iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VIII **Month of publication:** August 2020

DOI: https://doi.org/10.22214/ijraset.2020.31243

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 VIII Aug 2020- Available at www.ijraset.com

An Outline of Hadoop in Bigdata

Vandana Malik[1], Dr. Mukesh[2]
[1](Research Scholar), [2](Astt. Prof.), Dept. of Computer Science, BMU, Rohtak

Abstract: Adoption of big data technology has changed many business organizations' perspective on data and its value. Traditional data infrastructure has been replaced with big data platforms offering capacity and performance increases at a linear cost increase, compared with traditional infrastructure's exponential cost increase. This change in how businesses store and process their data has led them to derive more insight from their existing data by combining multiple datasets and sources to yield a more complete view of their customers and operations. The success of businesses using big data to change how they operate and interact with the world has made many other businesses prioritize big data rollouts as IT initiatives to realize similar results. Hadoop has been at the center of this big data transformation, providing an ecosystem with tools for businesses to store and process data on a scale that was unheard of several years ago. Two key components of the Hadoop ecosystem are Hadoop Distributed File System and Hadoop MapReduce; these tools enable the platform to store and process large datasets (terabytes and above) in a scalable and cost-effective manner.
Keywords: Hadoop, Big Data, HDFS, MapReduce,

I. INTRODUCTION

Hadoop is an open source software framework for processing huge volumes of distributed data using distributed processing capabilities. The framework supports distributed processing of large datasets distributed across clusters of computers using simple programming models. Scalability is a great feature the framework offers, it is designed to scale from few servers to 1000 of servers. The nodes can be added as needed without impacting the underlying applications. Instead of it Hadoop framework offers high degree of fault tolerance.

It does not rely on the hardware to provide high availability, instead it has mechanism to detect and handle the failures. Hadoop provides a solution for storing and analysing the enormous amount of data generated these days by the digital world. So basically it is an answer to the problem BIG DATA poses to the digital world.

On the whole, we significantly identify and describe the major factors, that Hadoop approach improves accessing large sets of data say "big data" to meet the rapid changing business environments.

We also provide a brief comparison Hadoop technique with traditional systems techniques, and discuss current state of adopting Hadoop techniques.

We speculate that from the need to satisfy the customer through time dependency. Hadoop is emerged as an alternative to traditional methods. The purpose of this paper is to provide an in-depth understanding, the major benefits of Hadoop approach to access, as well as provide a study report of Hadoop importance in the present scenario.

We all precisely know that Web applications have become an essential component of business, whereas social media also have its vital role in today's. Further to it, the web application helps to develop business and achieve its objectives much faster, resulting big data. Analyzing unstructured data typically involves complex algorithms. Sources of Big Data can be broadly classified into six different categories as *Enterprise Data* i.e. flat files, emails, Word documents, spreadsheets, presentations, HTML pages, pdf. etc. *Transactional Data* like Web Applications, Mobile Applications, CRM Systems, and many more. *Social Media* like Twitter, Facebook, etc.. *Activity Generated* include data from medical devices, censor data, surveillance videos, satellites, cell phone towers, industrial machinery, and other. *Public Data* published by governments, research data published by research institutes, data from weather and meteorological departments, census data, Wikipedia, sample open source data feeds, and other data which is freely available to the public.

This type can be broadly classified into two categories - *Structured Data* and *Unstructured Data.*

In present time Hadoop is one of the tools designed to handle big data. Hadoop and other software products work to interpret or parse the results of big data searches through specific proprietary algorithms and methods. Hadoop is an open-source program under the Apache license that is maintained by a global community of users. It includes various main components, including a MapReduce set of functions and a Hadoop distributed file system (HDFS).

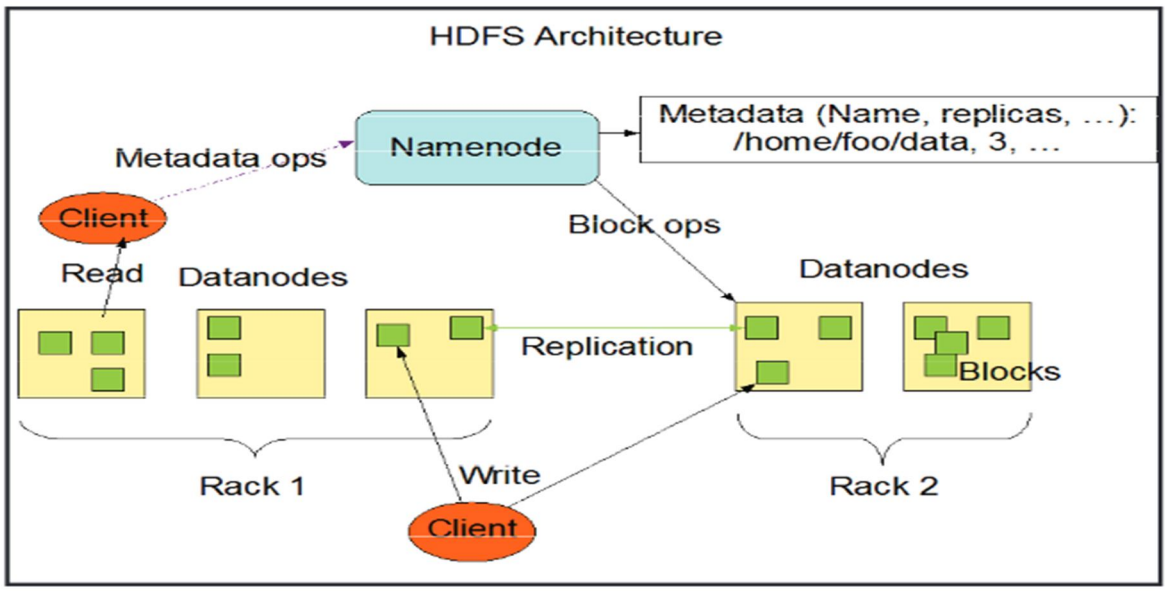
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

II. ARCHITECTURE OF HADOOP HELPS TO SECURE BIG DATA

The Hadoop framework application works in an environment that provides distributed *storage* and *computation* across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

A. At its core, Hadoop has two Major Layers Namely

1) Processing/Computation layer (MapReduce), and
2) Storage layer (Hadoop Distributed File System).



B. MapReduce

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.

C. Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules

1) *Hadoop Common:* These are Java libraries and utilities required by other Hadoop modules.
2) *Hadoop YARN:* This is a framework for job scheduling and cluster resource management.

D. How Does Hadoop Work

Hadoop helps to execute large amount of processing where the user can connect together multiple commodity computers to a single-CPU, as a single functional distributed system and have the particular set of clustered machines that reads the dataset in parallel and provide intermediate, and after integration gets the desired output.

Hadoop runs code across a cluster of computers and performs the following tasks

1) Data are initially divided into files and directories. Files are divided into consistent sized blocks ranging from 128M and 64M.
2) Then the files are distributed across various cluster nodes for further processing of data.
3) Job tracker starts its scheduling programs on individual nodes.
4) Once all the nodes are done with scheduling then the output is return back.

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

III. BENEFITS OF HADOOP IN CURRENT SCENARIO

Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.

Unlike traditional relational database systems that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data. It also offers a cost effective storage solution as the problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data. In Hadoop, on the other hand, is designed as a scale-out architecture that can affordably store all of a company's data for later use.

Hadoop also enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data.

In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection. Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing.

The most important key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use.

IV. THE MAIN CHARACTERISTICS OF HADOOP ARE

A. It is open source based on Java applications and hence compatible on all the platforms.
B. It can handle unstructured data and semi-structured data.
C. Hadoop clusters provide storage and distributed computing all in one.
D. Hadoop framework has built-in power and flexibility to do what was not possible earlier.
E. It give access to the user to rapidly write and test the distributed systems and then automatically distributes the data and works across the machines and in turn utilizes the primary parallelism of the CPU cores.
F. Servers can be added or removed from the cluster dynamically at any point of time.
G. It helps in distributing data on different servers and prevents network overloading.
H. Hadoop library are developed to find/search and handle the failures at the application layer.
I. Hadoop offers scalability, reliability and plenty of libraries for various applications at lower cost.
J. More storage and computing power can be achieved by addition of more nodes to Hadoop cluster. This eliminates need to buy external hardware. Hence it is cheaper solution.
K. HDFS layer in hadoop has self-healing, replicating and fault tolerance characteristics. It automatically replicates data if server or disk got crashed.

V. CONCLUSION

Hadoop, which is based on the Hadoop HDFS and MapReduce has provided a distributed data processing platform. The high fault tolerance and high scalability allow its users to apply Hadoop on cheap hardware. The MapReduce distributed programming mode allows the users to develop their own applications without the user shaving to know the bottom layer of the MapReduce. Because of the advantages of Hadoop, the users can easily manage the computer resources and build their own distributed data processing platform. Above all, it is obvious to notice the convenience that the Hadoop has brought in Big Data processing.

It also should be pointed out that since Google published the first paper on the distributed file system till now, the history of Hadoop is only 10-year old.

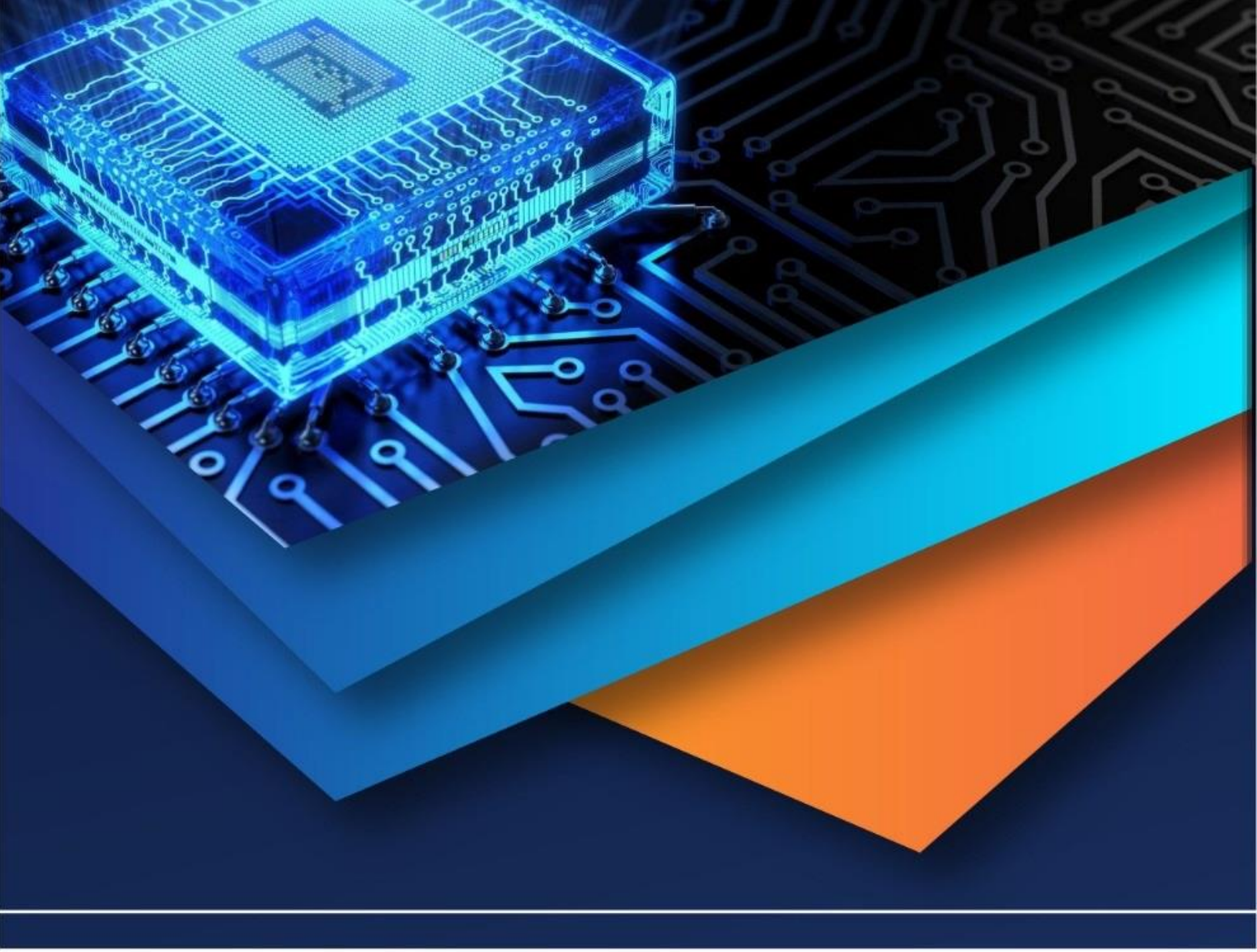
With the advancement of the computer science and the Internet technology, Hadoop has rapidly solved key problems and been widely used in real life.

In spite of this, there are still some problems in facing the rapid changes and the ever increasing demand of analysis. To solve these problems, Internet companies, such as Google also introduced then ever technologies.

It is predictable that with the key problems being solved, BigData processing based on Hadoop will have a wider application prospect. Hence, the Hadoop technology provide mature, stable and feature rich platforms for global vertical organizations to implement complex Digital projects.

REFERENCES

[1] S.Vikram Phaneendra, E.Madhusudhan Reddy,"Big Data- solutions for RDBMS problems- A survey", In 12thIEEE/IFIP Network Operations & Management Symposium.
[2] Hewlett-Packard Development Company, Big Security for Big Data. L.P.: Hewlett-Packard Development Company.
[3] Aveksa Inc. Ensuring "Big Data" Security with Identity and Access Management. Waltham, MA: Aveksa.
[4] Kaisler, S., Armour, F., Espinosa, J. A., Money, W.. Big Data: Issues and Challenges Moving Forward. International Confrence on System Sciences (pp. 995-1004). Hawaii: IEEE Computer Society.
[5] Katal, A., Wazid, M., Goudar, R. H. Big Data: Issues, Challenges, Tools and Good Practices. IEEE.
[6] Marr, B. The Awesome Ways Big Data is used Today to Change Our World. Retrieved.
[7] https://hadoop.apache.org/http://opensource.com/life/14/8/intro-apache-hadoop-big-data http://www.ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506229.pdf.
[8] http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html.
[9] http://searchcloudcomputing.techtarget.com/definition/Hadoophttp://searchdatamanagement. techtarget.com/definition/Apache-Hive.
[10] http://www.experfy.com/blog/hadoop-market-size-adoption-growth-2020.
[11] http://www.alliedmarketresearch.com/hadoop-market.
[12] http://www.transparencymarketresearch.com/hadoop-market.html.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY