e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

International Journal Of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704094 |

Tools and Technologies for Big Data Analysis with Hadoop: A Review

Vaibhav Dashora[1], Dr. Vishal Shrivastava[2], Dr. Akhil Pandey[3]

B.Tech. Scholar, Department of Computer Science & Engineering, Arya College of Engineering & I.T., Jaipur, India [1]

Professor, Department of Computer Science & Engineering, Arya College of Engineering & I.T., Jaipur, India [2,3]

ABSTRACT: The term 'big data' refers to a substantial amount of data that may exist in organized, semi-structured, or unstructured formats. The generation of Big Data occurs at a substantial scale, posing challenges for processing using conventional methodologies. The conventional big data management system is inadequate for efficiently handling the substantial volume of data created. Hadoop is a framework that has been specifically developed to handle and analyse extensive datasets, offering exceptional speed and fault tolerance over a wide range of computing resources, ranging from individual servers to large-scale clusters including thousands of workstations. This article presents a comprehensive examination of big data and Hadoop, including an extensive analysis of several tools and technologies associated with big data.

I. INTRODUCTION

The exponential growth in the volume of unstructured data being produced every day has led to the widespread use of the term "big data" in the 21st century. Data that conforms to a predetermined standard and is easily represented as a grid of columns and rows is considered structured. Structured data includes things like ERP and CRM systems. There is no predetermined organization for unstructured data. Audio, video, picture, etc. are the example of unstructured data. Semi-structured data is a kind of information that combines aspects of both unstructured and structured data. XML, emails, JSON, etc., are all instances of semi-structured data.

1.1 The 6 Vs are used to define big data: -

First, Petabytes and zettabytes are used to describe the massive amounts of data generated by industries like healthcare, education, and finance.

Second, Data may be organized, unstructured, or semi-structured, and this is what we mean by "variety."

Third, velocity describes the pace at which huge data is produced. The daily data output from telecommunications is 35 terabytes.

Fourth, data must be reliable, which means defining what constitutes a bias, a noise level, or an outlier.

Fifth, Infinite data that can be transformed into useful information is valuable.

The sixth concept is "variability," which describes information whose interpretation is in flux.

1.2. Big-Data Origins Many different entities provide data for big data analysis, but the three most common are:
1.Community groups.

2. Commercial model.

3.Internet of Things.

These datasets may include any mix of organized, semi-structured, and unstructured information. Social media platforms like Twitter, Facebook etc., provide data gathered by real people. Commercial transactions, electronic commerce, bank records, credit cards, medical records, and the Internet of Things. are all examples of services and

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704094 |

goods provided to clients via conventional business systems. Safety, pictures, and footage from surveillance cameras Satellite pictures, computer data (server logs, website statistics, etc.

1.3. Big-Data Use Cases: -

Many industries, including e-health, the Internet of Things (IoT), transportation and logistics, may benefit from big data.

E-Health: The Big Data Health System is Crucial to Individual and Community Health and Wellness in Both the Medical and Public Sectors. The Internet of Things (IoT) is a system that combines big data sensing networks with artificial technologies to provide a finished product or service in the event of a catastrophic collapse.

Transportation and logistics: When big data and real-time processes are taken into consideration, transportation and logistics companies are able to deploy maintenance personnel in advance of a critical failure.

1.4. Restrictions in Current Infrastructure: -

There are significant shortcomings in the current systems that hinder their widespread implementation. These are the restrictions:

1. Dishonesty
2. Availability and consistency issues
3. Inaccuracy
4. Existing systems have vertical scalability, which is a key feature.
5. New methods and software are part of Big Data. The current methods of big data management are inadequate for processing such massive amounts of data.

It processes the mountain of information. To address the shortcomings of the standard large data management infrastructure, developers have turned to Hadoop.

II. BIG DATA

Why Hadoop?

Big Data poses challenges, including the storage, computation and analysis of amounts of data. In the past limited technology hindered corporations, from managing data. However with the advent of Hadoops framework for handling large scale data decision making has greatly improved. Hadoop has emerged as an scalable solution for addressing these challenges. Its versatility, scalability, performance and affordability have contributed to its popularity in the realm of Big Data.

An essential component of Hadoop is MapReduce, a programming system that enables cost processing and analysis of datasets. Within Hadoops data analytics ecosystem exist functionalities for storage, processing, access control, administration and privacy management. The hierarchical structure of HDFS (Hadoop Distributed File System) consists of a master name node for managing the file system namespace and directories in clusters.

The data node within an HDFS cluster comprises two files; one providing the data while the other serves as a blocking generation stamp. Hardware used in conjunction with HDFS is both fault tolerant and cost effective. Each file within HDFS undergoes authentication and authorization processes to ensure security.

YARN (Yet Another Resource Negotiator) facilitates processing capabilities by enabling graph processing, streaming operations batch tasks alongside yarn processing. In MapReduce – a part of Hadoop – mapping involves breaking down data into parts for translation purposes while reducing combines map outputs, into key value pairs to produce concise tuples.

Hadoop Common encompasses directories and libraries that further enhance its functionality.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704094 |

III. CONCLUSION

We have conducted research on both Big Data and Hadoop. In this post, we will explore the concepts of Big Data and Hadoop, as well as the six V's and several big data technologies. Conventional big data management systems possess several limitations that hinder their ability to effectively process vast quantities of data. These deficiencies encompass imprecision, lack of honesty, absence of confidentiality, and similar issues. There have been some significant difficulties associated with the conventional large-scale data management method. We utilize Hadoop to effectively manage large data collections, hence mitigating the associated challenges. Hadoop has exceptional scalability, enabling us to easily expand our infrastructure from a single server to a large cluster of thousands of computers, as required. The technology called Map Reduce was developed.

Google initially developed this technology to internally handle extensive data harvests. This article has discussed many big data technologies capable of handling diverse data kinds. Apache Spark is an alternative tool utilized for analyzing large-scale data. While Hadoop is known for its ability to store and analyze massive datasets, it is not the sole technology that provides faster performance than MapReduce.

REFRENCES

1] Kaur, Iqbal Deep, 2. K. (2016). Write a paper about Big Data and Hadoop. International Journal of Technology and Computer Science, 4.
[2] P. Acharya and D. P. Acharya (2016). A look at the problems, unanswered questions, and tools in big data analytics. IJACSA, which stands for the International Journal of Advanced Computer Science and Applications, 8.
[3] In 2018, Farhan, M. N., and Ali, M. A. A study and comparison of how well Map Reduce and Apache Spark work with Twitter data on a Hadoop system. https://www.mecs-press.org/, 10.
[4] J. D. Ghemawat (2004). MapReduce makes it easier to work with large amounts of data. This is Google, Inc.
5]: Glouchkov, D., Jovanovic, P., & Abelló, A. (2017). Performance Models for MapReduce in Hadoop 2.x. ICDT/EDBT, 10.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com