E-ISSN No : 2581-6195 | Volume : 5 | Issue : 3 | May-June 2022

EXPLORING THE NEXUS OF BIG DATA AND HADOOP: AN IN-DEPTH REVIEW

Maitri Rajesh Gandhi

ABSTRACT

Within the field of data management and analytics, the convergence of Big Data and Hadoop represents a force that is capable of bringing about significant change. The purpose of this review paper is to investigate the complexities of this dynamic relationship by analyzing its consequences, difficulties, and possible applications. This article provides a complete overview of the landscape by synthesizing existing research and perspectives from industry practitioners. It also sheds light on key concepts, methodology, and developing trends in the industry. It elucidates the enormous influence that the integration of Big Data and Hadoop has had across a variety of industries by means of critical analysis and empirical evidence, so paving the way for improved decision-making, innovation, and competitive advantage.

KEYWORDS: Big Data; Hadoop; Map Reduce; Hdfs; Data Mining

Bhuj, Kachchh

HOW TO CITE THIS ARTICLE:
Maitri Rajesh Gandhi. (2022).Exploring the Nexus of Big Data and Hadoop: An In-Depth Review, International Educational Journal of Science and Engineering (IEJSE), Vol: 5, Issue: 3, 01-04

INTRODUCTION

The challenges of storing, processing, and extracting relevant insights from massive datasets are becoming increasingly difficult for enterprises to manage in our era, which is characterized by the exponential growth of data sources. The introduction of Big Data, which is distinguished by its volume, velocity, and variety, has ushered in a new era that is both complex and full of opportunities. During this time, Hadoop has evolved as a dominant framework for distributed storage and processing. It provides scalability, fault tolerance, and cost-effectiveness, among other benefits. At the junction of Big Data and Hadoop lies a rich area for investigation, which holds the potential to change several industries, propel innovation, and enable decision-making that is driven by data.

The goal of this research is to examine how Big Data and Hadoop interact with one another by reviewing the relevant literature and conducting empirical studies. The study's goals are to (1) assess the connection's efficacy and (2) identify research gaps by examining its most salient features. By bringing together theoretical frameworks, practical insights, and real-world implementations, this review aims to give a complete knowledge of the synergy between Hadoop and Big Data. This understanding will be used to influence strategic initiatives and to stimulate collaboration across disciplinary lines.
The concept of "big data" is imprecise, and there is no single definition that is universally accepted by all parties involved. The term "Big Data" is typically used to refer to information that is extremely huge in volume, originates from a wide variety of sources, originates in a wide variety of forms, and arrives at us with a high velocity. Large amounts of data can be structured, unstructured, or semi-structured, and these types of data cannot be processed using the traditional methods of data management. On the internet, data can be generated in a variety of formats, including words, photos, videos, and posts on social networking platforms. It is necessary to make use of parallelism in order to process these massive amounts of data in a manner that is both economical and effective [1].

The concept of large data may be broken down into four features. Volume, Velocity, Variety, and Veracity are the four components.

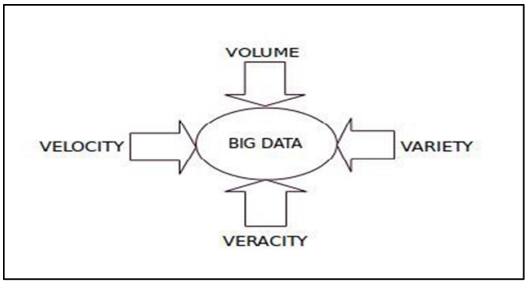

Fig. 1. 4 V's of Big Data.

The word "volume" refers to the amount of data generated every second or the scale of that data. One example of this type of data is data generated by computers. Data production is currently escalating from gigabytes to petabytes [2]. There will be thirty times as much data created in 2020 as there was in 2005, or 40 zettabytes [3]. Fast

Research Paper

processing of real-time sent data is the second defining feature of Big Data.

Data generation and processing rates are collectively known as velocity. For example, postings on several social media sites [2].

Another essential quality of big data is its wide range of different varieties. The type of data is what is meant by this. Data can be presented in a variety of formats, including text, numerical values, photos, audio, video, and data from social media platforms [2]. 400 million tweets are put out on Twitter every single day, and there are 200 million users that are actively using the platform [3].

Uncertainty or the accuracy of data is what veracity refers to. As a result of the inconsistency and incompleteness of the data, there is uncertainty [2].

2. CHALLENGES AND OPPORTUNITIES

In the vast expanse of the Internet, 800 million web pages offer details about Big Data. Data will be the next big thing after cloud [11]. Despite the many potential applications of big data in healthcare, academia, environmental protection, and industry, dealing with data sets that are so massively large becomes exceedingly difficult when using conventional methods. In order to effectively analyze data, it is necessary for us to consider the issues posed by large data and to develop certain computing models [13].

A. Challenges with Big Data: [12]

1) Heterogeneity and Incompleteness:

Even though Big Data can involve both structured and unstructured data, the former is preferable for evaluation purposes. But the data needs to be formatted for analysis. In the field of data analysis, heterogeneity is the most significant obstacle, and analysts are required to overcome it. Take, for instance, a patient who is situated in a hospital. For each and every medical examination, we shall create a record. In addition to that, we will also create a record of the hospital stay. The responses of each individual patient will be unique. The construction of this design is not very reliable. Consequently, it is necessary to deal with heterogeneous and imperfect information. An effective data analysis ought to be applied to this situation.

2) Scale:

The term "Big Data" refers to the presence of massive data sets, as the name suggests. For decades, one of the most significant challenges has been managing massive data volumes. Once upon a time, this issue was resolved by the processors becoming more efficient; however, the data volumes are now increasing enormous, and the processors are remaining unchanged.

Data is being produced at an extremely rapid rate as a result of the change that is occurring in the world toward the technology of the cloud. For the data analysts, this rapid increase in the amount of data is becoming an increasingly difficult problem to solve.

In order to store the data, hard disks are utilized. Their I/O performance is slower than average. On the other hand, solid-state drives and other technologies have mostly supplanted hard disks in recent years. Therefore, a new storage system ought to be built because these are not operating at a slower rate than hard drives.

3) Timeliness:

Speed is yet another difficulty that comes with stature. The amount of time required to evaluate the data will increase proportionately with the size of the data sets. In terms of speed, any system that can efficiently handle the size should do really well. Nevertheless, there are instances where we need the analysis's results immediately. For instance, it is important to analyze any fraudulent transactions before they are finalized. Consequently, this data processing problem necessitates the creation of a new system.

4) Privacy:

Data privacy is yet another significant issue that arises with big data. In certain nations, there are stringent rules that govern the privacy of data; for instance, in the United States of America, there are stringent laws that govern health records. However, in other nations, the laws are not as stringent. We are unable to obtain the private posts of users in social media platforms, for instance, in order to do sentiment analysis.

5) Human Collaborations:

There are a great number of patterns that a computer is unable to recognize, despite the fact that there are powerful computational models. A novel approach to utilizing human intellect to find solutions to problems is known as crowd-sourcing. For the best illustration, see Wikipedia. Even while we can trust the strangers' word for it, most of the time what they say is wrong. But there are other people out there who might be giving incorrect information or misleading information for other purposes. To be able to deal with this, we require a technological model. As human beings, we are able to read reviews of books and determine that some of them are favorable while others are negative. Based on this information, we may then decide whether or not to purchase the book. In order for us to make decisions, we require system intelligence.

B. Opportunities to Big Data: [14]

Now we are in the midst of the Data Revolution. Businesses are being presented with a multitude of chances to expand their operations and achieve higher levels of profitability as a result of Big Data. Nevertheless, big data is not just making a significant impact in the computer industry; it is also permeating every sector of society, from healthcare to finance to politics.

1) Technology:

Facebook, IBM, and Yahoo are just a few of the highly successful organizations that have invested in and successfully used big data. Facebook oversees fifty billion photographs of individuals. Every every month, Google handles 100 billion searches. From these numbers, we may infer that the internet and social media present a plethora of chances.

2) Government:
The challenges that the government is currently facing can be addressed with the use of big data. A research and development project for big data was announced by the Obama administration in the year 2012. The Bharatiya Janata Party (BJP) used big data analysis to great effect in the 2014 elections, and the present government in India is doing the same thing with the Indian electorate.

3) Healthcare:
Unstructured data makes up eighty percent of all medical information, as stated by IBM Big Data for Healthcare. There is a growing trend among healthcare institutions to adopt big data technology in order to obtain comprehensive information about patients. In order to enhance healthcare and reduce costs, it is necessary to do statistical analysis on large amounts of data and to adapt particular technologies.

4) Science and Research:
Big data is one of the newest fields of study. Massive datasets are the subject of current investigations. Many different types of articles are covering the issue of big data right now. At the NASA center for climate simulation, the data is stored in a system with a capacity of 32 petabytes [15].

5) Media:
The media is capitalizing on consumers' interests in online content through the use of big data in product advertising. Data analysts, for example, will look at a user's interest level once they've counted the number of postings on social media. Another option is to ask for people's thoughts on social media, whether they're good or negative.

3.. HADOOP FRAMEWORK

In order to process large amounts of data, Hadoop is a piece of open-source software. It is widely utilized by companies and researchers for the purpose of doing Big Data analysis. The architecture of Google, the Google File System, and MapReduce are all factors that have affected Hadoop. Within a distributed computing system, Hadoop is responsible for processing the enormous data sets. The Hadoop Kernel, MapReduce, and HDFS are the components that make up an Apache Hadoop ecosystem. Other components include Apache Hive, Base, and Zookeeper. (1) [1]

A. Hadoop consists of two main components:

1) Storage: The Hadoop Distributed File System (HDFS): A distributed file system with fault tolerance, designed to run on commodity hardware, is what this system is all about. When it comes to applications that have enormous data sets, HDFS is an excellent choice because it offers high throughput access to the data of the application. Through the use of thousands of servers, HDFS is able to store data. The master/slave design is utilized by HDFS [5]. When files are added to HDFS, they are partitioned into blocks of a predetermined size. The block size can be customized, but it is always set to 64 megabytes by default.

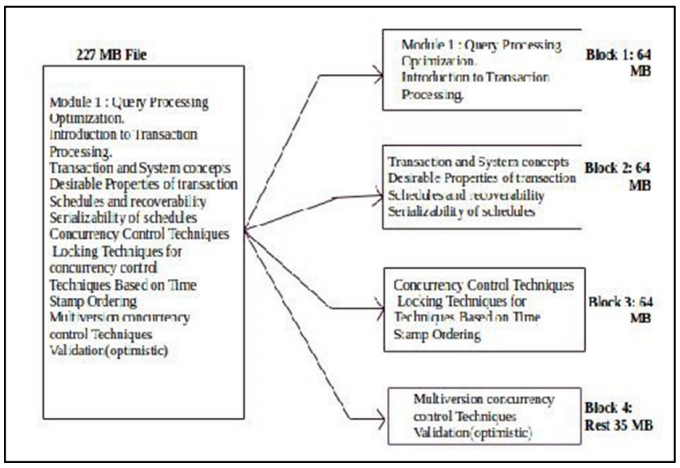

Fig. 2. HDFS Blocks.

2) Processing: MapReduce [4]: Google created this programming style in 2004 to facilitate the development of fault-tolerant applications capable of processing large datasets in parallel on large hardware clusters. The operation is executed in parallel on a large data set once the problem and data sets are divided.

Two functions that are part of MapReduce are:

1. **Map:** It is customary to begin with the Map function, which is often utilized to filter, alter, or parse the data initially. When Reduce is applied, the output from Map is used as the input.
2. **Reduce:** In most cases, the Reduce function (which is not required) is utilized to summarize the data obtained via the Map function.

4. APPLICATIONS IN DATA MINING

Businesses and academics alike can benefit greatly from big data since it allows them to search through massive datasets in search of trends. The phrase "data mining" describes the steps used to glean useful insights from large datasets. There is an abundance of data available online, including text, figures, social media posts, images, and videos. There will be thirty times as much data created in 2020 as there was in 2005, or 40 zettabytes [3]. We need to establish a new, effective data mining system to assess this data and extract information that could be useful for many things, like education, health, and security. Data mining techniques can be utilized with large amounts of data in a variety of ways, some of which include the following:

A. Classification Analysis:

Through the use of a methodical approach, it is possible to acquire significant information on data and metadata. Additionally, classification might be utilized in order to cluster the data.

B. Cluster Analysis:

Finding data sets that are comparable to one another is the technique that is being referred to here. In order to determine the similarities and differences that exist within the data, this is done. On social media, for instance, it is possible to target groups of clients who share similar preferences [6].

C. Evolution Analysis:
Information extraction from DNA sequences is the main goal of the procedure, which is sometimes called genetic data mining. But financial institutions can use it to predict stock market movements using historical time series data [7].

D. Outlier Analysis:
In a data set, there are some observations and identifications of things that are carried out, but they do not constitute a pattern. In the context of financial and medical issues, this is utilized.

5. LITERATURE REVIEWS

After conducting research on Indian recipes, Anupam Jain, Rakhi N K, and Ganesh Bagler came to the realization that the inclusion of particular spices significantly reduces the likelihood that a dish will have components that share flavors. As part of their investigation, Jain and a few other individuals selected the website TarlaDalaa.com and downloaded more than two thousand five hundred recipes. Over the course of these recipes, 194 distinct components were discovered. After that, they investigated the network of connections that existed between these recipes. They came to the conclusion that Indian cuisine is distinguished by a remarkable negative food pairing that is even more pronounced than any other cuisine. In their words, "Our study reveals that spices occupy a unique position in the ingredient composition of Indian cuisine and play a major role in defining its characteristic profile." This is the conclusion that can be drawn from their findings. In conclusion, Jain and his colleagues state that their research has the potential to result in the development of methodologies for the creation of innovative Indian hallmark dishes, healthy recipe modifications, and recipe recommender systems. It is [8,9].

An investigation of Big Data and the Hadoop system was carried out by Vidyasagar S. D., who discovered that businesses require the ability to analyze and manage petabytes of data sets in a manner that is both efficient and cost-effective. It is possible that we will lose some information if there is a failure of any of the nodes, according to him. Using the Apache License, Hadoop is an efficient and dependable open-source software. Hadoop is utilized for the purpose of managing big data sets. The author elaborated on its requirements, applications, and uses. At the present time, Hadoop is playing a significant part in the Big Data movement. At the end of the day, Vidyasagar S.D. came to the conclusion that "Hadoop is designed to run on cheap commodity hardware, it automatically handles data replication and node failure, it does the hard work – you can focus on processing data, Cost Savings, and efficient and reliable data processing" [10].

6. CONCLUSION

The intersection of Big Data and Hadoop is that which is a fundamental component of contemporary data management and analytics. Hadoop has made it possible for enterprises to derive value from their data assets by democratizing access to Big Data technology. This is accomplished through Hadoop's distributed architecture, parallel processing capabilities, and ecosystem of tools. On the other hand, this convergence is not without its difficulties, which include worries about the privacy of data, issues about governance, and concerns about data security. In order to successfully traverse this complicated environment, it is absolutely necessary for stakeholders to solve these problems while also capitalizing on the revolutionary potential of Big Data and Hadoop combinations.

When looking to the future, it is necessary to conduct additional study in order to investigate upcoming trends within the Big Data and Hadoop ecosystem. Some examples of these trends include the integration of machine learning, real-time analytics, and edge computing. Through the promotion of collaboration between the academic world and the business world, we can propel innovation, address societal concerns, and unlock the full potential of insights that are powered by data. At the end of the day, the intersection of Big Data and Hadoop holds the potential to propel long-term growth, encourage digital transformation, and mold the future of decision-making that is driven by data.

REFERENCE

1. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.
2. SMITHA T, V. Suresh Kumar "Application of Big Data in Data Mining" in International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 7, July 2013).
3. IBM Big Data analytics HUB, www.ibmbigdatahub.com/infographic/four-vs-big-data
4. Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N "Analysis of Bidgata using Apache Hadoop and Map Reduce" in International Journal of Advance Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.
5. Apache Hadoop Project, http://hadoop.apache.org/, 2013.
6. Smitha.T, Dr.V. Sundaram, "Classification Rules by Decision Tree for disease prediction" International journal for computer Application, (IJCA) vol 43, 8, No-8, April 2012 edition. ISSN0975- 8887; pp- 35-37
7. Mucherino A. Petraq papajorgji P. M. Paradalos 1998. A survey of data mining techniques alied to agriculture CRPIT.3(3): 555560.
8. Anupam Jain, Rakhi N K and Ganesh Bagler, arxiv.org/abs/1502.03815 Spices Form The Basis Of Food Pairing In Indian Cuisine.
9. MIT Technology Review, http://www.technologyreview.com/view/535451/data-mining- indian-recipes-reveals-new-food-pairing-phenomenon/.
10. Vidyasagar S. D, A Study on "Role of Hadoop in Information Technology era", GRA - GLOBAL RESEARCH ANALYSIS, Volume: 2 | Issue: 2 | Feb 2013 • ISSN No 2277 – 8160.
11. BIG DATA: Challenges and opportunities, Infosys Lab Briefings, Vol 11 No 1, 2013.
12. Divyakant Agrawal, Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the United States.
13. Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challenges and Solutions in International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.
14. Big Data, Wikipedia, http://en.wikipedia.org/wiki/Big_data Webster, Phil. "Supercomputing the Climate: NASA's Big Data Mission". CSC World. Computer Sciences Corporation. Retrieved 2013-01-18.