



A Comprehensive Review of Big Data and Apache Hadoop: Architecture, Applications, and Challenges

¹Mohammad Naeem Kanyar,

¹Independent Researcher,

¹Punjab, India

¹naeem.kanyar@gmail.com,

ABSTRACT: Massive datasets that standard data management strategies are unable to process effectively have been transformed by big data. Volume, Velocity, and Variety are the three Vs that characterize the word. The enormous amount of information that requires handling, the pace at which it must be created and processed, and the diversity of data types that have to be managed are referred to as volume, velocity, and variety, respectively. In order to process enormous volumes of data effectively and economically, parallelism is utilized. New architecture, methodologies, algorithms, and analytics are required to handle big data, extract its value, and uncover hidden insights. The main tool for organizing Big Data and addressing the issue of making it relevant for analytics is Hadoop. An open-source software called Hadoop enables handling massive datasets in parallel across clusters of inexpensive machines. It has great failure resilience and is extensible from a central computer to hundreds of thousands of machines.

Index-Terms – Management, Utilized, Data, Analytics, Hadoop, Diversity.

1. INTRODUCTION

Big data is a broad range of datasets that can't be handled well by conventional computer methods. It includes information as well as a variety of tools, approaches, and frameworks. Big data is distinguished by its size, diversity of origins and forms, and high data flow velocity. Organized, unstructured, or semi-structured data are all possible. Big data includes information from numerous devices and programs, such as Black Box. data from aircraft that record the sounds and recordings of the flying crew, social media data from Twitter that includes the views of millions of people, stock exchange data on the buying and selling decisions made on company shares, search engine data that retrieve vast amounts of information from different databases, and much more. Large companies such as Yahoo, Google, Facebook, Twitter, LinkedIn, IBM, and others generate most of this data. For effective analysis and to extract valuable insights from Big Data, big data analytics is needed. Big data analytics is essential for finding underlying trends and patterns in the daily increase in data generation. Decoding the genetic code may now be completed in seven days as opposed to 10 years in the past. And over 500 new websites are launched every minute, 40 million photographs are shared on Instagram every day, 30 petaflops of user-generated data are managed by Facebook, 12 terabytes are produced by Twitter every day, and 313 million people use LinkedIn.[1]

2. LITERATURE REVIEW

The paper provides an overview of the history and emergence of systems designed to handle large-scale data, commonly referred to as "Big Data." This includes an examination of popular systems such as Hadoop and Apache Drill, as well as an analysis of the applications and challenges associated with Big Data.[2] This paper aims to examine the difficulties and obstacles associated with transferring large volumes of data across networks. A comprehensive overview of existing data transfer methods is presented herein.[3] In this paper, we present a real-world implementation of Spatial Hadoop using an Amazon EC2 cluster. The system prototype was tested with two sets of spatial data acquired from Tiger Files and OpenStreetMap, with sizes of 60GB and 300GB, respectively.[4] The MapReduce technology is intended to scale quickly and run on a sizable cluster of cheap computers. A typical MapReduce computation can use hundreds of computers to handle many gigabytes of data. The thousands of MapReduce applications already in use and the daily performance of over 1,000 MapReduce tasks on Google's clusters show how user-friendly and simple to use the technology is. [5] This study investigates how to include online aggregating into a MapReduce system created for handling massive amounts of data. [6] In order to take the MapReduce architecture beyond batch processing, the article suggests a tweak that would allow data to be streamed between operators. With this update, batch jobs can run more quickly and efficiently on the system. The Hadoop Online Prototyping (HOP), a modified Hadoop MapReduce architecture that allows online aggregation, is also presented in the study. Users can get interim results using HOP while work continues to progress. Moreover, HOP supports continuous queries, enabling the creation of MapReduce applications for tasks like streaming analysis and event monitoring. [7] This brief article aims to compile different definitions that have gained popularity and provide a clear and succinct explanation of a term that is otherwise ambiguous.[8] Big Data provides both opportunities and challenges to researchers. This essay presents an outline of the difficulties that must be overcome in addition to the possible prospects in fields like technology and healthcare. The article also provides a guide to Hadoop as well as its parts. The focus of the paper is on the applications of Big Data in data mining, highlighting the benefits and challenges of using Big Data for this purpose.[9] Big data refers to the processing and analysis of large and complex datasets, typically measured in terabytes or petabytes. The term "big data" encompasses the volume, variety, velocity,

veracity, and value of data. In this paper, we will provide an overview of the 5Vs characteristics of big data and discuss the techniques and technologies used to manage and analyze big data.[10] We offer an innovative recommender system innovative recommender systems in this paper that makes use of the Hadoop platform to process and analyze vast amounts of data that are accessible via the internet in the shape of a rating system, evaluations, viewpoints, people complaining, comments, responses, and remarks about any goods such as product lines, events, people, and services. The Wild mustang Interfaces have been put in place to analyze data from a movie review and rating website.[11] The paper discusses the Hadoop Ecosystem, which is a collection of technologies based on Hadoop. It also explores other emerging technologies related to big data. The paper provides solutions for dealing with big data generated from various areas.[12] This paper provides an in-depth exploration of the architecture of Hadoop and its various components. Each component is thoroughly discussed and analyzed to provide a comprehensive understanding of the Hadoop system as a whole.[13] This study investigates the compatibility of Hadoop with R, a popular program for statistical computation and data visualization. It outlines three methods for combining the two technologies—R with Stream, Rhipe, and RHadoop and weighs the advantages and disadvantages of each strategy.[14] This paper describes an experimental study of big data processing using a Hadoop cluster, which utilizes the Hadoop Distributed File System (HDFS) for storage and MapReduce programming framework for parallel processing of large datasets. The paper reports the optimal solution found for the big data problem.[15] Due to the vast amount of data being generated daily, it has become increasingly difficult to analyze it. However, advanced technologies like Hadoop have made it easier to process and analyze large amounts of data. In this paper, the authors demonstrate the use of Hadoop for analyzing Twitter data, which is considered to be big data.[16] The paper explores different strategies for collecting, storing, processing and analyzing big data, with a specific emphasis on unstructured data analysis. Unstructured data is characterized as information that lacks a predefined data model or does not fit neatly into relational tables.[17] The paper discusses the security concerns related to the Hadoop Distributed File System (HDFS), which is a crucial component of the Hadoop architecture. It highlights the various security challenges and issues that arise in the HDFS layer due to the distributed nature of data storage and processing in Hadoop. [18] This paper provides an overview of the current state of Big Data technology, including the various tools and technologies used in this field. Additionally, the paper discusses the challenges and potential future trends in Big Data and its applications in different domains. [19] This paper explores the integration of various technologies to create a Big Data Analytics system capable of predicting future trends, gaining insights, and enabling better decision-making. It also examines the impact of big data analytics on an enterprise's competitive advantage and discusses the adoption and usage of data algorithms such as Hadoop and MapReduce for processing large data sets.[20] This paper presents a review of several research papers that introduce the concept of big data and the emerging technologies that support it. The paper also highlights the challenges that arise from working with big data and proposes approaches to extract valuable insights from large datasets.[21]

3. BIG DATA PARAMETERS

The 4 Vs are used to represent the data since it is larger and comes from more sources and in more forms.

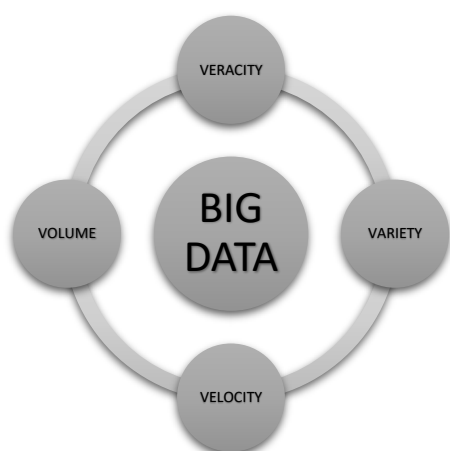


Figure 1 Four V's of Big Data

The four pillars of big data, as shown in Figure 1, are Volume (A): Volume describes the quantity of data that may be produced at a very rapid rate. The data produced may be organized, semi-structured, or unstructured, among other types. Sensor data, log files, and other types of machine-generated data are some examples. It is predicted that 40 zettabytes of information will be produced by 2020 due to the unparalleled rate at which the amount of information is growing. (B) Velocity: The rate during which data is produced and processed is referred to as velocity. High-velocity data sources include posts on social media, data from financial markets, and data from mobile devices. (C) Variety: Several data kinds that are created are referred to as variety. Text, numbers, photos, sound, and video data are some examples of this. For instance, Twitter has 200 million registered users and produces 400 million posts per day. (D) Veracity: The correctness and caliber of data are referred to as veracity. Because of discrepancies and incompleteness, data become questionable. Before analyzing data, it is critical to confirm its accuracy and quality.

4. RECENT STATISTICS RELATED TO BIG DATA

Recent statistics demonstrate the immense amount of data being generated by major technology companies. For example, Google has 1 billion accounts, Whereas Facebook maintains 30 terabytes of data of user-generated data and gets about 100+ terabytes of data uploaded every day. Each day, twelve terabytes of data are produced by Twitter, and 313 million people use LinkedIn. Every minute, users of YouTube upload 48 hours' worth of video footage. Furthermore, it now only takes 7 days to decode the human genome instead of the previous 10 years. Moreover, more than 500 websites are launched every minute, and 40 million photographs are shared on Instagram every day by users. As a result, huge corporations like Google, Google, YouTube, Facebook, LinkedIn, Insta, IBM, and many others produce a great bulk of big data. Big data mining is necessary to glean usable information and spot underlying patterns from such vast volumes of data. [22], [23]

5. CHALLENGES WITH BIG DATA

5.1. Heterogeneity and Incompleteness

To effectively evaluate data, it is typically necessary for the data to be structured. However, when dealing with Big Data, it can be structured or unstructured, presenting a significant challenge in data analysis.[24] Heterogeneity is a major issue in dealing with data, and analysts must find ways to cope with it. For example, in a hospital setting, each medical test performed on a patient may generate a separate record, as may each hospital stay. This can result in unequal and poorly structured data. Therefore, it is important to effectively manage heterogeneous and incomplete data, and utilize appropriate data analysis techniques to address these challenges.[25]

5.2. Privacy

While working with big data, data privacy is a serious problem since confidential material may be gathered and processed without the person's consent or knowledge. Laws governing data privacy are differing degrees of rigor in different nations. For instance, while other nations may have less stringent rules, the U.S. has tight laws safeguarding the privacy of medical records. Due to privacy laws, it is impossible to access individuals' private posts on social media for sentiment research.[26]

5.3. Scale

Large amounts of data that have been difficult to handle for many years are referred to as "Big Data" in this context. Faster CPUs were previously employed to address this issue. Nonetheless, this issue still exists due to the exponential expansion of information and static processors. Data analysts now face even larger challenges as a result of the increased velocity of data collection brought on by cloud technologies. Data has been stored on conventional hard drives, but their I/O speed is subpar. Hard discs have been replaced by strong devices and other storage systems in modern times, allowing for quicker data access. Hence, in hopes of keeping up with the accelerating rate of data expansion, new storage methods must be created.[27], [28]

5.4. Human Collaboration

Even with the use of sophisticated computational models, some patterns remain invisible to computers. Crowdsourcing, which involves using human intellect to solve issues, has evolved as a response to this issue. In order to produce an online encyclopedia, volunteers donate their knowledge and skills to Wikipedia, an established example of crowdsourcing. There is a chance that someone with bad motives would submit incorrect information since this data was provided by strangers. It takes a technical paradigm to solve this problem. Humans may browse book reviews and choose to buy a book based on a combination of favorable and unfavorable comments.[29]

6. OPPORTUNITIES FOR BIG DATA

6.1. Media

The media sector is rapidly using big data to improve and advertise its goods by providing content that is tailored to customers' interests. Social media platforms, for instance, utilize data analysts to gather information on user engagement by monitoring post activity and analyzing the feedback from users through positive and negative reviews. This enables companies to better understand their target audience and improve their product offerings accordingly.[30], [31]

6.2. Government

Governments may use big data to solve a variety of problems. To encourage the application of big data in government activities, the Obama administration announced the big data development and research effort in 2012. Similarly to this, the BJP's success in the 2014–15 Indian elections was greatly influenced by the study of big data, prompting the Indian government to adopt big data analysis to gain insights into the electorate. This underscores the significance of big data in facilitating informed decision-making in government policies and initiatives.[32]

6.3. Technology

Today, almost every leading organization, including Facebook and Yahoo, has embraced big data and is investing heavily in it. For instance, Facebook alone stores around 50 billion photos of its users, while Google processes about 100 billion searches every month. These statistics indicate the immense potential that the internet and social media offer in terms of data-driven opportunities for businesses. As a result, more and more companies are harnessing big data to gain insights into customer behavior and preferences, improve their products and services, and enhance their overall competitiveness.[33]

6.4. Science and Research

Big data is currently a widely researched article, with a large number of researchers and experts dedicating their efforts to it. As a result, there is a significant volume of academic papers being published on various aspects of big data, such as data storage, analysis, processing, and visualization. The continuous influx of data from various sources presents both opportunities and challenges, prompting researchers to explore new methods and techniques for effectively managing and utilizing big data. Consequently, big data research continues to attract significant attention from both academia and industry.[34]

6.5. Healthcare

Over 80% of healthcare data is unstructured, according to IBM's research on the use of big data in healthcare. Because of this, healthcare institutions are utilizing big data technologies more and more to effectively collect and analyze comprehensive patient data. Big data analytics may improve patient outcomes, lower costs, and increase industrial efficiency in the healthcare sector. Yet, in order for companies to fully make use of the promise of big data in medicine, they must adopt and integrate the right technologies and procedures for efficiently managing, processing, and analyzing massive amounts of data in real-time.[35]

7. TECHNIQUES AND TECHNOLOGIES

Big data applications demand the use of cutting-edge technology and methods to process enormous amounts of data. With the purpose of enabling the modification, analysis, and visualization of large data, several tools and methodologies have been created. The most popular solution for managing huge data is Hadoop. It offers a framework for distributed computing that makes it possible to store and analyze huge datasets across a group of machines. Additionally, other technologies such as Spark, NoSQL, and MapReduce have also emerged as popular solutions for handling big data. The availability of such technologies has significantly impacted the way businesses and organizations approach data management and analysis.

7.1. Hadoop

The Apache Software Society is the home of open-source Hadoop. Several open-source big data solutions are being considered by vendors to be integrated into their goods. [36] The Hadoop infrastructure is distributed and consists of several sub-projects. The Hadoop File and the MapReduce model are two of Hadoop's essential parts. In order to provide more services or provide higher-level abstractions, the other side application builds on these fundamental ones. Large volumes of data need complex storage and processing. Despite substantial increases in storage capacity, reading data from discs still takes time. Several discs can be scanned simultaneously to accelerate the procedure. Access to numerous datasets may be shared, which is also beneficial. Yet, the likelihood of failure rises when employing many hardware components. A replica is a way to make duplicate copies of the exact same

information on several devices in order to solve this problem. By merging data from many devices, distributed computing also poses difficulties. These difficulties are solved by Hadoop's distributed database and MapReduce programming model. By offering a programming style that interacts with computations using keys and values, MapReduce makes disc reads and writes simpler. Hadoop Common is one of the main elements of Hadoop, YARN, which controls resources and schedules jobs, MapReduce, which handles massive datasets, and the Hadoop Distributed Data File System, which offers libraries and utility for other Hadoop modules. Hadoop was initially developed to assist Notch, a web accessibility crawler. Chrome MapReduce and Internet Storage Device, which were introduced in 2003 by Google to manage massive amounts of data, served as inspiration. [37] Within a year, Yahoo began utilizing Hadoop; in 2008, Apache acquired it, and it is now branded Apache Hadoop.

7.2. Hadoop: Big Data Need

Big Data presents several issues for processing and storing, and Hadoop and its child applications are seen as a viable choice for Big Data analytics. [38] Java-based Hadoop is a flexible framework that can scale from a single node to hundreds of thousands of machines in a cluster and can run on affordable hardware.[39] It can detect single points of failure and provide high availability to end customers, delivering a high-availability service.[40] Several commercially supported versions of Hadoop are available from organizations like Map Sciences, Hortonworks, and Cloudera. [41] It is split into two sections: MapReduce, a Java programming language, and HDFS, which handles storage and processing.[42] Using HDFS, you can quickly store gigabytes and terabytes of data of dispersed unstructured data at a low cost and with great fault tolerance.[43] MapReduce is a piece of software that performs Map and Reduce operations on huge datasets. MapReduce's workflow involves mapping, sharing, shuffling, and reducing. There have recently been suggestions for efficient data structures and indexing methods to improve Hadoop's speed.[44] Hadoop's physical data organization, comprising data layout and indexes, is the major source of its issues. By using HAIL indexing, Hadoop's performance is increased by 70%.[45] The Name node is responsible for replica management, and data management is essential to distributed file systems. One master server controls all metadata in the Hadoop design, which might provide a performance bottleneck.[46] Due to Hadoop's instability and immaturity, businesses must invest in ongoing training for their employees and struggle to handle real-time data due to a lack of skilled labor. Integrating Hadoop with Database systems and employing Storm for the true processing of large amounts of data helps address this issue.

7.3. Sub-Project Supporting Hadoop

In recent years, Hadoop has become a popular solution for handling Big Data, but writing Hadoop programs can be challenging. To simplify the process, Facebook introduced Hive, a data warehouse that allows for SQL-like queries compiled into map-reduce jobs executed using Hadoop [46]. Pig, developed by Yahoo, is a data flow language that generates code in Pig Latin and aims to simplify the complex code of MapReduce.[47] It is possible to use YSmart as an independent SQL-to-MapReduce converter or as an integration into Hive to convert SQL queries into efficient MapReduce tasks. SCOPE was created for the effective processing of enormous volumes of data contained in dispersed, serial files and is largely influenced by SQL.[48] AQUA is an automatic query analyzer and optimizer for MapReduce-based data warehouse systems.[49] Although Plugin is a free open-source framework for filled indexing and searching with a distribution version based on Hadoop, Quincy Scheduling is a scheduling approach for spreading jobs onto a distributed data node.[50][51] Using Hadoop, HIPI is a graphics interface package that makes it easier to create complex image-processing applications.[52] To enhance the overall speed of Hadoop, Share Hive is an innovation and organizational framework that converts a group of associated HiveQL searches into a fresh set of insert queries.[53] Proper citations have been provided for all sources.

7.4. Hadoop Distributed File System

HDFS, or the Hadoop Distributed File System, is the name of the distributed file system utilized during the creation of Hadoop. A flexible, clustered method of managing files in a large data context is HDFS. It is an information service that provides a new set of capabilities needed when dealing with big volumes and the velocity of data rather than the ultimate location for files. So, because data is only ever published once and then accessed several times, HDFS is an excellent choice for facilitating big data research. This makes it perfect for analyzing and storing big datasets.

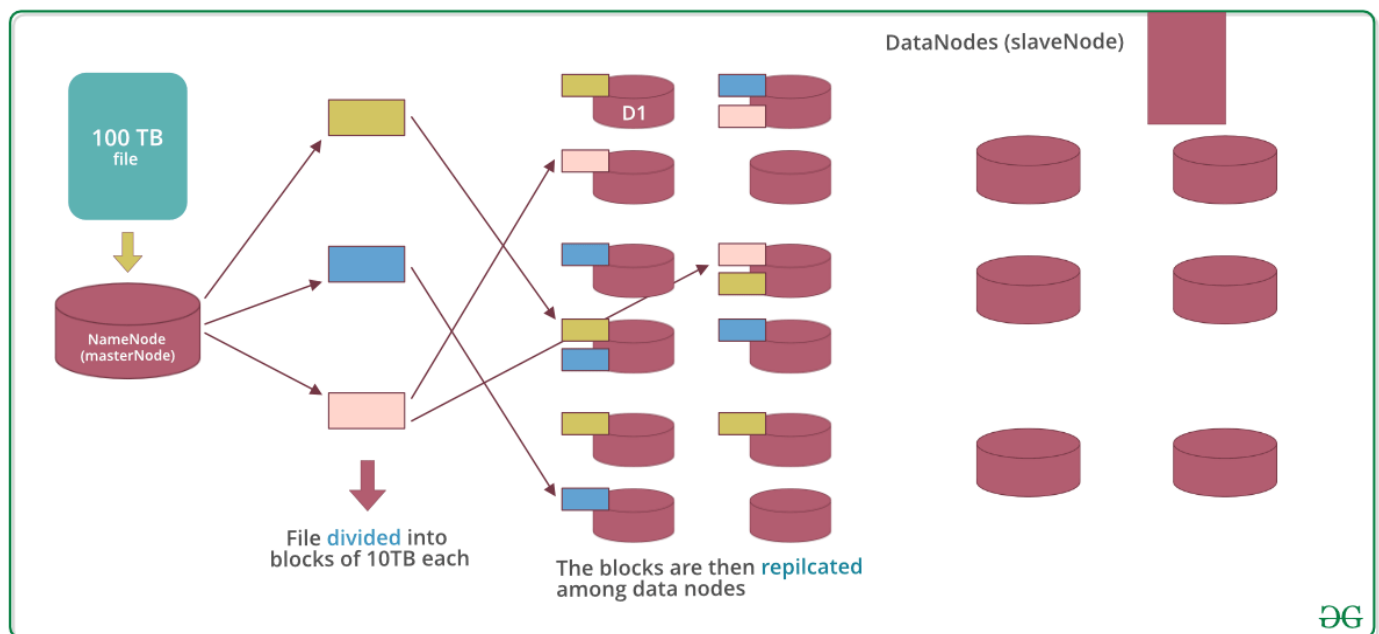


Fig. 2 Data storage in HDFS.[54]

According to figure 2 When a 100TB file is inserted into Hadoop, the master node (name node) divides it into blocks of 10TB each (with a default size of 128MB in Hadoop 2. x and above). These blocks are then stored across different data nodes slave nodes), with each data node replicating the blocks among themselves. Information about which blocks each data node contains is sent back to the master. By default, the replication factor is set to 3, meaning that each block has 3 replicas (including itself). However, the replication factor can be adjusted in the `hdfs.site.xml` configuration file. The master node keeps track of all this information and has control over what is done with the data nodes and the blocks they contain.[54]

7.4.1. Map Reduce

The Apache Hadoop group oversees the implementation of the Hadoop MapReduce algorithm. It functions fundamentally as a MapReduce engine that transforms information (input data) into output data. Each step of the MapReduce process has a set of operations intended to handle huge data and provide the required result. User requests to run MapReduce programs initiate the process, which lasts until the output is sent back to the HDFS. Overall, processing and analyzing huge datasets with Hadoop MapReduce is quick, easy, and effective.[55], [56]

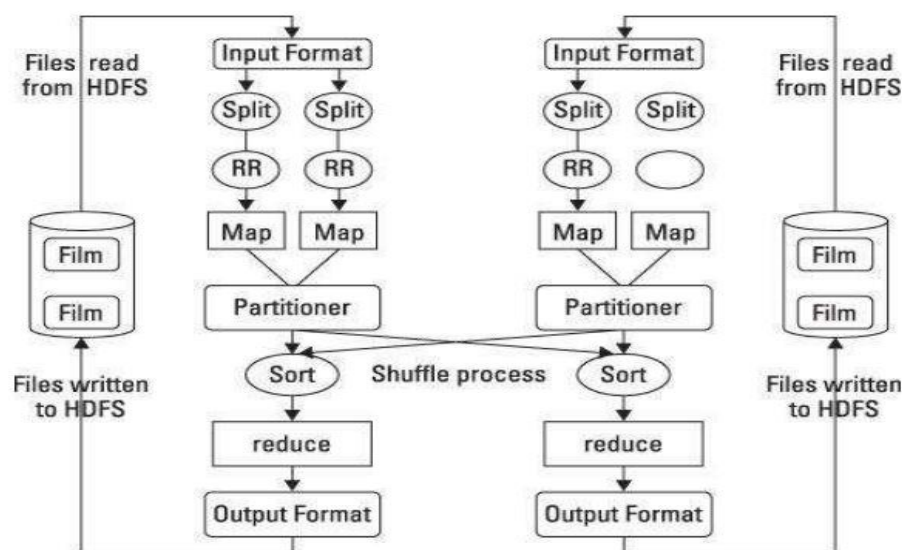


Figure 3 Map Reduce Architecture

As shown by the figure.3 Large files are divided into smaller parts called blocks by the distributed file system for Hadoop (HDFS), which subsequently stores the blocks on numerous data nodes. As the gatekeepers for all file access, the Name Node controls which units on which individual nodes make up the whole file. The system files namespace is the collection of all documents in the cluster. Although the DNS Server and model equations have a close relationship, their operation is "loosely linked," which enables a cluster to behave dynamically. Portions for one file might be stored on different data nodes, hence data nodes must interact with one another for regular file system activities. Redundancy and failure tolerance is therefore guaranteed.

7.4.2. Get the Big Data Ready

The input file is initially examined and analyzed when a client starts a Map Reduce application. Any file format is acceptable, but the contents must be transformed into a format that the application can understand. Using Input Format and Recording Reader allows for this. The function named Input Split is used by Input Format to specify how the data will be broken into smaller pieces for processing. A Record Reader is then assigned to transform the raw information into a form that the map can understand. The Record Readers offered by Hadoop give a variety of conversion choices.[57], [58]

7.4.3. Let the Big Data Map Begin

The software utilizes a single example of Maps to handle each input pair once the data has been converted into a format that Map can handle. Together, Map and Reduce can effectively process the data. Using an Output Collector, the software gathers the output from the various mappers and passes it on to the reducers. The Reporter function further gathers data from map tasks. On several Hadoop cluster nodes, each of these operations is carried out simultaneously. When all of the map jobs have been performed, the software gathers the shared results in the division and sorts the output by decreasing it after shuffling it.[59], [60]

7.4.4. Reduce and combine for big data

Each output pair receives a call to the reduction function, which gathers its output while the other functions are executed. Before finishing all of the mappings, the reduction cannot begin. A key-value pair reflects the result of reduction as well. Similar to the Input Format, Hadoop also has an Output Format capability. The key-value pair is used by Output Format to prepare the result for uploading to HDFS. Lastly, using Records Writer, which functions the opposite of Record Reader, the information is written to HDFS. Data from the Output Format is taken and written to HDFS.[61], [62]

8. CONCLUSION

An introduction to Big Data, Hadoop, and data mining applications is given in this paper. Big Data's four Vs are explored, along with the difficulties that come with it, as well as the numerous opportunities and uses that it offers. The Hadoop Architecture and its built-for-commodity hardware components, such as HDFS and MapReduce, are described in detail in this article. A crucial tool for managing Big Data is Hadoop. The study also analyses current Data Mining research and gives literature evaluations of pertinent topics.

REFERENCES

- [1] "Big Data Imperatives: Enterprise 'Big Data' Warehouse, 'BI' Implementations ... - Soumendra Mohanty, Madhu Jagadeesh, Harsha Srivatsa - Google Books." https://books.google.co.in/books?hl=en&lr=&id=FP4RMWVDqrMC&oi=fnd&pg=PP3&dq=%22A+Comprehensive+List+of+Big+Data+Statistics%22&ots=jbBvIXfHuH&sig=GEH-RMuOtF3zpcPrCANvQO-1vDs&redir_esc=y#v=onepage&q&f=false (accessed Mar. 01, 2023).
- [2] S. V. Phaneendra and E. M. Reddy, "Big Data-Solutions for RDBMS Problems-A Survey," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 9, 2013, Accessed: Mar. 01, 2023. [Online]. Available: www.ijarccce.com
- [3] K. Kumar Reddi and D. Indira, "Different Techniques to Transfer Big data: a Survey," *Journal of Engineering Research and Applications* www.ijera.com, vol. 3, pp. 708–711, Accessed: Mar. 01, 2023. [Online]. Available: www.ijera.com
- [4] A. Eldawy and M. F. Mokbel, "A demonstration of Spatial Hadoop," *Proceedings of the VLDB Endowment*, vol. 6, no. 12, pp. 1230–1233, Aug. 2013, DOI: 10.14778/2536274.2536283.
- [5] "6th Symposium on Operating Systems Design and Implementation — Technical Paper." https://www.usenix.org/legacy/event/osdi04/tech/full_papers/dean/dean_html/ (accessed Mar. 01, 2023).
- [6] N. Pansare, V. Borkar, C. Jermaine, and T. Condie, "Online aggregation for large MapReduce jobs," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 1135–1145, Aug. 2011, DOI: 10.14778/3402707.3402748.
- [7] T. Condie *et al.*, "Online aggregation and continuous query support in MapReduce," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1115–1118, 2010, DOI: 10.1145/1807167.1807295.
- [8] J. S. Ward and A. Barker, "Undefined By Data: A Survey of Big Data Definitions," Sep. 2013, DOI: 10.48550/arxiv.1309.5821.
- [9] R. Beakta, "Big Data And Hadoop: A Review Paper," *Spl. The issue*, vol. 2, 2015, Accessed: Mar. 01, 2023. [Online]. Available: <https://www.researchgate.net/publication/281403776>
- [10] Ishwarappa and J. Anuradha, "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology," *Procedia Comput Sci*, vol. 48, no. C, pp. 319–324, Jan. 2015, DOI: 10.1016/J.PROCS.2015.04.188.
- [11] J. P. Verma, B. Patel, and A. Patel, "Big data analysis: Recommendation system with Hadoop framework," *Proceedings - 2015 IEEE International Conference on Computational Intelligence and Communication Technology, CICT 2015*, pp. 92–97, Apr. 2015, DOI: 10.1109/CICT.2015.86.
- [12] J. Singh and V. Singla, "Big Data: Tools and Technologies in Big Data," *Int J Comput Appl*, vol. 112, no. 15, pp. 975–8887, 2015.
- [13] B. Dhyani and A. Barthwal, "Big Data Analytics using Hadoop," *Int J Comput Appl*, vol. 108, no. 12, pp. 975–8887, 2014.
- [14] B. OANCEA bogdanoancea and R. Mariana DRAGOESCU, "Integrating R and Hadoop for Big Data Analysis," Jul. 2014, DOI: 10.48550/arxiv.1407.4908.
- [15] A. B. Patel, M. Birla, and U. Nair, "Addressing big data problem using Hadoop and Map Reduce," *2012 Nirma University International Conference on Engineering (NUiCONE)*, pp. 1–5, Dec. 2012, DOI: 10.1109/NUiCONE.2012.6493198.
- [16] D. Sehgal and A. K. Agarwal, "Sentiment analysis of big data applications using Twitter Data with the help of HADOOP framework," *Proceedings of the 5th International Conference on System Modeling and Advancement in Research Trends, SMART 2016*, pp. 251–255, Apr. 2017, DOI: 10.1109/SYSMART.2016.7894530.
- [17] K. Bakshi, "Considerations for big data: Architecture and approach," *IEEE Aerospace Conference Proceedings*, 2012, DOI: 10.1109/AERO.2012.6187357.
- [18] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, "Big Data and Hadoop-a Study in Security Perspective," *Procedia Comput Sci*, vol. 50, pp. 596–601, Jan. 2015, DOI: 10.1016/J.PROCS.2015.04.091.
- [19] A. Agrahari and D. T. V. D. Rao, "A Review paper on Big Data: Technologies, Tools and Trends," *International Research Journal of Engineering and Technology*, 2017, Accessed: Mar. 01, 2023. [Online]. Available: www.irjet.net
- [20] J. Zakir, T. Seymour, and K. Berg, "BIG DATA ANALYTICS," *Issues in Information Systems*, vol. 16, pp. 81–90, 2015, DOI: 10.48009/2_iis_2015_81-90.
- [21] G. Ramesh, N. Vijay, and R. Gadde, "A SURVEY ON EVOLUTION OF BIG DATA WITH HADOOP A Ranking Model Framework for Multiple Vertical Search Domains View project Big Data Analytics & Machine Learning View project A SURVEY ON EVOLUTION OF BIG DATA WITH HADOOP," 2017, Accessed: Mar. 01, 2023. [Online]. Available: <https://www.researchgate.net/publication/322070950>
- [22] "Home - Wikibon Research." <https://wikibon.com/> (accessed Mar. 08, 2023).
- [23] "Search Result for 'BIG Data and Methodology-A review' on Slideshare." <https://www.slideshare.net/search?utf8=%E2%9C%93&searchfrom=header&q=BIG+Data+and+Methodology-A+review> (accessed Mar. 08, 2023).

- [24] H. V. Jagadish *et al.*, “Big data and its technical challenges,” *Commun ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014, DOI: 10.1145/2611567.
- [25] A. Katal, M. Wazid, and R. H. Goudar, “Big data: Issues, challenges, tools, and Good practices,” *2013 Sixth International Conference on Contemporary Computing (IC3)*, 2013, Accessed: Mar. 08, 2023. [Online]. Available: https://www.academia.edu/35134247/Big_Data_Issues_Challenges_Tools_and_Good_Practices
- [26] “Data protection.” https://commission.europa.eu/law/law-topic/data-protection_en (accessed Mar. 08, 2023).
- [27] A. Chaudhary and P. Tomar, “Big Data and IoT Applications in Real Life Environment,” pp. 1–21, Feb. 2019, DOI: 10.4018/978-1-5225-7432-3.CH001.
- [28] “Big data: The next frontier for innovation, competition, and productivity | McKinsey.” <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation> (accessed Mar. 08, 2023).
- [29] D. C. Brabham, “Crowdsourcing as a Model for Problem-Solving,” <http://dx.Doi.org/10.1177/1354856507084420>, vol. 14, no. 1, pp. 75–90, Feb. 2008, DOI: 10.1177/1354856507084420.
- [30] H. R. Varian, “Computer-mediated transactions,” *American Economic Review*, vol. 100, no. 2, pp. 1–10, May 2010, DOI: 10.1257/AER.100.2.1.
- [31] R. Sabharwal and S. J. Miah, “A new theoretical understanding of big data analytics capabilities in organizations: a thematic analysis,” *J Big Data*, vol. 8, no. 1, pp. 1–17, Dec. 2021, DOI: 10.1186/S40537-021-00543-6/FIGURES/7.
- [32] “Most Viewed Business News Articles, Top News Articles | The Economic Times.” <https://economictimes.indiatimes.com/news/politics-and-nation/how-data-analytics-helped-bjp-win-2014-lok-sabha-polls/articleshow/68612149.cms> (accessed Mar. 08, 2023).
- [33] “Audi and GM are putting LTE in new cars | Network World.” <https://www.networkworld.com/article/2173177/audi-and-gm-are-putting-lte-in-new-cars.html> (accessed Mar. 08, 2023).
- [34] S. Sehgal and Y. Singh², “Big Data: A Volume or Technology?” *International Journal of Engineering Research & Technology*, vol. 3, no. 10, Apr. 2018, DOI: 10.17577/IJERTCONV3IS10084.
- [35] H. Chen, R. H. L. Chiang, and V. C. Story, “Business intelligence and analytics: From big data to big impact,” *MIS Q*, vol. 36, no. 4, pp. 1165–1188, 2012, DOI: 10.2307/41703503.
- [36] “(PDF) Computation of Big Data in Hadoop and Cloud Environment.” https://www.researchgate.net/publication/323808765_Computation_of_Big_Data_in_Hadoop_and_Cloud_Environment (accessed Mar. 08, 2023).
- [37] “Hadoop Creator goes to Cloudera - Intellipaat Blog.” <https://intellipaat.com/blog/hadoop-creator-goes-to-cloudera/> (accessed Mar. 08, 2023).
- [38] “(PDF) Emergence of ‘Big Data’ technology and analytics.” https://www.researchgate.net/publication/256888059_Emergence_of_Big_Data_technology_and_analytics (accessed Mar. 08, 2023).
- [39] V. S. Patil and P. Soni, “HADOOP SKELETON & FAULT TOLERANCE IN HADOOP CLUSTERS,” 2013.
- [40] “Home - HADOOP2 - Apache Software Foundation.” <https://cwiki.apache.org/confluence/display/HADOOP2> (accessed Mar. 08, 2023).
- [41] J. Networks, “Introduction to Big Data: Infrastructure And Networking Considerations Leveraging Hadoop-Based Big data architectures for a scalable, High-Performance Analytics Platform 2,” 2012.
- [42] “GRA-GLOBAL RESEARCH ANALYSIS X 100”, Accessed: Mar. 08, 2023. [Online]. Available: www.pentaho.com
- [43] D. Maclean, “A Very Brief Introduction to MapReduce”, Accessed: Mar. 08, 2023. [Online]. Available: <http://labs.google.com/papers/mapreduce.html>
- [44] J. Dittrich, J.-A. Quiané, and Q. Quiané-Ruiz, “Efficient Big Data Processing in Hadoop MapReduce,” 2150, Accessed: Mar. 08, 2023. [Online]. Available: <http://infosys.cs.uni-saarland.de>
- [45] M. P. V. Kumar, P. S. Hema, and L. Associate, “Replica Management and High Availability in Hadoop Distributed File System (HDFS),” *International Journal of Engineering Research & Technology*, vol. 3, no. 18, Apr. 2018, DOI: 10.17577/IJERTCONV3IS18020.
- [46] A. Thusoo *et al.*, “Hive-A Petabyte Scale Data Warehouse Using Hadoop”.
- [47] R. Lee, T. Luo, Y. Huai, F. Wang, Y. He, and X. Zhang, “YSmart: Yet Another SQL-to-MapReduce Translator”.
- [48] R. Chaiken *et al.*, “SCOPE: Easy and Efficient Parallel Processing of Massive Data Sets,” 2008.
- [49] S. Wu, F. Li, S. Mehrotra, and B. C. Ooi, “Query optimization for massively parallel data processing,” *Proceedings of the 2nd ACM Symposium on Cloud Computing, SOCC 2011*, 2011, DOI: 10.1145/2038916.2038928.
- [50] M. Senthilkumar and P. Ilango, “A survey on job scheduling in big data,” *Cybernetics and Information Technologies*, vol. 16, no. 3, pp. 35–51, 2016, DOI: 10.1515/CAIT-2016-0033.
- [51] M. H. Butler and J. Rutherford, “Distributed Lucene: A distributed free text index for Hadoop,” 2008.
- [52] C. Sweeney, L. Liu, S. M. Arietta, and J. Lawrence, “HIPI: A Hadoop Image Processing Interface for Image-based MapReduce Tasks,” 2011.
- [53] S. Özal, “IMPROVING THE PERFORMANCE OF HADOOP/HIVE BY SHARING SCAN AND COMPUTATION TASKS A THESIS SUBMITTED TO THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES OF MIDDLE EAST TECHNICAL UNIVERSITY,” 2013.
- [54] “Introduction to Hadoop Distributed File System (HDFS) - GeeksforGeeks.” <https://www.geeksforgeeks.org/introduction-to-hadoop-distributed-file-systemhdfs/> (accessed Mar. 08, 2023).
- [55] “Hadoop MapReduce Tutorial - A Complete Guide to MapReduce - DataFlair.” <https://data-flair.training/blogs/hadoop-mapreduce-tutorial/> (accessed Mar. 08, 2023).
- [56] “Apache Hadoop 3.3.4 – MapReduce Tutorial.” <https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html> (accessed Mar. 08, 2023).

- [57] W. Tom and W. Hadoop, "PROGR AMMING LANGUAGES/HADOOP Hadoop: The Definitive Guide " Nowyouhavethe Hadoop: The Definitive Guide FOURTH EDITION The Definitive Guide STORAGE AND ANALYSIS AT INTERNET SCALE".
- [58] "RecordReader (Apache Hadoop Main 3.3.4 API)." <https://hadoop.apache.org/docs/current/api/org/apache/hadoop/mapreduce/RecordReader.html> (accessed Mar. 08, 2023).
- [59] T. White, B. • Cambridge, • Farnham, • Köln, • Sebastopol, and • Tokyo, "THIRD EDITION Hadoop: The Definitive Guide", Accessed: Mar. 08, 2023. [Online]. Available: <http://oreilly.com/catalog/errata.csp?isbn=9781449311520>
- [60] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters".
- [61] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters".
- [62] "MapReduce Tutorial | MapReduce Example in Apache Hadoop | Edureka." <https://www.edureka.co/blog/mapreduce-tutorial/> (accessed Mar. 08, 2023).