Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 08 Issue 01
2024

Review Article
https://doi.org/10.56979/801/2024

Evaluation of Big Data Tools: A Comparative Study

Prince Hamza Shafique[1*], Mubashar Hussain[2], Salahuddin[1], Meiraj Aslam[1], Muhammad Sufyan[1], and Syed Shahid Abbas[1]

[1]Department of Computer Science, NFC Institute of Engineering and technology, Multan, Pakistan.
[2]Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan.
[*]Corresponding Author: Salahuddin. Email: msalahuddin8612@gmail.com

Received: April 26, 2024 **Accepted:** September 29, 2024

Abstract: Due to increasing usage of internet huge volume of data is available online. Main source of this gigantic volume of data are social networking sites like Facebook and tweeter etc. It is difficult to handle this huge volume of data. This growing data affects business badly. This data is called Big Data. There are many tools for Big data analytics in this research our focus is on four Big data tools 1) Hadoop, 2) IBM InfoSphere BigInsights, 3) High Performance Computing Cluster (HPCC) and 4) Apache Spark. In this research I have studied architectures, file systems, shortcomings and solutions of those problems. In future this research could be enhanced by running an algorithm on all these tools and then comparing the results. These tools can also be compared by setting some parameters.

Keywords: Java Execution Environment; Apache Software Foundation; IBM Workbooks; Computing Cluster.

1. Introduction

Information is rapidly evolving and it is very difficult to process such a large amount of information. Since the value of data is rapidly increasing, it is becoming difficult to process this data. This volume increase has a significant impact on business. The main source of this information is the sharing on social media sites such as Twitter and Facebook. This data is called big data [1]. Equipment used [2]. Diversity:

The generated data is diverse and partly generated from various sources such as web directories, web pages, emails, social media sites, documents, and sensor equipment. All data is completely different and includes old data, partial data, structured data, and such information. Volume:

The word "volume" in the word volume means volume. SNs (social networks) provide terabytes of data every day, and it is difficult to manage such large data using existing systems [1,3]. Speed:

When it comes to big data, speed refers to the speed of data through various methods. This feature is not only related to the amount of incoming data and the amount of data flow [1,3]. Variability:

Variability refers to changes in the data flow. As the processing power of social networking sites increases, data loading becomes more difficult. The flow of information causes the information to be loaded until the end when a clear decision is made [1]. Complexity:

It is difficult to change, match, combine and clean the data in different parts of the system. It is also important to connect and interact with multiple data connections, hierarchies and relationships [1].

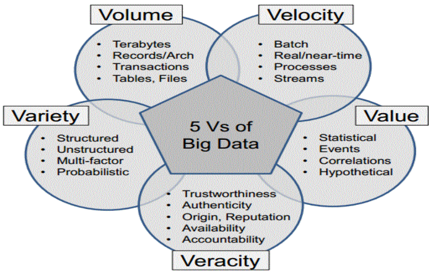

Figure 1. 5Vs of Big Data

Meaningful results can be obtained from clean data by running many queries on the stored data and can also be ranked according to the desired range. Such reports help to reveal organizational differences and people can change their minds accordingly [1]. Quantitative Analysis of Data In this study, we focus on four famous big data sources such as Hadoop, IBM InfoSphere BigInsights, HPCC and Apache Spark.

1.1. Hadoop

Hadoop is an open source software architecture used to store data and process large volumes of hardware. Apache Hadoop is a trademark of Apache Software Foundation. Document) System): - Document that provides information about products and provides more bandwidth from the team. And use this source to configure the user application. Since the theory of universal facts is accepted, these sins need to be evaluated according to these standards[2]. Cut was working at Yahoo! at the time. and named it after his beloved son who played elephant.) MapReduce Engine [4]

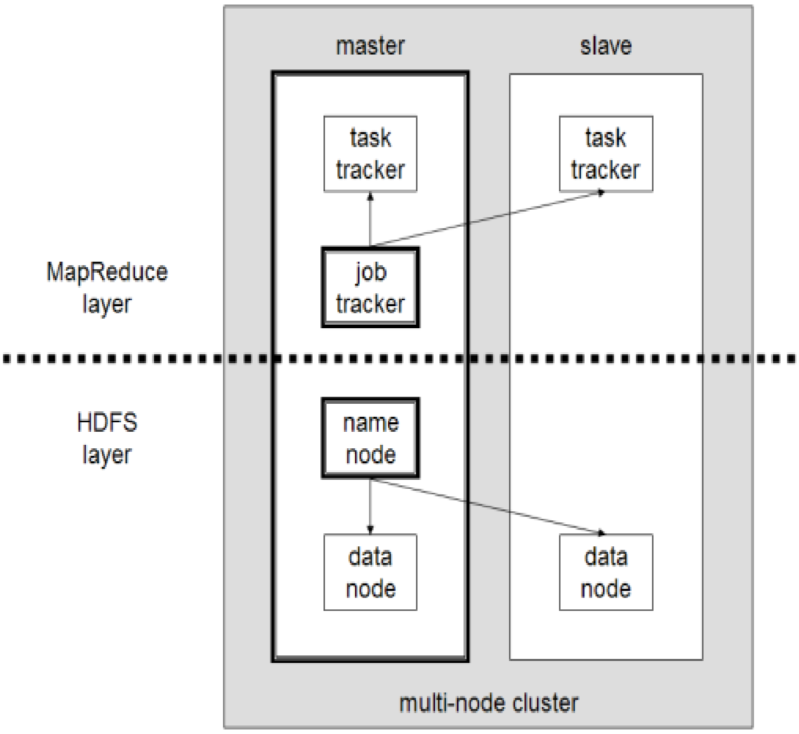

Figure 2. A multi-node Hadoop Cluster

The operational level abstraction and data system are provided by the Hadoop Common Package. It contains the Java Archive (JAR) scripts and files needed to install Hadoop. This package provides support information and source code for projects involving the Hadoop community [2]. A single master contains data nodes, name nodes, task trackers, and task trackers. A slave or worker node serves as both a data source and a processing node, but in non-standard implementations it is also possible to have only a data slave and only counting slaves. Hadoop requires JRE (Java Execution Environment) 1.6 or later. Secure Shell (ssh) must be installed on nodes in the group [5] during startup and shutdown. System failure and data loss HDFS creates a view of the name of the memory structure. In this way, a job tracking server can manage job scheduling [2]. GraphX extends Spark RDD by creating flexible distributed attribute graphs. It is a multi-graph representation with custom attributes for each edge and vertex. GraphX also includes a growing set of graph algorithms and generators to facilitate graph analysis tasks.

It is a query engine that can be used to perform interactive SQL queries on large databases. Allows users to switch between correct questions and response times.

1.2. Tachyon

A memory-oriented decentralized archive system that allows reliable archive distribution at the memory speed of a cluster, just like MapReduce and Spark. The working configuration file is cached in memory, thus avoiding disk hits and multiple reads of the file. This way, the cached data can be accessed from memory quickly by different operations/queries and transactions. Spark can be used to monitor data stored in the Cassandra database via the Cassandra Connector and perform data analysis on this data [18].

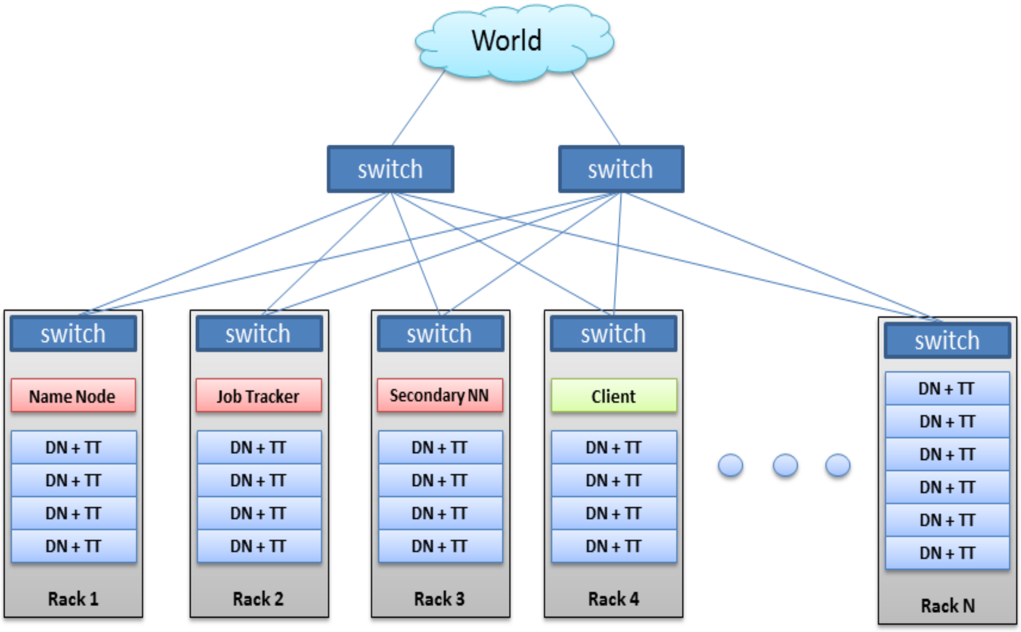

Figure 3. HDFS (Hadoop Distributed File System)

2. Materials and Methods

This section describes the history, function and background of big data. Software framework. Big data processing and faster processing are two important functions of Hadoop. information. In its early days, scientific results were overwhelmingly negative. However, as the number of web pages increased from a few to millions, automation was needed. Search engines began to grow rapidly (Yahoo, AltaVista, etc.) and web browsers were born. The main purpose is to use distributed information and computing on different computers to quickly reach web search results. In this way, many tasks can be done in parallel. At that time, Google was also successful. It is based on the same concept as Nutch. , The web browser, processing and distribution computing part was developed as Hadoop, and the web browser part was kept as Nutch. The name Hadoop is derived from Cuting's son's toy elephant. In 2008, Yahoo released Hadoop as an open source project, and the Apache Software Foundation (ASF) now manages and maintains the Hadoop family of systems and technologies. It is an international community of software contributors and developers [19]. information. As the volume and variety of data continues to increase, Hadoop has become indispensable to many organizations thanks to the use of social media and computer technology.

Other factors include

Low cost. It opens up and uses hardware to store large amounts of data. Thanks to the decentralized business model, a lot of information can be processed quickly. You can use additional codes to make the work more powerful. The system can be expanded by adding nodes, which requires very little management. Unlike traditional relational data, data does not need to be preprocessed before being stored. This can include intangible information such as text, video, and images. As much data as needed can be stored and we can choose how to use it later. Applications and data processing are protected by the inconsistency of device collisions. If a node fails, tasks are transferred to other nodes to ensure that the partition calculation does not fail, maintaining multiple copies of all data. Self-healing is implemented by the strangely known NameNode.

Hadoop is also considered as the next big data in many organizations, but the main purpose of Hadoop is to access more millions of web pages and return the relevant results[19]. It is built on Apache Hadoop, which uses IBM Big Data, and has parallelism, security, performance and management capabilities. IBM Workbooks is an excellent reference language for data exploration. BigInsights can analyze data with its natural structure, without insisting on a model/idea, and allows for rapid analysis. divided. The Basic edition is the entry-level free edition that can help companies learn big data. When users are ready to analyze data, it is best to upgrade to the Enterprise Edition; a Hadoop cluster can be set up in about 30 minutes with a minimum usage of $0.60 per hour per cluster. Both the Enterprise and Basic editions include a developer sandbox where users can create new business transaction systems.

2.1. Apache Spark

It is an open-source big data model. Its main features are speed, ease of use, and advanced analysis. It was first developed by UC Berkeley's AMPLab in 2009 and later developed by open source Apache in 2010. To meet the needs of processing big data, a large number of datasets (image files, text files, etc.) and data (time-course data and batch data) have different properties and combinations here. Complete the framework. Write quickly in Python or Java. It has over 80 active workers and you can request information in the shell. These functions can be used offline or they can be linked to run on a reference file. . Spark's performance is faster than previous big data technologies for two main reasons: near-term processing capability and in-memory data storage. . It provides higher level APIs to improve developer performance and reliable design for big data.

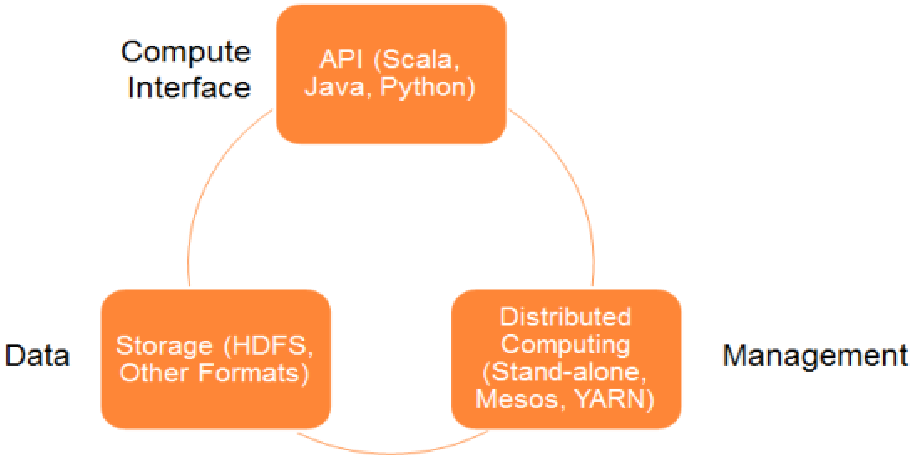
Compute Interface
API (Scala, Java, Python)
Data
Storage (HDFS, Other Formats)
Distributed Computing (Stand-alone, Mesos, YARN)
Management

Figure 4. Spark Architecture

3. Discussion

Big data has many strengths but also some disadvantages. This section discusses the shortcomings of big data. Also, the recipe and process need to be repeated every time, even if there are minor changes in the ideas.

1. Incremental Computing: Many online collections evolve gradually as new items are added and existing items are removed or released. It seems that the results of such results and the rapid results in processing additional information (like Google's PageRank) are leading to the latest changes today, but the gains from simple projects have not increased at the expense of social structure. MapReduce and, more importantly, require programmers to use special dynamic implementations, increasing the algorithms, ultimately increasing the complexity of the algorithm and code. For example, someone collecting data and information on LinkedIn is sufficient in MapReduce. If a person is following a group, now the user’s profile is only updated, so the reference will be counted again, which is not a good way [20].
2. Low-level procedures and standards. Using this tool requires skill in low-level operations.
3. Strict batch
4. Hadoop processes the job serially, thus reducing its speed [21].
5. Time Drift

Time skew is the phenomenon that the clock signal sent by the clock in the synchronous circuit arrives at different instances at different times [21].

Due to the distributed environment of Hadoop, it is difficult for a cluster to distinguish between faulty and failed nodes. Many innovations have been made in areas such as error tracking and analysis.

Hadoop MapReduce framework has both heterogeneous problems and failure problems. However, the hypothetical execution mechanisms and built-in fault tolerance are useful to some extent. The lack of suitable algorithms is the cause of different problems, and searching for slow and failing jobs requires restarting the failing job from scratch. For analysis:

Multiple copies of large files: Since HDFS is not designed with performance in mind, multiple copies of files will be created. There are usually at least three copies of the document. Due to the need to maintain data in the field while maintaining performance, we often see 6 copies of the required data, which is by definition "large". SQL support is very limited. They often lack basic SQL functions such as "group by" analysis subqueries, etc.

Error: HDFS has no idea about the query optimizer and therefore cannot choose the optimization plan based on the value. Therefore, Hadoop clusters are usually larger than required for similar data.

Process Challenges: The MapReduce framework is very complex to handle beyond simple transformations. There are some open-source mods that try to simplify this, but they also use proprietary languages.

No skills required: Using beautiful data is only part of the Hadoop project conflict and in any case requires expertise on the decentralized MapReduce highway to grow with the knowledge of the algorithm itself [23,32].

3.1. HPCC (High Performance Computing Cluster)

HPCC (High Performance Computing Cluster) is an open data center from Lexis Nexis Risk Solutions. Some of its shortcomings are:

1. HPCC is a word-based design, if one fails the whole cluster fails. ECL is a complicated version of SQL. There is a lot of syntax that simply doesn't make sense or is unnecessary. Modern software practices (like unit testing) are not available.
2. ECL has many ills. Too much time is wasted. There are no unit tests and real-time checks in the interface.
3. It loses its "introduction" feature when you have to explain every step you want to take.
4. The price is high. It has a free version but it has many problems where you have to buy a license. Nobody wants to develop like this because most developers prefer to use their own preferred IDE (Eclipse, VIM, emacs, etc.).
5. Performance is a product. ECL code is based on Hadoop which goes beyond the original ECL code with 20% less hardware. This does not mean that Hadoop is faster but it means that the performance is mostly up to the developer.) without public support. Looking at Jira systems there are 10 developers on the list.
6. There is no unity. No one is writing new technologies for HPCC. All new work is still being done on Hadoop [26].

4. Comparison of Tools

Table 1. Tools Properties

Tools	Properties				
	Recovery	**Failure Tolerance**	**Security**	**Processing**	**File System**
Apache Hadoop	It has rack-aware system to recover the crashed nodes [16]	It uses speculative execution for failure tolerance [17]	By default, non-secure mode but authentication can be provided by secure Data Nodes and end user accounts [30]	It has serial processing but batch processing is also possible with its MapReduce [2]	It has distributed file system called HDFS (Hadoop Distributed File System) [4]
IBM InfoSphere BigInsights	For recovery it have InfoSphere Information Server [13]	Infosphere Streams helps in failure tolerance [13]	Security is enhanced by LDAP(light-weight directory access protocol [13]	It supports batch and real time processing [13]	It is based on HDFS [13]

5. Conclusion

The previous section discussed the shortcomings of big data and in this section we will provide solutions to these problems.

5.1. Hadoop

Hadoop is a big data infrastructure and is used by many large organizations. In the previous section we discussed some shortcomings of Hadoop, in this section we will provide solutions to these problems. It shows the growth rate of many time records. It seems that the value of the results is to make the latest changes in fast and continuous data work (like Google's PageRank) today, but the results are relevant at the expense of simple operational standards that cannot be used at the cost of social relationships. . MapReduce and more importantly requires programmers to use special dynamic applications, increase the algorithms, ultimately increasing the complexity of the algorithm and code. For example, someone collecting data and information on LinkedIn is enough in MapReduce. If someone follows a group, the reference will be recalculated because the user's profile has just been updated, and this is not a good way [21].

6. Recommendations

Some caching should be done at the secondary level. But it is only a cache when the user is using the data or data in very large blocks or data with many connections and objects. This is true because the feature map will be fed by the job tracker. Since the cache is made for large chunks of data that change very little, we can stop the map and reduce the job every time in the same way. This will definitely save resources and time.

This cache will be divided into three levels like performance, performance-based and hardware-based (worst). Job-based cache is like a job based on the number of times a particular job is completed [21]. . Using this tool requires skill in low-level operations.

Solution: Learning and working with Hadoop is no longer difficult. There is a tutorial on Hadoop’s website that provides all the details of the tool and online training is also available to teach users how to use Hadoop better. as it works, its speed decreases.

Solution: Hadoop’s MapReduce framework overcomes the problems of batch processing. In MapReduce, tasks are set to different variables and then reduced, so the processing is instantaneous. Alarm clocks reach different nodes.

Solution: Unequal processing time of Hadoop cluster nodes causes time skew. This problem can be solved by Skew Tune. Skew Tune essentially rebalances the load across nodes. Skew Tune is ideal for reducing physical skew in MapReduce systems.

Skew Tune continuously checks the user’s work and finds conflicts that delay the completion of the job. If such a task is detected, the task is stopped and its unprocessed objects are returned to their default state. Repartitioning is only possible when there are empty nodes. If there is no face time skew, there is no additional overhead in skew adjustment. If there is no fault, the overhead it carries is negligible [23]

Fault Tracking: Due to the distributed environment of Hadoop, it is difficult to group to distinguish fault from fault. Many innovations have been made in areas such as fault tracking and analysis. and logging. Once the processing is complete and the data is collected, the batch will stop. This recorded data is used as input to the engine analysis module and used for analysis. Once the job is complete, the test engine generates a detailed human-readable report. This report provides complete information on the total duration, number of operations, and total bytes transferred. All jobs are represented in a completed form that reports their start time, associated HDFS jobs, job type, and additional details. The report shows the details of the entire process. It shows the total number of reduce operations, map operations, and the total number of bytes transferred. If a failure occurs, the results of the engine analysis are used for fault monitoring and analysis to understand the failure mode. This system actually uses a comparative model. First, search for the token of interest, then analyze the logs and match patterns to identify the token and its related information.

Limitations of Big Data: Today’s business needs are to process big data and provide rich analytics at a scale that businesses can afford, at a balance of speed and cost. Automated configuration is designed to command some small files. While Hadoop is useful in handling and organizing large data sets, it does not follow the rules of analysis. As the size and breadth of data increases, the business potential of big data analytics also increases. Companies are aware of this and are looking at analytics platforms as a balance between Hadoop’s data intelligence and processing capabilities.

References

1. Avita Katal, Mohammad Wazid and R H Goudar, "Big Data: Issues, Challenges, Tools and Good Practices" ©2013 IEEE.
2. Anam Alam and Jamil Ahmed, "Hadoop Architecture and Its Issues", IEEE 2014 International Conference on Computational Science and Computational Intelligence.
3. Stephen Kaisler, Frank Armour, J. Alberto and William Money, "Big Data: Issues and Challenges Moving Forward" 2013 46th Hawaii International Conference on System Sciences.
4. Han Hu, Yonggang Wen, Tat-Seng Chuai, and Xuelong, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial" © 2014 IEEE.
5. Seref SAGIROGLU and Duygu SINANC, "Big Data: A Review" ©2013 IEEE
6. Javier Conejero, Blanca Caminero and Carmen Carrio´n, "Analysing Hadoop Performance in a Multi-user IaaS Cloud" ©2014 IEEE
7. Guanghui Xu, Feng Xu*, Hongxu Ma, "Deploying and Researching Hadoop in Virtual Machines" International Conference on Automation and Logistics Zhengzhou, China, August 2012.
8. Divya M and Annappa B, "Workload Characteristics and Resource Aware Hadoop Scheduler" 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS).
9. Megha Sharma Amity, Nitasha Hasteer, Anupriya Tuli and Abhay Bansal, "Investigating the Inclinations of Research and Practices in Hadoop: A Systematic Review" ©2014 IEEE.
10. Kala Karun. A, Chitharanjan. K, "A Review on Hadoop – HDFS Infrastructure Extensions" Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).
11. Khan, S. R., Asif Raza, Inzamam Shahzad, & Hafiz Muhammad Ijaz. (2024). Deep transfer CNNs models performance evaluation using unbalanced histopathological breast cancer dataset. Lahore Garrison University Research Journal of Computer Science and Information Technology, 8(1).
12. Jean-François Weets, Manish Kumar Kakhani, Anil Kumar, Ecole des Mines de Nantes, "Limitations and Challenges of HDFS and MapReduce" ©2015 IEEE.
13. Xiaopeng Li, Wenli Zhou Beijing, "Performance Comparison of Hive Impala and Spark SQL" 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics.
14. IBM Software white paper "The value of IBM InfoSphere BigInsights".
15. Anthony M. Middleton, Ph.D. LexisNexis Risk Solutions white paper on HPCC (High Performance Computing Cluster)
16. http://hpccsystems.com/Why-HPCC/HPCC-vs-Hadoop/HPCC-vs-Hadoop-Detail
17. Khan, S.U.R.; Raza, A.;Waqas, M.; Zia, M.A.R. Efficient and Accurate Image Classification Via Spatial Pyramid Matching and SURF Sparse Coding. Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol. 2023, 7, 10–23.
18. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X. Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. Int. J. Imaging Syst. Technol. 2024, 34, e22975.
19. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X.; Zhu, Y. GLNET: Global–local CNN's-based informed model for detection of breast cancer categories from histopathological slides. J. Supercomput. 2023, 80, 7316–7348
20. Madhury Mohandas, Dhanya P M "An Approach for Log Analysis Based Failure Monitoring in Hadoop Cluster, 861978-1-4673-6126-2/13/$31.00 c 2013 IEEE".
21. Chin-Li-Yin, Ting-Hau Chen, and Yi-NoCheng "On improving fault Tolerance for Heterogeneous Hadoop MapReduce Clusters, 2013 International Conference on Cloud Computing and Big Data".
22. PARACCEL white paper on "Hadoop's Limitations for Big Data Analytics".
23. YongChul Kwon1, Kai Ren, Magdalena Balazinska, and Bill Howe1 University of Washington, Carnegie Mellon University, "Managing Skew in Hadoop".
24. http://www.hpccsystems.com
25. Khan, S. U. R., & Asif, S. (2024). Oral cancer detection using feature-level fusion and novel self-attention mechanisms. Biomedical Signal Processing and Control, 95, 106437
26. https://www.01.ibm.com/support/knowledgecenter/SSPT3X_2.1.2/com.ibm.swg.im.infosphere.biginsights.trb.doc/doc/troubleshooting_problems_console.html
27. http://www.quora.com/What-are-the-limitations-of-Apache-Spark
28. http://www.quora.com/What-is-the-difference-between-Apache-Spark-and-Apache-Hadoop-Map-Reduce

29. Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica University of California, Berkeley ” Spark: Cluster Computing with Working Sets”
30. Khan, S.U.R.; Asif, S.; Bilal, O.; Ali, S. Deep hybrid model for Mpox disease diagnosis from skin lesion images. Int. J. Imaging Syst. Technol. 2024, 34, e23044.
31. Raza, A.; Meeran, M.T.; Bilhaj, U. Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers. VFAST Trans. Softw. Eng. 2023, 11, 80–92.
32. Dai, Q., Ishfaque, M., Khan, S. U. R., Luo, Y. L., Lei, Y., Zhang, B., & Zhou, W. (2024). Image classification for sub-surface crack identification in concrete dam based on borehole CCTV images using deep dense hybrid model. Stochastic Environmental Research and Risk Assessment, 1-18.
33. Meeran, M. T., Raza, A., & Din, M. (2018). Advancement in GSM Network to Access Cloud Services. Pakistan Journal of Engineering, Technology & Science [ISSN: 2224-2333], 7(1).
34. Khan, U. S., & Khan, S. U. R. (2024). Boost diagnostic performance in retinal disease classification utilizing deep ensemble classifiers based on OCT. Multimedia Tools and Applications, 1-21.
35. Farooq, M. U., Khan, S. U. R., & Beg, M. O. (2019, November). Melta: A method level energy estimation technique for android development. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-10). IEEE
36. Shahzad, I., Khan, S. U. R., Waseem, A., Abideen, Z. U., & Liu, J. (2024). Enhancing ASD classification through hybrid attention-based learning of facial features. Signal, Image and Video Processing, 1-14.
37. Mahmood, F., Abbas, K., Raza, A., Khan, M. A., & Khan, P. W. (2019). Three dimensional agricultural land modeling using unmanned aerial system (UAS). International Journal of Advanced Computer Science and Applications, 10(1).
38. Khan, M. A., Khan, S. U. R., Haider, S. Z. Q., Khan, S. A., & Bilal, O. (2024). Evolving knowledge representation learning with the dynamic asymmetric embedding model. Evolving Systems, 1-16.
39. Wajid, M., Abid, M. K., Raza, A. A., Haroon, M., & Mudasar, A. Q. (2024). Flood Prediction System Using IOT & Artificial Neural Network. VFAST Transactions on Software Engineering, 12(1), 210-224.
40. HUSSAIN, S., RAZA, A., MEERAN, M. T., IJAZ, H. M., & JAMALI, S. (2020). Domain Ontology Based Similarity and Analysis in Higher Education. IEEEP New Horizons Journal, 102(1), 11-16.
41. Raza, A., & Meeran, M. T. (2019). Routine of Encryption in Cognitive Radio Network. Mehran University Research Journal of Engineering and Technology [p-ISSN: 0254-7821, e-ISSN: 2413-7219], 38(3), 609-618.
42. Al-Khasawneh, M. A., Raza, A., Khan, S. U. R., & Khan, Z. (2024). Stock Market Trend Prediction Using Deep Learning Approach. Computational Economics, 1-32.
43. Khan, U. S., Ishfaque, M., Khan, S. U. R., Xu, F., Chen, L., & Lei, Y. (2024). Comparative analysis of twelve transfer learning models for the prediction and crack detection in concrete dams, based on borehole images. Frontiers of Structural and Civil Engineering, 1-17.