IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 12, December 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112021 |

Study of Big Data Analysis and Hadoop Ecosystem

Prathamesh Hindurao Patil , Sumit Shivaji Patil, Dr. Sampada Gulavani, Prof. Keerti Kadam

Department of Master of Computer Application, Bharti Vidyapeeth (Deemed to be University) Pune Institute of Management, Kolhapur, India

ABSTRACT: The term, Big Data' has been coined to refer to the gargantuan bulk of data that cannot be dealt with by traditional data-handling techniques. A huge repository of terabytes of data is generated each day from modern information systems and digital technologies. With the advent of Internet of Things (IoT) and Web 2.0 technologies, there has been a tremendous growth in the amount of data generated. The collection of large datasets from these devices are difficult to process using traditional database management tools or data processing applications. Therefore, big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of big data challenges, open research issues, and various tools associated with it This chapter presents an overview of big data analytics, its application, advantages, and limitations.

KEYWORDS : Big Data, Analytics, Hadoop, Hadoop Ecosystem, MapReduce.

I. INTRODUCTION

Over the past two decades, there is a tremendous growth in data. This trend can be observed in almost every field. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone , GPS signals to name a few. In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. These are available in structured, semi-structured, and unstructured format in petabytes and beyond. Formally, it is defined from 3Vs to 4Vs where 3Vs refers to volume, velocity, and variety. By extension, the platform, tools and software used for this purpose are collectively called big data technology. Currently, the most commonly implemented technology is Hadoop which is the culmination of several other technologies like Hadoop Distribution File Systems, Pig, Hive and HBase. Big data is about how these data can be stored, processed, and comprehended such that it can be used for predicting the future course of action with a great precision and acceptable time delay. This paper gives an overview of big data ranging from its sources to dimensions, limitations of existing data processing approaches need for big data analytics and development, testing of new approaches for storing and processing bigdata.

II. WHAT IS BIG DATA, HOW IT WORKS

A data which is beyond storage and processing capacity of conventional database systems is called big data. Big Data is the term for collection of data sets of large and complex that it becomes difficult to process using conventional data bases system.

Sources of Big data

- **Social Media :** Facebook, Twitter, Linkedin etc
- **Mobile Devices :** Call, text, location, and app activity etc.
- **Internet transactions :**Banking, transactions, ecommerce, transactions etc.
- **Network devices / sensors :** Sensors (Proximity, Temperature), Internet connection to different devices etc.

Characteristics of Big Data :

We can categorize any data is big data based on 5 V's

1.Volume (Size of data)
2.Velocity (Processing speed of data)
3.Variety (Types of data)
4.Value (Worth of data being extracted)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112021 |

5.Veracity (How much accurate is the data being extracted)
If any data having above 5 characteristics called as BIG DATA.

Classification of Big Data:
It has been classified in three different types of data

i) Structured Data :Data which is stored in tabular format and which is having proper structure and has relationship between the tables. For example - Table format data like employee table.
ii) Semi-Structured Data : Data Which has some structure but cannot save in tabular format is known as semi-structured data. For example-.xlsx, email, .txt, .doc etc.
iii) Unstructured Data : Data which is not having any structure and cannot save in tabular format is known as unstructured data. For example- video files, call recordings, audio files etc.

IV. HADOOP ECO-SYSTEM

Big data concept has been implemented by using Hadoop eco-system. Hadoop is a free, Java-Based programming framework that supports processing of large data sets in a distributed computing environment. It has maintained by Apache software foundation. Hadoop used commodity (cheap) hardware and clusters concepts. Hadoop is recommended for large datasets. Hadoop Architecture 1.0 consist of MapReduce and HDFS.

Hadoop Components:
i)MapReduce: Processing of data which is stored on **HDFS (Hadoop Distributed File System).**
ii)HDFS is used to store huge amount of data sets. HDFS is specially designed file system for storing data set with cluster of commodity hardware with streaming access pattern.
iii)Cluster: Group of systems connected to LAN.
iv)Commodity H/W: Cheap hardware.
v)Streaming access pattern: Streaming access pattern means you can write once, read number of times but cannot change the content of file once it is kept in HDFS. WORM (write once read many times) Normal file system block size = 4KB. In HDFS 1.0 block size 64MB and HDFS 2.0 size is 128MB.
vi)HDFS Services (Daemons/ nodes) : HDFS provides 5 services with Master services and Slave services.
a) Master Services : Name Node, Secondary Name Node, Job Tracker (Resource Manager)
b)Slave Services : Data Node, Task Tracker (Node Manager)

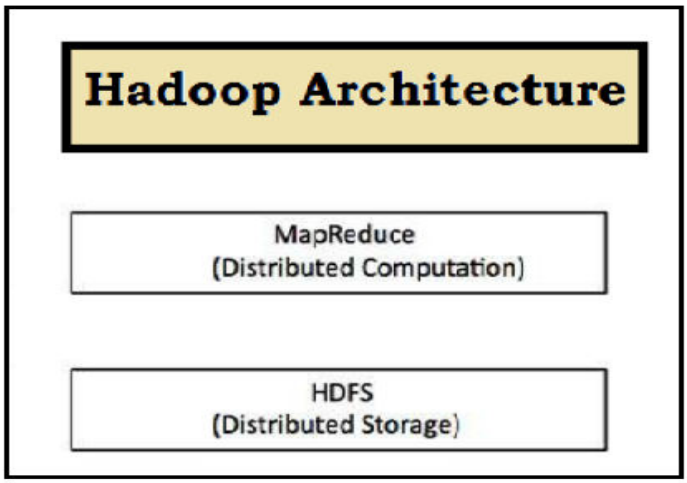

Fig 1 : Hadoop Architecture 1.0

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112021 |

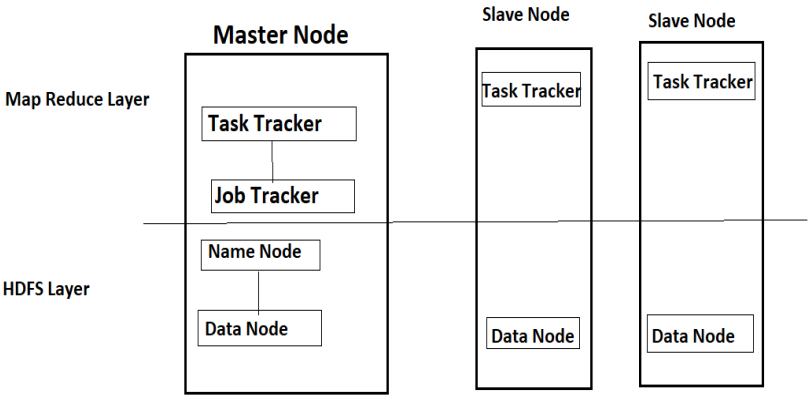

Fig 2 : High Level Architecture of Hadoop

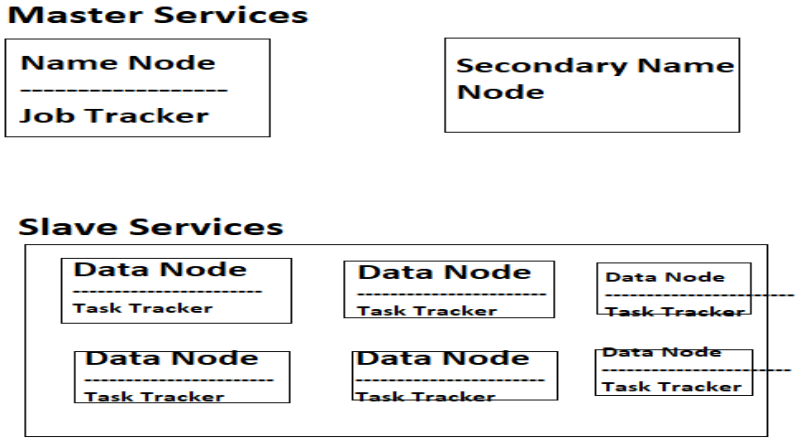

Fig 3 : Master and Salve Services

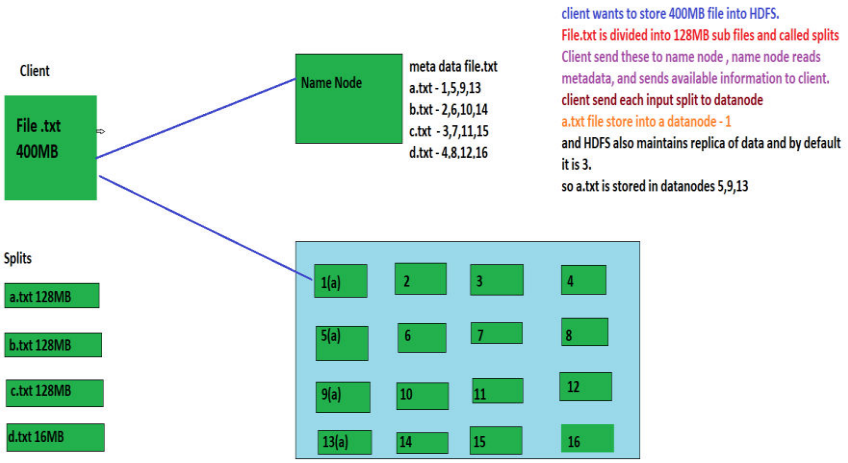

Fig 4 : Storing data into HDFS

4. Map Reduce (Retrieving of data)

- If we want to process data then we write simple program for processing, it is called as **MAP.**
- Whenever we want to process data, then we send command/ program to Job Tracker for processing job/file/data.
- Job tracker asks to Name Node and name node will reply metadata to job tracker.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112021 |

- Job tracker knows on which blocks data is stored on which data nodes.
- Using meta data, Job tracker communicates to Task tracker and assign the task and task will perform or complete by task tracker and same will be informed to Job tracker once it completes job.
- Note that Map reduce task is always performed by bigdata and the scripts are generally written in JAVA of Python.

Example: Sum of 1000 numbers
Suppose we have 10 persons to add 1000 numbers
Map Reduce Process
Step -I: Input splitting: 1-> 1 to 100, 2->101 to 200,3->201 to 300,…….. 10->901 to 1000
Step-II: Mapping Method: 1 -→ 1+2 +… 100 =5050
Map 1→ 5050 ,2→ 15150…….10→ 95950
Step-III: Reduce Method→ after addition performed by all 10 persons, reducer will collect and again add all the results.
5. Implement Map Reduce :
There are two ways to implement map reduce :

1. Splitting→Mapping → Reducer
2. Splitting → Mapping → Shuffling → Reducing

Map Reduce ex: word Count

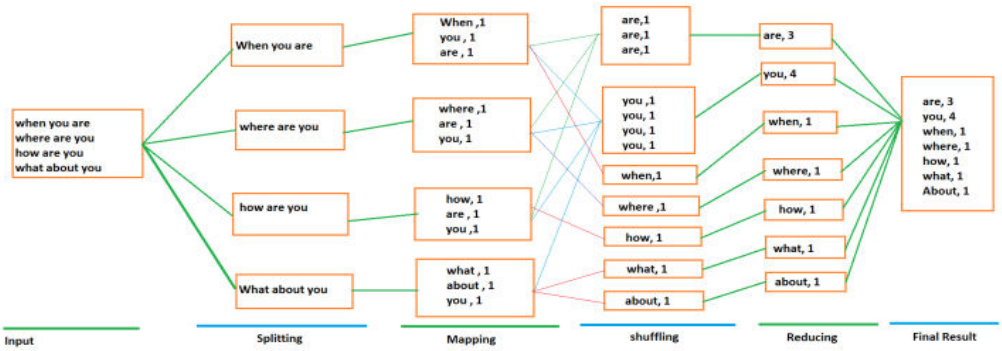

Fig.5 : Implementation of Map Reduce

If we want to use Unix commands on Hadoop Server then we must and should include either
Hdfs dfs -unix commands or Hadoop fs -unix commands
For ex: hdfs dfs -ls -l **or** hadoop fs -ls -l
Hadoop Tool – Language – It will convert language into map-Reduce script and extract data from HDFS.
6.Hadoop Ecosystem :
In Hadoop 2.x from onwards supports YARN (Yet another Resource Navigator), which helps to connect different application like Pig, Hive etc. Because of YARN it is possible to built an Apache Hadoop Ecosystem.

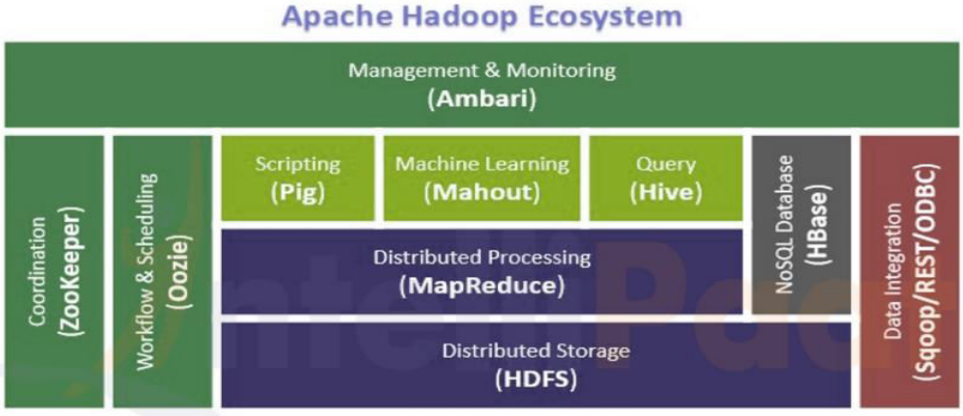

Fig. 6 : Hadoop Ecosystem

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112021 |

i)Hadoop : Hadoop provides scalable solution to store and process huge data sets in distributed fashion.
ii)Apache Hive: Hive is data warehousing tool that allows us to perform big data analytics using Hive query language and which is very similar to SQL.
iii)Apache Pig: Apache pig is used to analyze large data sets.
iv)Apache Spark: Spark is used for data processing Engine , that allows us to efficiently execute streaming or SQL workloads.
v)Apache HBase : HBase is NoSQL database that allows us to store unstructured and semi-structured data with real time read/write access.
vi)Apache Sqoop : Sqoop is used for transferring data in between Hadoop and relational data base servers.
vii)Apache Oozie: Apache Oozie is used to schedule hadoop jobs. It will integrate multiple jobs sequence- wise and it is integrated with Hadoop by using YARN
viii)Apache Zookeeper: Zookeeper is used for maintaining configuration like synchronization among all the tools present on Hadoop ecosystem. Apache Kafka is also used to manage configuration.
IX)Ambari : Ambari is used for managing and monitoring Apache Hadoop ecosystem.

V. CONCLUSION

In recent years data are generated at a dramatic pace. Analyzing these data is challenging for a general man. We survey the various research issues, challenges, and tools used to analyze these Big data. We have also tried to compare different platforms for addressing big data storage and tools for handling big data. Also different libraries and packages have been highlighted. From
this survey, it is understood that every big data platform has its individual focus. We believe that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

REFERENCES

1.Agneeswaran, V., (2012). Big-data - Theoretical, engineering and analytics perspective.
2. Srinivasa, S., Bhatnagar, V. (Eds.), Big Data Analytics. Vol. 7678 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 8-15.
3.https://en.wikipedia.org/wiki/Big_data
4.https://www.oracle.com/in/big-data/what-is-big-data/
5.https://www.bigdataframework.org/big-data-architecture/
6.https://www.sanfoundry.com/best-reference-books-big-data-analysis/
7.https://www.javatpoint.com/hadoop-tutorial
8.https://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm
9.https://www.google.com/search?q=mapreduceimage&oq=mapreduceimage&aqs=chrome..69i57.4359j0j7&sourceid=chrome&ie=UTF-8#imgrc=bbDU-80v_ZKI9M

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details