Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Distributed Systems for Data-Intensive Computing in Cloud Environments: A Review of Big Data Analytics and Data Management

Zeravan Arif Ali [1*], Subhi R. M. Zeebaree[2]
zeravan.ali@dpu.edu.krd, subhi.rafeeq@dpu.edu.krd
[1]IT Dept., Technical College of Duhok, Duhok Polytechnic University, Duhok, Iraq
[2]Energy Eng. Dept., Technical College of Engineering, Duhok Polytechnic University, Duhok, Iraq

Article Information

Submitted : 8 Mar 2024
Reviewed: 14 Mar 2024
Accepted : 15 Apr 2024

Keywords

Distributed Systems, Cloud Computing, Big Data Analytics, Data Management, Data-Intensive Computing, Cloud Environments

Abstract

Because of the increasing increase of data, which is frequently referred to as "big data," many different businesses have been severely impacted in recent years, necessitating the implementation of sophisticated data management and analytics solutions. By virtue of the fact that it provides scalable resources for applications that are data-intensive, cloud computing has emerged as an indispensable platform for the management of these enormous databases. The evolving landscape of distributed systems in cloud settings is the primary emphasis of this study, which is situated within the framework of big data analytics and data management. With the purpose of providing a comprehensive overview of distributed systems that are used in cloud settings for data-intensive computing, the review article seeks to offer. Furthermore, it evaluates the many ideas, techniques, and technical improvements that have been established in order to properly manage, store, and analyse large amounts of data. A comprehensive literature evaluation of recently published scientific references was successfully completed by our team. The analysis takes into account the theoretical foundations, as well as the research that has already been conducted on distributed computing systems, cloud-based data management, and enormous data analytics. The study places an emphasis on the significant role that distributed computing plays in ensuring the success of big data analytics. The interplay between distributed systems and cloud computing paradigms has resulted in the development of solutions that are robust, scalable, and economical for activities that need a significant amount of data. It is still a huge problem to ensure that data security, privacy, and interoperability are maintained across the many cloud services.

A. Introduction

In data-intensive computing, distributed systems have become essential, especially in cloud contexts. These systems efficiently handle the difficulties associated with large-scale data management and processing by dividing up data and computing workloads among several processors[1]. Distributed systems have the capacity to scale resources up or down in response to demand, which is particularly useful in cloud situations where scalability and resource management are critical. Managing big data workloads, which frequently call for significant processing and storage capacity, need this flexibility[2]. Additionally, distributed systems improve data availability and fault tolerance, guaranteeing data integrity and uninterrupted operation even in the event of network or hardware malfunctions[3]. Additionally, they make parallel processing possible, greatly speeding up processes like data processing and analysis. These characteristics are particularly crucial in cloud settings, which are frequently the foundation of data-driven services and applications in a variety of sectors, including retail, telecommunications, healthcare, and finance. Thus, a fundamental component of contemporary data-intensive computing techniques is the integration of distributed systems with cloud computing, which enables businesses to fully utilize big data in an effective, scalable, and dependable manner[4].

The use of big data analytics and data management in distributed cloud systems is revolutionizing the way large amounts of data are handled. Big data processing and analysis is made possible by the scalability, effectiveness, and affordability of cloud computing integrated with distributed systems[5]. These systems allow for speedier data analytics and parallel processing because the data is dispersed over several sites, possibly even worldwide[6]. Without having to make substantial upfront financial investments in infrastructure, enterprises can store and analyse massive datasets thanks to the cloud's adaptable and resource-efficient platform. This method improves data governance, security, and compliance while strengthening data management capabilities[7]. Furthermore, real-time data processing and analytics are made easier by cloud-based distributed systems, which is crucial for dynamic decision-making in a variety of industries, including e-commerce, healthcare, and finance. This paradigm change highlights the relevance of distributed computing in the era of big data, underlining the necessity for resilient and flexible data management solutions[8]. This review aims to analyse how big data is becoming more and more prevalent. Understanding distributed computing's role in managing such large amounts of data is critical given the exponential growth of data, and technological advancements offer insights into cutting-edge distributed computing technologies that are necessary in today's data-driven world[9]. It also helps to understand future trends in big data analytics, which will guide researchers and practitioners. Finally, addressing security in distributed systems is crucial. At last, our addition is to identify the pros, cons, and gaps regarding each approach, revealing disagreements related to the previous work.

B. Background Theory

2.1. Distributed Systems

Distributed systems are networks of independent computers that work together to achieve a common goal. This architecture allows for the sharing of resources and data among computers, enhancing performance and reliability[10].

A substantial change in the computing paradigm has been brought about by the development of distributed systems and their integration with cloud computing. Distributed systems were first created to connect multiple computers and enable data sharing and processing[11]. These systems developed over time to become more sophisticated and powerful, able to handle bigger datasets and more demanding applications as processing requirements increased[12]. This environment underwent even more transformation with the introduction of cloud computing. Utilizing the ideas of distributed computing, cloud computing provides scalable, on-demand computer resources via the internet[13]. Distributed systems offered the fundamental architecture for cloud services through this integration, creating a synergy that improved cost-effectiveness, accessibility, and efficiency[14]. These days, many current apps are built on cloud-based distributed systems, which provide scalable, reliable, and adaptable solutions for both individual users and companies, meeting a broad range of computing requirements[15]. This evolution aims to satisfy the ever-expanding data and computational needs of our digital world, reflecting the ongoing progress in technology[16].

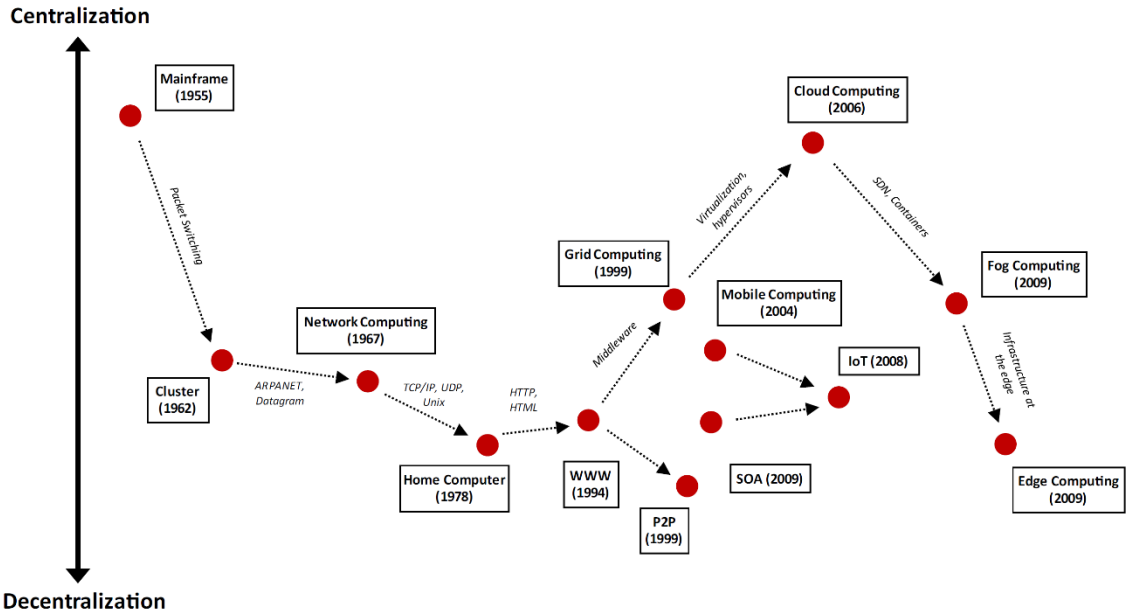

Figure1: Evolution of Distributed System Paradigm [16]

2.2. Cloud Computing

Cloud computing is a revolutionary technology that uses the internet (often known as "the cloud") to provide computing services, such as servers, storage, databases, networking, software, and more. Due to its pay-per-use model, which does not require large upfront capital expenditures on hardware and software, this invention offers flexible resources, rapid scaling, and cost efficiency[17]. Cloud

computing is transforming the way consumers and companies access and manage computational resources by supporting a wide range of applications and services. It improves accessibility, collaboration, and operational agility by allowing users to view and store data remotely[18].

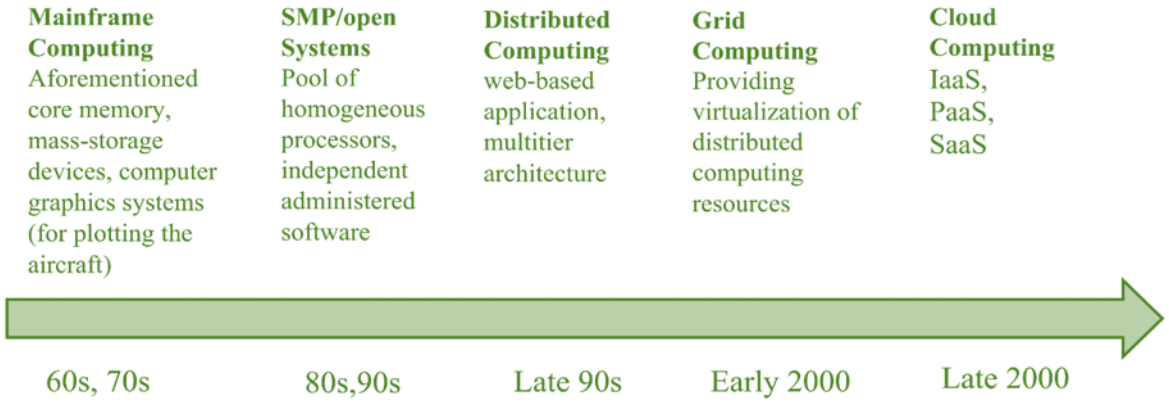

Figure2: Evolution of Cloud Computing Paradigm [18]

2.3. Big Data Analytics

Large and complex data sets are gathered, processed, and analysed to reveal hidden patterns, correlations, and insights. This process is known as big data analytics. To manage data that is too huge or complex for conventional data processing software, this field integrates big data technologies with advanced analytics approaches. Businesses and organizations may maximize operations, spot patterns, forecast future behaviour, and make better decisions with the help of big data analytics. It encourages creativity and efficiency in a variety of sectors, including healthcare, banking, retail, and telecommunications[19]. Significant progress in big data analytics for distributed systems has had a revolutionary effect. Key technologies that make it possible to handle massive datasets across distributed networks efficiently are Hadoop and Spark[20]. Advanced data analysis and predictive modelling are made possible by the integration of machine learning algorithms into these systems. Latency in data analysis has been greatly decreased by real-time processing capabilities, especially with the help of stream processing frameworks like Apache Storm. Data storage and retrieval efficiency have been improved with the usage of distributed databases like Cassandra and MongoDB. When combined, these technologies improve big data analytics' scalability, resilience, and speed in dispersed contexts[6].

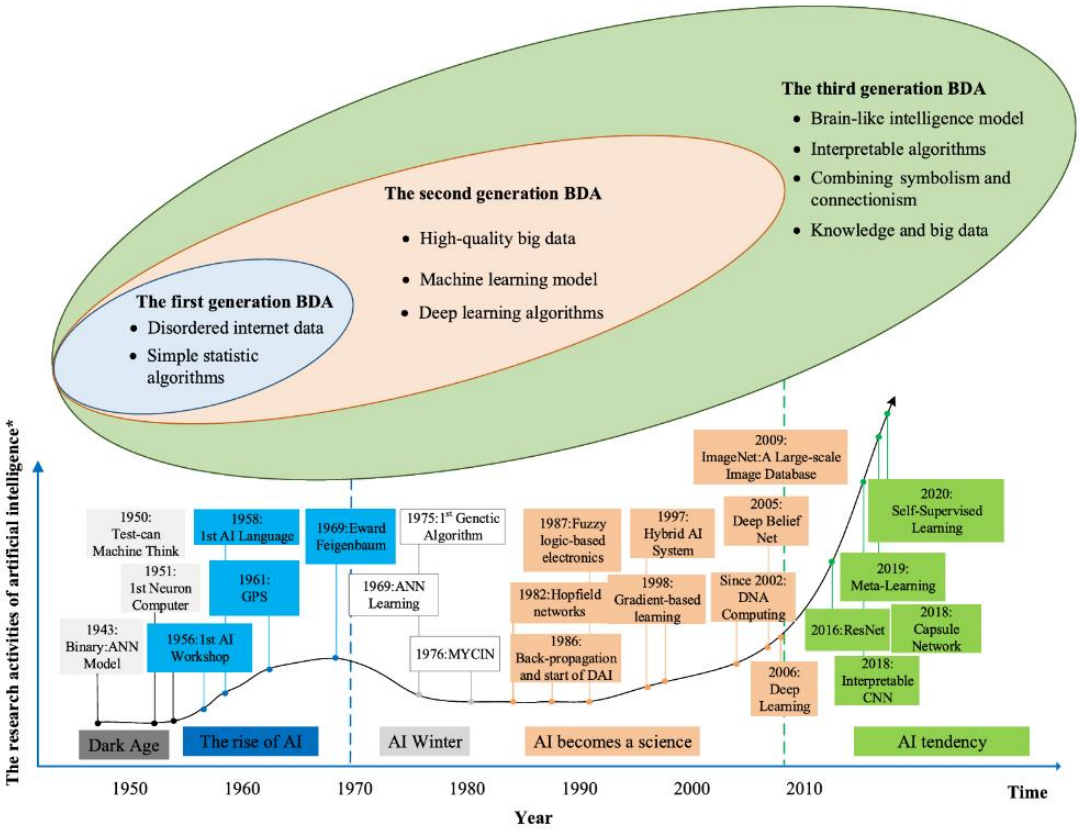

Figure 3: The development of big data analytics [6]

2.4. Data Management

Innovations in data management for distributed systems in the cloud focus on enhancing efficiency, scalability, and security. These advancements include automated data replication and synchronization across multiple cloud servers, ensuring data availability and fault tolerance. Big data technologies, like Hadoop and Spark, have been adapted for cloud environments to handle massive datasets efficiently. Additionally, there's a growing emphasis on implementing robust security measures and compliance protocols to protect sensitive data in distributed cloud systems, catering to the complex demands of modern data management[21].

C. Literature Review

3.1. Distributed Computing Principles

In [22] research focused specifically on the developing challenges that are brought about by the increasing amounts of big data, as well as the role that cloud computing services play in addressing these issues. Specifically, the study focused on the potential solutions to these problems. With regard to the United States of America, the study was carried out specifically within that setting. This article's objective is to give a comprehensive study of the definition, classification, and features of big data based on the information that is presented. In addition, the research offers an analysis of a number of cloud services, including Microsoft Azure, Google Cloud, Amazon Web Services, IBM cloud, Hortonworks, and MapR, as

well as an analysis of the relevance of these services in big data frameworks. In addition to that, the study includes an investigation of the significance of frameworks for the collection of large amounts of data. In addition to that, it contains a comparative study of a number of different big data frameworks that are hosted on the cloud.

[23] focused that the primary emphasis will be on describing the most recent advancements in parallel and distributed processing methods for handling huge amounts of data that have been remotely sensed. During the course of this session, the primary topic of conversation will take place. The purpose of this research is to analyse a broad range of different approaches to parallel and distributed repositories that are implemented on a computer network, which is also referred to as a network. These strategies include solutions that are based on cloud computing, grid computing, and cluster computing. These solutions are included in the scope of these strategies.

[24] explained that in order to achieve our objective of enhancing the parallelism of remote data processing, we conducted study on advanced development methodologies. This was undertaken in order to achieve our aim. Methods of parallel processing are utilised in the area of cloud computing, and this article provides an overview of the several ways that are utilised in relation to these approaches. It emphasises the need of using scheduling strategies in order to promote parallelism in settings that are either data-level or task-level, depending on the context. This is an extra point of attention that is brought to light by this.

The paper [25] An inquiry of the use of big data analytics and cloud computing in power management systems was carried out as part of this research project's scope of work. Taking into mind how successful the frameworks and technologies that are presently being used are in the administration and processing of enormous amounts of data for the purpose of power system monitoring and control, the objective of this research is to give an analysis of the frameworks and technologies that are currently being utilised.

The work [26] accomplished the goal of cloud manufacturing, the strategy that is being provided involves the creation of a framework that is defined by software. The framework in question is not only capable of being readily adjusted and programmed, but it also has the ability to adapt to a wide variety of diverse circumstances on its own. The company intends to accomplish this aim in a variety of different methods, including the use of big data analytics, the optimisation of systems, and the implementation of continuous improvements. This is all part of its mission, which is to promote the development of industrial systems. Building a manufacturing system that is capable of monitoring, reacting, optimising, and adapting on its own is one of the major aims that has to be accomplished. In the long run, this will result in increased reactivity, intelligence, resilience, energy conservation, and efficiency across the whole manufacturing process.

[27] The study was conducted with the intention of doing an examination of the methods that large-scale analytical organisations use in order to handle vast volumes of data. In the context of enterprises, it serves as a foundation for the development of data frames and processing models for the aim of CI (Corporate Intelligence). In order to show the approach and potential of Big Data Analytics (BDA) in terms of leveraging data, case studies of users who make use of BDA are

included in the research projects. These case studies are an integral part of the research projects. Big Data Analytics (BDA) has identified many components that are essential to its operation. These components include the coordination of service systems, data sources, and end-users. These components have been identified as being present.

The document [28] An investigation was conducted to look at the difficulties that arise as a result of the rapid development of large amounts of data and the massive quantity of data. These difficulties are associated with the visual depiction of enormous volumes of data. When it comes to the management of enormous amounts of data, traditional methods of data visualisation are not feasible options. As a consequence of this, the solutions that are now available often display a smaller portion of the data in attempt to address these issues. It is not easy to display enormous volumes of data in real time since there are several obstacles to overcome. Batch processing and real-time data processing are two examples of the sorts of big data computing methods that are necessary since different applications use different kinds of data. Other examples include data processing techniques such as batch processing and real-time data processing.

The fundamental principles of distributed computing, particularly relevant to cloud-based systems and big data analytics, involve several key concepts. Firstly, distributed systems rely on the idea of decentralization, where processing and storage tasks are spread across multiple nodes to enhance performance and reliability[29]. This approach is crucial for handling the large-scale, complex data sets typical in big data analytics. Secondly, distributed computing principles emphasize scalability, allowing systems to easily expand and contract resources in response to varying demand. This scalability is essential in cloud environments, where fluctuating workloads are common. Thirdly, fault tolerance and redundancy are integral, ensuring that the failure of one node doesn't compromise the entire system's functionality[30]. Finally, distributed computing includes efficient data distribution and parallel processing techniques, enabling rapid processing and analysis of big data, which are core to the functionality of cloud-based analytics platforms. These principles collectively ensure that distributed systems in cloud environments are robust, flexible, and capable of handling the intensive demands of big data analytics[31].

3.2. Cloud Computing Models

Infrastructure as a Service (IaaS): This model offers virtualized computing resources over the internet. It provides the essential infrastructure - servers, storage, networking capabilities - on a pay-as-you-go basis. IaaS is vital for big data analytics as it offers scalable resources to store and process vast amounts of data, without the need for physical hardware[32].

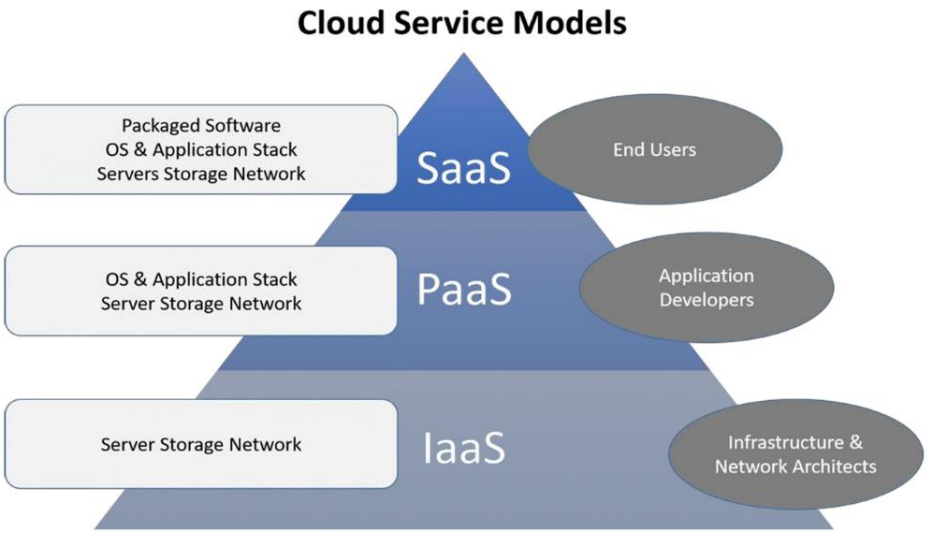

Fig 4. Framework for evaluation and ranking IaaS, PaaS and SaaS cloud service models [33]

Platform as a Service (PaaS): PaaS delivers a framework for developers to build upon and create customized applications. It includes operating systems, middleware, and runtime environments. For big data analytics, PaaS offers a platform with built-in tools and services, enabling developers to quickly create and deploy analytics applications without worrying about underlying infrastructure[29].

Software as a Service (SaaS): SaaS provides software applications over the internet, on a subscription basis. It's user-friendly and doesn't require the installation of applications on individual computers. In big data analytics, SaaS can include analytics tools and software that are accessible from anywhere, making data analysis more accessible and reducing the need for extensive in-house hardware and software maintenance[29].

Table 1: Commonality between Distributed Computing and Cloud Computing

Feature	Distributed Computing	Cloud Computing
Definition	Involves a network of autonomous computers that work together to perform a task.	Provides shared computing resources (like servers, storage, and applications) over the internet.
Resource Management	Distributed amongst multiple nodes, often requires more complex management due to the diverse nature of the resources.	Centrally managed, often more streamlined and user-friendly.
Scalability	Can be highly scalable, but depends on the architecture and the network.	Highly scalable, often a key feature, allowing for easy adjustment of resources.
Cost	Can be cost-effective, particularly for tasks that can be distributed across cheaper or existing systems.	Generally, follows a pay-per-use model, which can be cost-effective for variable workloads.
Control and Customization	High level of control and customization possible, as the infrastructure is often owned and managed by the user.	Less control over the infrastructure, as it is owned and managed by the cloud provider. Customization is possible to an

		extent.
Data Security and Privacy	Depends on the individual security measures of each node in the network.	Managed by the cloud provider, often with robust security measures, but raises concerns about data privacy.
Accessibility	Access can be limited to the network, may require specialized setups.	Easily accessible from anywhere via the internet.
Example Use Cases	Scientific research, complex simulations, peer-to-peer networks.	Web hosting, data storage, on-demand computing services.

3.3. Big Data Technologies

Big Data Technologies like Hadoop and Spark are crucial in managing and processing large datasets.

- Hadoop: An open-source framework that enables distributed storage and processing of big data sets using simple programming models. It consists of the Hadoop Distributed File System (HDFS) for storage and MapReduce for processing. Hadoop is designed to scale up from single servers to thousands of machines, each offering local computation and storage[34].

- Spark: Another open-source, distributed computing system that provides a fast and general-purpose cluster-computing framework. Unlike Hadoop's two-stage disk-based MapReduce paradigm, Spark's in-memory processing can optimize certain operations, making it faster for certain applications. Spark supports SQL queries, streaming data, machine learning, and graph processing[35].

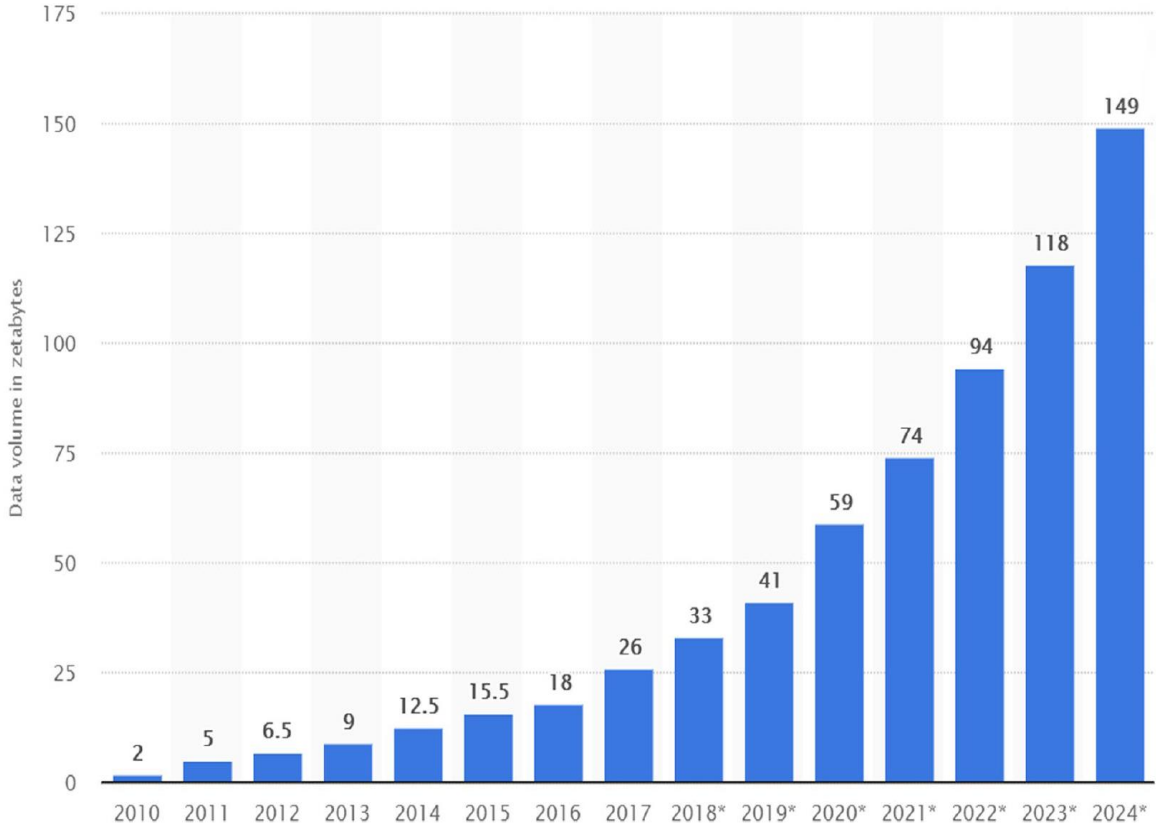

Figure 4. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024 (estimated) [36]

3.4. Data Management Strategies

Data Management Strategies in distributed systems encompass several key areas:

Data Storage: Involves selecting appropriate storage solutions that cater to the scale, accessibility, and type of data. This includes databases (SQL or NoSQL), data lakes, and distributed file systems like Hadoop Distributed File System (HDFS). The choice depends on data volume, velocity, and variety[37].

Data Processing: Strategies here focus on efficiently processing large datasets across distributed resources. Techniques include parallel processing, stream processing, and batch processing, using tools like Apache Spark and Kafka[38].

Data Security: Essential in distributed systems, this involves implementing robust encryption, access control, and network security measures. Regular audits, compliance with data protection regulations, and employing advanced security protocols like secure sockets layer (SSL) and transport layer security (TLS) are crucial for safeguarding data integrity and privacy[39].

D. Discussion and Comparison

Cloud-based distributed systems for big data analytics are an advanced technological integration that facilitates the handling, examination, and control of enormous volumes of data across several networked computing resources. The volume, variety, and velocity define the complexity of big data, which these systems are designed to manage. The vast amount of data produced by numerous sources, including social media, IoT devices, and workplace apps, is referred to as volume. Variety encompasses the various types of data, including as text, photos, and video, as well as structured, semi-structured, and unstructured formats. Velocity indicates how quickly this data is being generated and how quickly it needs to be analysed[40].

Data is stored across several cloud servers in the domain of distributed systems, offering resilience and guaranteeing data availability even in the case of hardware failures. Additionally, this distribution enables parallel processing, which drastically cuts down on the amount of time needed for data analytics operations. Technologies such as Hadoop and Spark are frequently used; Hadoop offers a framework for distributed processing (by MapReduce) and storage (via the Hadoop Distributed File System, or HDFS); Spark, on the other hand, provides an in-memory data processing capability that improves speed[41].

These systems also take use of the elasticity and scalability of cloud computing. They may dynamically assign resources according to the workload, guaranteeing economical and effective use. Through the use of artificial intelligence and machine learning techniques, which can be easily expanded and enhanced over time, the interface with cloud services also makes advanced analytics capabilities possible[42].

Distributed systems in the cloud also include a range of tools for data processing, visualization, and ingestion. These ecosystems would not function without technologies like Apache Kafka for real-time data intake, Apache Flink for stream processing, and business intelligence tools for data visualization[24].

Given the sensitivity of the data in these settings, security and privacy are critical. To protect data, distributed systems use strong security mechanisms such

network security protocols, access control, and encryption. Table2 as shown below illustrate some works or case studies concerning distributed systems used for big data analytics in the cloud.

Table 2. Distributed Systems work for Big Data Analytics in Cloud Environments

Reference	System Used	Cloud Platform	Data Size	Application Domain	Key Findings
[43]	Hadoop Distributed File System (HDFS) and MapReduce	unified cloud platform	20 GB to 1 TB	bioinformatics and weather data analysis	It is noteworthy that although Hadoop has limitations that make Spark necessary, Spark also has drawbacks that make flink necessary. Data is processed and stored in memory for later stages by Spark, an enhancement to MapReduce in Hadoop, while MapReduce processes data on disk.
[44]	Apache Hadoop and Apache Spark systems	unified cloud platform	600 GB	WordCount and TeraSort	Proper parameter selection and the amount of incoming data are critical factors that affect the performance of both Hadoop and Spark systems. According to the study, Spark outperforms Hadoop in terms of execution time, throughput, and speedup in general.
[45]	Hadoop Distributed File System (HDFS)	Hyperledger Fabric platform	1 MB to 100 GB	enhancing the security of HDFS for big data analysis	According to the performance evaluation, BlockHDFS adds minimal cost to the basic HDFS scenario regarding execution time and memory consumption.
[46]	FRTSPS	Apache Flink	large-scale data	handling data streams	It implies that integrating prominent data analytics aspects such as heterogeneity, scalability, fault tolerance, and query optimization in practical applications is complex, particularly in the data processing, storage, and analysis stages.
[47]	Apache Hadoop with MapReduce	Various	256, 512, and 1024 for small, medium, and large cities respectively	solving large-scale combinatorial problems, especially (ATSP)	High-speed processing and analysis of financial market data for predictive insights.
[48]	framework for sentiment analysis and classification of Twitter opinions	Google Cloud	many messages as 'tweets' a day on a variety of considerable issues	social media analytics, specifically focusing on Twitter	A critical component of the study's conclusions in this context is the mention of the sentiment analysis used to classify tweets as good, harmful, or neutral.
[49]	K-means	Google	Dataset	Bitcoin	The study provided insights into

	clustering algorithm	BigQuery	related Bitcoin	blockchain	the transactional behaviors within the Bitcoin network by effectively tying pseudonymous crypto addresses to miners with comparable hash rates.
[50]	Formal Concept Analysis	AWS S3	WS-DREAM dataset contains data on 4500 web services	QoS-aware big service composition	The study highlighted the utilization of various datasets and frameworks in the process and compared its approach with current extensive service composition methodologies.
[51]	machine learning (ML) models	Microsoft Azure ML Studio	3.63 GB	predicting Amazon product ratings using Big Data	The study showed how big data platforms may be used effectively for predictive analysis. Different machine learning models were implemented and contrasted, demonstrating the viability and scalability of big data solutions for e-commerce applications with substantial datasets.
[52]	Apache Spark MLlib	Microsoft Azure HDInsight	Spark-Pref dataset	focusing on performance analysis	The study produced formulas for the best Apache Spark parameter settings, a methodology for software and information assistance, and recommended metrics for evaluating Apache Spark calculations.
[53]	Hadoop Ecosystem System	Various by using using the Hadoop Ecosystem	Big Data	extracting, processing, and analyzing data from various cloud services	A study illustrates the effectiveness of the Hadoop ELT approach for cloud data integration
[54]	Azure Synapse Analytics	Microsoft Azure	Not Mentioned	focus on various components of Azure Synapse Analytics like Synapse SQL, Spark, Pipeline, Link, Workspace, and Studio	outlining the main elements and intricate architecture. This passage does not include any specific conclusions or outcomes from the book.
[48]	collection, sentiment analysis, and classification of Twitter opinions	BigQuery and Google App Engine	handles a vast amount of Twitter data, including many daily tweets	analysis of social media content, specifically Twitter, for various purposes including community, financial, manufacturing, and administrative policies	It highlights how important it is for people to use Twitter hashtags to share their opinions about current trends.
[55]	Big Data Storage	Cloudera Distribution	handling large	emphasis on data storage and	This paper demonstrates the transformation from a generic Big

		Storage layer	volumes of data	retrieval, and the extraction of intelligence from massive data sets	Data Storage layer meta-model to a Cloudera Distribution Storage layer meta-model using ATL transformation language, conforming with the MDA architecture's Platform-Specific Models (PSM).
[56]	Apache Spark	Apache Spark	2.5 quintillion bytes daily	focusing on the characteristics of big data, tools, application areas, and the Apache Spark Ecosystem	The paper aims to assist programmers in selecting the most appropriate programming language for Apache Spark-based projects

E. Extracted Statistics

The chart below visually represents the comparison between Distributed Computing and Cloud Computing across various characteristics such as Resource Management, Scalability, Cost, Control and Customization, Data Security and Privacy, and Accessibility.

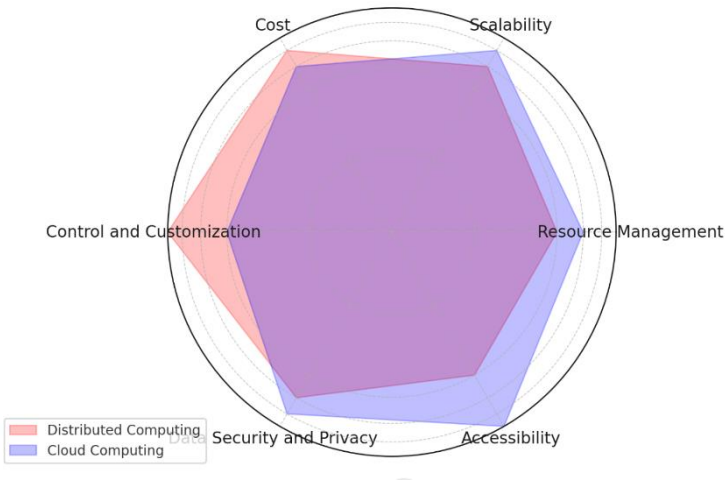

Figure 5. Visual Representation Of Comparing Between Distributed Computing And Cloud Computing

Every axis symbolizes a feature, where values nearer the centre signify a lesser presence or preference and values farther from the centre a more significant presence or preference. The graph highlights the fundamental differences between the two computing paradigms.

F. Recommendations

The integration of sophisticated distributed computing frameworks with robust data analytics capabilities is recommended in light of the review that has been presented. It is imperative to prioritize enhancing data management strategies and optimizing cloud infrastructure to facilitate extensive big data

analytics. Addressing the growing volume, velocity, and variety of big data entails utilizing scalable storage systems, practical data processing algorithms, and clever resource allocation tactics. Ensuring security, privacy, and adherence to rules is crucial in these distributed systems. Future studies might investigate novel approaches to improve these systems' efficiency, dependability, and affordability when managing intricate and ever-changing large data workloads.

G. Conclusion

Throughout the whole of this examination, a comprehensive investigation of the current state of distributed systems in cloud environments is carried out, with a specific focus on the uses of these systems in data-intensive computing. The significance of these systems in the administration of large-scale data processing operations is emphasised, and the remarkable advancements that have been achieved in big data analytics and data management are brought to light. Following the completion of the investigation, the researchers arrived at the conclusion that distributed systems in cloud environments are necessary in order to efficiently manage the growing quantity and complexity of big data. They offer solutions that are scalable, adaptable, and cost-effective; nevertheless, there are still challenges in the areas of data privacy, system interoperability, and security that need to be addressed via future development. These challenges need to be solved.

In addition, the investigation reveals that distributed systems that are running in cloud environments are necessary for the growth of big data analytics and data management. Currently, these systems are going through a process of development, which is driving breakthroughs in both technology and methodology. This is being done in order to meet the ever-increasing needs of applications that rely heavily on data. Enhancing the efficiency, security, and morality of data management in these systems need to be the primary focus of research in the years to come. It is because of this that they will continue to be powerful tools for the discovery of new information and the making of choices in a wide range of industries.

H. References

[1] N. T. Muhammed, S. R. Zeebaree, and Z. N. J. Q. Z. J. Rashid, "Distributed Cloud Computing and Mobile Cloud Computing: A Review," vol. 7, no. 2, pp. 1183-1201, 2022.

[2] S. A. Mostafa et al., "Applying Trajectory Tracking and Positioning Techniques for Real-time Autonomous Flight Performance Assessment of UAV Systems," vol. 54, no. 3, 2019.

[3] H. Wu, Z. Zhang, C. Guan, K. Wolter, and M. J. I. I. o. T. J. Xu, "Collaborate edge and cloud computing with distributed deep learning for smart city internet of things," vol. 7, no. 9, pp. 8099-8110, 2020.

[4] H. Yuan, M. J. I. T. o. A. S. Zhou, and Engineering, "Profit-maximized collaborative computation offloading and resource allocation in distributed cloud and edge computing systems," vol. 18, no. 3, pp. 1277-1287, 2020.

[5] H. Shukur et al., "Cache coherence protocols in distributed systems," vol. 1, no. 3, pp. 92-97, 2020.

[6] J. Wang, C. Xu, J. Zhang, and R. J. J. o. M. S. Zhong, "Big data analytics for intelligent manufacturing systems: A review," vol. 62, pp. 738-752, 2022.
[7] Z. N. Rashid, S. Zeebaree, and A. J. T. Sengur, "Novel remote parallel processing code-breaker system via cloud computing," 2020.
[8] Y. J. A. B. D. A. Shi, "Advances in big data analytics," 2022.
[9] S. R. Zeebaree, H. M. Shukur, L. M. Haji, R. R. Zebari, K. Jacksi, and S. M. J. T. R. o. K. U. Abas, "Characteristics and analysis of hadoop distributed systems," vol. 62, no. 4, pp. 1555-1564, 2020.
[10] H. I. Dino et al., "Impact of load sharing on performance of distributed systems computations," vol. 3, no. 1, pp. 30-37, 2020.
[11] A. Khole, A. Thakar, A. Kulkarni, H. Jadhav, S. Shende, and V. J. a. p. a. Karajkhede, "A Compendium on Distributed Systems," 2023.
[12] Z. N. Rashid, K. H. Sharif, and S. J. I. J. S. T. R. Zeebaree, "Client/Servers clustering effects on CPU execution-time, CPU usage and CPU Idle depending on activities of Parallel-Processing-Technique operations," vol. 7, no. 8, pp. 106-111, 2018.
[13] S. Zebari and N. O. J. J. U. A. P. S. Yaseen, "Effects of parallel processing implementation on balanced load-division depending on distributed memory systems," vol. 5, no. 3, pp. 50-56, 2011.
[14] M. Sedighidoost and M. J. E. I. Akbari, "Reduce task execution time in heterogeneous distributed systems using improved coa algorithm," pp. 1-19, 2022.
[15] H. M. Zangana, & Zeebaree, S. R. M., "Distributed Systems for Artificial Intelligence in Cloud Computing: A Review of AI-Powered Applications and Services," International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), vol. 5, no. 1, pp. 1-20, 2024.
[16] D. Lindsay, S. S. Gill, D. Smirnova, and P. J. C. Garraghan, "The evolution of distributed computing systems: from fundamental to new frontiers," vol. 103, no. 8, pp. 1859-1878, 2021.
[17] P. Y. Abdullah, S. Zeebaree, K. Jacksi, and R. R. J. I. J. o. R.-G. Zeabri, "An hrm system for small and medium enterprises (sme) s based on cloud computing technology," vol. 8, no. 8, pp. 56-64, 2020.
[18] H. Tabrizchi and M. J. T. j. o. s. Kuchaki Rafsanjani, "A survey on security challenges in cloud computing: issues, threats, and solutions," vol. 76, no. 12, pp. 9493-9532, 2020.
[19] Z. M. Khalid, S. R. J. I. J. o. S. Zeebaree, and Business, "Big data analysis for data visualization: A review," vol. 5, no. 2, pp. 64-75, 2021.
[20] S. Zeebaree and K. J. I. J. C. E. R. T. Jacksi, "Effects of processes forcing on CPU and total execution-time using multiprocessor shared memory system," vol. 2, no. 4, pp. 275-279, 2015.
[21] D. A. Hasan et al., "The impact of test case generation methods on the software performance: A review," vol. 5, no. 6, pp. 33-44, 2021.
[22] A. K. J. B. D. M. Sandhu and Analytics, "Big data with cloud computing: Discussions and challenges," vol. 5, no. 1, pp. 32-40, 2021.
[23] Z. Wu, J. Sun, Y. Zhang, Z. Wei, and J. J. P. o. t. I. Chanussot, "Recent developments in parallel and distributed computing for remotely sensed big data processing," vol. 109, no. 8, pp. 1282-1305, 2021.

[24] A. Rafique, D. Van Landuyt, E. H. Beni, B. Lagaisse, and W. J. I. S. Joosen, "CryptDICE: Distributed data protection system for secure cloud data storage and computation," vol. 96, p. 101671, 2021.
[25] A. H. A. Al-Jumaili, R. C. Muniyandi, M. K. Hasan, J. K. S. Paw, and M. J. J. S. Singh, "Big Data Analytics Using Cloud Computing Based Frameworks for Power Management Systems: Status, Constraints, and Future Recommendations," vol. 23, no. 6, p. 2952, 2023.
[26] C. Yang, S. Lan, L. Wang, W. Shen, and G. G. J. I. a. Huang, "Big data driven edge-cloud collaboration architecture for cloud manufacturing: a software defined perspective," vol. 8, pp. 45938-45950, 2020.
[27] Y. Niu, L. Ying, J. Yang, M. Bao, C. J. I. P. Sivaparthipan, and Management, "Organizational business intelligence and decision making using big data analytics," vol. 58, no. 6, p. 102725, 2021.
[28] H. Alshammari, S. A. El-Ghany, and A. J. J. o. I. P. S. Shehab, "Big IoT healthcare data analytics framework based on fog and cloud computing," vol. 16, no. 6, pp. 1238-1249, 2020.
[29] C. M. Mohammed, S. R. J. I. J. o. S. Zeebaree, and Business, "Sufficient comparison among cloud computing services: IaaS, PaaS, and SaaS: A review," vol. 5, no. 2, pp. 17-30, 2021.
[30] D. M. ABDULQADER, S. R. ZEEBAREE, R. R. ZEBARI, S. A. SALEH, Z. N. RASHID, and M. A. J. J. o. D. U. SADEEQ, "SINGLE-THREADING BASED DISTRIBUTED-MULTIPROCESSOR-MACHINES AFFECTING BY DISTRIBUTED-PARALLEL-COMPUTING TECHNOLOGY," vol. 26, no. 2, pp. 416-426, 2023.
[31] A. Rostami, "Cloud Service Models-IaaS, PaaS, and SaaS," 2021.
[32] K. Benhssayen and A. Ettalbi, "An Extended Framework for Semantic Interoperability in PaaS and IaaS Multi-cloud," Cham, 2022, pp. 415-424: Springer International Publishing.
[33] T. H. Noor, S. Zeadally, A. Alfazi, Q. Z. J. J. o. N. Sheng, and C. Applications, "Mobile cloud computing: Challenges and future research directions," vol. 115, pp. 70-85, 2018.
[34] M. A. Amanullah et al., "Deep learning and big data technologies for IoT security," vol. 151, pp. 495-517, 2020.
[35] S. Tang, B. He, C. Yu, Y. Li, K. J. I. T. o. K. Li, and D. Engineering, "A survey on spark ecosystem: Big data processing infrastructure, machine learning, and applications," vol. 34, no. 1, pp. 71-91, 2020.
[36] B. Berisha and E. Mëziu, Big Data Analytics in Cloud Computing: An overview. 2021.
[37] M. I. J. I. J. o. I. M. Bellgard, "ERDMAS: An exemplar-driven institutional research data management and analysis strategy," vol. 50, pp. 337-340, 2020.
[38] E. Badidi, Z. Mahrez, and E. J. F. I. Sabir, "Fog computing for smart cities' big data management and analytics: A review," vol. 12, no. 11, p. 190, 2020.
[39] I. Yaqoob, K. Salah, R. Jayaraman, Y. J. N. C. Al-Hammadi, and Applications, "Blockchain for healthcare data management: opportunities, challenges, and future recommendations," pp. 1-16, 2021.
[40] S. Adhikary and S. J. P. C. S. Banerjee, "Introduction to distributed nearest hash: On further optimizing cloud based distributed knn variant," vol. 218, pp. 1571-1580, 2023.

[41] S. Ketu, P. K. Mishra, and S. J. C. y. S. Agarwal, "Performance analysis of distributed computing frameworks for big data analytics: hadoop vs spark," vol. 24, no. 2, pp. 669-686, 2020.
[42] Q. Fang and S. J. J. o. B. O. Yan, "MCX Cloud—a modern, scalable, high-performance and in-browser Monte Carlo simulation platform with cloud computing," vol. 27, no. 8, pp. 083008-083008, 2022.
[43] P. R. Giri, G. J. S. S. Sharma, and I. Devices, "Apache Hadoop Architecture, Applications, and Hadoop Distributed File System," vol. 4, no. 1, pp. 14-20, 2022.
[44] N. Ahmed, A. L. Barczak, T. Susnjak, and M. A. J. J. o. B. D. Rashid, "A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench," vol. 7, no. 1, pp. 1-18, 2020.
[45] V. Mothukuri, S. S. Cheerla, R. M. Parizi, Q. Zhang, K.-K. R. J. B. R. Choo, and Applications, "BlockHDFS: Blockchain-integrated Hadoop distributed file system for secure provenance traceability," vol. 2, no. 4, p. 100032, 2021.
[46] B. G. Deepthi, K. S. Rani, P. V. Krishna, V. J. M. T. Saritha, and Applications, "An efficient architecture for processing real-time traffic data streams using apache flink," pp. 1-17, 2023.
[47] G. Saxena et al., "Auto-WLM: Machine learning enhanced workload management in Amazon Redshift," in Companion of the 2023 International Conference on Management of Data, 2023, pp. 225-237.
[48] S. Tamrakar, B. Madhavi, V. J. H. o. I. C. Mohan, and O. f. S. Development, "Democratizing Sentiment Analysis of Twitter Data Using Google Cloud Platform and BigQuery," pp. 287-304, 2022.
[49] M. Jeyasheela Rakkini and K. Geetha, "Detection of Bitcoin Miners by Clustering Crypto Address with Google BigQuery Open Dataset," in Soft Computing: Theories and Applications: Proceedings of SoCTA 2021: Springer, 2022, pp. 25-32.
[50] M. Sellami, H. Mezni, M. S. J. J. o. N. Hacid, and C. Applications, "On the use of big data frameworks for big service composition," vol. 166, p. 102732, 2020.
[51] J. Woo, M. J. W. I. R. D. M. Mishra, and K. Discovery, "Predicting the ratings of Amazon products using Big Data," vol. 11, no. 3, p. e1400, 2021.
[52] S. Minukhin, N. Brynza, and D. Sitnikov, "Analyzing performance of apache spark mllib with multinode clusters on azure hdinsight: Spark-perf case study," in International Scientific Conference “Intellectual Systems of Decision Making and Problem of Computational Intelligence”, 2020, pp. 114-134: Springer.
[53] H. K. Lin and T.-J. Liao, "Cloud Hadoop for Enterprise Collaboration System," in Handbook of Smart Materials, Technologies, and Devices: Applications of Industry 4.0: Springer, 2021, pp. 1-10.
[54] L. R. de Carvalho, M. A. da Cruz Motta, and A. P. F. de Araújo, "Performance Analysis of Main Public Cloud Big Data Services Processing Brazilian Government Data," in Latin American High Performance Computing Conference, 2020, pp. 49-61: Springer.
[55] A. Erraissi, M. Banane, A. Belangour, and M. Azzouazi, "Big data storage using model driven engineering: From big data meta-model to cloudera PSM meta-

model," in 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-5: IEEE.
[56] Y. K. Gupta and S. Kumari, "A study of big data analytics using apache spark with Python and Scala," in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 471-478: IEEE.