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 12, Issue 5, May 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

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: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205468 |

A Review on Bigdata Using Hadoop Framework

Ms. Shweta Arjun[1] , Dr.Pradnya Muley[2]

MCA Department, PES Modern College of Engineering, Pune, India[1,2]

ABSTRACT:-Big Data and Hadoop are two interconnected technologies that have revolutionized the way organizations handle and process vast amounts of data. Big Data refers to the massive volume, variety, and velocity of data being generated in today's digital age, which traditional data processing tools are unable to effectively manage. Hadoop, on the other hand, is an open-source framework designed to store, process, and analyze large datasets across distributed computing clusters.[1]

Additionally, the abstract highlights the advantages of using Hadoop for Big Data processing, such as scalability, fault tolerance, and cost-effectiveness. It emphasizes the ability of Hadoop to handle structured, semi-structured, and unstructured data from diverse sources, enabling organizations to derive valuable insights and make data-driven decisions.

Finally, the abstract briefly touches upon some real-world applications of Big Data and Hadoop, including customer analytics, fraud detection, recommendation systems, and scientific research. It concludes by emphasizing the growing significance of Big Data and Hadoop in various industries and the need for skilled professionals to leverage these technologies effectively[2]

I. INTRODUCTION

Big data Hadoop is a powerful and widely used framework for processing and analyzing large volumes of data. It is designed to handle massive amounts of information that cannot be efficiently processed by traditional data processing systems. Hadoop enables the distributed processing of large datasets across clusters of computers, providing scalability, fault tolerance, and high throughput.

At its core, Hadoop consists of two main components: the Hadoop Distributed File System (HDFS) and the MapReduce programming model. HDFS is a distributed file system that stores data across multiple machines, allowing for data replication and fault tolerance. It breaks down large files into smaller blocks and distributes them across the cluster.[3]

MapReduce is a programming model that allows for parallel processing of data across the Hadoop cluster. It divides data processing tasks into two stages: the map stage and the reduce stage. The map stage involves processing data in parallel across multiple nodes to generate intermediate results, while the reduce stage aggregates and summarizes those intermediate results to produce the final output.

Hadoop also provides a range of ecosystem tools and frameworks that enhance its capabilities. For example, Apache Hive provides a SQL-like query language for data analysis, Apache Pig offers a high-level scripting language, and Apache Spark enables faster data processing and analytics. These tools make it easier to work with big data and perform various tasks, such as data ingestion, transformation, analysis, and visualization.

The key advantages of Hadoop include its ability to handle large-scale data processing, its fault tolerance through data replication, its scalability by adding more nodes to the cluster, and its cost-effectiveness compared to traditional data processing systems. It is particularly useful for processing unstructured and semi-structured data, such as social media posts, log files, sensor data, and text documents[4]

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205468 |

II. LITERATURE SERVAY

Big data and Hadoop have significantly transformed the field of data processing and analytics. Here's a brief literature survey highlighting key aspects and references related to Big Data and Hadoop:

1.Hadoop:

The Definitive Guide" by Tom White (2015): This book provides a comprehensive introduction to Apache Hadoop and covers its architecture, components, and various ecosystem tools. It also discusses best practices for developing and deploying Hadoop applications.[5]

2.Big Data:

A Revolution That Will Transform How We Live, Work, and Think" by Viktor Mayer-Schönberger and Kenneth Cukier (2013): This book explores the potential of big data and its impact on various industries. It discusses the challenges and opportunities associated with analyzing large datasets and provides insights into the implications for society and business[6]

3.Hadoop Operations:

A Guide for Developers and Administrators" by Eric Sammer (2012): This book focuses on the operational aspects of running Hadoop clusters. It covers topics such as cluster planning, installation, configuration, and troubleshooting. It also provides guidance on monitoring, security, and performance tuning.

4.Pro Hadoop:

by Jason Venner (2012): This book delves into the advanced concepts and techniques of Apache Hadoop. It covers topics such as Hadoop's distributed file system (HDFS), MapReduce programming, and data warehousing with Hive.

5.Hadoop in Practice:

by Alex Holmes (2014): This book offers a collection of practical examples and techniques for working with Hadoop. [7]

III. HOW TECHNOLOGY WORKS:

1.Define your problem: Determine the specific business problem or use case you want to address using big data analytics. This will help guide your data collection and analysis efforts.

2.Data collection: Gather the relevant data from various sources, including structured and unstructured data. This may involve data extraction, transformation, and loading (ETL) processes to ensure the data is in a suitable format for analysis.[8]

3.Setting up Hadoop: Install and configure a Hadoop cluster. Hadoop is an open-source framework that allows for distributed processing and storage of large datasets across a cluster of computers.

4.Data ingestion: Transfer the collected data into the Hadoop cluster. You can use various methods for data ingestion, such as using Hadoop's command-line interface (CLI), Hadoop APIs, or third-party tools like Apache NiFi.

5.Data storage: Store the data in Hadoop's distributed file system, HDFS. HDFS divides the data into blocks and replicates them across multiple nodes in the cluster to ensure fault tolerance and high availability.

6.Data processing: Use the MapReduce framework or higher-level processing tools built on top of Hadoop, such as Apache Spark or Apache Hive, to process and analyze the data.

7.Working with Hadoop

Big data Hadoop technology is a powerful framework designed to handle and process large volumes of data across distributed computing clusters.[9]

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205468 |

IV. METHODOLOGY

Hadoop is an open-source framework used for distributed storage and processing of large datasets. It provides a reliable, scalable, and fault-tolerant platform to handle big data. The methodology of working with Big Data using Hadoop typically involves the following steps:

1.Data Acquisition:

The first step is to collect the data from various sources such as databases, web applications, sensor networks, social media, etc. The data can be structured, semi-structured, or unstructured.

2.Data Preparation:

Once the data is acquired, it needs to be cleaned and transformed into a suitable format for analysis.

3.Data Storage:

Hadoop uses a distributed file system called Hadoop Distributed File System (HDFS) to store large datasets across multiple machines. [10]

4.Data Processing:

Hadoop provides a programming model called MapReduce for distributed processing of data. MapReduce involves two main steps: Map and Reduce.

5.Data Analysis:

After processing the data, the results are analyzed to extract meaningful insights. This can involve tasks like statistical analysis, data mining, machine learning, and visualization.

6.Data Presentation:

The analyzed data is presented in a meaningful way to end-users or stakeholders. This can include reports, dashboards, charts, or any other form of data visualization.

Here's a step-by-step methodology for working with Big Data using Hadoop:

1)Define the Problem: Clearly define the problem you want to solve using Big Data and identify the specific goals and objectives.[11]

2)Data Collection: Identify the data sources relevant to your problem and collect the necessary data.

3) Data Extraction and Transformation: Cleanse and preprocess the collected data to remove any inconsistencies, errors, or irrelevant information.

4)Data Storage: Determine the appropriate data storage infrastructure for your needs. Hadoop Distributed File System (HDFS) is commonly used in Hadoop environments for scalable and fault-tolerant storage.

5)Data Processing: Use Hadoop's MapReduce paradigm or higher-level processing frameworks like Apache Spark to perform distributed processing on the data. This step involves breaking down the processing tasks into smaller units and distributing them across a cluster of computers.[12]

6)Data Analysis: Apply various analytical techniques and algorithms to derive meaningful insights from the processed data. This could include tasks such as data mining, machine learning, statistical analysis, or graph analysis, depending on your problem and objectives.

7)Data Visualization: Present the analyzed data in a visually appealing and understandable format. Use visualization tools like Tableau, Power BI, or matplotlib to create charts, graphs, or dashboards that effectively communicate the insights obtained from the data.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205468 |

1.Install Hadoop: Begin by installing Hadoop on each site where you want to set up your cluster.

2.Configure Hadoop: Once installed, you need to configure Hadoop on each site. The configuration files you need to modify are typically found in the etc/hadoop directory of your Hadoop Installation.

3.Set up NameNode: Identify one site as the primary NameNode, which will store the metadata about the Hadoop cluster. Configure the hdfs-site.xml file on the primary NameNode to specify the location of the NameNode data directory.[13]

4.Set up DataNodes: Each site should have DataNodes to store and process data locally. Configure the hdfs-site.xml file on each DataNode to specify the location of the DataNode data directory and the IP address or hostname of the primary NameNode.[14]

5.Set up ResourceManager: Identify one site as the primary ResourceManager, responsible for managing resources and scheduling tasks in the cluster.

6.Set up NodeManagers: Each site should have NodeManagers to manage resources and execute tasks locally.

7.Configure network and connectivity: Ensure that all sites are networked together and can communicate with each other. Verify that the necessary ports are open for inter-site communication.[15]

8.Distribute Hadoop configuration: Copy the modified configuration files from the primary NameNode and primary ResourceManager to all other nodes in the cluster, ensuring consistency across the sites.

9.Start Hadoop services: Start the Hadoop services, including the NameNode, DataNodes, ResourceManager, and NodeManagers, on each site.

10.Submit and run jobs: With the Hadoop cluster set up, you can now submit big data jobs for processing. Use the Hadoop command-line tools, such as hdfs and yarn, to interact with the distributed file system and submit jobs to the cluster.

11.Monitor and manage the cluster: Continuously monitor the health and performance of your Hadoop cluster. Use tools like the Hadoop Resource Manager web UI and monitoring tools like Ganglia or Ambari to monitor various aspects of the cluster, such as resource utilization, job status, and overall performance.

By following these steps, you can effectively set up and use Hadoop for big data processing across multiple sites[16].

V. CHALLENGES

Big data Hadoop is a powerful framework for processing and analyzing large volumes of data. While it offers numerous benefits, it also poses several challenges. Some of the common challenges associated with Big data Hadoop include:

1)Data Variety: Big data often comes in various formats, such as structured, semi-structured, and unstructured data.

2)Data Volume: Big data implies dealing with enormous volumes of data that exceed the capacity of traditional data processing systems. Storing and processing such massive amounts of data requires a scalable and distributed infrastructure, which Hadoop provides.

3)Data Velocity: The speed at which data is generated and needs to be processed has increased exponentially. Real-time and near-real-time data processing requires low-latency systems, but Hadoop's batch-oriented processing approach may not be suitable for real-time analytics[17].

4)Fault Tolerance: Hadoop's distributed architecture ensures fault tolerance by replicating data across multiple nodes.

5)Skill Gap: Hadoop and its ecosystem consist of several components like HDFS, MapReduce, Hive, Pig, etc. These technologies often require specialized skills and knowledge. Finding professionals with expertise in Hadoop and related tools can be challenging, limiting the availability of skilled resources.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205468 |

6.Security: Big data often contains sensitive information, and ensuring data security and privacy is a significant challenge. [18]

7.Scalability: As data volumes continue to grow, scaling Hadoop clusters to accommodate the increased load becomes crucial. Adding more nodes and managing the cluster's performance, resource allocation, and data distribution require careful planning and monitoring to ensure efficient scalability.

8)Data Quality: Big data sets may contain inconsistencies, errors, or missing values. Ensuring data quality and reliability is challenging, as Hadoop primarily focuses on processing large volumes of data rather than data quality management. [19]

VI. APPLICATIONS

Big data Hadoop is a powerful framework for storing, processing, and analyzing large volumes of structured and unstructured data. Here are some common applications of Big Data Hadoop:

1.Data Warehousing: Hadoop can be used as a cost-effective alternative to traditional data warehousing solutions. It can store vast amounts of structured and unstructured data, enabling businesses to perform complex analytics and generate reports.

2.Log Analysis: Hadoop is widely used for log analysis in various industries such as IT, cybersecurity, and telecommunications. It can process and analyze log files from different sources, helping organizations identify patterns, detect anomalies, and troubleshoot issues.

3.Recommendation Systems: Hadoop facilitates the implementation of recommendation systems by processing large datasets and generating personalized recommendations for users.

4.Fraud Detection: Hadoop can be leveraged to detect fraudulent activities by analyzing large volumes of transactional data in real-time. [20]

5.Sentiment Analysis: Hadoop is employed for sentiment analysis, which involves analyzing text data (such as social media posts, customer reviews, and surveys) to determine the sentiment expressed.

6.Internet of Things (IoT): With the rise of IoT devices generating massive amounts of data, Hadoop provides a robust infrastructure for storing and processing IoT data.

7.Genomics and Bioinformatics: Hadoop is used in genomics and bioinformatics research to handle large-scale DNA sequencing and analysis. [21]

VII. ADVANTAGES

Big Data and Hadoop offer several advantages for organizations dealing with large volumes of data. Here are some key advantages:

1)Scalability:

Hadoop enables the distributed processing of large datasets across a cluster of computers. It allows organizations to scale their data storage and processing capabilities easily by adding more nodes to the cluster.

2)Cost-effective:

Hadoop runs on commodity hardware, which is significantly less expensive compared to traditional high-end servers.[22]

3)Fault tolerance:

Hadoop has built-in fault tolerance mechanisms that ensure data reliability and availability. It automatically replicates data across multiple nodes in the cluster, allowing the system to continue functioning even if some nodes fail.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205468 |

4)Flexibility:

Hadoop can handle a wide variety of data types, including structured, semi-structured, and unstructured data. It is not limited to a specific data format, making it suitable for diverse data sources such as social media data, log files, sensor data, etc.

5)Processing speed:

Hadoop's distributed computing model enables parallel processing, which significantly speeds up data processing tasks. It divides the data into smaller chunks and processes them simultaneously across multiple nodes, reducing the overall processing time. [23]

6)Scalable storage:

Hadoop's Hadoop Distributed File System (HDFS) provides a scalable and distributed storage solution. It allows organizations to store vast amounts of data across multiple nodes in the cluster. As the data volume grows, additional nodes can be added to the cluster to accommodate the increased storage requirements.

7)Data processing capabilities:

Hadoop provides powerful tools and frameworks, such as MapReduce and Apache Spark, which enable efficient data processing and analysis.

8)Data locality:

Hadoop's data locality principle aims to minimize data movement across the network during processing. It brings the computation close to the data, ensuring faster data access and reduced network congestion.

9)Advanced analytics:

Hadoop integrates with various analytics tools and libraries, such as Apache Hive, Apache Pig, and Apache Mahout, to perform advanced analytics tasks.[24]

VIII. CONCLUSION

In conclusion, Big Data Hadoop is a powerful and widely adopted framework for processing and analyzing large volumes of data. It provides a scalable and distributed computing environment that enables organizations to efficiently store, process, and derive insights from massive datasets.

Hadoop's core components, such as the Hadoop Distributed File System (HDFS) and MapReduce, allow for distributed storage and parallel processing across a cluster of commodity hardware. This architecture ensures high fault tolerance, scalability, and cost-effectiveness compared to traditional data processing systems.

By leveraging the Hadoop ecosystem, which includes various tools and technologies like Hive, Pig, Spark, and HBase, organizations can perform a wide range of data processing tasks, including batch processing, real-time analytics, machine learning, and data exploration. These tools enable data engineers, data scientists, and analysts to extract valuable insights, make data-driven decisions, and uncover hidden patterns or trends within large datasets.[25]

In summary, Big Data Hadoop remains a significant and influential technology in the world of big data analytics. Its ability to handle massive amounts of data and provide distributed computing capabilities has made it a cornerstone of many organizations' data infrastructure. However, as technology continues to evolve, it's essential to evaluate the specific requirements and consider other alternatives before adopting any big data processing framework.

IX. FUTURE ENHANCEMENT

1.Performance Optimization: Future enhancements may focus on improving the performance of Hadoop by optimizing resource allocation, reducing data transfer overhead, and enhancing task scheduling algorithms. This could involve advancements in data locality, parallel processing, and intelligent task assignment.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205468 |

2.Real-time Analytics: While Hadoop has primarily been used for batch processing of large datasets, future enhancements may enable it to handle real-time data processing and analytics more efficiently. This could involve integrating Hadoop with streaming frameworks like Apache Kafka or implementing in-memory processing techniques.

3.Advanced Security and Privacy: As data privacy and security concerns continue to grow, future enhancements may focus on strengthening the security features of Hadoop. This could include improvements in authentication, authorization, data encryption, and access control mechanisms.

4.Machine Learning Integration: Big Data and machine learning often go hand in hand. Future enhancements may involve tighter integration between Hadoop and popular machine learning frameworks like Apache Spark or TensorFlow. This could enable seamless data processing, model training, and inference on large-scale datasets.

5.Containerization and Cloud Native Support: Hadoop deployments could evolve to embrace containerization technologies like Docker and orchestration frameworks like Kubernetes. This would enable easier deployment, scalability, and management of Hadoop clusters in cloud environments.

6.Simplified Management and Monitoring: Enhancements could focus on providing better tools and interfaces for managing and monitoring Hadoop clusters. This may involve user-friendly dashboards, automated cluster tuning, and proactive fault detection and recovery mechanism.[26]

REFERENCES

Research Paper:-

1)https://www.researchgate.net/publication/281403776_Big_Data_And_Hadoop_A_Review_Paper
2)https://www.researchgate.net/publication/339676259_A_REVIEW_PAPER_ON_BIG_DATA_AND_HADOOP
3)http://www.ijcst.com/vol74/1/11-iqbaldeep-kaur.pdf
4)https://link.springer.com/chapter/10.1007/978-3-319-08976-8_16
5)https://www.academia.edu/36834863/A_Review_Paper_on_Big_data_and_Hadoop
6)https://www.jetir.org/papers/JETIRB006016.pdf
7)http://ijarse.com/images/fullpdf/1519302484_SVCET2097ijarse.pdf
8)https://www.ijert.org/big-data-analytics-with-hadoop
9)https://www.researchgate.net/publication/264555968_Big_Data_Analytics_A_Literature_Review_Paper
10)https://www.researchgate.net/publication/346445572_Using_Hadoop_Technology_to_Overcome_Big_Data_Problems_by_Choosing_Proposed_Cost-efficient_Scheduler_Algorithm_for_Heterogeneous_Hadoop_System_BD3
11)https://arxiv.org/ftp/arxiv/papers/1610/1610.04572.pdf
12)https://www.simplilearn.com/challenges-of-big-data-article
13)https://ieeexplore.ieee.org/document/8668136
14)https://www.xenonstack.com/insights/big-data-challenges
15)https://thesai.org/Downloads/Volume7No2/Paper_67-A_Survey_on_Big_Data_Analytics_Challenges.pdf
16)https://thesai.org/Downloads/Volume7No2/Paper_67-A_Survey_on_Big_Data_Analytics_Challenges.pdf
17)https://www.sciencedirect.com/science/article/pii/S187705091500592X
18)https://www.sciencedirect.com/science/article/pii/S187705091500592X
19)https://www.educba.com/what-is-big-data-and-hadoop/
20)**A STUDY OF WORKING METHODOLOGIES AND ARCHITECTURES BIG DATA AND HADOOP**
July 2018
Author:Kirandeepkour
*21)**https://www.researchgate.net/publication/281403776_Big_Data_And_Hadoop_A_Review_Paper***
22)https://www.researchgate.net/publication/281403776_Big_Data_And_Hadoop_A_Review_Paper
23)A REVIEW PAPER ON BIG DATA AND HADOOP
March 2020
Conference: International Journal of Advance and Innovative Research Volume 7, Issue 1 (VI): January - March, 2020 Part - 1 ISSN-2394-7780
https://www.researchgate.net/publication/339676259_A_REVIEW_PAPER_ON_BIG_DATA_AND_HADOOP
24)Big Data Hadoop Tools and Technologies: A Review
Hussain, Toshifa and Sanga, Anirudh and Mongia, Shweta, Big Data Hadoop Tools and Technologies: A Review (October 1, 2019). Proceedings of International Conference on Advancements in Computing & Management (ICACM) 2019, Available at SSRN: https://ssrn.com/abstract=3462554 or http://dx.doi.org/10.2139/ssrn.3462554

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205468 |

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3462554

25)Big Data Analysis Using Apache Hadoop Shankar Ganesh Manikandan Department of Information Technology, Dhanalakshmi College of Engineering Tambaram, Chennai, India

https://ieeexplore.ieee.org/document/7021746/authors#authors

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

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

www.ijirset.com

Scan to save the contact details