iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI:

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

Review: Big Data and Hadoop

Harsh Kaler[1], Hemant Kumar Sharma[2]
[1, 2]*Department of CSE, Arya Institute of Engineering and Technology, Jaipur*

Abstract: During this world of technology the term of 'Big Data' is refers to new techniques and technologies to capture, store, manage, process, distribute and different data sets. 'Big Data' may be an information that contains an oversized size. It is collection of knowledge that is large and increasing exponentially with time. Data is so extensive and complicated that none of conventional data management tools are able to store it or process it efficiently. In 'Big Data' data generally may be in three categories: structured, unstructured and semi structured. Data is also generated from different sources and arrives with within the system with different rates. Big Data may be an information whose ratio, variety, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is that the core platform for structuring Big Data, and solves the matter of creating it useful for analytics purposes. Hadoop is an open source software project that permits the distributed processing of massive data sets across clusters of commodity servers
Keywords: Bigdata, Hadoop framework, HDFS, MapReduce, Hadoop Component.

I. INTRODUCTION

Big Data may be group of giant dataset which is able not be processed using traditional computing techniques. Big Data is not merely an information rather it is become a full subject which involve various tools, techniques and framework. The need of giant data generated from the massive companies like Facebook, Yahoo, Google, YouTube etc. for the aim study of enormous amount of information also google contains the large amount of information Big Data may be a term that refers to dataset whose size, complexity and rate of growth make them too difficult to captured, managed, processed or analyzed by traditional technology and tools like relational database. There are various technologies within the market from different vendors including Amazon, IBM, Microsoft etc. to handle big data. There are five characteristics for big data. They are Volume, Velocity, Variety, Veracity and Value.

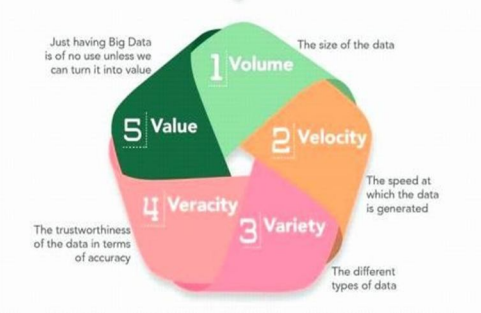


Fig. 1 Characteristic

1) *Volume:* The quantity of information that companies can collect is basically anormous and hence the degree of information becomes a critical factor in big data analytics.
2) *Velocity:* The speed at which new data is being generated all due to our dependence on the web, sensor, machine to machine data is additionally important to parse big data in a very timely manner.
3) *Variety:* The information is generated is totally heterogeneous within the sense that it can be in various formats like video, text, database, numeric, sensor data so on and hence understanding the sort of huge data is essential factor to unlocking it is value.
4) *Veracity:* Knowing whether the information that is available from a reputable source is vital before deciphering and implementing big data for business needs.
5) *Value:* Data by itself is of no price unless it's processed to induce the information victimization that one might initiate actions, the large volume of data makes process tough. As luck would have it, computing power and storage capability have conjointly inflated hugely.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

II. HADOOP: DATA PROCCESING

Hadoop is an open source programming framework which written in java that allows processing of large data sets in a distributed computer environment. Hadoop was developed by google on the MapReduce system and it applies concepts of functional programming. Hadoop follows horizontal scaling instead of vertical scaling. In horizontal scaling, you we add new nodes to HDFS (Hadoop Distributed File System) cluster on the run as per requirement, instead of increasing the hardware stack present in each node. An Apache Hadoop ecosystem which mainly consists two components in Hadoop kernal:

HDFS and MapReduce

Other components like YARN, HBase, Pig, Oozie, ZooKeeper, Sqoop, Flume.

A. HDFS (Hadoop Distributed File System)

The Hadoop Distributed File System (HDFS) is a file system that runs on the standard and low end software. It developed by apache Hadoop. HDFS works like a standard distributed file system but provides better data and throughput and access through the MapReduce algorithm, high fault tolerance and native support of large data sets.

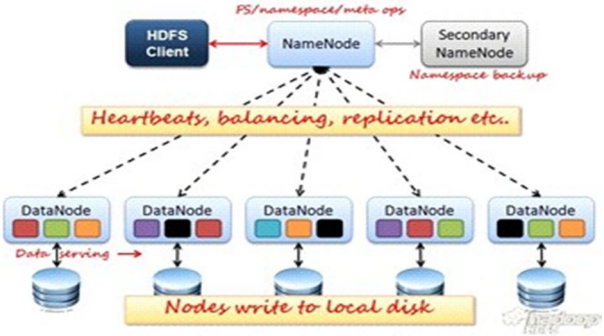


Fig. 2 Architecture of HDFS

The HDFS stores an outsized amount of information placed across multiple machines, typically in hundreds and thousands of simultaneously connected nodes and provides data reliability by replicating each data instance as three different copies-two in one group and one in another. These copies are also replaced within the event of failure. The HDFS architecture consists of clusters, each of which is accessed through one NameNode software tool installed on a separate machine to watch and manage the that cluster's classification system and user access mechanism. The opposite machines install one instance of DataNode to manage cluster storage. Because HDFS is written in JAVA, it's native support for JAVA application programming interface(API) for application integration and accessibility. It is also accessed through standard wed browsers.

B. MapReduce

MapReduce may be a programming model introduced by google for processing and generating large data sets on clusters of computers. Google first formulated the framework for the aim of serving Google's web content indexing, and also the new framework replaced earlier indexing algorithms. Beginner developers find the MapReduce framework beneficial beacause it's library routines are often accustomed create parallel programs with none worries about infra-cluster communication, task monitoring or failure handling processes. MapReduce runs on an oversized cluster of commodity machines and his highly scalable. it's several styles of implementation provided by multiple programming languages, like JAVA, C# and C++.

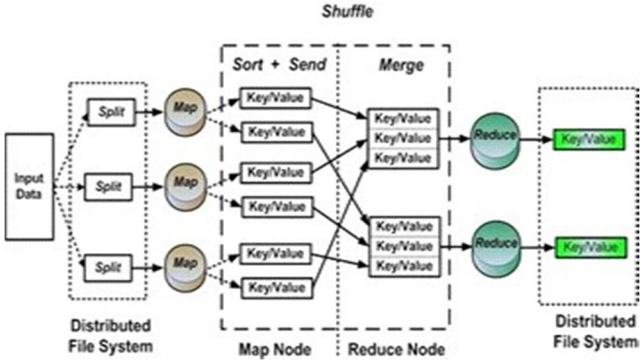


Fig. 3 Architecture of MapReduce

The MapReduce framework has two parts: A function called "Map," which work allows different points of the distributed cluster to distribute their work and another one function called "Reduce," which is meant to scale back the ultimate variety of the clusters' results into one output The main advantage of the MapReduce framework is that its fault tolerance, where periodic reports from each node within the cluster are expected when work is completed. A task is transferred from one node to another node. If the master node notices that a node which has been silent for an extended interval than expected, the most node performs the reassignment process to the frozen/delayed task. The MapReduce framework is inspired by the "Map" and "Reduce" functions employed in functional programming. Computational processing occurs on data stored during a filing system or within a database, which takes a group of input key values and produces a group of output key values.

C. YARN

YARN (Yet Another Resource Nagotiator) is manage the processing in Hadoop as HDFS is manage the storage part. YARN has three components:

1) *Resource Manager:* The Hadoop there's one resource manager in one hadoop cluster. It works what it's name sys, managing the resource.
2) *Node Manager*: Each node manager takes intructions from the resource manager and reports and handles containers on one node.
3) *Application Manager:* It takes the tasks in variety of application that are submitted to the cluster and assigns it to the information nodes.

D. HBASE

Hbase is distributed column oriented non-relational database where as HDFS is file system. It is written in Java and runs on the top of HDFS system. It can serves as input and output for the MapReduce.

E. Pig

Pig is providing platform for big data analysis and processing. Pig adds another level that is abstraction in data processing and it makes writing and maintaining data processing jobs very easy. Pig also can process tera bytes of data with half dozen lines of code.

F. Hive

Hive is sort of a data ware housing framework that's on the highest of Hadoop. Hive allows to write down SQL like sql queries to process and analyze the large data stored in HDFS system. it's Data warehousing application that gives the SQL interface and relational model.

G. Sqoop

Hive is sort of a data ware housing framework that's on the highest of Hadoop. Hive allows to write down SQL like sql queries to process and analyze the large data stored in HDFS system. it's Data warehousing application that gives the SQL interface and relational model.

III. ADVANTAGES OF HADOOP

1) *Varied Data Sources:* The data can come from a different sources such as mail conversation, YouTube, Facebook, Twitter, video, text and picture etc. These data can be in the form of structured and un structured. Hadoop also can accept data in various file like XML, CSV, JASON etc.
2) *Cost-Effective:* Hadoop also provide a cost-effective storage solution for financial exploding data sets. The problem with traditional electronic database management systems is that it's extremely cost prohibitive to scale to such a degree so as to process such massive volumes of knowledge. The data would be deleted, because i**t** would be too cost-prohibitive to stay.
3) *Flexible:* Hadoop enables businesses to simply access new data sources and tap into differing kinds of knowledge (both structured and unstructured) to come up with value from that data. this suggests businesses can use Hadoop to derive valuable business insights from data sources like social media, email conversations.
4) *Fast:* Hadoop's has own storage method is based on a distributed file system that basically check data wherever it is located on a cluster. The tool which data processed are often on the same server where the data is located. In the result faster data processed.

IV. SCOPE OF HADOOP IN FUTURE

The scope of Hadoop is increasing within the future. Most of the large companies are working in Hadoop for Data analysis like Google, Facebook, Linkedin and lots of more.

But presently Hadoop isn't stable, new technologies during this field are coming day by day. Even Java took time to induce stable within the market and as we all know now java is stable language. within the coming years Hadoop will get stable and can be the simplest technology for Data Analysis.

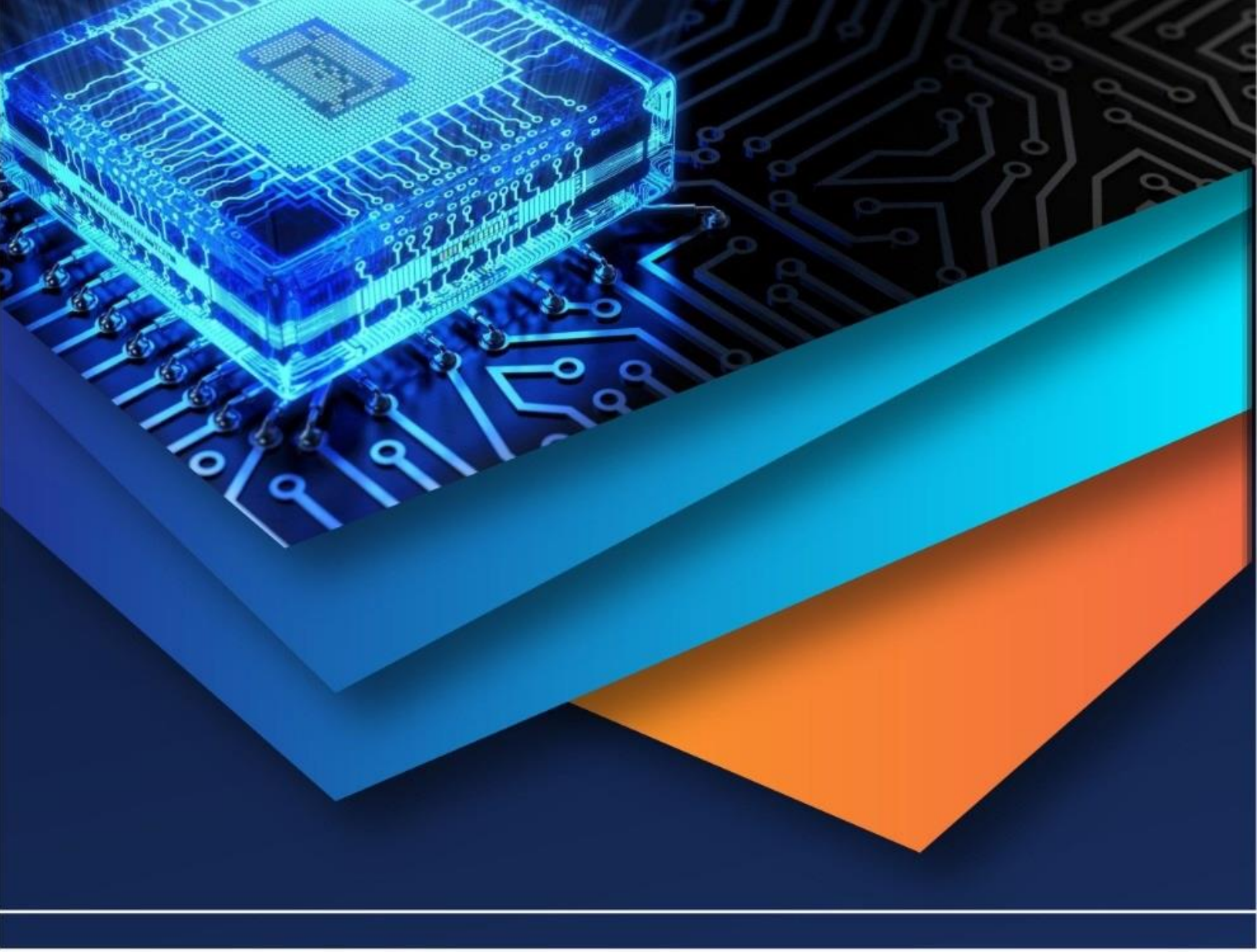
V. CONCLUSIONS

In this paper we studied about BigData and Hadoop, The availability of massive Data, low-cost commodity hardware, and new information management and analytic software have produced a novel moment within the history of information analysis. The convergence of those trends implies that we've got the capabilities required to investigate astonishing data sets quickly and cost-effectively for the primary time in history. These capabilities are neither theoretical nor trivial. They represent a real breakthrough and a transparent opportunity to appreciate enormous gains in terms of efficiency, productivity, revenue, and profitability.

The Age of massive Data is here, and these are truly revolutionary times if both business and technology professionals still work together and deliver on the promise.

REFERENCE

[1] Kiran kumara Reddi & Dnvsl Indira “Different Technique to Transfer Big Data : survey” IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}
[2] Kiran kumara Reddi & Dnvsl Indira “Different Technique to Transfer Big Data : survey” IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}
[3] Jonathan Paul Olmsted “Scaling at Scale: Ideal Point Estimation with ‘Big-Data” Princeton Institute for Computational Science and Engineering 2014.
[4] Apache Hadoop Project, http://hadoop.apache.org/ ,2013.
[5] C Lam, “Introducing Hadoop”, in Hadoop in Action, MANNING,2011.
[6] J.Venner and S.cyrus , “The Mapreduce”, in Pro Hadoop, vol 1, New York, APRESS,2009.
[7] https://www.pdfdrive.com /big-data-analytics
[8] https://searchdatamanagement.techtarget.com/definition/Hadoop- Distributed-File-System-HDFS



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY