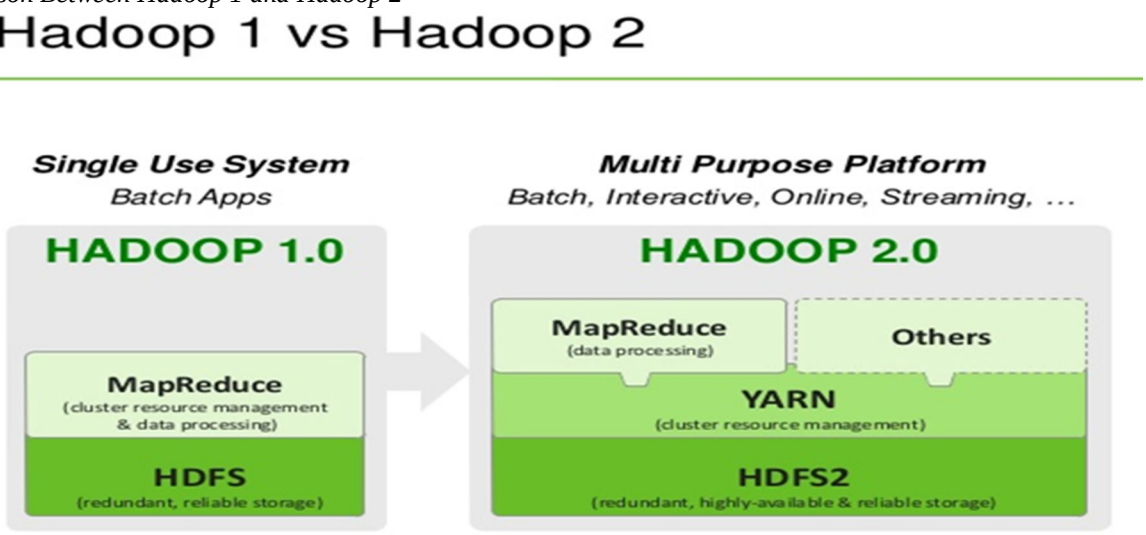
c) *Application Master (AM):* It manages application lifecycle and task scheduling [5].

Fig 3: YARN Architecture

C. *Main Advantages of Hadoop YARN*

1) It provides high cluster use [5].
2) It is highly scalable [5].
3) It is beyond Java [5].
4) It is modern programming models and services [5].
5) It is agility [5].

D. *Comparison Between Hadoop 1 and Hadoop 2*



The above fig. shows the version of Hadoop i.e. Hadoop 1.0 and Hadoop 2.0. Where, the Hadoop 1.0 is a single use system and the Hadoop 2.0 is a multi-purpose platform system. Hadoop 1.0 includes two components i.e., MapReduce and HDFS and Hadoop 2.0 includes four components i.e., MapReduce, YARN, HDFS and other components. Hadoop 1.0 has only one purpose i.e., for batch apps and Hadoop 2.0 has multi-purpose platform i.e., for batch, interactive, online, streaming and so on. In Hadoop 1.0 the MapReduce used for both large scale data processing and cluster resource management. But in Hadoop 2.0 the MapReduce component only used for large scale data processing and YARN component is separately used for cluster resource management. In

Hadoop 1.0 HDFS component is use for redundant and reliable storage. But in Hadoop 2.0 HDFS component is use for redundant, reliable storage and also highly available.

E. Advantages of Hadoop

1) *Scalable:* Hadoop is a highly scalable storage platform [6].
2) *Cost Effective:* Hadoop also offers a cost effective storage [6].
3) *Flexible:* Hadoop enables to easily access new data sources and tap into different type of data to generate value from the data [6].
4) *Fast:* Hadoop resulting in much faster data processing [6].
5) *Resilient to Failure:* A main advantage of using Hadoop is fault tolerance [6].

F. Disadvantages of Hadoop

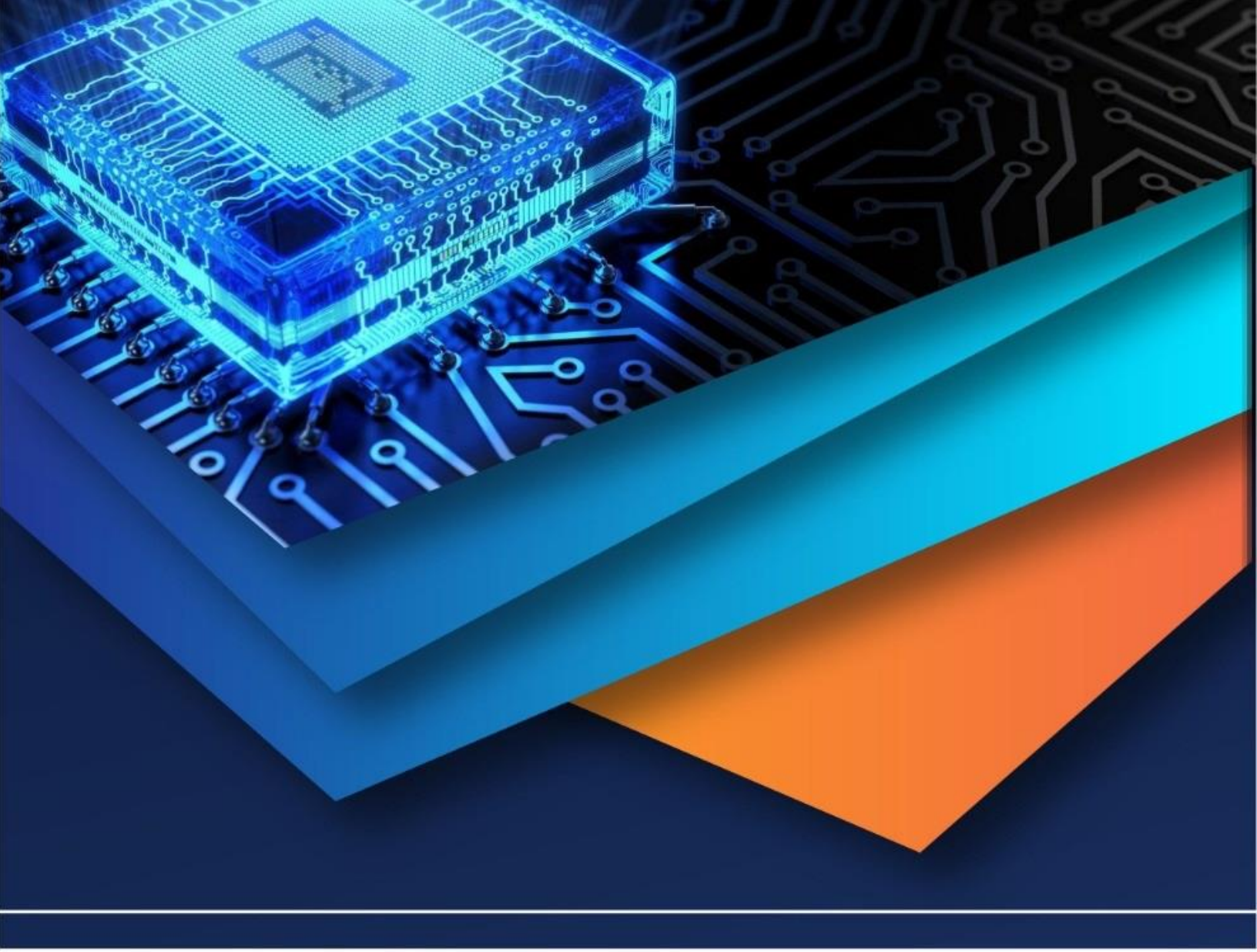
1) *Security Concerns:* Managing a complex application such as Hadoop can be challenging [7].
2) *Vulnerable by Nature:* The framework is written in Java. Java has been heavily exploited by cybercriminals and as a result, involve in numerous security violation [7].
3) *Not Fit for Small Data:* It is not recommended for organization with small quantities of data [7].
4) *Potential Stability Issues:* Hadoop has its fair share of stability issues [7].

II. CONCLUSION

In Hadoop at a time large quantity of data will be stored. It is not recommended small quantities of data. It mainly includes four components used to store the large amount of data i.e., Hadoop common, HDFS, MapReduce and YARN. In Hadoop, there are some advantages and disadvantages of Hadoop. In the comparison of Hadoop version Hadoop 1 is single Name node to manage the entire namespace but in Hadoop 2 multiple Name node servers manage multiple namespace [8].

REFERENCES

[1] http://www.seminarsonly.com
[2] Varsha B. Bobade “Survey paper on Big Data and Hadoop”, IRJET 2016
[3] http://www.collegelib.com
[4] http://www.zapmeta.co.in
[5] https://www.dezyre.com
[6] http://www.itproportal.com
[7] http://blogs.mindsmapped.com
[8] http://acadgild.com
[9] C Lakshmi, V. V .Nagendra Kumar “Survey paper on Big Data”, ijarcsse 2016
[10] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar “Review paper on Big Data and Hadoop”, ijsrp 2014
[11] Ms. Vibhavari Chavan, Prof. Rajesh N. Phursule “Survey paper on Big Data”, ijcsit 2014
[12] Andrew Pavlo, “A Comparison of Approaches to Large-Scale Data Analysis”, SIGMOD, 2009.
[13] Apache Hadoop: http://Hadoop.apache.org
[14] Dean, J. and Ghemawat, S., “MapReduce: a flexible data processing tool”, ACM 2010.
[15] DeWitt & Stonebraker, “MapReduce: A major step backwards”, 2008.
[16] Hadoop Distributed File System, http://hadoop.apache.org/hdfs
[17] Hadoop Tutorial: http://developer.yahoo.com/hadoop/tutorial/module1.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY