Volume: 5, Issue: 2
Page: 31-41
2021

International Journal of
Science and Business
Journal homepage: ijsab.com/ijsb

IJSAB International

Improvised Distributions framework of Hadoop: A review

Baydaa Hassan Husain & Subhi R. M. Zeebaree

Abstract:
HADOOP is an open-source virtualization technology that allows the distributed processing of large data sets across standardized server clusters. With two modules, HADOOP Distributed File System (HDFS) and MapReduce framework, it is designed to scale single servers to thousands of computers, providing local computation and storage. Over a decade after HADOOP emerged on the forefront as an open system for Big Data analysis. Its growth has prompted several improvisations for particular data processing needs, based on the type of processing conditions at various periods of computation. This paper, through reviewing several kinds of research provides the basic HADOOP system structure and the description of the MapReduce, HDFS Efficiency. Explaining how the HADOOP framework can overcome the "5Vs" challenges in Big Data. However, in addition to the many benefits of the HADOOP system, like fault tolerance, reliability, high availability, scalable, decreases execution time, reduces latency, improve the security issues, improving the quality of data analysis, better scheduling model, and cost-efficiently. On the other hand, there were some barriers and challenges regarding adjusting data regularly, security issues, and load balancing. Finally, the certainly benefit and challenges of the HADOOP system have been represented paving the way for the future research to find solutions to these challenges.

Keywords: HADOOP, HDFS, MapReduce, Big Data.

IJSB
Literature Review
Accepted 20 January 2021
Published 25 January 2021
DOI: 10.5281/zenodo.4461761

About Author (s)

Baydaa Hassan Husain, ISE Department, Erbil Polytechnic University, Erbil - Kurdistan Region - Iraq, Baydaa.mei20@epu.edu.iq.
Subhi R. M. Zeebaree (corresponding author), Information Technology Department, Duhok Polytechnic University, Duhok - Kurdistan Region – Iraq, subhi.rafeeq@dpu.edu.krd.

1. Introduction

HADOOP enables massive data sizes sets to be processed, offers high-performance computing skills, built to operate on several capacity-providing machines. HADOOP's multidimensional function allows for various data types (structured, unstructured, semi-structured) from many sources (Alzakholi et al., 2020; Seay et al., 2015; Zeebaree et al., 2020). Countless communications entries such as Facebook, WhatsApp, Twitter, Google, etc. are now available daily where consumers can ask their friends and followers about any object, service, event, and problems before making any decision (Zeebaree et al., 2020). To manage this massive volume of data (also known as Big Data) we used the HADOOP framework (Verma et al., 2015). Many distributions allow a Big Data scheme to be manipulated and its key components to be managed: HortonWorks, Cloudera, MapR, IBM BigInsights Infosphere, Pivotal, Microsoft HDInsight, etc. (Erraissi et al., 2017b; Zebari et al., 2019; Zeebaree et al., 2019). Nearly all businesses have moved their data as well as applications to the cloud because of the popularity of the Internet. Controlling massively distributed data such as the cloud is a hard task (Haji et al., 2020; Shukur et al., 2020; Vijay). In the Big Data revolution, several publishers present ready-to-use packages to maintain a Big Data structure, including HortonWorks, Cloudera, MapR, IBM Infosphere, BigInsights, and Pivotal HD (Erraissi et al., 2017a; Shukur et al., 2020; Zeebaree et al., 2020). The challenges of working with Big Data can always be interpreted as Big Data's "5Vs": (Volume) representing amount of data, (Velocity) Speed for processing data, (Value) value of the Big Data, (Variety) makes the content so large, and (Veracity) refers to noise (Abdullah et al., 2020; Bobade, 2016; Haji et al., 2020). HADOOP ecosystem is an open-source technology for Store and processes massive data sets. It's got comprehensive power of computing and it consists of large computer cluster networks. HADOOP makes it possible for hundreds of terabytes to be handled (Abdullah et al., 2018; Zeebaree et al., 2019). The system automatically manages hardware failures. There are four major forms of HADOOP: HADOOP Distributed File System (HDFS), HADOOP MapReduce, HADOOP YARN('Yet Another Resource Negotiator'), and HADOOP Common Resource Negotiator (Jader et al., 2019; Ravichandran, 2017). The key components of the Hadoop system are HDFS and MapReduce. HDFS runs on modern hardware and it provides easy accessibility to and storing of, semi-structured, and unstructured cluster data and it can implement CRUD (Create, Read, Update, and Delete) on files. It gets hard work done. It is automatically handles node failure and data replication (Dwivedi & Dubey, 2014; Zeebaree et al., 2020; Zeebree et al., 2020). While to manage vast amounts of data on supply clusters in a cost-effectively way, MapReduce is used (Dino et al., 2020; Sallow et al., 2020; Singh et al., 2018). HDFS includes a node name and data nodes, and act in Master-slave architecture. There is a single node name that behaves as the master server that governs the namespace of the file system as well as, the directories in the hierarchy format, and Node Data contains two directories. The first file contains data and the second file is the stamp generation block Data Mining, which basically means taking a critical content from a vast and wide range of data (Dino et al., 2020; Hussain et al., 2019; Zebari et al., 2020). There are countless data mining techniques that can be used for big data, some of which are: Arrangement Analysis, Group Analysis, Evolution Analysis, and Outlier Analysis (Bhosale & Gadekar, 2014; Ibrahim et al., 2019; Pujari et al., 2016; Zeebaree et al., 2017). Although the ecosystem of HADOOP is a useful option for big data distribution, nonetheless, HADOOP doesn't sound fine for adjusting data regularly. HADOOP's key challenge is its physical arrangement of data, including data structure and indexes (Gadde & Vijay, 2017). Another downside of the HADOOP system is the need to enhance Map Reduce. In addition to being the backbone of HADOOP storage, HDFS requires the ability to easily access files of various sizes, and storage protection efficiency needs to be enhanced as well (Feng l., 2016; Rashid et al., 2018). Another issue is load balancing which means the flow of data between systems creates a problem when the data is

spread through several device clusters (Abduallah & Zeebaree, 2017; Kumar et al.,). Finally, HADOOP is the perfect option for cloud and big data computations. But in the near future, more research must be done to improve its effectiveness, so that full use can be made of it. This paper brings to light numbers of research about improvised HADOOP system and show the research position in the context of components and structure, showing the advantage and challenges of this system.

2. Literature Survey

In 2016 Gourisaria et al. (Gourisaria et al., 2016)proposed improvisation of name node results by HADOOP framework enabled by the aggregator to achieve further data processing and analysis by optimization, and the workload at the name node has been reduced. They suggested the MapReduce approach of the HADOOP system for data processing using different algorithms for data analysis, such as clustering, fragmentation, and aggregation. the goal was to reduce the workload on the Name node by assisting through the aggregator node, which serves as an interface between the Name Node and the data node.

In the same year, Kawises and Vatanawood (2016) provided a model for developing the transferring data and processing queries in the HADOOP system. They suggest a supportive tool to transfer RDF data to XML which then translated to N-triple form and transferred to HADOOP system and SPARQL query executed to obtain the results. They present the map and decrease algorithms in a standardized shape. SPARQL is then translated into BGP (Basic Graphic Pattern) automatic query using Jena algebra. When they create maps, these BGPs are addressed. A map function defines a parameter key and a decrease function to help post-processing combine and get the final answers in XML type words. In order to improve the MapReduce algorithm, Manipulate the RDF graph accordingly. Sehgal & Agarwal (2016) used the HADOOP system with Sentiment Analysis techniques to analyze Twitter data as a big data application. Sentiment Analysis approach used by a methodical analysis using a mood value formula based on the proximity of the terms with adjectives such as 'excellent',' worse',' bad' etc. They use the Nai've-Bayes approach and a HADOOP cluster for distributed, all-type computing the data. They depended on two types of parameters, first, Latent Semantic Analysis, Which is a language possessing strategy deals with the relationship among different data types. The second parameter is known as a real mathematical statement which is used to calculate the relationship between two objects. In the end, the goal of this model system was to evaluate the sentiment, improving the quality of data analysis, and expanding to the social media site and film reviews such as blogs and feedback every day and likes with accuracy increased. Verma & Pandey (2016) suggested a representation for Big Data approach in analyzing the grade of the student using the Hadoop MapReduce system dependent upon the cloud system. The model takes big data represented in a huge set of student records from a specific university and stores them in tuples; this is called the mapping stage. The data is sorted in the sorting stage. Then the similar tuples are merged in the shuffle stage. Finally removed the extra data in the reduce stage. The output of all these stages will be mapped to HDFS (HADOOP Distributed File System).depending on the stored records the system can predicate the rating average of the students. In 2017, Dick, Geun, and Kwon addressed practical challenges and anomalies for systems that used HADOOP frameworks for processing big data in LINUX CENT OS 6.6 as an example of a heterogeneous cluster where data nodes have a variety of computing capabilities. The first challenge was the unexpected Pause or stop in MapReduce performance. One of the greatest challenges is accumulating all the processing information from dissimilar sources into one place for association and analysis. Within the configuration of a HADOOP framework, the major issue is the running and completion of work. Through the tests carried out, there was a breaking down of the impact of the cluster

since of irregular failure of MapReduce forms on a portion of information nodes. As it is not a physical problem it is not easy to fix it. Another challenge represented in data balancing performed by a HADOOP cluster to overcome the skewing of the data, as it can produce some types of anomaly. Erraissi et al. (2017) proposed them study about the architectural exploration of the HADOOP framework for processing Big Data. An architectural study was required to understand the method of operation. They carried on a study on a variety of companies that manufactured their own HADOOP applications like Pivotal HD, Cloudera, HortonWorks, HortonWorks, IBM Infosphere BigInsights, and so on and examined their individual properties. The goal of this study was to exam and distinguish the common features and characteristics of the main HADOOP distributions of Big Data to standardize Big Data principles. Reza et al. (2017) prepared a study on the configuration characteristic of HADOOP Cluster to achieve improvement of usage efficiency. Some of the essential configuration properties related to block size, memory allocation, CPU allocation, MapReduce number of occupations, job scheduling, and JVM are outlined, which should be optimized for the advanced output of the HADOOP application. Finally, the output has been analyzed by the Benchmark equation. The goal of this paper was to indicate the parameters by which we can improve the HADOOP application performance. Cao et al. (2017) proposed a HADOOP-based platform for harmonics big data analysis. A harmonic function approximation algorithm was implemented and an observable THD measurement algorithm based on MapReduce programming was displayed. Compared to traditional SHSE, the results show the significant efficiency of the proposed MHSE (Mean Hash Signature Estimation algorithm) measurement. In extension, using distinctive dataset sizes, the computation times of the proposed MapReduce-based algorithm are examined concerning the number of VMs. Jena et al. (2017) presented a paper explaining an approach to minimize the memory waste that typically occurs in the block placement method and to optimize the execution of the HADOOP framework by sufficient memory allocation. The MapReduce technique is used to evaluate data using various data processing calculations such as clustering, fracture, and aggregation. The effect of this model was to free memory wastage by correctly allocating the split data sets between distinctive data nodes.

Bhathal & Dhiman (2018) carried on a detailed study indicating that because of its low cost, flexible data processing capabilities, and high fault-tolerant, HADOOP has developed rapidly. The Apache Hadoop open-source center release was improvised by HADOOP sellers and made it ready for investment. Abound together item was shown by HADOOP distributions: an included apache market, modern software, security and central organization all together so that organizations did not have to spend time integrating all these basics into a single utilitarian product. They conclude that the administration component used for centralized organization in both Cloudera (Cloudera Manager) and Hortonworks (Apache Ambari) is distinct, after contrasting distinctive highlights and evaluating the advantages and disadvantages of distinctive information solution providers. The distribution of Hortonworks is open source and cost-free, while Cloudera provides two versions, regular free of cost and premium charged. The distribution of MapReduce is the quickest out all distributions, but the UI (user interface) is better distributed through Cloudera.

Last year Bhathal & Singh (2019) provided a comprehensive study about HADOOP framework security challenges and attacks. Due to the immense volume, fast development, and different data quality characteristics, these are invincible and established security solutions are not adequate. HADOOP is sometimes a collection of individual applications for Pig, Hive, Flume, Oozie, HBase, Start, and Strom. Each one of these products includes the environmental security capabilities for Big Data specifications and data scaling functionality.

In this study the researchers suggest some security tools naming: Apache Sentry, Apache KNOX, Apache Ranger, Project Rhino, and Kerberos to improve the security of the HADOOP framework. Lei et al. (2019) proposed an approach with HADOOP-based Big Data architecture. The approach extends traditional HADOOP to add data on maritime locations to analyze AIS (Automatic Identification System) data. The framework he proposed can be used in spatial statistics, grouping, clustering, and architecture mining, and visualization applications for maritime traffic surveillance. In the storage layer, this method presents a two-layer spatial index structure. Include two new components in computational layer programming in MapReduce. With this approach, we can create different spatial analysis operations on broad maritime position data based on this function provided by the framework. Jiang (2019) presented HADOOP-based basic processing of a Big Data analysis study. The application of Big Data is to mine possible Big Data value by using methods and techniques for data measurement and analysis that analyzes relevance and functional meaning logically. HADOOP was used to build the application environment and the case of WordCount was combined to evaluate the calculation preparation of Outline and Reduce. In this study, HDFS (HADOOP Distributed File System), MapReduce parallel computing model, and Hive data warehouse were evaluated and the MapReduce parallel programming model was practiced and linked in practice. To get reference value for the Big Data stage creation and the analysis and processing of Big Data.

Shah & Padole (2019) measured and analyzed the execution of various scheduling algorithms on multiple applications of big data using the HADOOP/MapReduce model, and taking into account the time of latency, time of completion, and location of data. For this reason, three distinct planning queue configurations, such as single line, multi-queue, and mixed multi-queue were used to run six big data applications. Different experiments were carried out for the HADOOP environment by fine-tuning scheduling arrangements. The study relied on the big data framework that needs to be processed, to establish a better scheduling model. In 2019 Meena and Sujatha proposed a framework for weather forecasting prediction using the HADOOP MapReduce system to deal with this scenario of Big Data. This approach proposed climate classification and expectation observational strategies by following the Co-EANFS (Co-Effective and Flexible Neuro-Fuzzy Framework) method for data managing. This method contains three stages first, collecting the weather data for processing, second, divided the data into groups according to the season, finally, implement the model and getting the output from the Co-EANFS method. This model achieved less time for execution and Strong precision of estimation. Kong (2020) published an article on how to use the HADOOP distributed framework in the development of the distributed Land Engineering Data Management Process with thematic data management, data recovery, simulation, decision analysis, and detailed administration system counting. By integrating the distributed HADOOP database, MySQL social database, Geodatabase space style database, setting up various modular data management, data in a structured and unstructured data and spatial information exchange grouping display for an advanced interface, distributed storage structure is adopted to build up the data distribution center and the land engineering data management system (LE-DBMS). The final result was the land engineering database is set up and a new land engineering application mode is created. Kumar et al. (2020) did a research on a distributed system based on the HADOOP MapReduce framework applying the Random Forest Algorithm. The implementation of the Random Forest algorithm in the HADOOP framework is planned. The Random Forest algorithm will be executed on four HADOOP cluster hubs, which input a variety of size data at a time. The algorithm's output will be calculated by analyzing execution time, measured by precision, kappa, reliability, and standard deviation. The result was, the Random Forest algorithm is able to perform

effectively with large Data sets in the HADOOP system, maintaining an excellent execution time and accuracy. In 2020 Machina, and Songjiang suggest a model for carrying on an assessment of crime and exchanging intelligence information using Big Data. The research used a survey, and an interview report is used to obtain data from various law authorization workers. Microsoft Excel is used to generate reliable results and visualize the outcome in the form of a pie chart, whereas UML models are used to define the proposed model's logical and physical schema. This study hypothesis is confirmed by the release of the manual and standard police and crime investigation methods. Aung & Zaw (2020) proposed an approach to applying the HADOOP YARN tuning parameters for improving the scheduling and execution time. For all types of clusters and applications, Apache HADOOP offers more than 200 default parameter configuration options. HADOOP YARN breaks down the functionality of the distributed application open-source system and performs job scheduling and tracking, along with the storage, handling, and review of big data on production equipment. Leading to improving implementation time and effective scheduling of jobs. Shah & Padole (2020) proposed an algorithm for block reorganization which significantly improves the processing time in mixed configurations through successful file system management has been proposed by Shah and Padole in a scheme for improving the optimization for processing of big data in a homogeneous and heterogeneous environment. HADOOP HDFS algorithm enables large data to be stored and offers practical support for the vast volume of data to be transmitted and distributed. A significant improvement in the placement of data replicas and processing is expected by the default HDFS block placement policy. This work helps to contribute to a block rearrangement algorithm that optimizes device (i.e. memory, CPU) resource allocation for block rearrangement. Results show that approximately 90 percent of the proposed legislation offers compared to the 68-70 percent data position in the normal case, which effectively decreases work execution time and reduces the latency.

3. Discussion

It is obvious from previously stated literature reviews that numerous research studies have concentrated on the HADOOP framework because of its importance. This study showed that researchers suggest HADOOP as a solution for Big Data processing as it enables distributed processing of large amounts using simple datasets through clusters of computers model for programming and has many important features like fault tolerance, reliability, high availability, scalable, and cost-effective. In Table 1, the statistical assessment between these studies is shown.

Table 1: statistical assessment of previous researches.

Year	Author	Objective	Methodology	Result/Goal
2016	Jena et al. (2016)	Improvisation of Name Node and reduce the workload	MapReduce approach of HADOOP system	Reduce the workload on the Name node by assisting through the aggregator node
2016	Kawises, Vatanawood (2016)	Developing the transferring data and processing queries in the HADOOP system	A used tool to transfer RDF data to XML which translated to N-triple form and SPARQL query executed to obtain the results, then SPARQL is then translated into BGP	Improve the MapReduce algorithm, Manipulate the RDF graph.
2016	Sehgal, and Agarwal (2016)	Analyses Twitter data as a big data application	Sentiment Analysis techniques represented by methodical analysis using a mood value formula	Evaluate the sentiment, improving the quality of data analysis. And expanding to the

				social media site and film reviews.
2016	Verma, and Pandey (2016)	Representation for big data approach in analyzing the grade of the student.	Using HADOOP MapReduce system dependent upon the cloud system	Predication for the rating average of the students
2017	Dick et al. (2017)	Addressing challenges and anomalies for systems that used HADOOP frameworks	Using LINUX CENT OS 6.6 as an example of a heterogeneous cluster where data nodes have a variety of computing capabilities.	_____
2017	Erraissi et al.(2017)	Exam and distinguish the common features and characteristics of the HADOOP distributions.	Study some supplier like Pivotal HD, Cloudera, HortonWorks, HortonWorks, IBM Infosphere BigInsights, and so on and examined them individual properties.	Standardized the concepts of Big Data in HADOOP system.
2017	Reza et al. (2017)	Improvement of usage efficiency by studying the configuration characteristic of the HADOOP Cluster.	The essential configuration properties related to block size, memory allocation, CPU allocation, MapReduce number of occupations, job scheduling, and JVM are outlined, then the output has been analyzed by the Benchmark equation.	Improve the HADOOP application performance
2017	Cao et al. (2017)	A HADOOP-based platform for harmonics big data analysis	Estimation algorithm and demonstrated a measurable THD calculation algorithm based on MapReduce programming.	Significant efficiency of the proposed MHSE measurement
2018	Jena et al. (2018)	Minimizing the memory waste that typically occurs in the block placement method and optimizing the execution of the HADOOP framework	The MapReduce technique is used to evaluate data using various data processing calculations	free memory wastage
2018	Singh, Bhathal & Dhiman (2018)	The Apache HADOOP improvisation	Studying the properties of UI MapReduce Cloudera (Cloudera Manager) and Hortonworks (Apache Ambari)	_____
2019	Bhathal, and Singh (2019)	Understanding HADOOP framework security challenges and attacks	Applying security tools like Apache Sentry, Apache KNOX, Apache Ranger, Project Rhino, and Kerberos	Improve the security of the HADOOP framework
2019	Lei (2019)	HADOOP framework extended to add data on maritime locations in order to analyze AIS (Automatic Identification System) data.	The storage layer provides two-layer hierarchical indexes. Include two new elements in computational layer programming in MapReduce	Create different spatial analysis operations on broad maritime position data based on HADOOP framework.
2019	Jiang (2019)	Using HADOOP Analyzation and processing of Big Dataset	HDFS, MapReduce parallel computing model, and Hive data warehouse, the MapReduce parallel programming model was practiced and linked in practice.	Get reference value for the Big Data stage creation.
2019	Shah and Padole (2019)	Measure and analyze the execution of various scheduling algorithms.	Three distinct planning queue configurations used to run six big data applications HADOOP environment by fine-	better scheduling model

			tuning scheduling arrangements.	
2019	Meena, and Sujatha (2019)	A framework for weather forecasting prediction.	HADOOP MapReduce system Used for climate classification and expectation observational strategies by following the Co-EANFS method.	Achieved less time for execution, and Strong precision of estimation.
2020	Kong (2020)	Using HADOOP distributed framework in the growth of distributed land engineering.	integrating HADOOP database, MySQL social database, Geodatabase space style database, setting up various modular data management, data in a structured and unstructured data system and spatial information, build up the data distribution center, and (LE-DBMS)	Land engineering database is set up and a new land engineering application mode is created.
2020	Kumar et al. (2020); Jena et al. (2020)	The implementation of the algorithm for the Random Forest in the HADOOP framework	The Random Forest algorithm was executed on four HADOOP cluster nodes. The algorithm's output is calculated by analyzing execution time, measured by precision, kappa, reliability, and standard deviation.	Perform effectively with Large Data sets.
2020	Machina & Songjiang (2020)	Assessment of crime and exchanging intelligence information.	Depends on surveys, interview report, Microsoft Excel is used for results, Proposed UML models for the logical and physical schema.	Releasing of manual and standard police and crime investigation methods.
2020	Aung, and Zaw (2020)	Applying HADOOP YARN tuning parameters for improving the scheduling and execution time	Apache HADOOP offers more than 200 Options for default parameter setup.	Improving the scheduling and execution time
2020	Shah & Padole (2020)	Improve optimization for processing of big data in a homogeneous and heterogeneous environment.	HADOOP HDFS algorithm to stored and offers practical support for the vast volume of data to be transmitted and distributed.	It brings down execution time and reduced latency.

4. Conclusion

In short, Remarkable advances in Social media networks, the internet, digital technologies, and mobile devices have led to an unprecedented increase in data that can be managed by all businesses. These technologies typically generate streams of data that can be gathered, classified, deployed, stored, and analyzed, and so on by data analysts. In this paper, the HADOOP system is represented as a platform which can be used to support a stable and scalable method of distributed Computing for Big Data. Firstly, define the basic HADOOP system structure and the description of the MapReduce, HDFS efficiency and explaining how the"5Vs" weaknesses in Big Data can be treated by the HADOOP model. Afterward, it showed uses of HADOP in various fields such as Twitter, Google, Yahoo data set, beside, using it in weather forecasting prediction, minimizing the memory wastage, and in the formation of the distributed Land Engineering. However, form reviewing some previous researches, in addition to the many benefits of this system, like fault tolerance, reliability, high availability, scalability, decreases execution time, reduces latency effectively decreases execution time, reduces for latency, improve the security issues and the quality of data analysis, better

scheduling model, and cost-efficiently. On the other hand, there were some barriers and challenges concerning adjusting data regularly, security issues, and load balancing. Finally, these benefits and challenges can pave the way for future studies to find solutions for these challenges.

5. References

Abduallah, W. M., & Zeebaree, S. R. M. (2017). New Data hiding method based on DNA and Vigenere Autokey. *Academic Journal of Nawroz University, 6(3), 83-88.*

Abdullah, P. Y., Zeebaree, S. R., Jacksi, K., & Zeabri, R. R. (2020). An hrm system for small and medium enterprises (sme) s based on cloud computing technology. *International Journal of Research-GRANTHAALAYAH, 8(8), 56-64.*

Abdullah, P. Y., Zeebaree, S. R., Shukur, H. M., & Jacksi, K. (2020). HRM System using Cloud Computing for Small and Medium Enterprises (SMEs). *Technol. Rep. Kansai Univ, 62(4), 1977-1987.*

Alzakholi, O., Shukur, H., Zebari, R., Abas, S., & Sadeeq, M. (2020). Comparison among cloud technologies and cloud performance. *Journal of Applied Science and Technology Trends, 1(2), 40-47.*

Aung, T. H., & Zaw, W. T. (2020). Improved Job Scheduling for Achieving Fairness on Apache Hadoop YARN. *Paper presented at the 2020 International Conference on Advanced Information Technologies (ICAIT).*

Bhathal, G. S., & Singh, A. (2019). Big data: Hadoop framework vulnerabilities, security issues and attacks. Array, 1, 100002.

Bhosale, H. S., & Gadekar, D. P. (2014). A review paper on big data and hadoop. *International Journal of Scientific and Research Publications, 4(10), 1-7.*

Bobade, V. B. (2016). Survey paper on big data and Hadoop. *International Research Journal of Engineering and Technology (IRJET), 3(01).*

Cao, Z., Lin, J., Wan, C., Song, Y., Taylor, G., & Li, M. (2017). Hadoop-based framework for big data analysis of synchronised harmonics in active distribution network. *IET Generation, Transmission & Distribution, 11(16), 3930-3937.*

Dick, M., Ji, J. G., & Kwon, Y. (2017). Practical Difficulties and Anomalies in Running Hadoop. *Paper presented at the 2017 International Conference on Computational Science and Computational Intelligence (CSCI).*

Dino, H., Abdulrazzaq, M. B., Zeebaree, S. R., Sallow, A. B., Zebari, R. R., Shukur, H. M., & Haji, L. M. (2020). Facial Expression Recognition based on Hybrid Feature Extraction Techniques with Different Classifiers. *TEST Engineering & Management, 83, 22319-22329.*

Dino, H. I., Zeebaree, S. R., Ahmad, O. M., Shukur, H. M., Zebari, R. R., & Haji, L. M. (2020). Impact of Load Sharing on Performance of Distributed Systems Computations. *International Journal of Multidisciplinary Research and Publications (IJMRAP), 3(1), 30-37.*

Dwivedi, K., & Dubey, S. K. (2014). Analytical review on Hadoop Distributed file system. *Paper presented at the 2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence).*

Erraissi, A., Belangour, A., & Tragha, A. (2017a). A big data hadoop building blocks comparative study. *International Journal of Computer Trends and Technology. Accessed June, 18.*

Erraissi, A., Belangour, A., & Tragha, A. (2017b). A Comparative Study of Hadoop-based Big Data Architectures. *Int. J. Web Appl., 9(4), 129-137.*

Erraissi, A., Belangour, A., & Tragha, A. (2017c). Digging into Hadoop-based big data architectures. *International Journal of Computer Science Issues (IJCSI), 14(6), 52-59.*

Feng, D., Zhu, L., & Zhang, L. (2016). Review of hadoop performance optimization. *Paper presented at the 2016 2nd IEEE International Conference on Computer and Communications (ICCC).*

Gadde, R., & Vijay, N. (2017). A SURVEY ON EVOLUTION OF BIG DATA WITH HADOOP. *International Journal of Research in Science and Engineering, 3, 92-99.*

Haji, L., Zebari, R., Zeebaree, S., Abduallah, W., Shukur, H., & Ahmed, O. GPUs Impact on Parallel Shared Memory Systems Performance. *Int. J. Psychosoc. Rehabil, 24(08), 8030-8038*.

Haji, L. M., Zeebaree, S., Ahmed, O. M., Sallow, A. B., Jacksi, K., & Zeabri, R. R. (2020). Dynamic resource allocation for distributed systems and cloud computing. *TEST Eng. Manag, 83, 22417-22426*.

Hussain, T., Sanga, A., & Mongia, S. (2019). Big Data Hadoop Tools and Technologies: A Review. *Available at SSRN 3462554.*

Ibrahim, R., Zeebaree, S., & Jacksi, K. (2019). Survey on Semantic Similarity Based on Document Clustering. *Adv. Sci. Technol. Eng. Syst. J, 4(5), 115-122.*

Jader, O. H., Zeebaree, S., & Zebari, R. R. (2019). A State of Art Survey for Web Server Performance Measurement and Load Balancing Mechanisms. *International Journal of Scientific & Technology Research (IJSTR), 8(12), 535-543*.

Jena, B., Gourisaria, M. K., Rautaray, S. S., & Pandey, M. (2016). Improvising name node performance by aggregator aided HADOOP framework. *Paper presented at the 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT).*

Jena, B., Kanaujia, P. K., Rautaray, S., & Pandey, M. (2017). Improvising block placement policy in HADOOP framework. *Paper presented at the 2017 International Conference on Computer Communication and Informatics (ICCCI).*

Jiang, H. (2019). Research and Practice of Big Data Analysis Process Based on Hadoop Framework. *Paper presented at the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).*

Kawises, J., & Vatanawood, W. (2016). A development of RDF data transfer and query on Hadoop Framework. *Paper presented at the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).*

Kong, H. (2020). Construction of distributed Data management platform for Land Engineering based on Hadoop Framework. *Journal of Electronics and Information Science, 5(1), 63-71.*

Kumar, A., tech Scholar, M., & Rai, K. A REVIEW PAPER ON LOAD BALANCING IN HADOOP CLUSTER OVER CLOUD COMPUTING ENVIRONMENT.

Kumar, K., Sharma, N. A., & Ali, A. S. (2019). Classification in a Distributed System-A Study of Random Forest in the Hadoop MapReduce Framework. *Paper presented at the 2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE).*

Lei, B. (2019). A Hadoop-Based Spatial Computation Framework for Large-Scale AIS Data. *Paper presented at the 2019 IEEE 2nd International Conference on Electronics Technology (ICET).*

Machina, A. A., & Songjiang, L. Crime Analysis and Intelligence System Model Design using Big Data. *International Journal of Computer Applications, 975, 8887.*

Meena, K., & Sujatha, J. (2019). Reduced Time Compression in Big Data Using MapReduce Approach and Hadoop. *Journal of Medical Systems, 43(8), 239.*

Pujari, V., Sharma, Y. K., & Rane, R. (2016). A Review Paper on Big Data and Hadoop.

Rashid, Z. N., Zebari, S. R., Sharif, K. H., & Jacksi, K. (2018). Distributed Cloud Computing and Distributed Parallel Computing: A Review. *Paper presented at the 2018 International Conference on Advanced Science and Engineering (ICOASE).*

Ravichandran, G. (2017). Big Data processing with Hadoop: a review. *Int. Res. J. Eng. Technol, 4, 448-451.*

Reza, M., Tripathy, B., Ranjan, H., & Kumar, G. A. (2017). Study and analysis of hadoop cluster optimization based on configuration properties. *Paper presented at the 2017 Innovations in Power and Advanced Computing Technologies (i-PACT).*

Sallow, A. B., Sadeeq, M. A., Zebari, R. R., Abdulrazzaq, M. B., Mahmood, M. R., Shukur, H. M., & Haji, L. M. An Investigation for Mobile Malware Behavioral and Detection Techniques Based on Android Platform. (2020). *IOSR Journal of Computer Engineering (IOSR-JCE), 22(4), 14-20.*

Seay, C., Agrawal, R., Kadadi, A., & Barel, Y. (2015). Using hadoop on the mainframe: A big solution for the challenges of big data. *Paper presented at the 2015 12th International Conference on Information Technology-New Generations.*

Sehgal, D., & Agarwal, A. K. (2016). Sentiment analysis of big data applications using Twitter Data with the help of HADOOP framework. *Paper presented at the 2016 international conference system modeling & advancement in research trends (SMART).*

Shah, A., & Padole, M. (2019). Performance analysis of scheduling algorithms in Apache Hadoop Data, *Engineering and Applications (pp. 45-57): Springer.*

Shah, A., & Padole, M. (2020). Saksham: Resource Aware Block Rearrangement Algorithm for Load Balancing in Hadoop. *Procedia Computer Science, 167, 47-56.*

Shukur, H., Zeebaree, S., Zebari, R., Ahmed, O., Haji, L., & Abdulqader, D. (2020). Cache coherence protocols in distributed systems. *Journal of Applied Science and Technology Trends, 1(3), 92-97.*

Shukur, H., Zeebaree, S., Zebari, R., Zeebaree, D., Ahmed, O., & Salih, A. (2020). Cloud computing virtualization of resources allocation for distributed systems. *Journal of Applied Science and Technology Trends, 1(3), 98-105.*

singh Bhathal, G., & Dhiman, A. S. (2018). Big Data Solution: Improvised Distributions Framework of Hadoop. *Paper presented at the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS).*

Singh, V. K., Taram, M., Agrawal, V., & Baghel, B. S. (2018). A literature review on Hadoop ecosystem and various techniques of big data optimization Advances in Data and Information *Sciences (pp. 231-240): Springer*.

Subhi R. M. Zeebaree, S. Y. A., M. Sadeeq Mohammed, A. (2020). Social Media Networks Security Threats, Risks and Recommendation: A Case Study in the Kurdistan Region. *nternational Journal of Innovation, Creativity and Change, 13(7), 349-365.*

Verma, C., & Pandey, R. (2016). Big Data representation for grade analysis through Hadoop framework. *Paper presented at the 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence).*

Verma, J. P., Patel, B., & Patel, A. (2015). Big data analysis: recommendation system with Hadoop framework. *Paper presented at the 2015 IEEE International Conference on Computational Intelligence & Communication Technology*.

Vijay, M. V. A Review on Cloud based Hadoop Environment.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends, 1(2), 56-70.*

Zebari, R. R., Zeebaree, S. R., & Jacksi, K. (2018). Impact Analysis of HTTP and SYN Flood DDoS Attacks on Apache 2 and IIS 10.0 Web Servers. *Paper presented at the 2018 International Conference on Advanced Science and Engineering (ICOASE).*

Zebari, R. R., Zeebaree, S. R., Jacksi, K., & Shukur, H. M. (2019). E-business requirements for flexibility and implementation enterprise system: A review. *International Journal of Scientific and Technology Research, 8(11), 655-660*.

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zeebaree, S. R. (2017). Combination of K-means clustering with Genetic Algorithm: A review. *International Journal of Applied Engineering Research, 12(24), 14238-14245.*

Zeebaree, S., Salim, B., Zebari, R., Shukur, H., Abdulraheem, A., Abdulla, A., & Mohammed, S. (2020). Enterprise Resource Planning Systems and Challenges. *Technol. Rep. Kansai Univ, 62(4), 1885-1894.*

Zeebaree, S. R., Haji, L. M., Rashid, I., Zebari, R. R., Ahmed, O. M., Jacksi, K., & Shukur, H. M. (2020). Multicomputer Multicore System Influence on Maximum Multi-Processes Execution Time. *TEST Engineering & Management, 83, 14921 - 14931.*

Zeebaree, S. R., Shukur, H. M., Haji, L. M., Zebari, R. R., Jacksi, K., & Abas, S. M. (2020). Characteristics and Analysis of Hadoop Distributed Systems. *Technol. Rep. Kansai Univ, 62(4), 1555-1564.*

Zeebaree, S. R., Shukur, H. M., & Hussan, B. K. (2019). Human resource management systems for enterprise organizations: A review. *Periodicals of Engineering and Natural Sciences, 7(2), 660-669.*

Zeebaree, S. R., Zebari, R. R., & Jacksi, K. (2020). Performance analysis of IIS10. 0 and Apache2 Cluster-based Web Servers under SYN DDoS Attack. *TEST Engineering & Management, 83, 5854 - 5863.*

Zeebaree, S. R., Zebari, R. R., Jacksi, K., & Hasan, D. A. (2019). Security Approaches For Integrated Enterprise Systems Performance: A Review. *International Journal of Scientific & Technology Research (IJSTR), 8(12), 2485-2489.*

Cite this article:

Baydaa Hassan Husain & Subhi R. M. Zeebare (2021). Improvised Distributions framework of Hadoop: A review. ***International Journal of Science and Business,*** *5*(2), 31-41. doi: https://doi.org/10.5281/zenodo.4461761

Retrieved from http://ijsab.com/wp-content/uploads/668.pdf

Published by