International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5
***Article Received:** 25 February 2023 **Revised:** 12 March 2023 **Accepted:** 30 April 2023*

To Paper and Design Different Big Data Techniques for Social Networks

Raghupathi K, Dr. Kailash Patidar
Department of Computer Science Engineering , Dr. A. P. J. Abdul Kalam University, Indore (M.P.), India 452016
Corresponding Author : - Raghupathi K

Abstract-Big data refers to datasets so enormous, diverse, and complicated that collecting them, storing them, analyzing them, and displaying them for further processing all present unique issues. To capture and analyze this data effectively, new technologies and systems are needed. Due to its large size and variety of formats, Big Data presents significant challenges for conventional data processing infrastructures. In today's high-tech society, information is being generated at an exponential rate because to the proliferation of di gital tools like smart phones, computers, and other mobile gadgets. The term "big data" refers to the massive volumes of information gathered from a variety of sources, including social networking and commercial websites like Face book and Amazon. The distributed architecture framework is where all the big data is kept. Hadoop is a free and open-source software environment for building scalable, distributed applications. Hadoop centralizes its features so they may be accessed from a single location. Some examples of information preparation procedures include highlighting, indexing, and searching. Data sets of this size are notoriously challenging for a single system to process. In this paper, we explored fundamental ideas and methods for handling Big Data.

Keywords: *Big Data, Hadoop, NoSQL, Data Privacy, Semantic Web, Distributed Queries, RDF, Cloud Computing.*

1. INTRODUCTION

A common definition of "big data" is a collection of methods that provide an integration strategy for revealing massive previously unknown values. These numbers are extremely complicated and enormous in size. Back in 1997, NASA researchers were the first to use the phrase "Big Data" in a formal publication. The concept of "big data" refers to data that is extremely recent, diverse, and large. The term "big data" has become a common expression used to describe extremely huge data sets. This information may be organized or unorganized. And it's tough to get good at using the standard equipment and methods. There is a large quantity of data (big data) in the IT industry that is shared across many divisions, meaning that a vast quantity of data is sitting idle in the backend of these companies without a suitable tool to manage it. Big data is the paper and processing of massive amounts of data collected from a variety of different sources. Big data may seem like a brand new field, but in reality, it has been maturing for quite some time. It has been a challenge for a long time to properly store and analyze massive datasets. It is extremely challenging to process and interpret Big Data due to its sheer size and complexity. Attempting these using standard methods is problematic. Conventional relational database management systems (RDBMS) are incapable of handling the data management needs of Big Data (Relational database management systems). It's challenging to extract useful information from Big Data because of its sheer volume. Information gained from analyzing massive data sets is of great value. In reality, information is derived from raw data. Big Data may be used to analyze a wide variety of data formats. It's possible for humans to come up with solutions, but it's computers that create the most predictable results. The term "human-generated data" refers to the information produced as a result of human use of technology. Structured data, video data, and textual data are all examples of human-generated information. Information produced by machines occurs when software and hardware react to external stimuli. For instance, a point-of-sale system generates an exchange against stock to represent items received by a customer, and a log document records an approval decision made by a security feature. Web logs, sensor data, and other examples of machine- generated information are incorporated into various models. Primary information includes:

- Structured data / Organized data
- Unstructured data
- Semi-structured data / semi-organized data

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5
***Article Received:** 25 February 2023 **Revised:** 12 March 2023 **Accepted:** 30 April 2023*

Organized Data (or) structured Data

Organized information fits into an information model or composition, and it shows how different entities relate to each other.

Unstructured Data

Unstructured data is information that doesn't fit into an information model or data schema. It is thought that 80% of the information in a random effort is unstructured. Organized information grows at a faster rate than unstructured information.

Semi-organized Data

Information that is only partly organized has some structure and consistency, but it is not social in nature. Information that is only partly organized is more like a progression or a chart. Usually, this kind of information is kept in documents that have text. Structured data is usually stored in a tabular format and fits into a data model.

Figure 1. Big Data Types

The most important thing to know about Big Data is that data must be used in a way that helps real- world outcomes that are profitable or helpful. Most people have only just begun to use Big Data. Many companies have been trying out methods that let them collect huge amounts of data so they can look for hidden patterns that could be an early sign of an important change. Data might show, for example, that customers' buying habits are changing or that the business needs to think about new factors. The term "Big Data" has been used in research since the 1970s, according to a paper called "The Evolution of Big Data as a Research and Scientific Topic." Today, the idea of "Big Data" is looked at from many different points of view, which shows how important it is. Big Data is important in a lot of ways. The large amount of data generated makes it possible for stakeholders to make quick decisions that can save money and improve operations in both the public and private sectors. For example, in the retail business, Big Data analysis can help understand how customers behave and what they like by looking at things like how they move around a store, how they use a website, what products they look for, and so on. Also, the US Healthcare Big Data World, which has records for more than 50 million patients, uses data-driven ideas to figure out problems in the healthcare field. Big Data has the potential to bring in more money while reducing risk and making it easier and cheaper to predict what will happen in the future. The Big Data management cycle is shown in the picture below.

2. LITERATURE REVIEW

Suvarnamukhi et al (2018), With the proliferation of digital technologies like smartphones, tablets, computers, and machines, data is expanding at an exponential rate. The term "big data" refers to a collection of many types of data, such as that found on online shopping and social networking sites. All the large data is stored in the context of distributed architecture. Hadoop is an open-source platform for developing programs that scale to run on a cluster of computers and analyze massive amounts of data. Hadoop centralizes all of their operations, making them easier to manage. Information can be prepared in a variety of ways, including highlighting, indexing, and searching. When dealing with a massive volume of data, it might be challenging to do so on a single computer. Big data processing was the primary focus of this article. **Hiba Jasim Hadi et al (2015),** Big Data refers to data sets that are so vast that they are challenging to manage, much alone store, analyze, and display for subsequent use in other procedures or to produce desired outcomes. By analyzing massive datasets, known as "Big Data," researchers hope to unearth previously unknown relationships and trends. The knowledge gained from this article may help any company or organization stay ahead of the curve. This means that proper Big Data implementation is crucial. Therefore, Big Data is a significant technological development that has the potential to significantly alter the ways in which organizations leverage data to enhance the customer experience and develop new lines of business. In addition, Big Data frequently involves achieving results that were previously out of reach for the majority of people due to limitations in available technology or financial constraints. Volume, Velocity, Variety, Value, and Veracity are only few of the qualities of Big Data discussed in this article, along with its content, kinds, architecture, technologies, and other distinguishing features.

Mrutyunjaya Panda et al (2018), As the popularity of social media platforms has skyrocketed over the past several years, so too has the volume of data being generated every day. The intricacy and variety of social networks also make social network mining a difficult task. There is a need for new analytic methodologies and tools since traditional algorithm software cannot handle such complicated and large volumes of data. This encyclopedia covers topics related to social networks and big data analysis. It discusses the concepts, methods, and difficulties of social networking. The topics covered in this book range widely, from security and intrusion detection to online learning in higher education and healthcare to neural networking and deep learning. **Muhammad Noor et al (2020),** People's opinions and input on a service or product are greatly valued by companies, and social media and networks facilitate this process. Businesses may either expand their consumer base and income by using this information, or they can lose both. It's possible that conventional methods of data analysis won't perform well with the social network's data due to its extreme variability. Machine learning methods facilitate this sort of organizational development. Machine learning is a field that offers methods to automatically learn and extract useful patterns from data without the need for explicit programming. This paper provides a forum for future scholars to identify unexplored areas of machine learning approaches for social network analytics and a summary discussion of useful tools for this and other areas of Big Data research.

Himanshu Shekhar et al (2015), The rapid development of new technologies in recent years has resulted in a deluge of data in many fields, including the sciences, the web, business intelligence, and medicine. To sum up this new way of thinking, the term "big data" was coined. Big data, in contrast to conventional data, reveals features beyond its sheer bulk. For example, huge data has no central location and is completely unstructured. A new infrastructure for collecting, sending, storing, and analyzing data is required by this developing trend. In this paper, we describe a framework for big data analytics in the form of scalable systems, with the goal of offering an overall dimension of the big data system and so advancing the big data issues. The term "big data" is first defined, and then some context is provided. Next, we lay forth a plan for characterizing huge data's underlying structure. Finally, we outline the problems and ongoing research for large data systems.

Jonathan Magnusson (2012), The amount of information available in today's world has grown exponentially in recent years. This is most obviously seen in the telecoms industry. Data transfer via mobile devices is expanding rapidly. Telecommunications companies may use this to learn more about their individual customers. Many other uses may be found for this, including marketing segmentation and the identification of churners (customers who are likely to transfer service providers). Consequently, the analysis and data extraction are really helpful. Social network analysis is one method used in this investigation. Using such techniques provides means of determining the worth of each network subscriber. The purpose of this thesis is to learn more about how social network analysis may be used to the field of telecommunications. Because of the potentially massive size of these networks, any analysis conducted on them must scale linearly with the size of the underlying network. Thus, identifying which algorithms for social network analysis have this scalability is a crucial element of the research.

Maruti M Arer et al (2022), Big data storage in the existing distributed file system has security, availability, and scalability challenges due to the presence of a single point of failure. Block chain is a novel distributed ledger system that has gained prominence with the cryptocurrency market. These days, massive amounts of data are being stored using Block chain as a distributed storage platform. Block chain has storage cost and scalability difficulties as data volumes grow. We propose a block chain-based distributed large data storage solution utilizing IPFS (Interplanetary File System) to address this issue. By breaking up huge files into smaller chunks and distributing them over numerous nodes, IPFS is able to decrease the amount of data that has to be stored. Efficient query retrieval also makes advantage of additional elastic search. The massive data storage, search latency, and precision can all be improved using the aforementioned block chain, IPFS, and Elastic search architecture. It also has a security feature that ensures only authorized people may access the data. Data may be stored, retrieved, and protected effectively inside this system.

Ajmera Rajesh, Siripuri Kiran (2018), There has been a significant increase in the use of long- distance Internet for social networking (communication (e.g., messaging, video collecting, and so on), social security, online business, bank transactions, and a variety of other administrations. These web-based programs need to be shielded and secured in a way that users are comfortable with. However, our digital infrastructure is vulnerable and frequently targeted by attackers. Attack and penetration tools for networks are becoming more widely available. Abnormal and illegal activities in social organizations are actions that are different from what is typically seen in a group of this type. This paper discusses the innovative ordering of anomalies as a function of a variety of characteristics. This paper provides an overview of many methods for spotting and avoiding anomalies, as well as the underlying assumptions and rationales for why they are so prevalent in the world around us. This paper presents a review of many anomaly detection data mining techniques. Abnormal behavior, irregularities, data mining methods, critical analysis, social media.

Go Muan Sang et al (2017), New information technology adoption has been a major issue for companies for decades.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5
***Article Received:** 25 February 2023 **Revised:** 12 March 2023 **Accepted:** 30 April 2023*

Big data is unquestionably an innovative and interesting business idea. New big data platforms are needed to increase data accessibility and enable big data analysis. However, there is still a dearth of big data reference architecture. To better explain the entire design, this paper proposes a new high- level abstract reference architecture for large data systems based on current reference architectures. One previously developed scenario is used in addition to a newly developed use case to validate the new reference architecture.

Hong-Yen Tran et al (2019), This paper provides a thorough overview of methods for doing big data analytics while protecting user privacy. We present well crafted taxonomies that categorize this complex area of paper and provide systematic perspectives. Recent research on hot problems in the field is discussed, providing useful context for readers. We also outline uncharted areas for further paper in the field of privacy-preserving big data analytics. Various privacy-related circumstances are likely to be faced in practice, and this paper can serve as a helpful reference source for the development of contemporary approaches to preserve privacy in these situations.

Dipti Sharma et al (2020), The world is gradually becoming more and more digital. An abundance of user-generated data on social media platforms, which can't be ignored. Since it is hard to read every word, sentiment analysis simplifies things by assigning a positive or negative tone to each section of text. The accuracy of a classification task might vary depending on the method used. The research set out to get an understanding of the several approaches to sentiment analysis. The evaluation also included a comparison of several sentiment analysis methods and how well they performed.

Ankit Didwania et al (2016), There is no guaranteed path from one user's device to another in a mobile social network, therefore users must be tolerant of delays. A store-carry-forward mechanism is at the heart of how it functions. Mobile social networks' emphasis on community is a major strength; being social creatures, people thrive in close quarters. Effective, infrastructure- free communication between gadgets carried by humans is made possible by such a community structure. We have researched the available community identification techniques and isolated the most promising ones for use in mobile social networks. Important characteristics, such as complexity and the kind of community found, have been used in our examination of the many available distributed community recognition methods in mobile social networks. This kind of paper will be useful for learning about the advantages and disadvantages of various existing algorithms. In order for the mobile social network to function as a self-organizing, real-world network on extremely resource-constrained mobile devices, it is crucial that each mobile device be able to independently recognize its own community with as little data, compute, and storage as possible. Very little progress is made in this area because of how difficult it is. Therefore, there is a huge window of opportunity for paper in this sector.

3. RESEARCH METHODOLOGY

The development of computing technology has allowed organizations to handle data quantities that previously required extremely expensive supercomputers. A lot of money has been saved on them recently. Consequently, cutting-edge methods of distributed computing are already commonplace. Companies like Yahoo!, Google, and Face book realized they needed assistance in monetizing the vast amounts of data their offers were collecting, and so Big Data became a top priority. Because of this, emerging businesses are on the lookout for novel ways to collect, store, and analyse massive volumes of data in real time. Performing analysis in real time is essential for making use of the vast amounts of user data being generated. The solutions they came up with have had an impact on the data management industry as a whole. In particular, MapReduce, Hadoop, and Big Table have shown to be the technologies that usher in a new era in data management. One of the most fundamental issues that these technologies will help organizations with is the capacity to process huge volumes of data rapidly, cheaply, and efficiently.

MapReduce

Google developed MapReduce as a means of rapidly processing massive datasets with several tasks in parallel. The "map" part of the system partitions the programming issue or task among several computers, balancing the workload, and allowing for the recovery of failed computers. When the distributed computation is finished, another function called "reduce" brings together all of the individual parts to form a final result. Find out how many pages of a book are written in each of 50 different languages using MapReduce as an example. With its high degree of parallelism, Map Reduce is well-suited for processing massive datasets. These applications are also scalable since they may be deployed to clusters of inexpensive computers. Java is the foundation of Map Reduce. Both a Map task and a Reduce job make up the Map Reduce algorithm. Figure 6 depicts the whole dataflow in a MapReduce environment. Key-value pairs (or tuples) are a fundamental unit of the Map job. These intermediate tuples are then used as input in the subsequent reduces operation. The data is then reduced via the Reduce job. To begin the Reduce work, the preceding Map task must first be finished. Most MapReduce applications are developed in Java. The likes of C++, Python,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5
***Article Received:** 25 February 2023 **Revised:** 12 March 2023 **Accepted:** 30 April 2023*

Ruby, R, etc. may also be used to programme them. Different file and database formats may be processed by these applications. For instance, MapReduce is built on top of the Google File System (GFS).

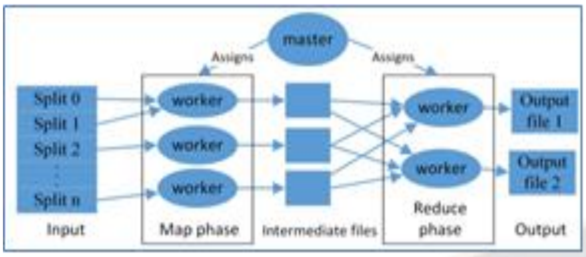

Figure 2. General Dataflow for MapReduce

Standard SQL used by relational DBMSs makes it difficult to create applications like indexing and search, graph analysis, text analysis, machine learning, data transformation, and many more. MapReduce's procedural design makes it simple for experienced programmers to grasp in such settings. The benefit of this approach is that developers do not need to worry about parallel computing because the system handles it automatically. Even though MapReduce was created with programmers in mind, the benefits of premade MapReduce applications and function libraries may be reaped by those who aren't familiar with the language.

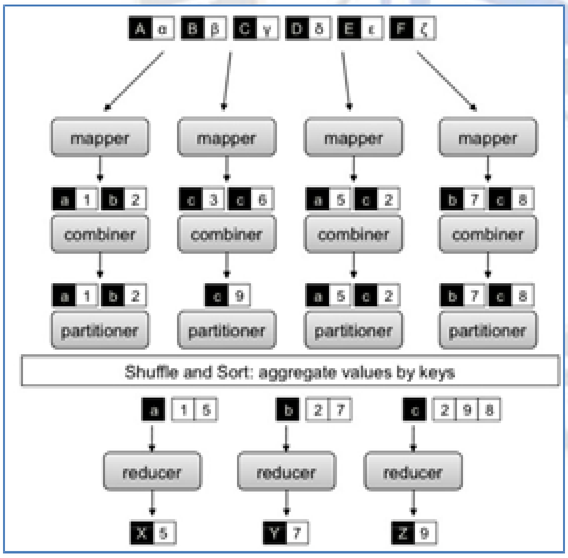

Figure 3. MapReduce Architecture

The MapReduce paradigm has greatly improved scalability by permitting highly parallelized and distributed execution across a large number of nodes. The MapReduce paradigm involves breaking down a single Map or Reduce operation into many smaller tasks that are then distributed among the network's nodes. Whenever a node fails, its tasks are automatically redistributed to other nodes, increasing the system's reliability. In particular, Hadoop is a popular open-source MapReduce implementation because it runs MapReduce atop the Hadoop Distributed File System (HDFS).

HADOOP

Hadoop was created as an open source project with the intention of making the management of large data sets easier. The Apache Software Foundation is in charge of the actual implementation of this. It is made up of numerous smaller paper that all fall under the umbrella of distributed computing infrastructure. The core components of Hadoop include:

- File System (The Hadoop File System [HDFS]) - Storage part
- Programming Paradigm (Map Reduce) – Processing part

All the other side paper either provide supplementary services or are constructed in tandem with the main project to provide additional abstraction layers. Managing the storage of massive amounts of data presents a number of challenges. Even though hard drives' storage capacities have increased dramatically, readers haven't kept pace with the expansion. Time is a major factor in both reading and writing because both are labor-intensive processes. Multi- disk reads can reduce the time needed to complete a single read. It may seem excessive to use only
1/100th of a disc. Only if each of the one hundred
datasets is a terabyte in size and access to them is made available to everyone. Utilizing many pieces of hardware raises the risk of failure, which is an issue in and of itself. Method of making several, identical copies of data on separate devices, with the intention that at least one of these copies will continue to function in the event of a device failure. The main issue is integrating information from multiple sources. Although there are various approaches in distributed computing that can solve this issue, doing so is still difficult. Hadoop easily handles every one of the aforementioned complications. Both the failure problem and the data-combination issue are identified by Hadoop's Distributed File System and are addressed by the Map reduce programming paradigm. Providing a programming approach for computation based on keys and values, Map Reduce simplifies disc records and writes. Hadoop's HDFS is a fault-tolerant distributed file system built to operate on inexpensive servers. It is well suited for applications that require access to big data volumes and has a high throughput. HDFS is built on a master-slave model. With Hadoop, large datasets are automatically partitioned into smaller pieces that can be handled by individual nodes. Figure 4 illustrates this. There is one Name Node in an HDFS cluster that acts as

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5
***Article Received:** 25 February 2023 **Revised:** 12 March 2023 **Accepted:** 30 April 2023*

the master node, keeping tabs on the file system's namespace and the clients' access to the file system; and several Data Nodes, often one per machine, that act as the cluster's storage space. Data Nodes are used to hold the input file's blocks.

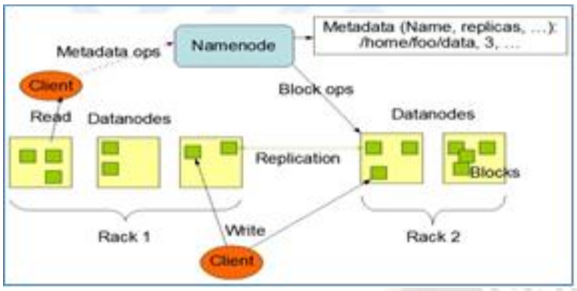

Figure 4. HDFS Architecture

4. RESULT AND ANALYSIS

The initial objective of this project was to compile a comprehensive count of tweets mentioning specific programming languages. The first order of business was to employ several big data techniques to produce the required output. It was picked because it can function with a wide variety of environments, from single-node to clustered computing, from small to large data sets, and from manual to automated methods. The next step was to use MapReduce and Spark to perform the computations on data sets of varying sizes, and record the outcomes. For the reasons described above, we combined Map Reduce and Apache Spark to determine how often specific programming languages were mentioned in tweets. Four of the most frequently mentioned languages in tweets were then selected, and a graph was created. Then, we generated a comparison graph based on the elapsed time and other relevant statistics for the Map Reduce and Apache Spark scripts that we ran on our data sets, collected from streaming on Twitter using Apache Flume. Thereafter, we ran the algorithms again, this time altering the size of our data sets for both Map Reduce and Spark. After then, the time required to complete each programme was determined and recorded. Figure 8 is a visual representation of the information in Table 1. It's also worth noting that the amount of blocks on HDFS is used to automatically generate map tasks in both MapReduce and Spark.

Table 1. Comparison of average runtimes between actual and estimated

	Word Count		**Sort**		**Inverted index**	
Inp ut Dat a Siz e(G B)	**Avg. Act ual runt ime (s)**	**Esti mat ed runt ime (s)**	**Avg. Actu al Run time (s)**	**Esti mat ed runt ime (s)**	**Avg.A ctual runti me(s)**	**Esti mat ed runt ime (s)**
1	168 72	164 84	94.3 6	103 45	268.3 6	240 36
2	360 21	361	164 44	169 3	635.8 1	650 41
3	784 96	712 84	367 12	346 19	1215	123 6.87
5	117 2	120 1.05	605 36	654	2678 85	305 1

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5
***Article Received:** 25 February 2023 **Revised:** 12 March 2023 **Accepted:** 30 April 2023*

10	225 4.06	238 9.64	128 7	130 6.23	6803 24	510 3.61
20	487 2.12	488 4	240 2.36	265 3.26	6869	781 6

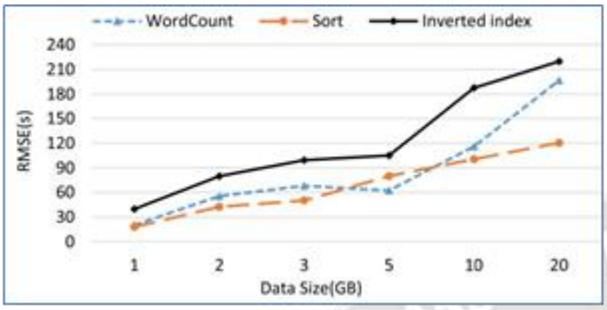

Figure 5. Comparison of average actual and expected duration

Apache Hadoop is a well-known open-source platform for handling large amounts of data. Estimating a job's execution time is crucial for better management in this architecture. In this research, we studied two scenarios for estimating the runtime of a job, one in which the job is running for the first time and no prior information about it is available, and another in which the job has ran before and a profile is provided. To begin with, we formulize each stage of the Hadoop execution pipeline and state them by mathematical formulas in order to compute runtime of a task, taking into account crucial and efficient parameters that have a bigger impact on runtime. In the latter situation, the anticipated runtime is arrived at by consulting the profile or history of a job in the database and applying a weighting algorithm. We choose RMSE and MAPE as our metrics because they are simple and straightforward. The results reveal that when a job's profile or history is available, the average error rate drops to under 8.5% from an initial estimate of less than 12%. We hope to use machine learning methods to further develop the proposed scheme, and we'll use Meta heuristic algorithms to further fine-tune the Hadoop settings. We also plan to investigate methods for predicting how long Spark jobs will run when analyzing streams of data.

5.CONCLUSION

There is a significant need in the market for Big Data at the moment. Hadoop is attractive because it can be deployed on inexpensive hardware and utilized by a wide variety of people with different datasets. The business world is sitting on a treasure trove of data that is now untapped. To put it simply, map reduce is the backbone of Hadoop. The efficient map reduce plan can then be used in conjunction with a number of other optimization approaches to speed up the underlying system or data retrieval. In order to speed up processing, the schedule has been optimized with the use of various techniques such as Quincy, Asynchronous Processing, Speculative Execution, Job Awareness, Delay Scheduling, and Copy Compute Splitting. There are significant obstacles for conventional methods of data processing and storage to keep up with the ever- increasing computational needs of Big Data. The focus of this work was on Map Reduce, one of the fundamental enabling technologies for handling Big Data by means of highly parallel processing on a large number of commodity nodes. Data storage, analytics, online processing, and security and privacy are the four basic categories of Big Data tasks, and each has its own unique set of problems and obstacles for Map Reduce to overcome. In addition, initiatives to enhance and expand Map Reduce to meet identified difficulties are described. This paper gives an overview of the area, helps with planning, and points to future research prospects by highlighting the difficulties encountered while using MapReduce for Big Data.

REFERENES

[1] Bijesh Dhyani, Anurag Barthwal, "Big Data Analytics using Hadoop", International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12, December 2014.

[2] Prof. Devendra P. Gadekar, Harshawardhan S. Bhosale, "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications [IJSRP], Volume 4, Issue 10, October 2014 1 ISSN 2250-3153.

[3] Ms. Gurpreet Kaur, Ms. Manpreet Kaur, "REVIEW PAPER ON BIG DATA USING HADOOP", International Journal of Computer Engineering & Technology (IJCET), Volume 6, Issue 12, Dec 2015, pp. 65-71, Article ID: IJCET_06_12_008 ISSN Print: 0976-6367 and ISSN Online: 0976–6375.

[4] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chandler, "The Hadoop Distributed File System", October 2010, 978-1-4244-7153, IEEE.

[5] Dr. Sanjay Srivastava, Swami Singh, Viplav Mandal, "The Big Data analytics with Hadoop", International Journal of Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue III, March 2016, ISSN: 2321-9653. [6] Varsha B. Bobade, "Survey Paper on Big Data and Hadoop", International Research Journal of

Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 03, Issue: 01, Jan-2016.

[6] Poonam S. Patil, Rajesh. N. Phursule, “Survey Paper on Big Data Processing and Hadoop Components”, International Journal of Science and Research (IJSR), Volume 3 Issue 10, O0ctober 2014, ISSN (Online): 2319-7064.

[7] Rahul Beakta, “Big Data And Hadoop: A Review Paper”, RIEECE -2015, Volume 2, Spl. Issue 2 (2015), e-ISSN: 1694-2329, p-ISSN: 1694-2345.

[8] Ivanilton Polatoa, Reginaldo Reb, Alfredo Goldman, Fabio Kona, "A Comprehensive View of Hadoop Research - A Systematic Literature Review", Journal of Network and Computer Applications, Volume 46, PP 1-25, November 2014.

[9] Dr. Madhu Goel, Suman Arora, “Survey Paper on Scheduling in Hadoop”, International Journal of Advanced Research in Computer Science and Software Engineering [IJARCSSE], Volume 4, Issue 5, May 2014, ISSN: 2277 128X.

[10] J. Heer and S. Kandel, "Interactive analysis of Big Data," XRDS: Crossroads, the ACM Magazine for Students, 19(1), pp. 50-54, 2012.

[11] P. Bhatotia, A. Wieder, R. Rodrigues, U. A. Acar and R. Pasquin, "Incoop: MapReduce for incremental computations," Proc. of the 2nd ACM Symposium on Cloud Computing, 2011.

[12] Y. Chen, S. Alspaugh and R. Katz, "Interactive analytical processing in Big Data systems: A cross-industry paper of MapReduce workloads," Proc. of the VLDB Endowment, 5(12), pp. 1802-1813,2012.

[13] Mrutyunjaya Panda, Ajith Abraham, Aboul Ella Hassanien, “Big Data Analytics: A Social Network Approach”, 2018.

[14] Noor, Muhammad & Haneef, Farah. (2020). A Review on Big Data and Social Network Analytics Techniques.

[15] Himanshu Shekhar, Manoj Sharma, "A Framework for Big Data Analytics as a Scalable Systems" (2015)

[16] Jonathan Magnusson, "Social Network Analysis Utilizing Big Data Technology" (2012) [19]. Maruti M Arer, Praveen M Dhulavvagol, S G Totad, "Efficient Big Data Storage and Retrieval in Distributed Architecture using Blockchain and IPFS" (2022)

[17] Ajmera Rajesh, Siripuri Kiran, “Anomaly Detection Using Data Mining Techniques in Social Networking” (2018)