The Importance of Hadoop in Big Data Analytics (BDA)

Swapnil Raj

SOEIT, Sanskriti University, Mathura, Uttar Pradesh, India

Email Id- swapnil.cse@sanskriti.edu.in

ABSTRACT: ***The phrase Big Data has arisen in the age of technology, bringing with it unprecedented possibilities and difficulties for dealing with enormous amounts of data. Big Data has risen to prominence and is increasingly being used in emerging scholarly projects. We must evaluate the data in order to get valuable knowledge from huge amounts of data for businesses. To extract knowledge from unorganized datasets available online, such as texts, pictures, videos, or social media postings, expertise of data processing is needed. The research work provides an overview of Big Data, including its benefits and potential for future study. For academics, Big Data presents both possibilities and difficulties. An introduction of healthcare, technologies, and other possibilities is provided. This article provides an overview of Hadoop and its many aspects. The study also looks at how big data may be used in knowledge discovery.***

KEYWORDS: ***Big Data, Big Data Analytics, Hadoop, HDFS, MapReduce.***

1. INTRODUCTION

The introduction of innovations such as mobile computation, cloud services, the Internet of Things (IoT), sensor-dependent systems, and the reach of the World Wide Web (WWW) in portable devices has led in the production of enormous amounts of data, both organized and unorganized, dubbed Big Data [1]. Companies, institutions, and businesses are seeing the potential to organize this enormous amount of data into relevant and useful information. However, the problem concerning massive datasets is that it would be impossible to properly manage such huge amounts of data using conventional techniques. To deal with large data, software equipment, techniques, concepts, and methods are used. Hadoop, an open standard platform, is widely deployed to handle large amounts of data [2]. It is a well-known shared repository and computing platform for collecting and transferring large amounts of data.

Table 1: Illustrates the comparison between Big Data Analytics (BDA) and Traditional Analytics.

	Traditional Analytics	Big Data Analytics
Type of Analysis	Diagnostic and Descriptive analysis	Predictive and Prescriptive Analysis
Data Source	Limited data sets, cleaned data, simple models	Large scale data sets, variety of data like, structured/unstructured/semi-structured data, unprocessed data, Complex data model
Analytical Domain	What happened and why?	Gain new insights, find trends, hidden patterns, correlations

The contrasts among Big Data Analytics (BDA) and traditional analytics are shown in Table 1. BDA may be used on a variety of data sources, including textual, picture, views, logging, and blogging, to provide insights into user behavioural patterns, optimize efficiency, make wise commercial choices, forecast future assets, prevent illnesses, fight criminality, reduce deception, and mitigate vulnerabilities.

2. DISCUSSION

2.1. Big Data:

Big Data is a large accumulation of data produced at an accelerating pace in a multitude of types that has proven difficult to manage through conventional data management methods [3].

There are five V's in big data theory as stated below [4].

i. Volume: Users, companies, devices, and other entities produce a large amount of data every moment.
ii. Velocity: The pace with which the data is produced.
iii. Variety: Data is accessible in a variety of forms, including texts, blogging, twitter and Facebook posts, videos, barcodes, databases, and so on.
iv. Veracity: Data completeness and reliability.
v. Value: Revelations or knowledge that may be derived through the use of analytics on large amounts of data.

Businesses are becoming more interested in big data as a result of the potential benefit it can provide for their companies and projects. Businesses would like to grow, improve their commercial choices, and develop new goods and infrastructure, and big data may help them do so. The data contains significant information with huge quantities of data ranging from consumer purchasing patterns to Facebook posts. Effective data processing and analysis may offer insights in the long term, allowing companies to make more lucrative marketing choices or generate meaningful information.

2.2. Big Data Analytics (BDA):

BDA is the practice of deploying sophisticated analytical methods to huge, diverse datasets in order to uncover underlying knowledge which might assist professionals, companies, and academicians in formulating quicker and wiser choices [5]. Conventional analytics works with organized, transactional statistics stored in data repositories over time to conduct data analytics.

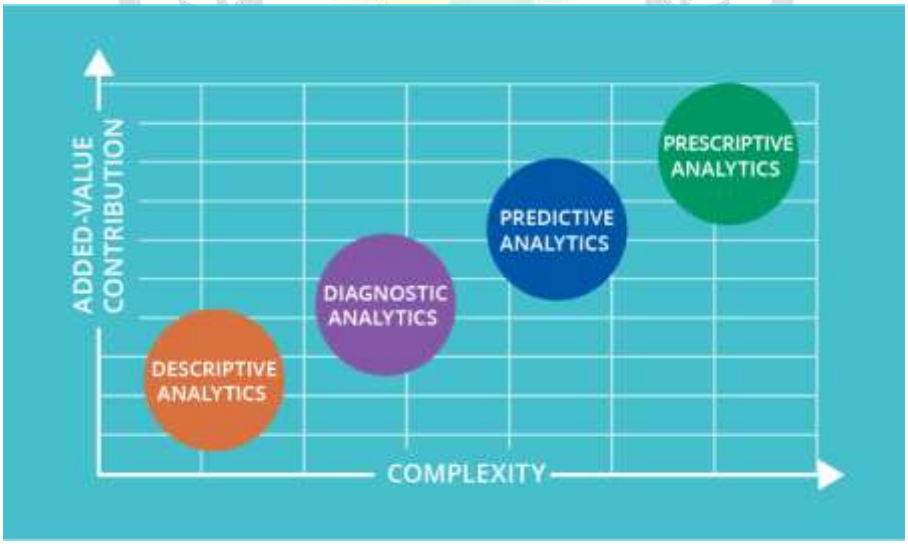

Figure 1: Depicts the types of Big Data Analytics (BDA).

A business intelligence analyst is responsible for identifying patterns, creating reports, and analyzing data visually. Data analysts, prognostic designers, as well as other analytical experts use BDA to analyze huge quantities of transactional and non-transactional data gathered from a variety of resources that are often overlooked by traditional data analytics tools. Web service log data, click streams records, social media posts, online activity reports, diagnosis health information, message from consumer email or text messages, feedback

forms, call logs, and data collected from sensing devices linked to the web of things are all examples of this type of data. Figure 1 depicts the many kinds of BDA.

Nowadays, proprietary and open source technologies such as BigInsights, Apache Storm, Hana, Lumify, RapidMiner, and others are capable of conducting numerous kinds of Big Data analyses. Hadoop is based on BDA and is a famous open source platform. Many businesses have developed big data systems on top of Hadoop, including IBM, Hortonworks, Cloudera etc.

2.3. Hadoop: An Overview:

Hadoop is an open platform developed by Apache for analysing and storage of large scale data volumes in a decentralized manner [6]. Google, Yahoo, LinkedIn, Facebook, Twitter, and a slew of other companies utilize it to improve customer engagement, receiver input, and develop innovative solutions and systems. Hadoop is well-known for its cloud-enabled design, which allows it to collect and analyze large amounts of information on computer clusters. It enables huge datasets to be processed over a cluster of computers in a distributed manner. A cluster is a collection of computers and other assets that work together to provide good scalability, synchronous replication, and parallelization. In a high performance computing context, a Hadoop cluster is a particular kind of computer cluster built especially for gathering and processing large quantities of unstructured data. It's built to expand from a central host to millions of nodes, each of which performs computing and storage locally.

Hadoop is not reliant on a large amount of hardware. At the higher layers, the Hadoop framework is intended to identify and manage problems. This ensures that the application is always accessible, although each unit is subject to error. Flexibility, cost-effectiveness, and dependability are the key advantages of utilizing Hadoop cluster in BDA. Hadoop divides the enormous data into manageable pieces and sends each piece to a separate node in the network for analysis. Hadoop relies on concurrency instead of relying on the efficiency of a given device. Hadoop easily manages growing data via virtualized environments and introducing more nodes to the network. Instead of engaging in strong, high-performance, and costly workstations, a Hadoop cluster may be built on desktop machines. Hadoop clusters are very failure resistant and robust to node and rack failures. Hadoop data is duplicated to many other clusters, ensuring that in the event of a node collapse, other multiple replicas are accessible. Once information is maintained in Hadoop, it is always accessible for assessment.

2.4. Hadoop 2.0 Basic Structure:

A Hadoop cluster is a collection of nodes i.e. host computers that are divided into racks [7]. The primary-replica architecture is used in the cluster. Name Nodes, a subset of nodes, serve as primary nodes. The Data Nodes on the other hand are the cluster's other nodes, which serve as replica nodes. The Hadoop's controller nodes are known as Name Nodes. They are in-charge of the whole storage service's namespace. The pieces of information in the records are stored in data or replica nodes. Hadoop 2.0 is made up of four major components: Hadoop Standard, HDFS 2, YARN, and MapReduce (MR). These modules are used in a variety of projects, including Pig and Hive. Hadoop 2.0's structure is shown in Figure 2.

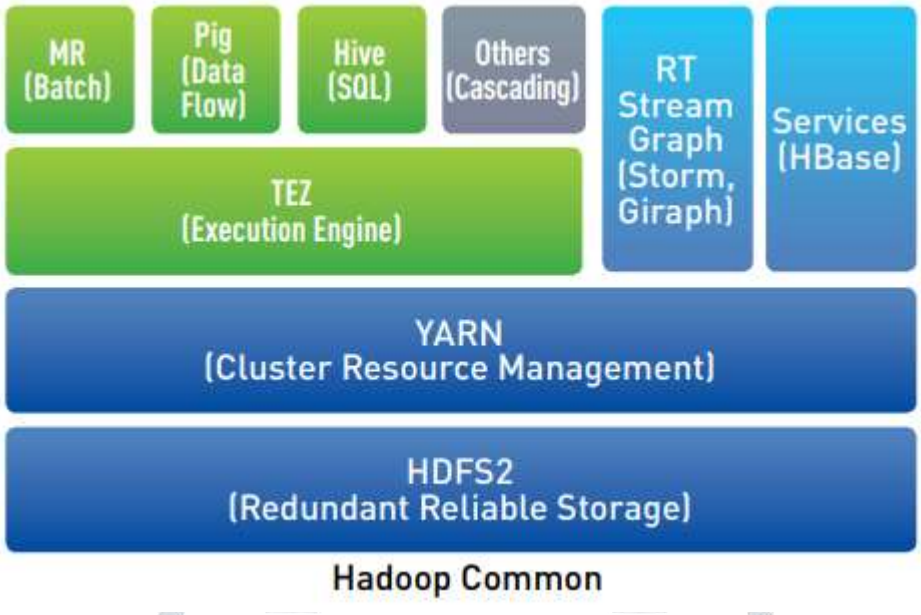

Figure 2: Illustrates the structure of Hadoop 2.0 [CSI].

The following are Hadoop 2.0's four major modules:

2.4.1. Hadoop Common:

It refers to the set of few Java libraries and tools required by other Hadoop modules.

2.4.2. HDFS2 (Hadoop Distributed File System):

A Java-based storage structure for scalability and reliability in information archiving, HDFS2 has a storage capacity of approximately 200 PB [8]. Further, to save a file system in HDFS2, it is first split into database files, which are subsequently stored in various Hadoop master nodes. The directories and information linked to the data structure are stored in name nodes. It also keeps track of file-to-block linkages and associated actual locations. Name Node may instruct Data Nodes to retrieve and fetch relevant data blocks. It is also in charge of block duplication over several Data Nodes. To ensure that it is operating properly, every storage node transmits a heartbeat and report to the Name Node on frequent basis. A block report is indeed a listing of all the blocks held by a Data Node. Block duplication with a preset replicating factor of three provides error detection in HDFS2. The key length and replicating factor are adjustable variables that may be changed to meet the needs of the system.

2.4.3. Hadoop YARN:

It is in charge of task scheduling, cluster planning and management, and monitoring [9]. It serves as a centralized system for ensuring that Hadoop 2.0 cluster management, privacy, and resource administration are standardized. YARN is a strategic planning system that includes a general processing system for stored data throughout a cluster. It can do batch computing with MapReduce, real-time analytics with Apache Storm, and graph analysis with Apache Graph.

2.4.4. Hadoop MapReduce:

It's a database paradigm for handling big data collections concurrently [10]. Individuals define a feature map as well as a reduction mechanism under this framework based on their needs. As an output, the Mapper function analyzes a key-value pair and returns a collection of arbitrary key-value pairs. To produce the final output, a reduction method takes the result of the Mapper function and combines all intermediate results linked with the very same key. Hadoop 2.0 is responsible for dividing the data input, allocating the system's implementation over a group of computers, dealing with hardware malfunction, and handling the necessary

cross-machine communications. Even developers with no prior expertise with concurrent or decentralized computing may take use of the capabilities of a Hadoop cluster.

2.5. Hadoop in Big Data Analytics (BDA):

Organisations that use Hadoop require a wide range of mathematical and statistical infrastructures and procedures to address their most pressing business questions. As well as the information upon which analytics ought to be conducted is often real-time or streaming content. When the data processing and monitoring needs are deterministic and also the customer is willing to wait for batch systems, MapReduce is the ideal choice. Spark is suggested if you need to perform analytical research on data streams, such as sensor information from a manufacturing floor, or deploy programs that need numerous actions on about the same set of data. Spark is free software used for cluster computing that is optimized for speed. It operates on top of a Hadoop framework and uses HDFS2 and HBase to extract content. As illustrated in Figure 3, this could even handle structured Hive data as well as semi structured data.

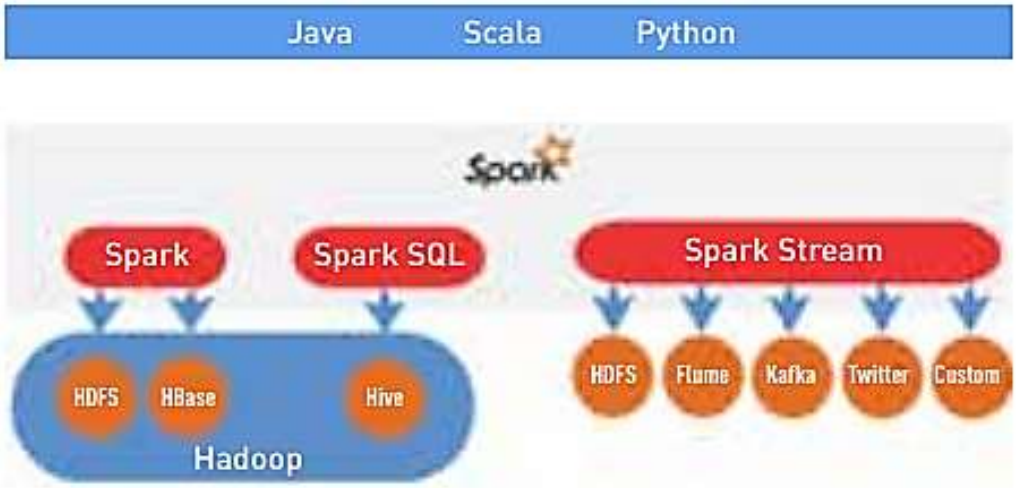

Figure 3: Illustrates the integration of Spark with a variety of data stores.

2.6. Implementation of Hadoop in BDA:

For BDA, Hadoop was being embraced by virtually every sector. Besides the usual benefits of boosting revenue, lowering expenses, expanding client base, and enhancing company effectiveness and productivity, it has been shown to offer certain unique benefits in several sectors. Following is a list of a few of the areas which have benefited from the usage of Hadoop in BDA.

2.6.1. Education:

BDA with Hadoop is often implemented in education sector to improve curriculum, increase student involvement, improve outcomes, provide career counselling, choose topics, and assess instructor effectiveness, among other things.

2.6.2. Healthcare:

BDA is widely utilized in individual medical records, DNA analysis for illness prevention and treatment, epidemic prediction, and improving overall living standards. DignityHealth Explorys etc. are analyzing the data and monitoring the health records of the youngsters to determine the appropriate medical strategy for them.

2.6.3. Retail:

Pattern analytics, media platforms evaluation, commodity basket evaluation, and product sentiment classification on Big data are being used by the general merchandise or industrial sector to improve customer

service, hold consumers, personalize purchasing, suggest products, optimize store layouts, anticipate requirement, and successfully promote products, among other things. Hadoop is used for BDA by retail behemoths like Walmart, as well as online behemoths like Amazon.

2.6.4. Entertainment:

BDA is being used by on-demand music and movie services to offer content-based suggestions and customize material for its customers in order to improve customer experience. Pandora makes music suggestions to its customers using Hadoop. Amazon does different kinds of analytics using Spark MapReduce, Hive, etc.

2.6.5. Banking:

BDA is used to identify, stop, and eradicate fraud using payment cards. Hadoop is used by Bank of America and JPMorgan to handle large quantities of data.

2.7. Challenges of Hadoop in BDA:

Despite its advantages, a data centre might not remain the ideal option for a company's data analytic needs. Enterprises having modest amounts of data may not benefit much from a Hadoop framework, even though intensive and sophisticated data processing is needed. Further disadvantage of the distributed system is that almost all of its mining methods are dependent on concurrent processes that operate on different clusters. Hadoop isn't always the best option if your research methodology does not fit in a concurrent computing scenario. The learning curve involved with building, running, and sustaining a Hadoop cluster is the most major impediment to utilizing it. It will be very challenging to conduct the necessary data processing unless companies have Hadoop specialists in their teams.

3. CONCLUSION

Hadoop is a prominent system for storing and analyzing large amounts of data. It offers decentralized computing and extensible storage space. Hadoop is used to deal with very big datasets that would be difficult to work with using conventional DBMS. The scale of the datasets exceeds the capacity of conventional computer programs and memory devices to collect, store, organize, and analyze the information in a reasonable amount of time. Apache Hadoop may be applied to gain information from data, which could also contribute to improved decisions and smart business actions. It does provide a framework under which IT firms may use advanced analytics techniques to offer product recommendations, website analysis, social analysis, and emotion assessment. BDA using Hadoop may help a company run more effectively, discover new possibilities, and gain a competitive edge. It allows you to experiment with new ideas while spending very little money.

REFERENCES

[1] Y. Wang, L. A. Kung, and T. A. Byrd, “Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations,” *Technol. Forecast. Soc. Change*, 2018, doi: 10.1016/j.techfore.2015.12.019.

[2] T. White, “Hadoop: The definitive guide 4th Edition,” *Online*, 2012, doi: citeulike-article-id:4882841.

[3] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” *Journal of King Saud University - Computer and Information Sciences*. 2018, doi: 10.1016/j.jksuci.2017.06.001.

[4] M. Viceconti, P. Hunter, and R. Hose, “Big Data, Big Knowledge: Big Data for Personalized Healthcare,” *IEEE J. Biomed. Heal. Informatics*, 2015, doi: 10.1109/JBHI.2015.2406883.

[5] L. Duan and Y. Xiong, “Big data analytics and business analytics,” *J. Manag. Anal.*, 2015, doi: 10.1080/23270012.2015.1020891.

[6] R. C. Taylor, “An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics,” *BMC Bioinformatics*, 2010, doi: 10.1186/1471-2105-11-S12-S1.

[7] R. C. M. Correia, G. Spadon, P. H. D. A. Gomes, D. M. Eler, R. E. Garcia, and C. O. Junior, “Hadoop cluster deployment: A methodological approach,” *Inf.*, 2018, doi: 10.3390/info9060131.

[8] M. R. Ghazi and D. Gangodkar, "Hadoop, mapreduce and HDFS: A developers perspective," 2015, doi: 10.1016/j.procs.2015.04.108.

[9] B. J. Mathiya and V. L. Desai, "Apache Hadoop Yarn Parameter configuration Challenges and Optimization," 2015, doi: 10.1109/ICSNS.2015.7292373.

[10] J. Dittrich and J. A. Quiané-Ruiz, "Efficient big data processing in Hadoop MapReduce," *Proc. VLDB Endow.*, 2012, doi: 10.14778/2367502.2367562.