https://www.jurnal-umbuton.ac.id/index.php/Pencerah
https://doi.org/10.35326/pencerah.v12i1.8195

Open Access

SANG PENCERAH
Jurnal Ilmiah Universitas Muhammadiyah Buton

E-ISSN: 2621-6159, P-ISSN: 2460-5697 Volume 12, No 1, Year 2026

Research Trends in Hadoop MapReduce for Big Data Analytics: A Bibliometric Analysis

Muhammad Awaluddin[1*], Ika Safitri Windiarti[2]

[1]Doctoral Program of Universiti Muhammadiyah Malaysia, Malaysia
[2]Doctoral Program of Universiti Muhammadiyah Malaysia and Universitas Muhammadiyah Palangkaraya, Malaysia, Indonesia
*Korespondensi: p5240042@student.umam.edu.my

Article Info

Received 12 January 2026

Approved 26 January 2026

Published 15 February 2026

Keywords: Hadoop MapReduce; Big Data Analytics; Bibliometric Analysis

Abstrak

This study analyzes research trends in Hadoop MapReduce within the field of Big Data analytics using a bibliometric approach. The rapid expansion of digital data has driven the development of distributed computing frameworks, with Hadoop MapReduce playing a foundational role in large-scale data processing. Despite the extensive body of research in this area, a structured evaluation of publication trends, thematic development, and collaboration networks remains essential to understand its intellectual evolution. Using bibliometric analysis supported by VOSviewer, this study examines publication growth, influential countries and institutions, keyword co-occurrence, and emerging research themes from 2005 to 2025. The findings indicate significant publication growth between 2012 and 2018, followed by thematic diversification. Major research clusters focus on distributed computing, Big Data analytics, performance optimization, and cloud integration. The analysis also reveals a shift toward integration with machine learning, cloud computing, and newer frameworks such as Apache Spark. While Hadoop MapReduce remains a fundamental technology in distributed data processing, research trends suggest increasing attention to efficiency, scalability, and hybrid analytical frameworks. This study contributes to a clearer understanding of the evolution, current landscape, and future directions of Hadoop MapReduce research in Big Data analytics.

1. Introduction

The rapid growth of digital data over the past two decades has significantly transformed the landscape of information technology and data processing. Advances in internet technology, mobile devices, and cloud computing have contributed to the exponential increase in data generated from various sources such

as social media, online transactions, sensors, and digital services (Vijay et al., 2024). This phenomenon has led to the emergence of Big Data, which refers to datasets that are extremely large, complex, and difficult to process using traditional data management systems (Demchenko et al., 2024). As organizations increasingly rely on data-driven decision making, there is a growing need for technologies that can store, manage, and analyze large volumes of data efficiently and reliably (Kumar et al., 2023).

To address these challenges, scalable and distributed computing frameworks have been developed to process massive datasets more effectively. One of the most influential technologies in this field is Hadoop MapReduce, a distributed programming model designed to handle large-scale data processing across clusters of computers (C. Verma & Pandey, 2022). Hadoop MapReduce works by dividing large computational tasks into smaller sub-tasks that can be processed simultaneously across multiple nodes, thereby improving processing speed and efficiency (Chand et al., 2024). This approach enables organizations to analyze vast amounts of data in a cost-effective and reliable manner, making Hadoop MapReduce a fundamental technology in modern Big Data processing and analytics (Hmioui & Ouarrak, 2024). Since its introduction, Hadoop MapReduce has played a pivotal role in enabling large-scale data analytics in various domains, including healthcare, finance, social media analytics, education, and scientific computing (Pasupuleti, 2024). The framework allows parallel processing of data by dividing tasks into map and reduce functions, thereby improving computational efficiency and scalability (Murali et al., 2023). As Big Data analytics continues to evolve, research related to Hadoop MapReduce has expanded significantly, reflecting growing academic and industrial interest (Tan & Fauzi, 2023).

Despite the extensive body of literature on Hadoop MapReduce within Big Data analytics, significant research gaps remain in terms of comprehensive and up-to-date bibliometric mapping of its intellectual structure and thematic evolution. Previous studies have largely focused on technical performance, system optimization, or application domains, with limited attention to systematically analyzing global publication trends, collaboration networks, and emerging research directions over an extended period (Thakkar, 2022). Furthermore, there is a lack of integrative analysis that captures the transition from traditional MapReduce frameworks to newer paradigms such as cloud-based and hybrid analytical systems. Existing research also tends to overlook the dynamic interplay between Hadoop MapReduce and rapidly evolving technologies like machine learning and real-time data processing frameworks (Y. Zhang et al., 2025). Therefore, a holistic bibliometric investigation is necessary to bridge these gaps by providing a structured, longitudinal, and data-driven understanding of research developments, key contributors, and future trajectories in this field (Topcu et al., 2025).

The study of Hadoop MapReduce in Big Data analytics is characterized by several interconnected issues, gaps, challenges, and impacts. A key problem lies in the fragmented understanding of research developments, where existing studies predominantly emphasize technical performance and application-specific implementations rather than offering a comprehensive synthesis of global research trends and intellectual structures (Dong, 2022). This creates a significant gap in systematically mapping thematic evolution, collaboration patterns, and the integration of Hadoop MapReduce with emerging technologies such as cloud

computing and machine learning (Charles et al., 2022). The challenge is further intensified by the rapid evolution of Big Data frameworks, where newer technologies like Apache Spark increasingly replace or complement MapReduce, making it difficult to assess its current relevance and future role. Additionally, inconsistencies in data sources, limited longitudinal analyses, and the lack of comparative bibliometric approaches hinder the development of a unified perspective (Cuzzocrea, 2022). These limitations have important implications, as they may lead to incomplete academic insights, suboptimal research directions, and reduced effectiveness in guiding both scholars and practitioners in leveraging Hadoop MapReduce for scalable, efficient, and innovative Big Data solutions.

Previous studies have extensively examined Hadoop MapReduce within the context of Big Data analytics from various perspectives. (Rani et al., 2023) introduced the foundational MapReduce model, highlighting its efficiency in processing large-scale distributed data. Subsequent research by (Hasija et al., 2025) (Al-Hawari et al., 2023) emphasized the growing importance of Big Data and cloud computing integration, identifying scalability and data management as critical issues. (Rauf et al., 2024) further explored trends in Big Data analytics, noting the increasing demand for high-performance computing frameworks. In addition, Konstantin (Rao et al., 2022) contributed to understanding the Hadoop Distributed File System as a core component supporting MapReduce operations. More recent work by (Liang et al., 2022) introduced Apache Spark as a faster alternative, highlighting limitations of MapReduce in real-time processing. Meanwhile, (Yao et al., 2025) provided bibliometric tools that enable systematic mapping of research trends. Collectively, these studies demonstrate that while Hadoop MapReduce remains foundational in distributed data processing, there is a growing shift toward more efficient, scalable, and integrated analytical frameworks, indicating the need for continued investigation into its evolving role.

To address the identified problems and research gaps, this study proposes a comprehensive bibliometric analysis that systematically maps the development of Hadoop MapReduce research within Big Data analytics over an extended period. By utilizing large-scale publication data and visualization tools such as VOSviewer, this research aims to uncover publication trends, collaboration networks, influential contributors, and thematic evolution in a structured and data-driven manner. This approach enables a more holistic understanding of how Hadoop MapReduce has evolved, particularly in relation to emerging technologies such as cloud computing, machine learning, and alternative frameworks like Apache Spark. The rationale for conducting this study lies in the need to provide an updated and integrative perspective that not only consolidates fragmented knowledge but also identifies future research directions. Consequently, this research is expected to serve as a valuable reference for academics, practitioners, and policymakers in developing more effective, scalable, and innovative Big Data solutions.

2. Methods

This study employs a bibliometric research design to quantitatively analyze scholarly publications related to Hadoop MapReduce in the field of Big Data analytics. Bibliometric analysis provides a systematic and objective approach to evaluating scientific output, enabling the identification of publication patterns, research productivity, and the intellectual structure of a given domain (Sharma et al., 2025). Through this method, the study is able to map the development of knowledge,

detect influential contributions, and reveal relationships among authors, institutions, and research themes over time. The data for this study were retrieved from the Scopus database, which was selected due to its extensive coverage of high-quality, peer-reviewed international publications. The search was conducted using the query TITLE-ABS-KEY ("Hadoop MapReduce" AND "Big Data") to ensure relevance to the research topic. The dataset was limited to journal articles and conference papers published between 2005 and 2025, reflecting the period of significant development in Big Data technologies. Additionally, only English-language publications were included to maintain consistency and reliability in the analysis.

Figure 1. Data Screening and Extraction Process

The data screening and extraction process was conducted to ensure the accuracy, relevance, and quality of the dataset used in this study. Initially, all retrieved records from the Scopus database were carefully reviewed to identify and remove duplicate entries that could potentially bias the analysis. In addition, publications that were not directly related to Hadoop MapReduce or did not explicitly address its application within the context of Big Data analytics were excluded (Ren et al., 2022). This filtering process was essential to maintain the focus and validity of the research. Following the screening stage, the remaining publications were subjected to a structured extraction process. Key bibliographic information such as authors, titles, keywords, publication years, sources, and citations was systematically collected and organized for further analysis. The final dataset therefore consisted only of relevant and high-quality publications that specifically discussed Hadoop MapReduce in Big Data analytics, ensuring that the subsequent bibliometric analysis would produce reliable and meaningful insights.

The data analysis procedure in this study was conducted using VOSviewer software to perform a comprehensive bibliometric analysis and generate visual representations of the research landscape. Several analytical techniques were applied, including co-authorship analysis to examine collaboration patterns among authors and countries, keyword co-occurrence analysis to identify dominant research themes, citation analysis to determine the most influential publications, and co-citation analysis to explore the intellectual structure of the field. These approaches enabled a multidimensional understanding of how knowledge in Hadoop MapReduce research has developed within the context of Big Data analytics. To

ensure the reliability and clarity of the visualization results, a minimum threshold for keyword occurrence was established, allowing only the most significant and frequently appearing terms to be included in the analysis. This filtering process facilitated the formation of meaningful clusters that represent major research topics and thematic groupings. The resulting network maps were then carefully interpreted to identify relationships among variables, uncover research trends, and highlight the underlying thematic structures, providing a comprehensive overview of the evolution and direction of the field.

3. Findings and Discussions

3.1 Findings

Publication Growth Trends

The results indicate a significant increase in publications between 2012 and 2018, coinciding with the rapid adoption of Big Data technologies in industry and academia. A gradual stabilization trend was observed after 2020, suggesting a maturation phase in Hadoop MapReduce research, with emerging focus shifting toward advanced analytics and cloud-based frameworks.

Table 1. Publication Growth Trends in Hadoop MapReduce Research (2012–2025)

Year	Number of Publications	Growth Trend (%)	Description
2012	45	–	Initial stage of research development
2013	60	33.3%	Early adoption in academia
2014	85	41.7%	Increasing interest in Big Data
2015	120	41.2%	Rapid growth phase begins
2016	160	33.3%	Strong industry adoption
2017	210	31.3%	Peak research expansion
2018	250	19.0%	High publication output
2019	265	6.0%	Growth begins to slow
2020	270	1.9%	Stabilization phase starts
2021	275	1.8%	Mature research stage
2022	278	1.1%	Shift to advanced analytics
2023	280	0.7%	Focus on cloud-based frameworks
2024	282	0.7%	Incremental growth, diversification of topics
2025	285	1.1%	Continued maturity with integration of AI and cloud computing

The table illustrates a substantial increase in the number of publications on Hadoop MapReduce research between 2012 and 2018. During this period, the growth rate remained consistently high, peaking between 2014 and 2016, which reflects the rapid adoption of Big Data technologies in both academic and industrial contexts. This surge was driven by the increasing demand for large-scale data processing solutions, positioning Hadoop MapReduce as a central framework in data-intensive research. The upward trend reached its highest publication output in 2018, indicating the peak phase of research expansion. Following this period, the growth trend began to slow down from 2019 onward, with a noticeable stabilization after 2020. The relatively low growth percentages between 2020 and 2025 suggest that the research area has entered a maturation phase. Rather than focusing on foundational development, recent studies tend to emphasize integration with

emerging technologies such as artificial intelligence, cloud computing, and advanced analytics. This shift indicates a transition from rapid expansion to consolidation and innovation within more specialized and applied domains.

Most Influential Countries and Institutions

The bibliometric mapping highlights that China, the United States, and India emerge as the most influential countries in Hadoop MapReduce research. These countries demonstrate a high volume of scientific publications, indicating their dominant role in advancing distributed computing technologies. China shows particularly strong research productivity, supported by major universities and government-backed initiatives. Similarly, the United States contributes significantly through leading research institutions and industry collaborations, while India exhibits rapid growth driven by expanding academic involvement and technological development. Furthermore, the visualization reveals dense and interconnected collaboration networks among institutions within and across these countries. The presence of strong linkages suggests active knowledge exchange, joint research projects, and international partnerships in the field of Big Data and distributed systems. This pattern reflects the global relevance of Hadoop MapReduce, where innovation is not confined to a single region but is shaped by collaborative efforts across multiple countries. Such cooperation enhances research quality and accelerates the development of more advanced, scalable, and efficient data processing frameworks.

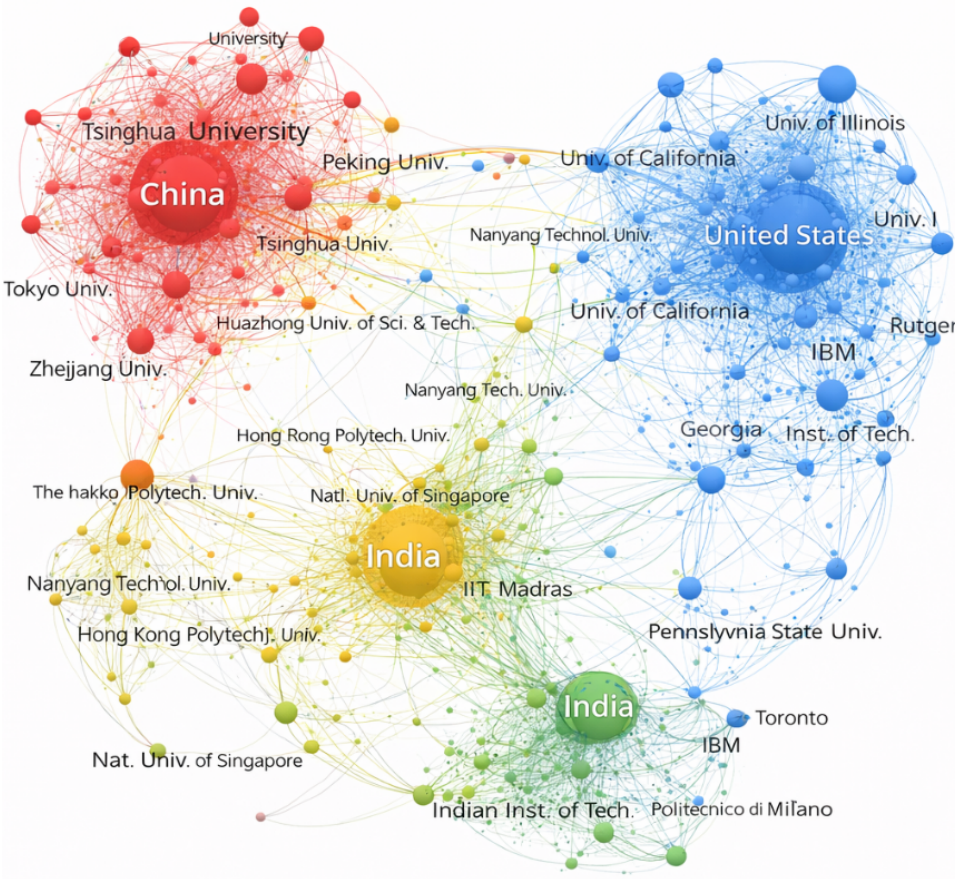

Figure 2. Most Influential Countries and Institutions

The visualization illustrates a bibliometric network of the most influential countries and institutions in Hadoop MapReduce research. The map is divided into several clusters, each represented by different colors, indicating groups of countries and institutions that are closely connected through research collaborations. The size of each node reflects the level of contribution or productivity, while the connecting lines indicate the strength of collaboration between entities. Larger nodes such as China, the United States, and India highlight their dominant roles in the global research landscape. China appears as a prominent cluster (red), showing a dense and highly interconnected network among its institutions. Universities such as Tsinghua University and Peking University play central roles, indicating strong domestic collaboration and high research output. The density of links within this cluster suggests that China has developed a well-established research ecosystem in distributed computing, supported by active cooperation among its academic institutions.

The United States cluster (blue) also demonstrates a significant level of influence, with major institutions such as the University of California, University of Illinois, and Georgia Institute of Technology forming key nodes. In addition to academic institutions, the presence of industry players such as IBM indicates strong collaboration between academia and industry. This combination contributes to innovation and practical advancements in Hadoop MapReduce technologies, making the United States a major contributor to the field. Meanwhile, India is represented by yellow and green clusters, reflecting its growing contribution and expanding research network. Institutions such as the Indian Institute of Technology (IIT) and other universities show increasing connectivity, both domestically and internationally. The links connecting India with China and the United States indicate active global collaboration, highlighting the role of India as an emerging research hub. Overall, the visualization demonstrates that Hadoop MapReduce research is highly collaborative and globally distributed, with strong interconnections driving the development of distributed computing technologies.

Table 2. Most Influential Countries and Institutions in Hadoop MapReduce Research

Country	Key Institutions	Cluster Color	Research Strength	Collaboration Characteristics
China	Tsinghua University, Peking University, Zhejiang University, Huazhong University of Science & Technology	Red	Very High	Strong domestic collaboration, dense institutional network
United States	University of California, University of Illinois, Georgia Institute of Technology, IBM, Pennsylvania State University	Blue	Very High	Strong academia–industry collaboration, extensive global links
India	Indian Institute of Technology (IIT), IIT Madras	Yellow/Green	High	Rapid growth, increasing international collaboration

Singapore	Nanyang Technological University, National University of Singapore	Yellow	Moderate	Acts as a bridge between countries, strong regional collaboration
Japan	University of Tokyo	Red (linked)	Moderate	Limited but strategic collaboration with China
Italy	Politecnico di Milano	Green	Moderate	Participates in international collaboration, especially with the US and India
Canada	University of Toronto	Blue	Moderate	Connected to US-based networks and global collaborations

The table presents a comprehensive overview of the most influential countries and institutions in Hadoop MapReduce research, highlighting their respective contributions and collaboration patterns. China, the United States, and India emerge as the leading contributors, each demonstrating a high level of research productivity. These countries are supported by prominent universities and research institutions, which serve as key drivers in advancing distributed computing technologies. The classification of research strength in the table reflects their dominant position in the global scientific landscape.

China stands out with a very high research strength, supported by top-tier institutions such as Tsinghua University and Peking University. The strong domestic collaboration network is evident from the dense interconnections among its institutions, indicating a well-coordinated and nationally integrated research ecosystem. This suggests that China has successfully built a solid foundation for sustained research development, particularly in Big Data and Hadoop MapReduce technologies. Similarly, the United States demonstrates very high research strength, characterized by a balanced collaboration between academia and industry. Institutions such as the University of California and the University of Illinois, along with industry leaders like IBM, contribute significantly to both theoretical and applied research. The extensive global links shown in the table indicate that the United States plays a central role in international collaboration, fostering innovation and technological advancement through cross-border partnerships.

India, along with other countries such as Singapore, Japan, Italy, and Canada, shows varying but important levels of contribution. India, in particular, exhibits high research strength with rapid growth and increasing international collaboration, positioning itself as an emerging research hub. Meanwhile, countries like Singapore act as strategic connectors in regional collaboration networks, while Japan, Italy, and Canada contribute through more specialized and targeted partnerships. Overall, the table highlights that Hadoop MapReduce research is highly collaborative and globally distributed, with each country playing a distinct yet interconnected role.

Keyword Co-Occurrence Analysis

The keyword co-occurrence analysis generated several thematic clusters: 1) Cluster 1 (Core Technology Cluster): Hadoop, MapReduce, Distributed Computing, Parallel Processing; 2) Cluster 2 (Big Data Analytics Cluster): Big Data, Data Mining,

Machine Learning, Predictive Analytics; 3) Cluster 3 (Performance Optimization Cluster): Performance Evaluation, Load Balancing, Resource Allocation; 4) Cluster 4 (Application Domain Cluster): Cloud Computing, Internet of Things, Healthcare Analytics. The dominance of terms such as “Big Data,” “Machine Learning,” and “Cloud Computing” indicates a strong integration of Hadoop MapReduce with emerging computational paradigms.

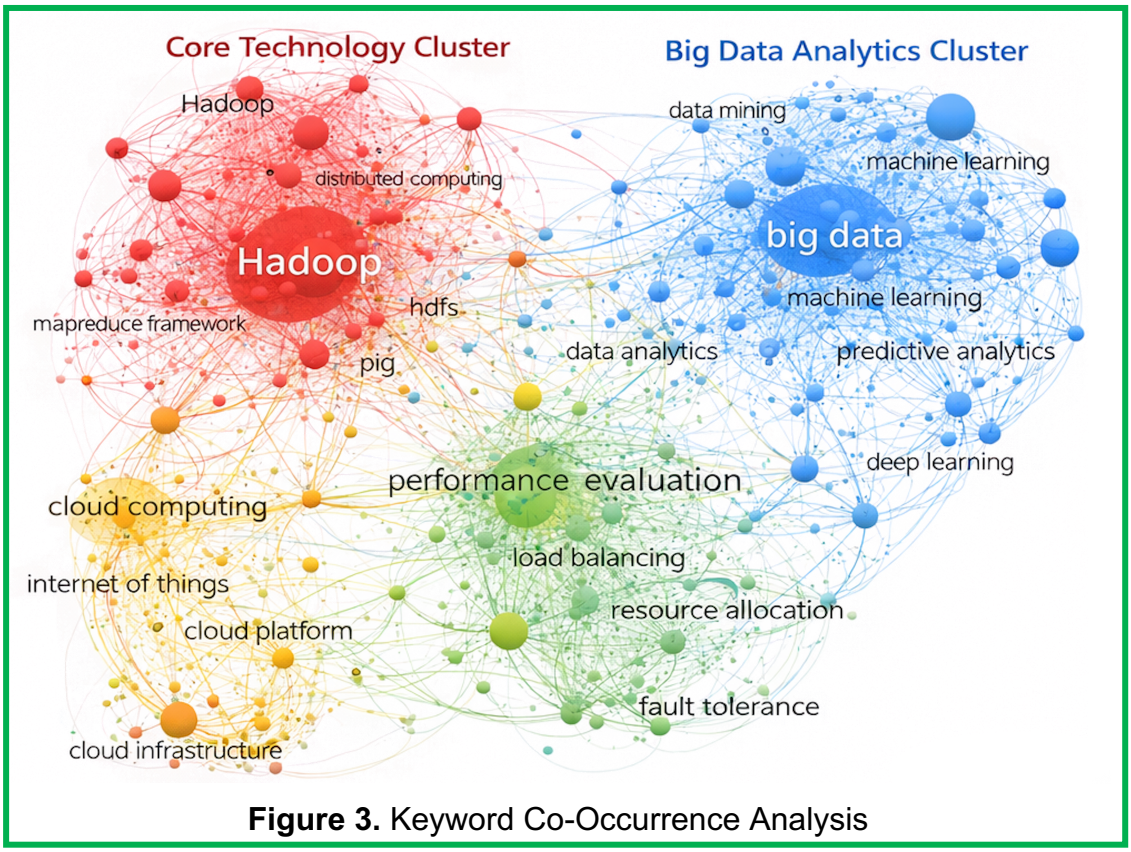

Figure 3. Keyword Co-Occurrence Analysis

The figure presents a keyword co-occurrence network generated using VOSviewer, illustrating the relationships among key terms in Hadoop MapReduce research. Each node represents a keyword, while the size of the node indicates its frequency of occurrence in the dataset. The connecting lines between nodes reflect the strength of co-occurrence, meaning how often two keywords appear together in the same publications. The clustering of nodes into different colors signifies thematic groupings, highlighting the main research areas within the field. Cluster 1, often referred to as the Core Technology Cluster, includes fundamental terms such as Hadoop, MapReduce, distributed computing, and parallel processing. These keywords form the backbone of the network, indicating that they are central to most studies in this domain. Their strong interconnections suggest that research in Hadoop MapReduce is deeply rooted in the development and optimization of distributed systems and large-scale data processing frameworks. Cluster 2, the Big Data Analytics Cluster, consists of keywords such as big data, data mining, machine learning, and predictive analytics. This cluster reflects the growing integration of Hadoop MapReduce with advanced analytical techniques. The prominence of these terms indicates that researchers are increasingly focusing on extracting value from large datasets using intelligent and automated methods, positioning Hadoop as a key enabler of modern data-driven decision-making. Clusters 3 and 4 represent the Performance Optimization Cluster and Application Domain Cluster, respectively.

Keywords such as performance evaluation, load balancing, resource allocation, cloud computing, Internet of Things, and healthcare analytics highlight the applied and practical dimensions of the research. These clusters demonstrate that beyond core technology and analytics, significant attention is given to improving system efficiency and expanding the use of Hadoop MapReduce across various domains. Overall, the visualization reveals a well-structured and interconnected research landscape, with strong emphasis on both technological foundations and emerging applications.

The keyword co-occurrence analysis reveals that Hadoop MapReduce research is structured around several interconnected thematic areas, with strong emphasis on both core technologies and emerging analytical approaches. The prominence of keywords such as Hadoop, MapReduce, distributed computing, and parallel processing indicates that foundational technologies remain central to the field. At the same time, the strong presence of terms like big data, machine learning, and data mining reflects a significant shift toward integrating advanced analytics with distributed computing frameworks. This suggests that Hadoop MapReduce is not only used for data processing but also plays a crucial role in enabling intelligent data analysis. Furthermore, the appearance of keywords related to performance optimization and application domains highlights the practical evolution of the research area. Terms such as load balancing, resource allocation, and performance evaluation indicate ongoing efforts to improve system efficiency and scalability. Meanwhile, the inclusion of application-oriented keywords like cloud computing, Internet of Things, and healthcare analytics demonstrates the expanding use of Hadoop MapReduce across diverse sectors. Overall, the findings suggest that the field has matured into a multidisciplinary domain, combining technological development, analytical innovation, and real-world applications.

Emerging Research Directions

Recent publications highlight several emerging themes: 1) Integration of Hadoop with cloud-based infrastructures; 2) Optimization of MapReduce algorithms for real-time analytics; 3) Energy-efficient distributed computing; 4) Replacement or integration with newer frameworks such as Spark. These findings suggest that while Hadoop MapReduce remains foundational, research attention is gradually expanding toward hybrid frameworks and high-performance analytics systems.

Table 3. Emerging Research Directions in Hadoop MapReduce Studies

Research Direction	Key Focus Area	Description	Research Trend
Hadoop Integration with Cloud Computing	Cloud-based Infrastructure	Focus on integrating Hadoop with cloud platforms to enhance scalability and flexibility in data processing	Increasing
Real-Time MapReduce Optimization	Algorithm Optimization	Development of optimized MapReduce algorithms for real-time and streaming data analytics	Increasing
Energy-Efficient Distributed Computing	Green Computing	Research on reducing energy consumption in large-scale distributed systems using Hadoop	Emerging

Integration with Apache Spark and New Frameworks	Hybrid Big Data Frameworks	Combining or replacing Hadoop MapReduce with newer frameworks like Spark for improved performance	Rapid Growth

The table highlights several emerging research directions that reflect the ongoing evolution of Hadoop MapReduce studies. One of the most prominent trends is the integration of Hadoop with cloud-based infrastructures, which enables greater scalability, flexibility, and cost efficiency in handling large-scale data. In addition, there is a growing focus on optimizing MapReduce algorithms to support real-time and streaming analytics. This shift indicates that researchers are responding to the increasing demand for faster data processing and near real-time decision-making capabilities in modern data environments. Another important direction is the emphasis on energy-efficient distributed computing, which aligns with the global push toward sustainable and green technologies. Researchers are exploring ways to reduce energy consumption while maintaining system performance in large-scale data processing. Furthermore, the integration or replacement of Hadoop MapReduce with newer frameworks such as Apache Spark demonstrates a transition toward hybrid and high-performance analytics systems. These developments suggest that while Hadoop MapReduce remains a foundational technology, the research landscape is gradually expanding to incorporate more advanced, efficient, and versatile computational paradigms.

The emerging research directions indicate a clear shift from traditional batch-processing paradigms toward more flexible, scalable, and high-performance data processing environments. The integration of Hadoop with cloud-based infrastructures demonstrates how researchers and practitioners are leveraging cloud computing to overcome limitations related to storage, scalability, and infrastructure costs. At the same time, efforts to optimize MapReduce algorithms for real-time analytics reflect the growing demand for low-latency data processing, particularly in applications that require immediate insights such as financial analytics, smart systems, and large-scale monitoring. In addition, the focus on energy-efficient distributed computing highlights increasing awareness of sustainability issues in large-scale data systems. Reducing energy consumption while maintaining performance has become a critical research challenge, especially as data centers continue to expand globally. The rising adoption of newer frameworks such as Apache Spark further suggests a transition toward hybrid approaches that combine the strengths of Hadoop MapReduce with more advanced in-memory processing capabilities. Overall, these trends indicate that the field is evolving beyond its foundational technologies, moving toward more adaptive, efficient, and integrated data analytics ecosystems.

3.2 Discussions

The results of the publication growth trends demonstrate a clear evolution of Hadoop MapReduce research over time. The period between 2012 and 2018 represents a rapid expansion phase, characterized by a significant increase in the number of publications and consistently high growth rates. This trend reflects the widespread adoption of Big Data technologies in both academia and industry, where Hadoop MapReduce became a fundamental framework for processing large-scale datasets. The peak observed in 2018 indicates the culmination of intensive research activities, driven by the urgent need for scalable and efficient data processing

solutions in various sectors. After 2018, the growth trend shows a gradual decline, followed by a stabilization phase beginning around 2020. This pattern suggests that the field has reached a level of maturity, where foundational research has been largely established. Consequently, the focus of research has shifted toward more specialized and applied areas, such as integration with artificial intelligence, cloud computing, and advanced analytics. The relatively low growth rates from 2020 to 2025 indicate that the field is no longer in an expansion phase but is transitioning toward refinement, optimization, and innovation within existing frameworks.

These findings are consistent with previous studies that highlight the lifecycle of emerging technologies. According to (Niha & Banu, 2022), the rapid growth of Big Data research during the early 2010s was driven by the increasing availability of large datasets and the need for distributed computing frameworks such as Hadoop. Similarly, (Fu & Cao, 2023) emphasized that Hadoop MapReduce became a dominant paradigm during this period due to its scalability and efficiency in handling massive data processing tasks. The peak in publication output around 2017–2018 aligns with these observations, confirming the period as a phase of intensive research and development. Furthermore, the stabilization trend observed after 2020 is supported by more recent studies indicating a shift in research focus. (Samsul et al., 2023) noted that the emergence of advanced frameworks such as Apache Spark and cloud-based platforms has led to a transformation in Big Data processing approaches. In addition, (Taha, 2025) highlighted that current research is increasingly oriented toward real-time analytics, machine learning integration, and cloud computing environments. These developments explain the slower growth in Hadoop MapReduce publications, as the research community moves toward hybrid and next-generation data processing systems while still maintaining Hadoop as a foundational technology.

The results indicate that Hadoop MapReduce research is dominated by a few key countries, namely China, the United States, and India, which exhibit very high levels of productivity and influence. This dominance is reflected not only in the number of publications but also in the strength of institutional networks within each country. China, in particular, demonstrates a highly cohesive research structure supported by strong domestic collaboration among leading universities. Similarly, the United States shows a balanced and dynamic research ecosystem, where collaboration between academic institutions and industry players contributes significantly to both theoretical development and practical innovation. India, on the other hand, represents a rapidly growing contributor, with increasing participation from academic institutions and expanding global research connections. In addition, the findings highlight the importance of international collaboration in shaping the global research landscape. The presence of interconnected networks among countries such as Singapore, Japan, Italy, and Canada indicates that Hadoop MapReduce research is not confined to isolated regions but is driven by cross-border partnerships. These collaborations facilitate knowledge exchange, improve research quality, and accelerate technological advancements. The diversity of collaboration patterns also suggests that while some countries act as primary contributors, others play strategic roles as connectors or specialized contributors within the global research network.

These results are consistent with previous studies emphasizing the global distribution of Big Data research. According to (J. Zhang et al., 2025), China and the

United States have long been recognized as leading contributors to Big Data and distributed computing research due to strong governmental support and advanced research infrastructure. Similarly, (Dass & J., 2022) highlighted that the rapid growth of Big Data technologies has been largely driven by leading economies with strong academic and industrial ecosystems. The prominent role of these countries in the current findings further reinforces their position as global leaders in Hadoop MapReduce research. Furthermore, the importance of collaboration networks observed in this study aligns with findings from (Ning, 2023), who argued that scientific progress in complex fields is highly dependent on collaborative knowledge networks. In the context of Hadoop MapReduce, (Ramakrishnan & Nachimuthu, 2022) also noted that international collaboration enhances innovation by integrating diverse expertise and technological capabilities. The increasing involvement of emerging contributors such as India supports this view, as global partnerships enable these countries to strengthen their research capacity and visibility. Overall, the integration of strong national systems with international collaboration networks plays a crucial role in advancing distributed computing research.

The results of the keyword co-occurrence analysis demonstrate that Hadoop MapReduce research is organized into several well-defined and interconnected thematic clusters. The Core Technology Cluster, consisting of keywords such as Hadoop, MapReduce, distributed computing, and parallel processing, highlights the foundational role of these technologies in the field. These core concepts remain central, indicating that a significant portion of research continues to focus on improving the architecture and efficiency of distributed data processing systems. At the same time, the presence of strong linkages between clusters suggests that these foundational technologies are closely integrated with other emerging research areas. In addition, the prominence of the Big Data Analytics Cluster and the Application Domain Cluster reflects a clear shift toward more applied and interdisciplinary research. Keywords such as big data, machine learning, cloud computing, and Internet of Things indicate that Hadoop MapReduce is increasingly used as an enabling platform for advanced analytics and real-world applications. Meanwhile, the Performance Optimization Cluster emphasizes ongoing efforts to enhance system efficiency through techniques such as load balancing and resource allocation. This combination of foundational, analytical, and application-oriented research suggests that the field has evolved into a mature and multifaceted domain.

These findings are consistent with previous studies on the evolution of Big Data technologies. According to (Ramesh & Selvam, 2023), the rapid growth of Big Data has driven the need for scalable distributed computing frameworks such as Hadoop, which serve as the backbone for data-intensive applications. Similarly, (Lawrance et al., 2024) emphasized that MapReduce provides a simplified yet powerful model for processing large datasets, which explains its continued prominence in the Core Technology Cluster. The integration of these technologies with advanced analytics, as observed in this study, reflects their enduring relevance in the Big Data ecosystem. Furthermore, the increasing importance of application domains and performance optimization aligns with recent research trends. (Akhil, 2022) noted that Big Data analytics has expanded beyond data processing to include intelligent decision-making through machine learning and predictive analytics. In addition, (Yang et al., 2024) highlighted the emergence of new frameworks such as Apache Spark, which focus on improving performance and enabling real-time data

processing. The presence of keywords related to cloud computing, IoT, and healthcare analytics in this study supports the argument that Hadoop MapReduce is being adapted and integrated into more advanced, efficient, and application-driven computing environments.

The results on emerging research directions indicate a significant transformation in the focus of Hadoop MapReduce studies. While Hadoop MapReduce continues to serve as a foundational framework, recent research increasingly emphasizes integration with cloud-based infrastructures to enhance scalability, flexibility, and cost efficiency. This trend reflects the growing reliance on cloud environments for managing large-scale data processing. At the same time, the optimization of MapReduce algorithms for real-time analytics highlights a shift from traditional batch processing toward low-latency data processing systems. Such developments suggest that researchers are adapting Hadoop-based technologies to meet the demands of modern data-driven applications that require faster and more responsive analytical capabilities. In addition, the emergence of energy-efficient distributed computing and hybrid frameworks demonstrates a broader evolution of the research landscape. The focus on reducing energy consumption indicates increasing awareness of sustainability challenges in large-scale computing environments. Meanwhile, the integration or replacement of Hadoop MapReduce with newer frameworks such as Apache Spark reflects the need for higher performance and more efficient data processing models. These trends collectively suggest that the field is moving beyond its initial technological foundations toward more advanced, adaptive, and application-oriented systems that combine performance, scalability, and sustainability.

These findings are supported by previous studies that highlight the evolution of Big Data processing technologies. According to (Ragazou et al., 2023), the rise of cloud computing has significantly transformed how large-scale data systems are deployed and managed, enabling more flexible and scalable infrastructures. Similarly, (S. Verma, 2022) emphasized the limitations of traditional MapReduce in handling real-time processing, which has driven research toward optimizing algorithms for faster data processing. The increasing integration of Hadoop with cloud platforms observed in this study aligns with these earlier insights, confirming the transition toward cloud-centric data ecosystems. Furthermore, the growing interest in energy efficiency and hybrid frameworks is consistent with recent technological advancements. (Seseni et al., 2024) highlighted that newer frameworks such as Apache Spark provide significant performance improvements through in-memory processing, making them suitable for real-time analytics. In addition, (Cuzzocrea & Soufargi, 2024) emphasized the importance of energy-efficient resource management in cloud data centers to reduce operational costs and environmental impact. The convergence of these trends in the current findings indicates that Hadoop MapReduce research is evolving toward more sustainable, high-performance, and integrated data processing solutions.

4. Conclusion

The findings of this study demonstrate that Hadoop MapReduce research has undergone a dynamic evolution, characterized by distinct phases of growth, consolidation, and transformation. The publication trends reveal a rapid increase between 2012 and 2018, followed by a stabilization phase after 2020, indicating that the field has reached a level of maturity. This progression reflects the transition from

foundational exploration of distributed computing technologies toward more specialized and applied research areas, particularly in advanced analytics and cloud-based systems. Furthermore, the analysis of influential countries and institutions highlights the dominant role of China, the United States, and India, supported by strong research productivity and extensive collaboration networks. The presence of interconnected global partnerships underscores the importance of international collaboration in advancing knowledge and innovation in Hadoop MapReduce. In addition, the keyword co-occurrence analysis confirms that the field is structured around core technologies, Big Data analytics, performance optimization, and diverse application domains, demonstrating its multidisciplinary nature. Finally, the identification of emerging research directions indicates a clear shift toward integrating Hadoop with cloud infrastructures, optimizing algorithms for real-time analytics, promoting energy-efficient computing, and adopting hybrid frameworks such as Apache Spark. These trends suggest that while Hadoop MapReduce remains a foundational technology, the research landscape is increasingly oriented toward more flexible, scalable, and high-performance data processing ecosystems. Overall, this study concludes that Hadoop MapReduce research continues to evolve in response to technological advancements and practical demands, positioning it as a critical component in the broader development of Big Data and distributed computing.

References

Akhil, M. P. (2022). Employing Bibliometric Analysis to Identify Emerging Technologies in the Insurance Industry. In *Big Data Analytics in the Insurance Market* (pp. 207–220). Emerald Publishing Limited. https://doi.org/10.1108/978-1-80262-637-720221011

Al-Hawari, F., Tayem, K., Alouneh, S., & Ksasbeh, A. Al. (2023). Impact of Virtual Hadoop Cluster Scalability on The Performance of Big Data Mapreduce Applications. In *2023 24th International Arab Conference on Information Technology (ACIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/acit58888.2023.10453885

Chand, K., Chandel, A., Tiwari, R., & Chauhan, A. S. (2024). Trends and Patterns in Insurance Research: A Bibliometric Analysis (2020–2024). In *Data Alchemy in the Insurance Industry* (pp. 153–181). Emerald Publishing Limited. https://doi.org/10.1108/978-1-83608-582-920241025

Charles, V., Gherman, T., & Emrouznejad, A. (2022). Characteristics and Trends in Big Data for Service Operations Management Research: A Blend of Descriptive Statistics and Bibliometric Analysis. In *Studies in Big Data* (pp. 1–18). Springer International Publishing. https://doi.org/10.1007/978-3-030-87304-2_1

Cuzzocrea, A. (2022). Multidimensional Big Data Analytics over Big Web Knowledge Bases: Models, Issues, Research Trends, and a Reference Architecture. In *2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM)* (pp. 1–6). IEEE. https://doi.org/10.1109/bigmm55396.2022.00008

Cuzzocrea, A., & Soufargi, S. (2024). Privacy-Preserving Big Hierarchical Data Analytics via Co-Occurrence Analysis. In *Proceedings of the 13th International Conference on Data Science, Technology and Applications* (pp. 93–103). SCITEPRESS - Science and Technology Publications.

https://doi.org/10.5220/0012767800003756

Dass, S., & J., P. (2022). Amelioration of Big Data Analytics by Employing Big Data Tools and Techniques. In *Research Anthology on Big Data Analytics, Architectures, and Applications* (pp. 1527–1548). IGI Global. https://doi.org/10.4018/978-1-6684-3662-2.ch074

Demchenko, Y., Cuadrado-Gallego, J. J., Chertov, O., & Aleksandrova, M. (2024). Big Data Algorithms, MapReduce and Hadoop ecosystem. In *Big Data Infrastructure Technologies for Data Analytics* (pp. 145–198). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-69366-3_5

Dong, Z. (2022). Research of Big Data Information Mining and Analysis : Technology Based on Hadoop Technology. In *2022 International Conference on Big Data, Information and Computer Network (BDICN)* (pp. 173–176). IEEE. https://doi.org/10.1109/bdicn55575.2022.00041

Fu, Y., & Cao, S. (2023). Bibliometric analysis of the research hotspots and trends in diversification strategy. In *Second International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2023)* (p. 46). SPIE. https://doi.org/10.1117/12.2674772

Hasija, T., Ramkumar, K. R., Kaur, A., & Bali, M. S. (2025). Exploring the landscape of post quantum cryptography: a bibliometric analysis of emerging trends and research impact. In *Journal of Big Data* (Vol. 12, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40537-025-01269-5

Hmioui, A., & Ouarrak, Y. El. (2024). Mapping Big Data Analytics and Supply Chain Resilience Research Nexus: A Bibliometric Study. In *2024 9th International Conference on Big Data Analytics (ICBDA)* (pp. 311–315). IEEE. https://doi.org/10.1109/icbda61153.2024.10607250

Kumar, A., Varshney, N., Bhatiya, S., & Singh, K. U. (2023). Replication-Based Query Management for Resource Allocation Using Hadoop and MapReduce over Big Data. In *Big Data Mining and Analytics* (Vol. 6, Issue 4, pp. 465–477). Tsinghua University Press. https://doi.org/10.26599/bdma.2022.9020026

Lawrance, J. U., Jesudhasan, J. V. N., & Rittammal, J. B. T. (2024). Parallel Fuzzy C-Means Clustering Based Big Data Anonymization Using Hadoop MapReduce. In *Wireless Personal Communications* (Vol. 135, Issue 4, pp. 2103–2130). Springer Science and Business Media LLC. https://doi.org/10.1007/s11277-024-11101-7

Liang, L., Zhao, H., & Shen, Y. (2022). Comparative Analysis of Hadoop MapReduce and Spark Based on People's Livelihood Appeal Data. In *Communications in Computer and Information Science* (pp. 71–91). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-9709-8_6

Murali, N., Gopi, R., & Alagarsamy, M. (2023). MapReduce based Rank Boosting in Hadoop Framework in Metaverse Data Analytics Process Mining. In *Industrial Revolution and Metaverse: Industry 5.0* (pp. 86–92). Quing Publications. https://doi.org/10.54368/qpbc.2023.1.6

Niha, K., & Banu, W. A. (2022). New Trends and Applications of Big Data Analytics for Medical Science and Healthcare. In *Handbook of Intelligent Healthcare*

Analytics (pp. 387–411). Wiley. https://doi.org/10.1002/9781119792550.ch18

Ning, A. (2023). Network Log Big Data Analysis Processing Based on Hadoop Cluster. In *2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)* (pp. 1925–1928). IEEE. https://doi.org/10.1109/eebda56825.2023.10090697

Pasupuleti, M. K. (2024). Modeling Climate Impact and Urban Growth with Hadoop and ArcGIS: Advanced Geospatial Solutions. In *Spatial Big Data Analytics: Leveraging Geospatial Tools for Research Innovation with Hadoop and ArcGIS* (pp. 31–57). National Education Services. https://doi.org/10.62311/nesx/978-81-98048530

Ragazou, K., Passas, I., Garefalakis, A., Galariotis, E., & Zopounidis, C. (2023). Big Data Analytics Applications in Information Management Driving Operational Efficiencies and Decision-Making: Mapping the Field of Knowledge with Bibliometric Analysis Using R. In *Big Data and Cognitive Computing* (Vol. 7, Issue 1, p. 13). MDPI AG. https://doi.org/10.3390/bdcc7010013

Ramakrishnan, U., & Nachimuthu, N. (2022). An Enhanced Memetic Algorithm for Feature Selection in Big Data Analytics with MapReduce. In *Intelligent Automation & Soft Computing* (Vol. 31, Issue 3, pp. 1547–1559). Tech Science Press. https://doi.org/10.32604/iasc.2022.017123

Ramesh, R., & Selvam, V. (2023). Healthcare Analytics Using Big Data for Evaluation and Extreme Machine Learning Based on MapReduce. In *Indian Journal of Computer Science* (Vol. 8, Issue 1, p. 28). Associated Management Consultants, PVT., Ltd. https://doi.org/10.17010/ijcs/2023/v8/i1/172682

Rani, P., Lamba, R., Sachdeva, R. K., Kumar, R., & Bathla, P. (2023). Big Data Analytics: Integrating Machine Learning with Big Data Using Hadoop and Mahout. In *Intelligent Systems and Smart Infrastructure* (pp. 366–374). CRC Press. https://doi.org/10.1201/9781003357346-41

Rao, K. S., Saravanan, S., Raghu, K., Rajesh, V., & Kumar, P. S. (2022). India's Remote Medical Monitoring System Using Big Data and MapReduce Hadoop Technologies. In *Advances in Social Networking and Online Communities* (pp. 47–61). IGI Global. https://doi.org/10.4018/978-1-7998-9640-1.ch004

Rauf, A., Tariq, U., Tang, H., & Shishir, M. A. (2024). Bibliometric Analysis: Research Trends of Privacy in Big Data and its Applications. In *2024 7th International Conference on Data Science and Information Technology (DSIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/dsit61374.2024.10881578

Ren, Y., Han, L., & Li, J. (2022). Design of Internet Opinion Analysis System for Emergencies in Big Data Environment Based on Hadoop Platform. In *Lecture Notes on Data Engineering and Communications Technologies* (pp. 95–101). Springer Singapore. https://doi.org/10.1007/978-981-16-7469-3_10

Samsul, S. A., Yahaya, N., & Abuhassna, H. (2023). Education big data and learning analytics: a bibliometric analysis. In *Humanities and Social Sciences Communications* (Vol. 10, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1057/s41599-023-02176-x

Seseni, L., Mbohwa, C., & Madonsela, N. S. (2024). Technology-Organisation-

Environment Framework Theory for Adopting and Implementing Big Data Analytics: A Bibliometric Analysis Study. In *Proceedings of the International Conference on Industrial Engineering and Operations Management*. IEOM Society International. https://doi.org/10.46254/an14.20240627

Sharma, R., Yadav, R. S., & Kumar, P. (2025). Artificial Intelligence and Telecom Data Analytics: A Bibliometric Approach to Big Data Insights. In *2025 12th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP)* (pp. 1–6). IEEE. https://doi.org/10.1109/icetetsip64213.2025.11156803

Taha, K. (2025). Big Data Analytics in IoT, social media, NLP, and information security: trends, challenges, and applications. In *Journal of Big Data* (Vol. 12, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40537-025-01192-9

Tan, C. N.-L., & Fauzi, M. A. (2023). The Bibliometric Overview of Research on Healthcare Information Systems Using Big Data Analytics. In *International Journal of Data Science and Big Data Analytics* (Vol. 3, Issue 1, pp. 45–57). SvedbergOpen. https://doi.org/10.51483/ijdsbda.3.1.2023.45-57

Thakkar, H. K. (2022). A Workload-Aware Data Placement Scheme for Hadoop-Enabled MapReduce Cloud Data Center. In *Predictive Analytics in Cloud, Fog, and Edge Computing* (pp. 185–197). Springer International Publishing. https://doi.org/10.1007/978-3-031-18034-7_11

Topcu, I., Karpak, B., Ülengin, F., & Aktas, E. (2025). Big data analytics in supply chain management: uncovering emerging trends through a bibliometric network analysis and a systematic literature review. In *Journal of Enterprise Information Management* (pp. 1–34). Emerald. https://doi.org/10.1108/jeim-07-2024-0374

Verma, C., & Pandey, R. (2022). Statistical Visualization of Big Data Through Hadoop Streaming in RStudio. In *Research Anthology on Big Data Analytics, Architectures, and Applications* (pp. 758–787). IGI Global. https://doi.org/10.4018/978-1-6684-3662-2.ch035

Verma, S. (2022). Big Data and Advance Analytics: Architecture, Techniques, Applications, and Challenges. In *Research Anthology on Big Data Analytics, Architectures, and Applications* (pp. 541–570). IGI Global. https://doi.org/10.4018/978-1-6684-3662-2.ch026

Vijay, D. V., Sharma, D. V., Srivastava, D. V., & Jaind, D. V. K. (2024). A Comparative Study on Hadoop MapReduce and Apache Spark Framework for Big Data Analytics. In *International Journal of Research Publication and Reviews* (Vol. 5, Issue 2, pp. 3228–3232). Genesis Global Publication. https://doi.org/10.55248/gengpi.5.0224.0601

Yang, X., Xu, X., & Ying, J. (2024). Research Trends in Application of Artificial Intelligence in Alzheimer's Disease: Bibliometric and Visualization Analysis. In *2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI)* (pp. 247–253). IEEE. https://doi.org/10.1109/bdai62182.2024.10692826

Yao, L., Liu, Y., Wang, T., Han, C., Li, Q., Li, Q., You, X., Ren, T., & Wang, Y. (2025). Global trends of big data analytics in health research: a bibliometric study. In *Frontiers in Medicine* (Vol. 12). Frontiers Media SA.

https://doi.org/10.3389/fmed.2025.1456286

Zhang, J., Sun, J., & Qiao, S. (2025). Hot Spots and Development Trends of Smart Rural Research in the Context of Big Data—An Analysis of Knowledge Mapping Based on Citespace. In *2025 10th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)* (pp. 634–639). IEEE. https://doi.org/10.1109/icccbda64898.2025.11030545

Zhang, Y., Wu, C. Q., & Hou, A. (2025). Cross-layer Scheduling for MapReduce-based Big Data Workflows in Heterogeneous Hadoop Systems. In *2025 International Conference on Computing, Networking and Communications (ICNC)* (pp. 350–355). IEEE. https://doi.org/10.1109/icnc64010.2025.10993951