International Journal of Leading Research Publication (IJLRP)

E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

Big Data Processing in Cloud: Hadoop vs. Spark

Neha Agrawal[1], Jasmeet Kaur [2], Chhaya Porwal[3], Shrishti Gupta[4], Puja Gupta[5]

[1,2,5]Asst. Professor, [1,2,5]Department of Information Technology, Shri G.S. Institute of Technology & Science, Indore.
[3]Asst. Vice President, [3]Barclays, USA.
[4]Research Scholar, [4]Department of Psychology, Algoma University, Brompton, Canada.

Abstract:
There are several sectors in the modern period that produce data on a daily basis, and the quantity of data that is produced is enormous, ranging from terabytes to petabytes. It is necessary to have big data technology in order to manage such a massive volume of data. This technology represents a significant revolution and has had an effect on the trends in applied science. The Hadoop system uses MapReduce in parallel across several nodes, which allows for the analysis of massive amounts of data. Both Map and Reduce are two of the most important functionalities of the MapReduce framework, which is used to store the vast amounts of data and information that HDFS contains. Spark was developed as a solution to the several shortcomings of MapReduce. It is capable of managing real-time data streams and performing queries in a short amount of time. DAG and RDD techniques form the foundation of the Spark framework. The purpose of this study is to make a comparison between the two fundamental characteristics of Hadoop and Spark, which will serve as the basis for the performance assessment that will be carried out.

Keywords: Big Data, Resilient Distributed Datasets, MapReduce, Spark, DAG, HDFS, Hadoop.

Introduction:

The use of traditional software tools and methodologies, including structured, semi-structured, or unstructured databases, as well as structured or semi-structured datasets, poses difficulties in the collection, organization, and analysis of substantial data volumes. Big data[1] refers to the magnitude, complexity, and growth rate of a database that makes it difficult to access, organize, and analyze the information. The phrase "data" encompasses a wide range of information kinds, including photos, videos, text documents, audio files, web pages, log files, and more. The phrase "big data" signifies a nascent paradigm in data management. Significant amounts of data provide volume, variety, and velocity, which are crucial for complying with data privacy rules. Confronting the obstacles and intricacies of big data necessitates solutions that surpass those used in the past, since former approaches are insufficient. Employing conventional software tools and methodologies, including structured, semi-structured, or unstructured databases, together with structured or semi-structured datasets, presents challenges in the collection, organization, and analysis of substantial data volumes. Big data denotes the substantial amount, intricate nature, and rapid expansion of a database that complicates data access, organization, and analysis. The word "data" encompasses a diverse array of information kinds, including photos, videos, text documents, audio files, web pages, log files, and more [2]. The phrase "big data" denotes an innovative paradigm of data. Substantial data quantities provide volume, variety, and velocity, which are essential to meet data privacy criteria. The solutions previously used are inadequate for addressing the challenges posed by big data.

An effective approach to addressing issues related to the big data model may be realized via the use of distributed and large-scale processing techniques. Large data approaches[3], both those used now and

those used in the past, might thus be useful in the process of resolving security issues. In order to create the politics and future work company, institutions, organizations, and government sectors have increased their attention on big data. In order to assist development for the best solution, future work, and advancement in advance politics, big data has been a possible solution concept for speed, deeper interference, and more accurate interference. Big data has also been a potential solution idea for more accurate interference.

In light of this, it is of the utmost need to acquire knowledge about the definition and comprehension of big data . The advanced way of database should be offered in order to safeguard spark vulnerabilities in comparison to conventional databases such as relational databases. This is due to the fact that relational databases have a shortcoming when it comes to incorporating huge data into NoSQL databases such as non-relational databases [4]. Big data is nothing more than a word that was created to scale out the current computing capabilities of the ordinal approach within a time constraint for massive, enormous, and complicated data that cannot be handled by traditional techniques [5, 6]. There is a wide body of literature that has shown the potential of big data in the face of challenges or obstacles to the digital world in terms of processing, storing, analyzing, recovering, and even visualizing the rapidly expanding information. The provision of a vast quantity of data presents a multitude of options with big data. The foundation of this work is comprised of big data technologies that allow for the identification of the programming language via the use of Apache Spark, which is entirely compatible with the application. The 5 V's of Big Data have been the subject of debate in the literature study that has been conducted.
already, the data that is troublesome is in the form of petabytes, but according to the projection, it may expand into zettabytes (ZB) in a few years. This is a significant increase from the information that is already available.

Volume - Predictions indicate that data may escalate to zettabytes (ZB) within a few years, from the already troublesome volume of petabytes [8].
Velocity refers to the pace of data acquisition and data transmission. The substantial volume of data and its continuous movement exacerbate the issues associated with prior analytics, resulting in heightened dependence on real-time data.
Variety - Data is derived from several sources rather than a one source, including messages, webpages, emails, sensors, etc.
Veracity - The primary objective of this dimension is to eliminate uncertainty in data, often caused by noise and abnormalities.

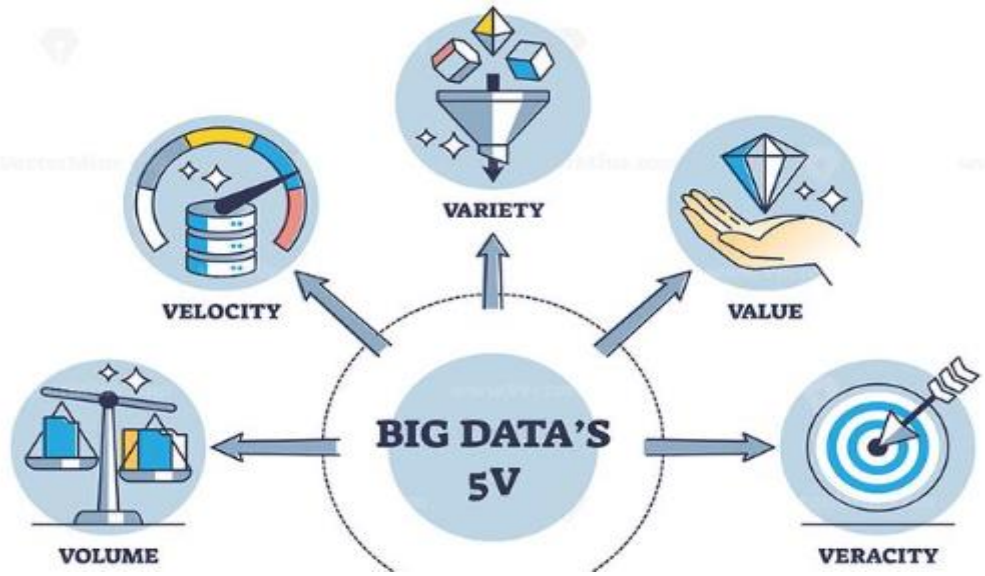

Figure 1: 5V of Bigdata

The sixth dimension of big data pertains to comprehending the advantages it offers to various stakeholders within a company, hence enhancing business value.

This Value dimension pertains to elements that address inquiries such as which business actions may capitalize on big data insights, the optimal timing for decision-making, and the direct beneficiaries of such decisions.

The analysis conducted encompasses the examination of features[9], the administration of storage and cloud computing, the processing of extensive data sets, and ultimately, the extraction of insights from the abundant data available. In contemporary society, the safeguarding of digital information is paramount. In the management of large data sets, security is paramount, since the data includes sensitive information, confidential keywords, and passwords, the compromise of which might result in severe consequences. Therefore, security is paramount when considering large data and cloud computing. A range of techniques, such as Node Authentication, encryption, access control, honeypot nodes, and others, may be used to achieve security. Implementing this system may encounter several challenges, including those related to data storage, speed, security, processing, transmission, visualization, architecture, integration, and quality, among others. The integration of cloud computing and big data has applications across several domains, including management and finance.

ANALYTICS OF BIG DATA

Big Data in the cloud refers to the extensive amount of information, perhaps including dozens of terabytes or even petabytes. Therefore, using big datasets in a conventional database management system[10] on a local computer poses considerable difficulties. The ability to augment storage, exhibit data, oversee, and acquire information is becoming progressively laborious and expensive; therefore, using cloud computing may be considered the most appropriate answer. Numerous major global corporations are consolidating all of their data in the cloud. By using integrated cloud capabilities or installing their own functions on the cloud, these organizations may analyze vast databases, facilitating the discovery of previously undiscovered information. Consequently, the cloud must provide a varied array of data structures, analytical methodologies, and instruments. Organizations inherently benefit from access to vast quantities of data that can be processed very instantaneously.

Bigdata Processing:

To process anything, there are four essential needs that must be met[11].

1. The capacity to load the data in a short amount of time is the most important criterion.
The processing of queries is quick.
3. The effective usage of available storage space 5.
4. High degree of adaptability to a task that is very dynamic

Providing MapReduce software is among the methods through which cloud service[12] providers support us in efficiently addressing all four requirements. Both Azure HDInsight and Amazon EWS offer MapReduce frameworks to their clients. As a parallel programming approach, the framework provides significant support in processing. Rather than increasing the storage capacity or processing capability of a server or computer, the MapReduce framework incorporates additional servers and computers. The fundamental concept is that we do not increase in scale but rather expand laterally. This serves as the fundamental principle.

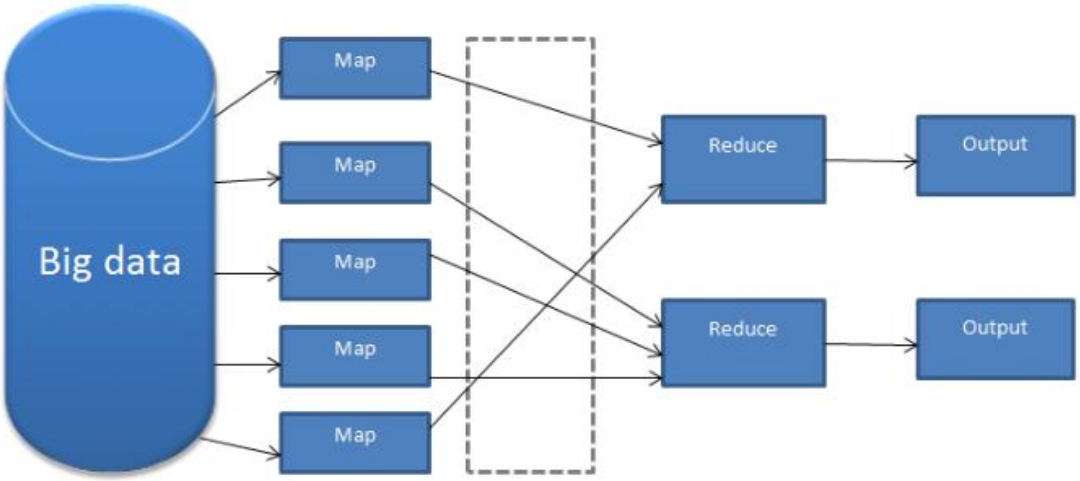

Figure 2: Map Reduce Architecture

The effectiveness of a job is increased by the use of Map Reduce, which involves breaking it down into stages that are then carried out in parallel. As the name implies, the operation is pretty straightforward; the first word Map is used to "map" the smaller jobs and give them the proper key value pair. If we have unstructured data, such as text, for instance, the key may be any word, and the value could be the number of times that word appears in the data. This brings us to the reduce function. For the purpose of providing the final result of the computational work, the reduce function is responsible for performing collection and combination of the output given by the "map" function. This is accomplished by combining all values that have the same key value. This presents a significant benefit due to the fact that cloud architecture is very quick, and when combined with parallel processing, the performance is unparalleled in comparison to that of a standard local computer. In situations when processing speeds are as fast as they are, we are able to do data analysis in real time while simultaneously receiving the result in real time. This kind of technology, when implemented on the cloud, is very beneficial, and Big Data with high velocity and large volume, as well as organizations and exchanges such as the NASDAQ, BSE, and NSE, may all profit from it. The storage, analytics, and processing are all carried out with more efficiency and at a lesser cost as compared to the conventional computers that are used today.

ANALYSIS BASED ON SPARK:

The Apache Spark technique is a model-based approach that is used for the purpose of analyzing massive datasets. The Hadoop MapReduce paradigm UC Berkeley AMPLAB served as the inspiration for the development of this framework.

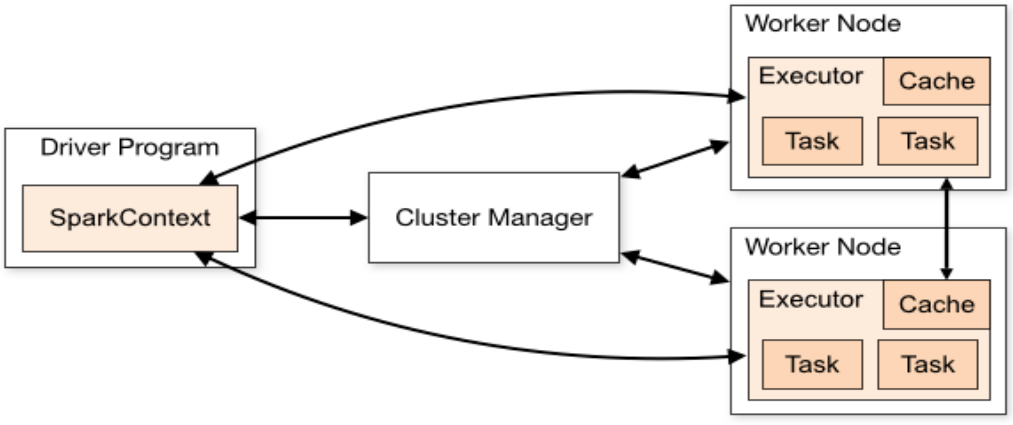

Figure 3: Architecture of Spark framework

Figure 2 illustrates the architecture of the Apache Spark framework. The Apache Spark framework is comprised of many components, which are as follows: the program of driver, initiators, cluster director, and the HDFS. The driver program is the most important software that the spark has. During the process of starting up the Spark application, a Spark Context is generated, which plays a vital part in the whole execution of the task [20]. Whenever the Spark Context application establishes a connection with the cluster manager, the resources are handled across the cluster. In order to store the information about the application and to execute the logic implementation, Cluster managers are given.
Within the Spark, there are two fundamental ideas that may be summarized as follows:

1. Resilient Distributed Datasets (RDD)
2. A Directed Acyclic Graph, often known as a DAG.

Resilient Distributed Datasets (RDD):

As the core notion of spark, "Resilient Distributed Datasets" are used to collect pieces that are able to withstand errors and continue in a parallel way. Once RDDs have been formed, they cannot be altered in any way. It is not even possible for them to modify this despite their power to transform and act. Reorganizing the calculations and improving the data processing are both made easier with the use of this data. The fact that they are resistant to flaws is due to the fact that they use RDD information to rebuild and decide themselves. It is [18] Changing the existing data lists or the files in HDFS might be used to generate this RDD. Both of these options are viable options. The default value is the value that a spark programmer will use in the event that some compartments of RDD contain incorrect information. Both of the following types of approaches are carried out by RDD:

- A new RDD is generated each time alterations are applied to an old RDD to get a singular value. This is termed transformation. The analysis speed is moderate, resulting in a lack of quick change in its computational analysis. They are deemed implemented when a program is performed on this. Transformation functions offered include Map, Filter, ReduceByKey, FlatMap, and GroupByKey.
- Action: When the action methods are used over RDDs, just a single value is taken for inspection and execution. This is because the action methods are more efficient. While an action method is being executed, the calculation of data processing is being carried out, and a resulting value is being sent. Take, reduce, collect, count, foreach, and count by key are some of the few action methods that are among the fundamental.

Directed Acyclic Graph(DAG):

Spark [23] is a sophisticated approach that works for cyclic data flow, and the DAG engine of Directed Acyclic Graph is the way that supports it. It is often believed that a DAG is made up of stages of jobs that need to be carried out at the same time throughout the Spark cluster. Sparks, on the other hand, generates DAGs that may be made up of any number of steps, while the DAGs that are generated by MapReduce only have two steps since (1) Map and (2) Reduce are the steps that are involved. When one of the phases is completed, it enables the completion of a simple job, as opposed to a complicated work that requires the completion of numerous steps in a single run.

An analogy may be drawn between the Spark model and Hadoop in terms of the common way of data analysis on single and distributed nodes. One of the benefits of using this approach is that it keeps the data on the memory disk, which in turn boosts the speed at which the data is processed by doing computations in memory. In order to get access to HDFS, Apache Hadoop is run on the node of Hadoop that is now sitting at the top of the hierarchy. In addition to this, it is able to handle data streaming as structured data addition in Hive by using twist on Twitter [21].
At its most basic form, Spark[23] is comprised of the following components, each of which performs a variety of activities, as seen in figure 4

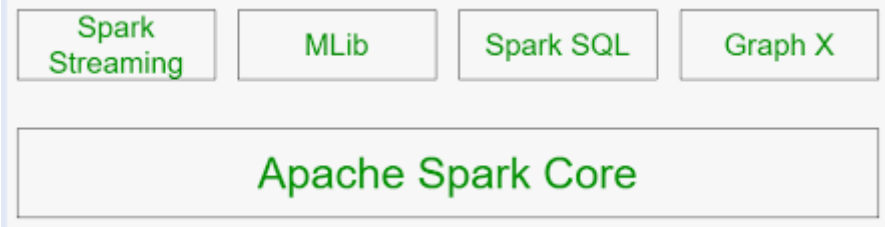

Figure 4: Apache Spark Components

Apache Spark is an open-source project that was developed by the company to solve the problem of massive amounts of information. It offers a decentralized processing platform that allows the handling of enormous amounts of information that is very complicated and extremely large-scale in a manner that is quick, efficient, fault-tolerant, and adaptable. A number of essential components, including Spark SQL, Spark Streaming, MLib, and GraphX, contribute to the overall magnificence of this system. These components are what make it so remarkable. Sparke SQL is used for the purpose of querying data using SQL, Spark Streaming is utilized to simplify stream processing, MLib is a library that contains machine learning methods, and GraphX is utilized for graph analysis [22].

The Spark Core, A.

In this approach, the Spark Core is the most fundamental and fundamental component.

A. The key characteristics of the core include the functions that are associated with input/output (I/O), dispatching, and the layout of the distributed task system.
B. Spark SQ**L** Spark SQL enables the ability to conduct SQL queries on Spark by allowing the usage of traditional business intelligence and visualization tools. In order to offer support for structured and semi-structured data, an enhanced RDD data idea approach has been devised.
C. Streaming using Spark: The processing of real-time data is often accomplished via the usage of Spark streaming. It makes use of DStream in order to facilitate the processing of real-time data from RDD.
D. The GraphX Spark : There is a new graph called "The Resilient Distributed Property Graph" that was presented by GraphX. This graph has the capability to connect with every edge and vertex. A number of operators are included in GraphX. These operators include aggregate messages, subgraph and join vertices, and an efficient form of the Pregel API. In addition, there are graph builders and algorithms that may ease the work associated with graph analytics.
E. MLlib : Machine learning library (MLlib) is a machine that has utilities and common learning techniques such as regression, arrangement, cooperative filtering, and dimensionality reduction respectively.

Spark's programming languages and languages

For the purpose of supporting the operation of massive data analysis, Spark makes use of many libraries of language. According to the illustration in figure 5a, Scala is used in order to develop code that enables Spark to be executed automatically.

In addition, as indicated in figure 5b [13], there are at least three more programming languages that exist in addition to Scala. These languages include Java, Python, and R. The size of the same amount of code written in Java is more than the size of Scala. According to [14], various ways are transitioning from Java to Scala in order to improve the efficiency and reliability of their operations.

Spark has the ability to read out information that is kept on many tools such as Hadoop HDFS, Mesos, Mongo DB, Cassandra, H-Base, Amazon S3, and the data source from the cloud. Spark is able to read out the data from every location. The image illustrates the many different entry points that are available to ignite [15].

Comparison with Hadoop:

For the goal of controlling the amplification, swiftness, and discrepancy of the data, the number of distributed systems in the market is quickly expanding with each passing year. This is being done for the purpose of managing the data. However, Hadoop and Spark are the two most prominent databases in the world. It is thus quite difficult to determine which option is the best. However, we are obligated to do so. There are higher optimizations in the case of Spark, which means that it will perform better than Hadoop. This is because the amount of time it takes to access the disk in one second is also more, and the amount of bandwidth that is used by memory is also greater. This is mentioned in [17][24].

According to the information shown in table 1, the comparison between Spark and Hadoop is based on the many characteristics of each of these technologies. The majority of the characteristics are those that have been picked from prior research that has compared Spark with Hadoop [23][25]. One of the most significant distinctions between Hadoop and Spark is the approach or method that is used to handle the data [19]. In the case of Spark, processing is carried out in the memory, but in the case of Hadoop, reading and writing are carried out from the disk. The speed at which processing is performed is concurrently impacted by the differences in the processing techniques they use. The fact that all of the processing in the case of Spark is done in memory is the reason why it requires less time (i.e., it requires 100 times less time), but in the case of Hadoop, more time is needed. It may be concluded that Hadoop is slower than Spark [17][26.

In contrast to Spark, which provides caching in data memory, Hadoop does not offer the functionality of data caching. Spark has a reduced latency of computing because it has a high level of data interactivity, while Hadoop has a greater latency of computing because it has less data interaction [19.

Consequently, when all of the points and facts that have been described above are taken into consideration, it is extremely difficult to select one over the other. However, both Spark and Hadoop are best in their own ways, such as Spark being the best in terms of speed and Hadoop being the best in terms of the quantity of data that can be handled [22].

FEATURES	HADOOP	SPARK
Processing Mode	Batch	Batch and Stream
Scalability	Horizontal	Horizontal
Message Delivery Guarantees	Exactly once	Exactly once
Computation Mode	Disk-based	In memory
Auto-scaling	Yes	Yes
Iterative Computation	Yes	Yes
SPEED	SLOW	FAST
AMOUNT OF DATA CAN BE PROCESSED	MORE	LESS
SECURITY	MORE SECURE	LESS SECURE
COST	HIGH	LOW
PERFORMANCE	FAIR	GOOD

On the other hand, I believe that Spark is superior than Hadoop in terms of its speed, ease of use, and simplicity. In addition, Spark delivers the findings in a very short amount of time, which is a highly helpful feature for firms that need to be able to witness development after every brief interval. In a similar vein, in the case of spark, no one is necessary to write each function, which is the most significant benefit of this technology. Additionally, spark displays a greater level of data interaction. Spark is more sophisticated than other clusters because it offers batch processing, streaming, and machine learning in the most efficient manner possible inside a single cluster.

CONCLUSION:

During this period of technological advancement, a variety of technologies, such as Hadoop MapReduce and Apache Spark, have been created in order to investigate the processing of large amounts of data. The importance of Apache Spark as an alternative to the MapReduce process to support logic and ad-hoc queries has increased in recent years. Spark has received a great deal of acclaim in a variety of domains, including data mining, information retrieval, machine learning, and image retrieval, among others.

In spite of this, the quantity of data that has to be processed increases, which means that older and more conventional techniques of data processing become less effective. In order to conduct an efficient analysis of large amounts of data stored in HDFS, this study investigates the novel approach of Apache Spark model, which serves as an alternative to the framework of Hadoop MapReduce. Spark has the potential to improve the speed of computation of iterative algorithms and reduce the amount of time it takes to do so in comparison to older approaches. For the purpose of analyzing the processing of large amounts of data, it also offers a framework that is highly accessible, extremely efficient in its working performance, and fault resistant. Based on the findings of prior studies that compared Spark with Hadoop, it was determined that Spark is a more superior technology to Hadoop in terms of both its speed and its memory use [23]. In addition, the results of our study demonstrate that Spark is capable of performing the machine learning job, as well as streaming and batch processing procedures. Spark is superior to other technologies in terms of security, speed, and memory use, as shown by the comparison presented above. In other words, it is the platform that allows for the processing of an enormous volume of data.

REFERENCES:

[1] Ranjan, A. and Ranjan, P., 2016, April. Two-phase entropy based approach to big data anonymization. In 2016 International Conference on Computing, Communication and Automation (ICCCA) (pp. 76-81). IEEE.

[2] Gupta, P. and Kulkarni, N., 2013. An introduction of soft computing approach over hard computing. International Journal of Latest Trends in Engineering and Technology (IJLTET), 3(1), pp.254-258

[3] Chaturvedi, A. and Gupta, P., 2021. The Cloud: Features, Challenges and Scope. International Journal of Progressive Research in Science and Engineering, 2(8), pp.466-471.

[4] Kushwaha, U., Gupta, P., Airen, S. and Kuliha, M., 2022, December. Analysis of CNN Model with Traditional Approach and Cloud AI based Approach. In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 835-842). IEEE.

[5] Rajput, A., Gupta, P., Ghodeshwar, P., Varma, S., Sharma, K.K. and Singh, U., 2023, June. Study of Cloud Providers (Azure, Amazon, and Oracle) According To Service Availability and Price. In 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN) (pp. 1177-1188). IEEE.

[6] Gupta, P., Arya, N., Singar, C.P., Chaudhari, A., Singh, U. and Gupta, S., 2025. Safety of Pedestrians in AI-Optimized VANETs for Autonomous Vehicles via Real-Time Vehicle-to-Vehicle Communication. In AI-Driven Transportation Systems: Real-Time Applications and Related Technologies (pp. 169-181). Cham: Springer Nature Switzerland.

[7] Agrawal, Divyakant & Das, Sudipto & Abbadi, Amr. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM International Conference Proceeding Series. 530-533. 10.1145/1951365.1951432
[8] Alsghaier, Hiba & Al-Shawakfa, Emad. (2018). An empirical study of cloud computing and big data analytics. International Journal of Innovative Computing and Applications. 9. 180. 10.1504/IJICA.2018.10014870
[9] Yadav S., Sohal A. (2017) "Review Paper on Big Data Analytics in Cloud Computing" in International Journal of Computer Trends and Technology (IJCTT) V49(3):156-160, July 2017. ISSN:2231-2803.
[10] Hariharan, U. & Kotteswaran, Rajkumar & Pathak, Nilotpal. (2020). The Convergence of IoT with Big Data and Cloud Computing. 10.1201/9781003054115-1.
[11] Alyass, Akram & Turcotte, Michelle & Meyre, David. (2015). From big data analysis to personalized medicine for all: Challenges and opportunities. BMC medical genomics. 8. 33. 10.1186/s12920-015-0108-y.
[12] Wang, Lizhe & von Laszewski, Gregor & Younge, Andrew & He, Xi & Kunze, Marcel & Tao, Jie & Fu, Cheng. (2010). Cloud Computing: A Perspective Study. New Generation Comput.. 28. 137-146. 10.1007/s00354-008-0081-5.
[13] Turcotte M., Alyass A. (2015) "Form Big Data Analysis to Personalized Medicine for all: Challenges and opportunities" Article in BMC Medical Genomics. DOI:10.1186/s12920-015-0108-y
[14] D. S. Terzi, R. Terzi, and S. Sagiroglu, "A survey on security and privacy issues in big data," in Internet Technology and Secured Transactions (ICITST), 2015 10th International Conference for, 2015, pp. 202-207.
[15] A. K. Jumaa, "Secured Data Conversion, Migration, Processing and Retrieval between SQL Database and NoSQL Big Data," DIYALA Journal for pure sciences, vol. 14, no. 4, p. 68, October 2018.
[16] N. P. KS and T. Pratheek, "Providing anonymity using top down specialization on Big Data using hadoop framework," in India Conference (INDICON), 2015 Annual IEEE, 2015, pp. 1-6.
[17] H. K. Patil and R. Seshadri, "Big data security and privacy issues in healthcare," in Big Data (BigData Congress), 2014 IEEE International Congress on, 2014, pp. 762-765.
[18] M. Al-Zobbi, S. Shahrestani, and C. Ruan, "Sensitivity-based anonymization of big data," in Local Computer Networks Workshops (LCN Workshops), 2016 IEEE 41st Conference on, 2016, pp. 58-64.
[19] M.A. Khan, M. F. Uddin, N. Gupta. Seven V's of Big Data Understanding Big Data to extract Value.In: Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education, 2014, 1–5.
[20] L. Hbibi and H. Barka, "Big Data: Framework and issues," in Electrical and Information Technologies (ICEIT), 2016 International Conference on, 2016, pp. 485-490.
[21] Marchiori, Massimo. (2017). Learning the way to the cloud: Big Data Park. Concurrency and Computation: Practice and Experience. 31. 10.1002/cpe.4234.
[22] Brevini, Benedetta. (2015). Book Review: To the Cloud: Big Data in a Turbulent World. Media, Culture & Society. 37. 1111-1113. 10.1177/0163443715596318a.
[23] Shaikh, E., Mohiuddin, I., Alufaisan, Y. and Nahvi, I., 2019, November. Apache spark: A big data processing engine. In 2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM) (pp. 1-6). IEEE.
[24] Rathore, P., Gupta, P., Jain, S. and Shrivastava, Y., 2022. A Study of the Automated Vehicle Number Plate Recognition System. i-manager's Journal on Pattern Recognition, 9(2), p.30.

[24] Singh, U, Gupta, P., Shukla, M., Sharma, V., Varma, S. and Sharma, S.K., 2023. Acknowledgment of patient in sense behaviors using bidirectional ConvLSTM. Concurrency and Computation: Practice and Experience, 35(28), p.e7819.
[25] Singh, U., Gupta, P. and Shukla, M., 2022. Activity detection and counting people using Mask-RCNN with bidirectional ConvLSTM. Journal of Intelligent & Fuzzy Systems, 43(5), pp.6505-6520
[26] Gupta, P., Shukla, M., Arya, N., Singh, U. and Mishra, K., 2022. Let the Blind See: An AIIoT-Based Device for Real-Time Object Recognition with the Voice Conversion. In Machine Learning for Critical Internet of Medical Things (pp. 177-198). Springer, Cham.