International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 2, Number: 2, Pages: 1 - 7, Year: 2023

Enhancing Agricultural Operations: Big Data Analytics Using Distributed and Parallel Computing

Syeda Alishba Fatima [1], Syeda Faiza Nasim [1] and Saad Ahmed [2]

[1] Computer Sciences Department, NED University of Engineering and Technology, Karachi, Pakistan
[2] Computer Sciences Department, Iqra University, Karachi, Pakistan
Corresponding author: Syeda Faiza Nasim (Email: sfnasim@cloud.neduet.edu.pk)

Received: 12/09/2023, Revised: 22/11/2023, Accepted: 15/12/2023

Abstract—**This comprehensive research investigates using distributed and parallel computing for big data analytics in agriculture to improve farming operations' sustainability, efficiency, and innovation. The paper emphasizes how big data analytics, cloud computing, and parallel distributed processing can revolutionize the agricultural industry. The research objectives include investigating the benefits and limitations of big data analytics in precision farming and crop monitoring, identifying the constraints of integrating big data analytics in agriculture and investigating the role of frameworks such as Hadoop and Spark in processing and analyzing agricultural data for informed decision-making and optimized farming operations. The methodology used in the paper is a literature review, which draws on various sources to provide insights into the topic matter. The findings indicate that big data analytics can considerably improve precision farming and crop monitoring; nevertheless, there are hurdles to incorporating big data analytics in agriculture, such as data privacy and security concerns. According to the study, frameworks such as Hadoop and Spark are crucial in processing and analyzing agricultural data for informed decision-making and better farming operations. Overall, this study offers useful insights into the possibilities of big data analytics and distributed and parallel computing in revolutionizing the agriculture industry.**

*Index Terms— **Big data, analytics, parallelization.***

I. Introduction

THIS paper provides a strong investigation of the transformative possibilities of distributed and parallel computing in agriculture big data analytics [1-4]. It emphasises the importance of harnessing sophisticated technologies to improve farming operations' sustainability, efficiency, and innovation [5-10]. This study intends to address the issues and opportunities in the Agriculture sector by diving into the application of big data analytics, cloud computing, and parallel distributed processing [11-16]. It focuses on the critical role of these technologies in precision farming, crop monitoring, yield prediction, and forecasting, as well as the impact of frameworks like Hadoop and Spark in processing and analysing agricultural data for informed decision-making and optimised farming operations [17-22].

A. Volume

Sensors, satellite imagery, weather stations, and other sources generate massive amounts of data, and big data analytics is utilised in agriculture to manage this data [23-26]. This includes data on crop condition, weather patterns, soil condition, and other topics. Volume emphasises the massive amount of data required for analytics. This feature emphasises the importance of handling huge data quantities that are beyond the capacity of typical systems.

B. Velocity

In the agriculture industry, the rate at which data is created and processed is crucial. The second V in the field of big data analytics is the rate at which information is generated, disseminated, and analysed. It is critical to obtain information from multiple sources in a timely manner, but it is also critical to put that information to use as soon as feasible. This speed is essential for quick decision-making because delayed processing may prohibit firms from keeping up with the volume of data. Farmers can make timely irrigation, pest management, and harvesting decisions by using real-time data from IoT devices, weather sensors, and machinery [27-29].

C. Variety

Agricultural data comes in a variety of formats, including semi-structured data from Internet of Things devices, unstructured data from satellite photos, and structured data from databases. By examining this diversity of data, it is able to anticipate crop yields, detect disease outbreaks, and enhance planting schedules. Variety emphasises the various forms and sources that can be employed to collect large data. In contrast to typical databases, which only contain orderly tables, big data can contain a variety of unstructured formats, such as information from social media, sensor data, and other sources. This is troublesome since arranging the data for processing may be difficult [30-32].

D. Veracity

Data quality is critical in agriculture to ensure reliable decision-making. Addressing issues with data accuracy, consistency, and completeness is required to ensure the veracity of agricultural data. Farmers can gain more reliable insights from higher-quality data analysis. The reliability and correctness of vast amounts of data are referred to as data veracity. When resolving problems, it is critical to ensure that the data is reliable so that judgments may be made with confidence. Because of the numerous duplicates, irregularities, and inconsistencies observed in large data, analysis and processing are required to improve data analysis accuracy [33].

E. Value:

The ultimate goal of big data analytics in agriculture is to produce relevant insights that benefit farmers. This includes reducing environmental impact, increasing crop yields, and allocating resources as efficiently as feasible. Social media and the internet have had an impact on the modern age, resulting in a large increase in digital data, or "big data." There are several platforms from which to generate this, such as wearable trackers, mobile devices, and sensors. The magnitude of big data poses significant challenges to the capability of existing storage, processing, and analysis technologies. To overcome these difficulties, it is vital to constantly develop new models, languages, systems, and algorithms for efficiently gathering, categorising, analysing, and learning from huge data [34].

II. Literature Review

In the application of big data in agriculture, four major phases can be distinguished: data collecting, data transformation, data storage, data analysis, and data marketing. The availability of free Big Data analysis tools encourages the growth of smart farming research. The new agricultural applications promise to enhance food production and security by empowering farmers to employ efficient agricultural methods with appropriate recommendations that are environmentally friendly. When Precision Agriculture (PA) is implemented, the agriculture industry generates massive amounts of diverse data. Data collection includes soil parameters, seeding rates, and crop yields, in addition to historical records on weather patterns, terrain, and crop performance [35].

The Internet of Things and social media platforms have generated an unprecedented amount of digital data, which is expanding at an unprecedented rate in the modern period [3]. Big Data refers to information that comes from a variety of sources, including security cameras, wearable trackers, mobile devices, and sensors. The sheer volume of Big Data, on the other hand, poses enormous difficulties to conventional processing, storage, and analytical capabilities. To properly obtain, preserve, assess, and learn from Big Data, new models, languages, systems, and procedures must be generated on a regular basis.

A. Big Data's Present Status in Agriculture

Big Data in agriculture is gaining traction as more people grasp how data-driven technology can alter farming processes, improve decision-making, and allow farmers to produce more food on less land.Big data is currently employed in a variety of areas, including business, healthcare, and agriculture, and it is rapidly expanding across all of these markets. Because of people's reliance on big data to remain ahead of the curve and keep their products and services up to date with developing trends, the big data market is also rapidly expanding. The volume of data also has a major impact on the conclusion and validity of the data [4].

B. Big Data's Accuracy in The agricultural sector

To assure the accuracy of big data analysis in agriculture, an emphasis on data quality, integration, precision agriculture practices, technology infrastructure, and the dependability of decision support systems is required. Taking care of these challenges can increase the precision and efficacy of big data use in agriculture. Low data volumes will diminish the model's accuracy, implying that predictions will be less accurate and may result in incorrect assumptions. The three basic Big Data frameworks—Velocity, Volume, and Variety—are covered in this section [4]. In the diagram below, we explore the accuracy of the data when any of the three variables is altered and how it affects the accuracy.

In agriculture, parallelization is required to manage data scalability, allow for real-time review, process data effectively, maximise resource consumption, and improve decision-making for lucrative and sustainable farming operations. Parallelization is critical in data analytics for increasing productivity and lowering costs. This method involves breaking the problem down into smaller pieces and running them concurrently on numerous cores, threads, or processors. This reduces calculation time and allows for the processing of larger datasets. This strategy works effectively when engaging with time systems or real-time data.

C. Advantages of Parallelization for Agricultural Matters: Time Reduction and Cost Effectiveness

Parallelization allows computational workloads to be divided into smaller, parallelizable subtasks. This significantly reduces the amount of time required to complete complex computational activities such as simulating, analysing enormous datasets, and optimising crop management strategies. Parallelization enhances resource utilisation, which reduces agricultural costs. By speeding up computational procedures, farmers and researchers can make better use of computing resources, reduce operational costs, and boost overall output. Parallelization has the potential to substantially reduce the amount of time necessary to manage datasets by allowing smaller actions to be completed in parallel. This accelerates processing and is valuable in instances where quick insights from real-time data are required.

Employing cloud computing to analyse big data concerning agriculture[24]Cloud computing provides powerful data storage and management capabilities, advanced analytics services, and scalable, accessible, and low-cost computational resources, making big data analytics easier to deploy in agriculture. Agricultural stakeholders may benefit from this link by better utilising big data to develop innovative farming practices, manage resources more effectively, and make better decisions. Security, processing, storing, and analysing big data present a variety of challenges. This aids in the management of massive amounts of data. Cloud computing, which offers scalable and cost-effective solutions, can help to tackle these issues.[5] Cloud computing enables quick access to a shared resource pool, which includes storage and applications. [6] It qualifies multitasking systems such as MapReduce, Hadoop, and Big Table for big data applications. Cloud computing successfully manages and analyses data by utilising cutting-edge techniques such as virtualization, indexing, and mining. Furthermore, authentication and encryption are used to preserve data privacy and security. The link between big data and cloud computing can significantly cut infrastructure costs and simplify data processing, allowing for faster analysis.

D. Big Data's Uncertainty in Agriculture

Big data applications in agriculture are fraught with unknowns and problems that must be overcome before they can be implemented successfully. Data quality and dependability, security and privacy concerns, data integration and interoperability, infrastructure for technology and accessibility, ethical and social implications, and regulatory and legislative frameworks are among the challenges surrounding big data in agriculture. A comprehensive approach that considers data governance, technology innovation, stakeholder participation, and ethical considerations is required to address these challenges. By proactively removing these impediments, the agriculture industry may be able to successfully leverage the benefits of big data while reducing associated uncertainties.

Uncertainty pervades every step of the process due to a variety of issues such as idea variance, limited data gathering, and the complexity provided by multimodal data. Particular situations of uncertainty were highlighted, such as a large number of missing links in social networks and missing values for temporal data elements. According to their forecast, 80 percent of the world's data would be unclear by 2015. When biases, incompleteness, or erroneous sampling pollute training data, uncertainty's negative influence on analytical outputs is increased. To solve these issues, a range of complicated approaches such as probability theory, fuzziness, fuzzy logic, Bayesian theory, and belief function theory are presented. Each of the aforementioned strategies provides a specific strategy for dealing with uncertainty, which reflects the state of the art [7].

E. Big Data's Implications for agriculture

Big data applications in agriculture offer a wide range of scenarios for use that make use of massive volumes of heterogeneous data to increase sustainability, efficiency, and innovation in agricultural operations. Precision farming, crop monitoring, yield prediction and forecasting, supply chain optimization, soil condition and nutrient management, market and price intelligence, disease and pest management, and other applications are some of the key uses of big data in agriculture. These examples demonstrate how big data is altering agriculture in a variety of ways, allowing stakeholders to make informed decisions, allocate resources wisely, and boost productivity while fostering resilience and sustainability in the agricultural business.

F. Intelligent Transportation System

Intelligent Transportation Systems (ITS) are critical in agriculture for boosting operational effectiveness, upgrading transportation and logistics systems, and fostering sustainable farming methods. By leveraging cutting-edge technologies and data-driven insights, ITS systems assist to optimise transportation operations, improve supply chain management, and connect transportation with precision agriculture practices. Big data techniques radically revolutionised the transportation business by allowing massive amounts of data produced by a number of sources, such as GPS, sensors, and monitoring devices, to be collected and analysed. The development of these algorithms seeks to improve the efficiency of the intelligent transportation system and provide data for traffic control. Big data in ITS has driven the development of increasingly complicated models to handle the rise of deep learning as an algorithm [8].

III. Methodology

The following research aims are highlighted in this comprehensive review paper:

1. To investigate the use of distributed and parallel computing for big data analytics in agriculture.
2. To investigate the potential benefits and drawbacks of using big data analytics in agriculture.
3. To investigate the role of several frameworks, such as Hadoop and Spark, in the processing and analysis of agricultural data.
4. To emphasise the importance of big data analytics, cloud computing, and parallel distributed processing in transforming agriculture and improving operational results and decision-making.

A. Key Features of the Parallel Architecture

The core of the design is its simplicity and independence from other file systems. It does not use static partitioning like standard configurations and instead partitions data dynamically during runtime. Each processing node works separately, handling a distinct data partition. This independence encourages balanced parallel processing across nodes, which is critical for speed optimization.

B. Node Interaction and Master Processing

The design takes a decentralised approach, with numerous processing nodes operating concurrently. Each node processes its allocated data partition separately, promoting parallelism in data computation. The partial findings from each node are then pooled at a master processing node. This master node is critical in aggregating scattered calculations and finalising overall results. This technique guarantees good coordination and synchronisation of parallel activities, which contributes to the architecture's overall efficiency.

C. Hadoop and Spark Frameworks for Agricultural Data Processing

The Hadoop framework, an open-source behemoth, and the adaptable Spark engine serve as the foundation of agricultural data processing. HDFS, YARN, and MapReduce are key components that contribute to scalable and effective data processing.

D. The Impact of Hadoop on Agricultural Innovation

This section delves into Hadoop's critical role in modern agriculture, examining how the architecture supplies critical computational resources. Key benefits for agricultural stakeholders include improved data analysis, real-time decision support, and the incorporation of cutting-edge technology.

E. Uncovering Hadoop's Potential: Storage and Processing

This section delves into Hadoop's dual functioning as a storage and processing solution. We deconstruct how Hadoop successfully handles and analyses huge data in the agricultural setting, from the distributed structure of the Hadoop file system to its scalability.

F. MapReduce: The Foundation of Agricultural Data Processing

This section investigates the popularity of MapReduce in parallel programming, focusing on its scalability, simplicity, throughput, and fault tolerance. MapReduce remains a preferred open-source framework for large-scale data processing in agriculture in both cloud computing settings and commodity clusters.

G. Online MapReduce Real-Time Analysis

This section focuses on the capabilities of MapReduce Online for online agricultural data analysis, utilising a customised version of Hadoop MapReduce. Its adaptability and aptitude for dynamic data processing, which distinguishes it from the original model, are critical aspects in its importance in the agricultural industry.

H. YARN for Agriculture: Cluster Resource Management

This section examines how YARN, as a key component of Hadoop's MapReduce 2.0 framework, alters cluster design. The elimination of bottlenecks and the flexibility to scale beyond 4000 nodes make YARN critical for providing computational capability to the agriculture business.

I. YARN in Action: Resource Management Optimization

This section discusses how YARN maintains optimal use by delving into its core role of efficiently allocating resources such as CPU and memory across diverse applications. The relationship between Application Master and Resource Manager, as well as the role of NodeManagers, is critical for improved agricultural resource management.

J. Agricultural Data Processing Using YARN and Spark

Integrating Spark with YARN for parallelism provides a powerful approach for improving agricultural data processing and analysis. This section looks at how to use YARN's resource management with Spark's cluster computing technologies for parallel and distributed processing, allowing for complex machine learning algorithms and real-time decision assistance. Spark's distributed computing capabilities to better understand agricultural data, allocate resources more efficiently, and make data-driven decisions that will increase production and sustainability in the industry by deploying Spark in agriculture.

K. Spark Executor Configuration and Dynamic Allocation

This section describes the deployment of Spark applications, with a focus on the development of Application Masters and Spark executors. It looks into configuration settings such as the number of executors, the number of cores per executor, and the amount of memory per executor, stressing appropriate parallelism and dynamic allocation for better resource usage.

L. MapReduce Optimization Techniques for Agricultural Data

This section discusses optimization techniques in MapReduce, including data location optimization, partitioning, combiner functions, and job parallelism. It stresses the efficiency attained by MapReduce Online, which allows for continuous and interactive queries on streaming data for better resource utilisation in agriculture.

M. Pig: Using High-Level Scripting to Simplify Agricultural Data Analysis

This section examines the merits of Pig, a high-level scripting language built on Hadoop MapReduce, for agricultural data processing. Pig's simplified scripting language makes it easier to communicate complex data conversions and analytics procedures, making it a vital tool for the agriculture business.

N. Agricultural Data Processing Map-Reduce Execution Model

This section delves into the Map-Reduce execution model, explaining how Pig Latin is translated into MapReduce tasks that are run on a Hadoop cluster. It highlights MapReduce's parallel and distributed processing capabilities, making it a viable method for managing enormous agricultural datasets and generating significant insights.

O. Agricultural Data Processing Scalability and Efficiency

This section emphasises the combined contribution of the discussed technologies to boosting resource efficiency, sustainability, and production in farming techniques by addressing the scalability and efficiency elements of the discussed technologies. The combination of YARN and Spark, optimization techniques in MapReduce, and the use of Pig pave the way for productive and insightful data analysis in the agriculture business.

Moreover, a wide range of tools and algorithms for creating prediction models, clustering analysis, and recommendation systems are available in Spark's machine learning library (MLlib). These tools and algorithms can be used for agricultural use cases like disease detection, crop yield prediction, and resource optimization. Additionally, Spark's GraphX library's graph processing capabilities make it possible to analyse intricate agricultural networks, including supply chains, logistics for transportation, and social connections among farming communities. Organisations can use

Apache Storm Precision agriculture, agricultural yield prediction, and weather forecasting can all benefit from Apache Storm's distributed real-time data processing capabilities. Storm's real-time handling of high-velocity data streams enables agricultural enterprises to monitor and respond to changing field conditions. Because of its support for machine learning libraries and complex event processing, predictive algorithms for agricultural applications can be constructed. Storm's fault-tolerant architecture, which provides continuous data processing and analysis, enables increased availability and reliability for agricultural data systems. By applying Apache Storm to the agriculture industry, organisations can improve efficiency and long-term profitability. This enables the study of real-time agricultural data, the efficient allocation of resources, and the formulation of data-driven decisions.

P. *Real-time Stream Processing*

Apache Storm is specifically designed for processing real-time, streaming data. [20]It enables the processing of continuous data streams with low-latency requirements, making it suitable for applications that demand real-time analytics.

Q. *Scalability*

Storm is horizontally scalable, which means that more nodes may be added to the cluster to manage increasing demands. Because of its scalability, it is well-suited to handling massive amounts of data and addressing the demands of Big Data applications. In the event of a node loss, Apache Storm provides built-in fault tolerance by shifting processing jobs to other nodes. This ensures that the calculation remains uninterrupted and contributes to the system's reliability. While Apache Spark is commonly used for batch and micro-batch processing, Apache Storm is a better solution for applications that require low-latency and continuous streaming data processing. The decision between Spark and Storm is frequently determined by the use case's specific requirements and the nature of the data processing burden.

R. *Comparison of Apache Storm and Spark in Agricultural Domain:*

Parameter	**Apache Storm**	**Apache Spark**
Processing Model	Real-time stream processing	Batch and real-time processing
Use Case	Ideal for real-time data processing in streams	Suitable for both batch and real-time processing
Latency	Low latency, suitable for time-sensitive tasks	Lower latency for real-time processing
Ease of Use	More complex, designed for stream processing	Simpler to use with a unified batch and stream processing model
Programming Language Support	Primarily Java	Supports multiple languages (Java, Scala, Python)
Fault Tolerance	Strong fault tolerance through message hacking	Fault tolerance through lineage information and data replication
Scalability	Scales well for high-throughput streams	Highly scalable, suitable for large-scale data processing
Data Processing Paradigm	Tuple-based processing model	Resilient Distributed Datasets (RDDs) and DataFrames
Integration with Ecosystem	Integration with Hadoop and various data sources	Tight integration with Hadoop, extensive ecosystem support
Batch Processing	Limited support, not the primary focus	Strong support for batch processing
Ease of Deployment	Requires more manual setup and configuration	Easier deployment with built-in cluster management (via Spark Standalone, YARN, or Mesos)
Community Support	Active community support	Large and active open-source community support
Use in Agriculture	Suitable for monitoring and reacting to real-time events in precision agriculture, weather monitoring, and sensor data	Well-suited for analysing large datasets in agriculture, including crop yield prediction, soil analysis, and farm management

S. *NoSQL Databases in Agriculture: Trade-off Analysis*

NoSQL databases such as MongoDB, CouchDB, and HBase promote availability and partition tolerance over tight consistency, addressing issues associated with maintaining varied agricultural data sets. Despite not having the data integrity of relational databases, they allow for excellent preprocessing and analysis, allowing machine learning techniques to be used for smart agricultural decision-making.

T. Knowledge of CAP Theory in Distributed Systems

The CAP theorem describes the trade-offs in distributed systems between consistency, availability, and partition tolerance. Exploring these characteristics clarifies the trade-offs involved in developing solutions for effective data management in agriculture.

U. Using NoSQL Databases to Accelerate Data Processing

Because of their distributed and horizontally scalable nature, NoSQL databases expedite data processing. Their ability to efficiently handle enormous volumes of data, particularly in real-time circumstances, distinguishes them as valuable instruments for fast data analysis in agricultural applications.

The Role of Kafka in Agricultural Data Flow Management

Kafka, a versatile platform, improves agricultural data flow efficiency across precision farming, smart agriculture, and agri-tech applications. Its fault-tolerant, scalable, and high-throughput qualities make it well-suited for a variety of activities such as sensor data collection, equipment telemetry, supply chain management, and real-time environmental monitoring.

Parallelism and Efficiency in Apache Kafka Data Pipelines Kafka's distributed streaming infrastructure excels at parallelism by partitioning pipelines for high throughput and fault tolerance. This section looks at how Kafka keeps message order within partitions, which helps with both speed and sequence preservation in data processing.

V. Kafka Fault Tolerance: Ensuring Uninterrupted Processing

Parallelism in Kafka contributes to its fault-tolerance characteristics. This section describes how Kafka automatically reassigns partitions in the event of consumer failures within a group, ensuring continuous data processing and avoiding single points of failure.

W. Kafka Streams: Taking Advantage of Parallelism in Stream Processing

Kafka Streams is a high-level stream processing toolkit that extends Kafka's partitioning mechanism to allow for parallel processing across several instances. This section describes how Kafka Streams helps agricultural stream processing applications be robust and scalable.

IV. CONCLUSION

Our review study provides a thorough examination of the use of distributed and parallel computing for big data analytics in the agriculture industry. It investigated the role of different frameworks, such as Hadoop and Spark, in processing and analysing agricultural data in order to promote informed decision-making and optimise farming operations. However, the report acknowledges the limitations and obstacles connected with the integration of big data analytics in agriculture, such as data quality, infrastructure, and privacy concerns. Overall, our paper has shown how big data analytics, cloud computing, and parallel distributed processing have the potential to alter the agriculture industry and improve operational performance and decision-making.

A. Limitations

Despite the potential benefits of using big data analytics in agriculture, there are significant constraints and problems involved with its implementation. One drawback is a lack of uniformity in data collection and management, which can result in discrepancies and mistakes in agricultural data analysis. Another constraint is the high cost of establishing and maintaining big data analytics infrastructure, which may be too expensive for small-scale farms. Concerns have also been raised concerning data privacy and security, as well as the possibility of unexpected outcomes such as greater environmental impact. These constraints and problems must be solved in order to fully realise the potential of big data analytics in agriculture.

B. Future Work

There are various areas for future research in the application of distributed and parallel computing for big data analytics in agriculture. One area of future research will be to solve the issues of data quality, privacy, and security in agricultural data. Another area of future work will be to investigate new prospects for innovation and sustainability in farming systems utilising big data analytics. Furthermore, additional study is required to optimise the performance of various frameworks such as Hadoop and Spark in processing and analysing agricultural data.

REFERENCES

[1] M. Y. Sokiyna, M. J Aqel, and O. A. Naqshbandi, “Cloud computing technology algorithms capabilities in managing and processing big data in business organizations: Mapreduce, hadoop, parallel programming,” Journal of Information Technology Management, vol. 12, no. 3, pp. 100–113, 2020.

[2] T. Hussain, A. Sanga, and S. Mongia, “Big data hadoop tools and technologies: A review,” in Proceedings of International Conference on Advancements in Computing & Management (ICACM), 2019.

[3] Y. Zhang, J. Ren, J. Liu, C. Xu, H. Guo, and Y. Liu, “A survey on emerging computing paradigms for big data,” Chinese Journal of Electronics, vol. 26, no. 1, pp. 1–12, 2017.

[4] H. Su, “How Accurate are Predictions Made Using Big Data?,” in 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022), Atlantis Press, 2022, pp. 806–810.

[5] H. Elazhary, “Cloud computing for big data,” MAGNT Res Rep, vol. 2, no. 4, pp. 135–144, 2014.

[6] D. Agrawal, S. Das, and A. El Abbadi, “Big data and cloud computing: current state and future opportunities,” in Proceedings of the 14th international conference on extending database technology, 2011, pp. 530–533.

[7] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, “Uncertainty in big data analytics: survey, opportunities, and challenges,” J Big Data, vol. 6, no. 1, pp. 1–16, 2019.

[8] S. Kaffash, A. T. Nguyen, and J. Zhu, “Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis,” Int J Prod Econ, vol. 231, p. 107868, 2021.

[9] Z. H. Munim, M. Dushenko, V. J. Jimenez, M. H. Shakil, and M. Imset, “Big data and artificial [10] intelligence in the maritime industry: a bibliometric review and future research directions,” Maritime Policy & Management, vol. 47, no. 5, pp. 577–597, 2020.

[10] C. Ordonez, S. T. Al-Amin, and X. Zhou, “A simple low cost parallel architecture for big data analytics,” in 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 2827–2832.

[11] L. Belcastro, R. Cantini, F. Marozzo, A. Orsino, D. Talia, and P. Trunfio, “Programming big data analysis: principles and solutions,” J Big Data, vol. 9, no. 1, pp. 1–50, 2022.

[12] S. Zeebaree, H. Shukur, L. Haji, R. Zebari, K. Jacksi, and S. Abass, “Characteristics and Analysis of Hadoop Distributed Systems,” Technology Reports of Kansai University, vol. 62, pp. 1555–1564, Apr. 2020.

[13] P. Natesan, V. E. Sathishkumar, S. K. Mathivanan, M. Venkatasen, P. Jayagopal, and S. M. Allayear, “A Distributed Framework for Predictive Analytics Using Big Data and MapReduce Parallel Programming,” Math Probl Eng, vol. 2023, 2023.

[14] R. K. Verma, S. Singh, and Y. Mohan, “Importance Of Big Data And Cloud Computing Techniques In Modern Scenario,” J Algebr Stat, vol. 13, no. 2, pp. 1024–1043, 2022.

[15] J.-Y. Kim and J. Kim, “Optimized data processing analysis using big data cloud platform,” Journal of Knowledge Information Technology and Systems (JKITS), vol. 16, no. 1, pp. 1–7, 2021.

[16] A. P. D. R. Hassan and J. N. Hasoon, “Big Data Techniques: A Survey,” Iraqi Journal of Information Technology Vol, vol. 9, no. 4, p. 2018, 2019.

[17] S. C. #1 and Z. Ansari, “Apache Pig-A Data Flow Framework Based on Hadoop Map Reduce,” International Journal of Engineering Trends and Technology, vol. 50, 2017, Accessed: Dec. 21, 2023. [Online]. Available: http://www.ijettjournal.org

[18] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, “Pig latin: a not-so-foreign language for data processing,” in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, pp. 1099–1110.

[19] N. Deshai, S. Venkataramana, B. Sekhar, K. Srinivas, and G. P. Saradhi Varma, “A Study on Big Data Processing Frameworks: Spark and Storm,” in Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2, Springer, 2020, pp. 415–424.

[20] R. Myung, H. Yu, and D. Lee, “Optimizing parallelism of big data analytics at distributed computing system,” Int J Adv Sci Eng Inf Technol, vol. 7, no. 5, pp. 1716–1721, 2017.

[21] A. Ali, S. Naeem, S. Anam, and M. M. Ahmed, “A state of art survey for Big Data processing and NoSQL database architecture,” International Journal of Computing and Digital Systems, vol. 14, no. 1, p. 1, 2023.

[22] B. Leang, S. Ean, G.-A. Ryu, and K.-H. Yoo, “Improvement of Kafka streaming using partition and multi-threading in big data environment,” Sensors, vol. 19, no. 1, p. 134, 2019.

[23] Á. B. Hernández, M. S. Perez, S. Gupta, and V. Muntés-Mulero, “Using machine learning to optimize parallelism in big data applications,” Future Generation Computer Systems, vol. 86, pp. 1076–1092, 2018, doi: https://doi.org/10.1016/j.future.2017.07.003.

[24] D. Waga and K. Rabah, "Environmental Conditions’ Big Data Management and Cloud Computing Analytics for Sustainable Agriculture," *World Journal of Computer Application and Technology*, vol. 3, no. 2, pp. 73-81, 2014.

[25] Abawajy, J. (2015). Comprehensive analysis of big data variety landscape. International journal of parallel, emergent and distributed systems, 30(1), 5-14

[26] Toshniwal, A., Taneja, S., Shukla, A., Ramasamy, K., Patel, J. M., Kulkarni, S., ... & Murthy, R. (2014). Storm@ twitter. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data (pp. 147-156). - Taylor, J. R., & Kumar, L. (2016). Big data analytics in agriculture: A review. Computers and Electronics in Agriculture, 123, 389-398

[27] Shankarnarayan, V.K. and Ramakrishna, H. "Paradigm change in Indian agricultural practices using Big Data: Challenges and opportunities from field to plate." Information Processing in Agriculture, 7(3), pp.355-368.

[28] Yadav, R., Rathod, J. and Nair, V. "Big data meets small sensors in precision agriculture." International Journal of Computer Applications, 975, p.8887.

[29] Anjanamma, C. and Rao, N.S. "An internet of things (IoT) system development and implementation of data analytics in agriculture production safety enhancement." Materials Today: Proceedings.

[30] Kaur, R., Garg, R. and Aggarwal, H. "Big data analytics framework to identify crop disease and recommendation a solution." 2016 International Conference on Inventive Computation Technologies (ICICT) (Vol. 2, pp. 1-5). IEEE.

[31] Rajeswari, S., Suthendran, K. and Rajakumar, K. "A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics." 2017 international conference on intelligent computing and control (I2C2) (pp. 1-5). IEEE.

[32] Priya, R., Ramesh, D. and Khosla, E. "Crop prediction on the region belts of India: a Naïve Bayes MapReduce precision agricultural model." 2018 international conference on advances in computing, communications and informatics (ICACCI) (pp. 99-104). IEEE.

[33] Bhosale, S.V., Thombare, R.A., Dhemey, P.G. and Chaudhari, A.N. "Crop yield prediction using data analytics and hybrid approach." 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE. ,[object Object],

[34] Ip, R.H., Ang, L.M., Seng, K.P., Broster, J.C. and Pratley, J.E. "Big data and machine learning for crop protection." Computers and Electronics in Agriculture, 151, pp.376-383.

[35] "International Journal of Scientific Research in Engineering and Management (IJSREM) Volume: 06 Issue: 04 | April - 2022.