International Journal of Applied Engineering & Technology

Challenges for Securing Big Data in Cloud Computing

Veerpal Kaur
Assistant Professor, IET Bhaddal Technical Campus, Ropar

ABSTRACT
As a result of the latest advances in the field of computers, there is an ever-growing amount of data available. Overabundance of data, however, presents significant difficulties for consumers. Thanks to cloud computing services, a great deal of data can be stored in a secure setting. They do away with different criteria, like setting off a certain area and keeping up with pricey computer software and hardware items. Large clusters of computers are needed to handle big data to ensure proper data processing and preservation. This paper discusses several cloud services, including MapR, Google Cloud, International Business Machine Cloud, Amazon Web Services, Hortonworks, and Microsoft Azure, as well as big data's means, categorization, and features. Additionally, a comparison of several cloud-based big data platforms is carried out. In visualization, distributed database storage, heterogeneity, and Data security describe several scientific challenges.

Keywords: Hadoop; big data; cloud computing; data analysis;

1. INTRODUCTION:
The latest advances in technology have led to a daily increase in the volume of data accessible. Social media platforms and networks of sensors, for instance, produce Enormous information flows. In another way, huge amounts of data are created quickly and through various sources in many formats [1]. Big data is currently a significant field of study. Big data are created rapidly, making it challenging to handle, store, or analyze them with conventional software. Technology for big data is instruments that can store important data in various formats. Many analytical tools have been available to help users analyze complex unstructured and structured information to satisfy user needs and research and store complicated data [2]. Many different hardware, software, and program designs had to be put out and created to retrieve the data from big data. These technologies' primary goal is to store accurate and dependable enormous amounts of data results [3]. Furthermore, big data necessitates cutting-edge technology to store and handle vast volumes of data effectively in a limited time. Participatory analysis tools, continuous processing tools, and batch processing tools are three categories of large-data platforms [4]. Data processing in interactive settings and real-time data interaction are accomplished using dynamic analytic tools. Real-time information storage frameworks include Google's Drill and Apache Drill. Information is continuously flown in and out of storage using tools for stream processing [5]. S4 and Electricity are the primary streaming data storage platforms. Material is stored in batches using Hadoop architecture. Numerous fields, including the processing of signals, statistics, visualization, social network analysis, neural networks, and data mining, use big data techniques [6]. An interactive gradient algorithm that accepts regulated messages from nearby nodes was created by Mohajer et al. [7]. The suggested approach makes use of a large data optimizing oneself framework [8].

2. Definitions of Big Data and its characteristics
Big Data refers to a vast and complex volume of data that exceeds the processing capabilities of traditional database systems [9]. It is characterized by the 3Vs: Volume, Velocity, and Variety, and often includes massive datasets from diverse sources [10]. The term encompasses the challenges and opportunities associated with collecting, storing, and analyzing such extensive and diverse datasets [11]. These characteristics highlight the need for advanced technologies, such as distributed computing, parallel processing, and machine learning, to effectively manage and extract value from Big Data in show in figure .1.

The term encompasses the challenges and opportunities associated with collecting, storing, and analyzing such extensive and diverse datasets.

Volume: Big Data involves a large amount of data, ranging from terabytes to petabytes. This sheer volume exceeds conventional databases' capacity and requires specialized storage and processing technologies.

Velocity: Refers to the speed at which data is generated, collected, and processed. Big Data applications often deal with real-time or near-real-time data streams, requiring rapid processing and analysis to derive valuable insights.

Variety: Big Data is not limited to structured data found in traditional databases; it also includes unstructured and semi-structured data, such as text, images, videos, and social media posts. The diversity of data types poses challenges in terms of storage, processing, and analysis.

Veracity: This aspect of Big Data emphasizes the quality and reliability of the data. With diverse sources and formats, ensuring the accuracy and trustworthiness of the data becomes crucial for making informed decisions.

Value: The ultimate goal of working with Big Data is to extract meaningful insights and value from the massive datasets. It may involve uncovering patterns, trends, correlations, or hidden knowledge that can inform business strategies, scientific research, or decision-making processes. Variability: Refers to the inconsistency in the data flow, both in terms of volume and velocity. Big Data environments must be capable of handling fluctuations in data patterns and adapting to changing circumstances.

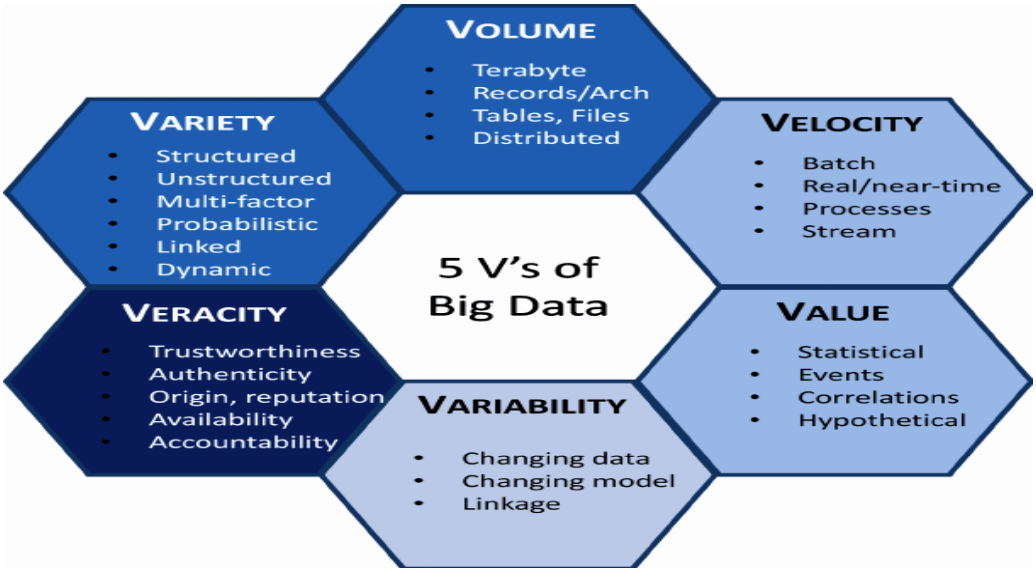

Figure 1 Five Big Data.

Complexity: Big Data involves intricate relationships and dependencies within the data. Analyzing and interpreting complex data structures require advanced tools, algorithms, and technologies to make sense of the information.

Accessibility: Big Data solutions aim to make data accessible to users across different levels of expertise. It involves user-friendly interfaces, visualization tools, and technologies that enable non-technical users to interact with and derive insights from the data [12].

These characteristics highlight the need for advanced technologies, such as distributed computing, parallel processing, and machine learning, to effectively manage and extract value from big data [13].

Table 1 Definitions of big data.

Author's name	Reference	Definition
D. A. Shafiq et al.	[15]	Load balancing in a cloud computing environment involves the effective distribution of computing workloads across multiple servers or resources to optimize performance, enhance resource utilization, and prevent any single resource from becoming overloaded. Various load balancing techniques are employed in cloud computing environments to achieve these objectives. These techniques aim to evenly distribute incoming requests among available resources, ensuring a balanced and efficient utilization of the cloud infrastructure. Examples of load balancing techniques include Round Robin, Least Connections, Weighted Round Robin,
S. Amamou et al.	[16]	Data protection in cloud computing refers to the comprehensive set of measures, policies, and technologies implemented to safeguard sensitive information and ensure the privacy, integrity, and availability of data stored, processed, and transmitted within cloud environments.
P. J. Sun et al.	[17]	Security and privacy protection in cloud computing are critical aspects that require ongoing discussions and solutions to address various challenges. In the context of cloud computing, where data and applications are hosted on

		remote servers and accessed over the internet, ensuring the confidentiality, integrity, and availability of data, as well as protecting user privacy, is of utmost importance.
R. Nachiappan et al.	[18]	The reliability of cloud storage for Big Data applications is a subject that has garnered significant attention, prompting a state-of-the-art survey to comprehensively explore the current landscape, challenges, and advancements in this domain.
A. O'Driscoll et al.	[19]	The integration of big data technologies, Hadoop, and cloud computing in genomics enables researchers to handle the unprecedented scale and complexity of genomic datasets. Large-scale genomic analyses, such as DNA sequencing, variant calling, and functional genomics studies, can be parallelized and distributed across cloud-based Hadoop clusters.

Table 2 Users in India as of February 2020.

Application name	**Count**
Twitter	1.75 Crore
Instagram	21 Crore
WhatsApp	53 Crore
Facebook	41 Crore
YouTube	44.8 Crore

3. Big Data with Machine Learning

Big Data and Machine Learning are two closely intertwined domains that, when combined, offer powerful capabilities for extracting meaningful insights, making predictions, and solving complex problems [13]. Big Data refers to the vast and diverse sets of data that exceed the capacity of traditional data processing systems. It encompasses large volumes of structured, unstructured, and semi-structured data, often generated at high velocity from various sources such as social media, sensors, and transaction records. The three key characteristics of Big

Data, commonly known as the 3Vs (Volume, Velocity, and Variety), highlight the scale, speed, and diversity of the data involved. Machine Learning, on the other hand, is a subset of artificial intelligence (AI) that focuses on developing algorithms and models capable of learning patterns from data and making predictions or decisions without explicit programming. It encompasses various techniques, including supervised learning, unsupervised learning, and reinforcement learning, allowing systems to improve their performance over time through experience. When Big Data and Machine Learning are combined, they form a synergistic relationship that leverages the capabilities of each to address complex challenges [14]. The abundance of data in Big Data environments provides the raw material for machine learning algorithms to learn patterns, relationships, and trends that may not be apparent through traditional analytical methods. Machine learning in the context of Big Data involves training models on large datasets to identify patterns and correlations, enabling systems to make accurate predictions or classifications on new, unseen data. This process often requires distributed computing frameworks and parallel processing to handle the immense scale of Big Data. Applications of Big Data with Machine Learning are widespread and diverse. In finance, it can be used for fraud detection and risk assessment. In healthcare, it aids in disease diagnosis and personalized medicine. In marketing, it enables targeted advertising and customer segmentation [15]. The integration of Big Data and Machine Learning is particularly valuable in optimizing business processes, enhancing decision-making, and unlocking valuable insights from massive datasets.

Challenges in combining Big Data with Machine Learning include data quality, scalability, and interpretability of complex models. However, the benefits of this integration are substantial, offering organizations the potential to gain a competitive edge, improve operational efficiency, and innovate across various industries. The continual advancements in both Big Data and Machine Learning technologies further contribute to the evolution of data-driven decision-making and predictive analytics in show in figure.2.

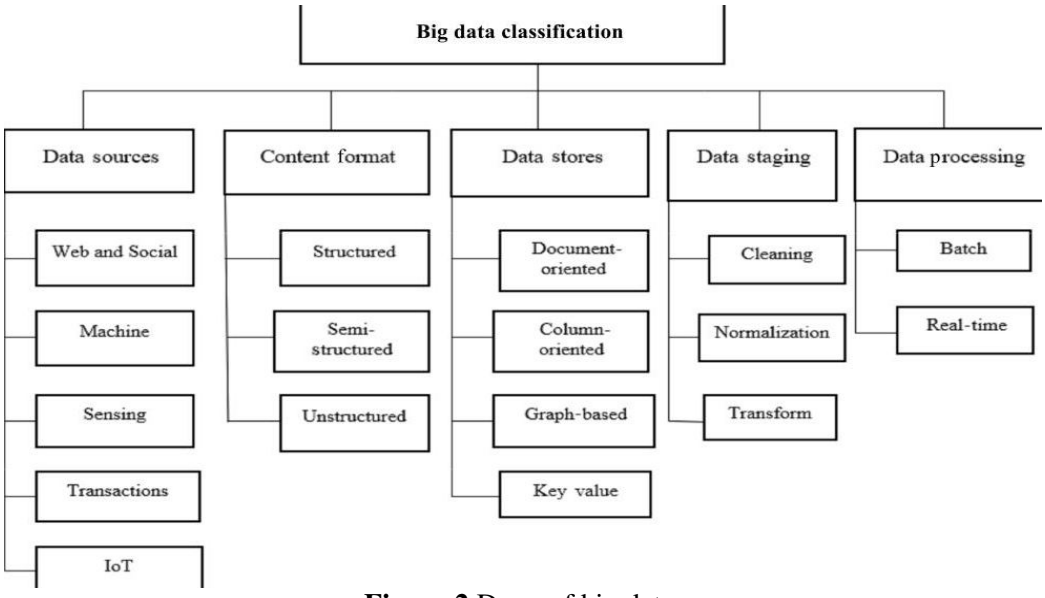

Figure.2 Draw of big data.

4. Cloud Computing

Cloud computing is a technology paradigm that involves delivering computing services over the internet [16-17]. It provides on-demand access to a shared pool of computing resources, including computing power, storage, and applications [18]. Cloud computing is classified into different types based on the services it offers and the deployment models used [19]. Here are the primary types of cloud computing:

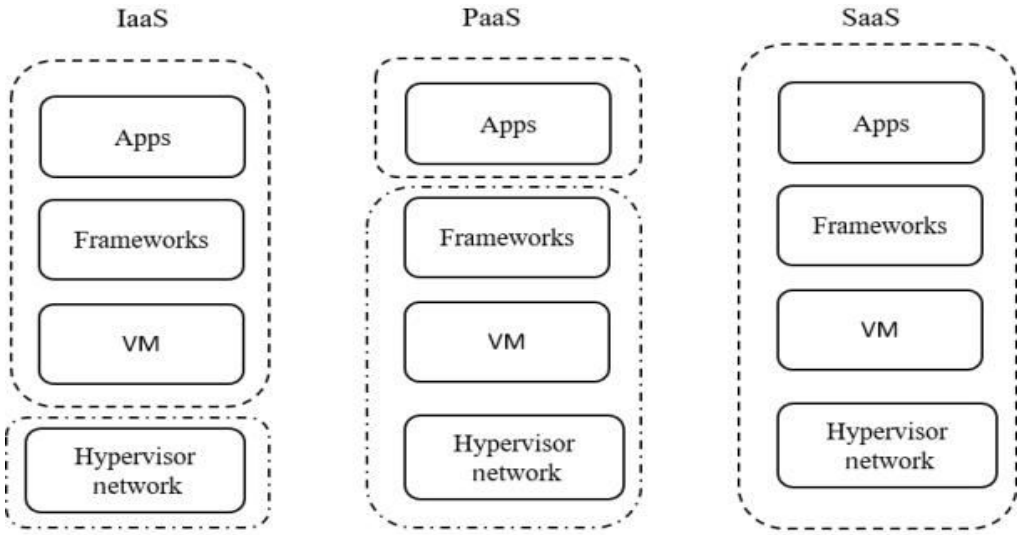

Figure.3 Cloud computing services

1. Infrastructure as a Service (IaaS):

- **Definition:** IaaS provides virtualized computing resources over the internet. Users can rent virtual machines, storage, and networking infrastructure on a pay-as-you-go basis.
- **Example:** Amazon Web Services (AWS) Elastic Compute Cloud (EC2), Microsoft Azure Virtual Machines.

2. Platform as a Service (PaaS):

- **Definition:** PaaS offers a platform allowing developers to build, deploy, and manage applications without dealing with the underlying infrastructure. It includes development frameworks, databases, and middleware.
- **Example:** Google App Engine, Heroku, Microsoft Azure App Service.

3. Software as a Service (SaaS):

- **Definition:** SaaS delivers software applications over the internet on a subscription basis. Users access applications through web browsers without worrying about software maintenance or infrastructure management.
- **Example:** Salesforce, Microsoft Office 365, Google Workspace.

4. Function as a Service (FaaS) or Serverless Computing:

- **Definition:** FaaS allows developers to execute individual functions or pieces of code in response to events without managing the entire infrastructure. It automatically scales based on demand.
- **Example:** AWS Lambda, Azure Functions, Google Cloud Functions.

Deployment Models:

1. Public Cloud:

- **Definition:** Public cloud services are provided by third-party cloud service providers and are accessible to the general public. Resources are shared among multiple users, offering scalability and cost efficiency.
- **Example:** AWS, Microsoft Azure, Google Cloud Platform.

2. Private Cloud:

- **Definition:** Private clouds are dedicated to a single organization. They can be managed on-premises or by a third-party provider and offer more control over resources and security.
- **Example:** VMware Cloud Foundation, OpenStack.

3. Hybrid Cloud:

- **Definition:** Hybrid clouds combine public and private cloud infrastructure, allowing data and applications to be shared between them. It provides flexibility, scalability, and the ability to optimize workloads.
- **Example:** Using a combination of AWS and on-premises data centers.

4. Community Cloud:

- **Definition:** Community clouds are shared by organizations with common interests, such as industry-specific regulatory requirements. They offer collaborative solutions while maintaining a level of isolation [19-20].
- **Example:** Cloud services shared by multiple healthcare organizations to comply with industry regulations.

Understanding these types of cloud computing helps organizations choose the most suitable services and deployment models based on their specific needs, requirements, and business objectives in show in figure.4.

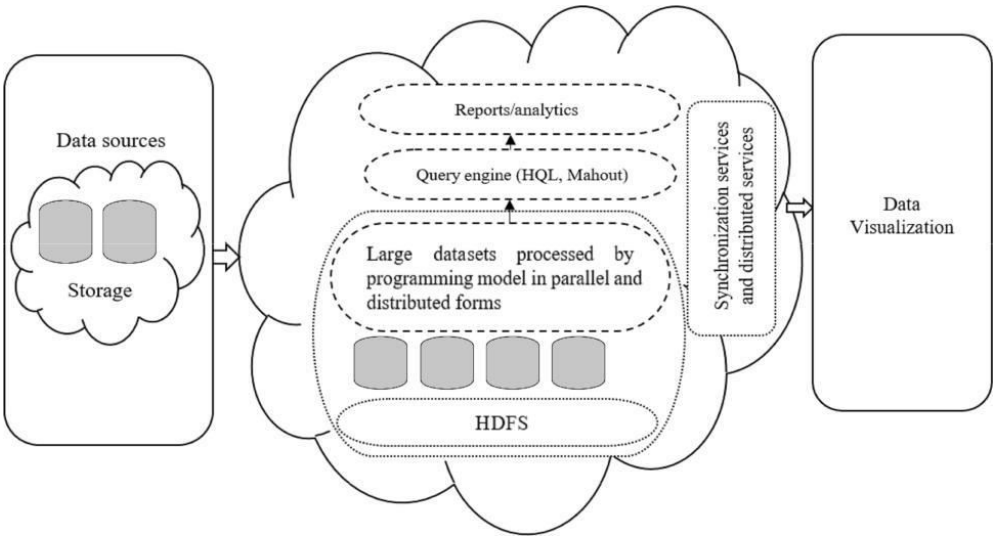

Figure.4 cloud computing and Big data.

5. Research Issues in Big Data

Research in the field of Big Data is a dynamic and evolving area, and various challenges or issues continue to be the focus of ongoing investigation. Some prominent research issues in Big Data include:

Scalability:

- **Issue:** As datasets continue to grow in size, handling and processing massive volumes of data become increasingly challenging. Scalability issues arise in both storage and processing, requiring efficient distributed computing frameworks.

Data Quality and Integration:

- **Issue:** Big Data often involves diverse and heterogeneous datasets from multiple sources. Ensuring data quality, resolving inconsistencies, and integrating disparate data types present ongoing challenges in research.

Privacy and Security:

- **Issue:** Preserving the privacy and security of sensitive information within large datasets is a critical concern. Research explores techniques for secure data sharing, encryption, and compliance with privacy regulations.

Data Storage and Retrieval:

- **Issue:** Efficiently storing and retrieving large volumes of data is a constant challenge. Research focuses on developing scalable storage solutions and optimized data retrieval mechanisms.

Data Analytics and Processing:

- **Issue:** Analyzing vast datasets in real-time or near-real-time requires advanced algorithms and distributed computing frameworks. Research aims to enhance the efficiency and speed of data analytics processes.

Machine Learning and Predictive Analytics:

- **Issue:** Integrating machine learning algorithms into Big Data analytics workflows poses challenges in terms of model complexity, interpretability, and ensuring the reliability of predictions.

Data Governance and Compliance:

- **Issue:** Establishing effective data governance frameworks and ensuring compliance with regulations (e.g., GDPR, HIPAA) are critical research areas. This involves addressing legal, ethical, and policy considerations.

Energy Efficiency:

- **Issue:** The energy consumption of large-scale data centers used for Big Data processing is a concern. Research explores techniques for optimizing energy usage and improving the overall sustainability of Big Data systems.

Semantic Understanding of Data:

- **Issue:** Extracting meaningful insights from unstructured or semi-structured data requires enhancing the semantic understanding of information. Research focuses on natural language processing and ontologies.

Distributed Computing Architectures:

- **Issue:** Designing and optimizing distributed computing architectures for Big Data processing involves addressing issues related to fault tolerance, load balancing, and resource allocation.

Data Stream Processing:

- **Issue:** Handling real-time data streams and developing algorithms for continuous analysis pose challenges. Research explores stream processing techniques for timely decision-making.

Data Privacy-preserving Techniques:

- **Issue:** Developing techniques that allow analysis on sensitive data without compromising individual privacy is a significant area of research. This involves methods such as homomorphic encryption and differential privacy.

Ethical Considerations:

- **Issue:** As Big Data analytics becomes more pervasive, ethical considerations surrounding the collection and use of data need careful examination. Research explores frameworks for ethical data practices and guidelines.

6 CONCLUSION

In conclusion, securing Big Data in cloud computing presents a multifaceted set of challenges that demand careful consideration and strategic solutions. The convergence of Big Data and cloud computing introduces complexities that go beyond traditional security paradigms. The preservation of data privacy, compliance with regulations, and the management of encryption keys remain focal points for security professionals. Additionally, the dynamic

nature of cloud environments, coupled with the shared responsibility model, necessitates a nuanced approach to security. The challenge extends to securing the entire data lifecycle, from its ingestion into the cloud to the intricate processes of analytics and decision-making. The scalability of security measures, without compromising performance, is a delicate balance that requires ongoing innovation and adaptation. The persistent threat of insider risks underscores the importance of robust identity and access management. As cloud environments evolve, incident response and forensics become integral components of a proactive security strategy. Detecting and mitigating security incidents in a timely manner is crucial, and the distributed nature of Big Data in the cloud adds complexity to forensic investigations. In addressing these challenges, collaboration between cloud service providers, enterprises, and the broader cybersecurity community is paramount. A collective effort is required to stay ahead of emerging threats, promote best practices, and foster an environment of continuous improvement in securing Big Data in the dynamic landscape of cloud computing. The journey toward comprehensive and effective security in this realm involves not only technological advancements but also a holistic understanding of the legal, regulatory, and ethical dimensions of data protection in the digital era.

REFERENCES

[1] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, The rise of 'big data' on cloud computing: Review and open research issues, Inform. Syst.,vol. 47, pp. 98–115, 2015.

[2] J. H. Yu and Z. M. Zhou, Components and development in big data system: A survey, J. Electr. Sci. Technol., vol. 17,no. 1, pp. 51–72, 2019.

[3] S. Kumar and K. K. Mohbey, A review on big data based parallel and distributed approaches of pattern mining, J. King Saud Univ. – Comput. Inform. Sci., doi: 10.1016/j.jksuci.2019.09.006.

[4] Y. N. Liu, N. Li, X. Zhu, and Y. Qi, How wide is the application of genetic big data in biomedicine, Biomed. Pharmacother., vol. 133, p. 111074, 2021.

[5] V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, Unstructured data analysis on big data using map reduce, Procedia Comput. Sci., vol. 50, pp. 456–465, 2015.

[6] S. Maitrey and C. K. Jha, MapReduce: Simplified data analysis of big data, Procedia Comput. Sci., vol. 57, pp. 563–571, 2015.

[7] A. Mohajer, M. Barari, and H. Zarrabi, Big data based self-optimization networking: A novel approach beyond cognition, Intell. Automat. Soft Comput., doi: 10.1080/10798587.2017.1312893.

[8] M. Batty, Big data, smart cities and city planning, Dialogues in Human Geography, vol. 3, no. 3, pp. 274–279, 2013.

[9] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, Fuzzy c-means algorithms for very large, IEEE Trans. Fuzzy Syst., vol. 20, no. 6, pp. 1130–1146, 2012.

[10] D. Fisher, R. Deline, M. Czerwinski, and S. Drucker, Interactions with big data analytics, Interactions, vol. 19, no. 3, pp. 50–59, 2012.

[11] The State Council of the People's Republic of China, Action plan for promoting big data development, (in Chinese), http://www.gov.cn/zhengce/content/2015- 09/05/content 10137.htm, 2015.

[12] M. A. Beyer and D. Laney, The importance of 'big data': A definition, Stamford, CT, USA: Gartner, G00235055, 2012.

[13] L. Rabhi, N. Falih, A. Afraites, and B. Bouikhalene, Big data approach and its applications in various fields: Review, Procedia Comput. Sci., vol. 155, pp. 599–605, 2019.

[14] F. Ridzuan and W. M. N. Wan Zainon, A review on data cleansing methods for big data, Procedia Comput. Sci., vol. 161, pp. 731–738, 2019.

[15] D. A. Shafiq, N. Z. Jhanjhi, and A. Abdullah, Load balancing techniques in cloud computing environment: A review, J. King Saud Univ. – Comput. Inform. Sci., doi: 10.1016/j.jksuci.2021.02.007.

[16] S. Amamou, Z. Trifa, and M. Khmakhem, Data protection in cloud computing: A survey of the state-of-art, Procedia Comput. Sci., vol. 159, pp. 155–161, 2019.

[17] P. J. Sun, Security and privacy protection in cloud computing: Discussions and challenges, J. Netw. Comput. Appl., vol. 160, p. 102642, 2020.

[18] R. Nachiappan, B. Javadi, R. N. Calheiros, and K. M. Matawie, Cloud storage reliability for Big Data applications: A state of the art survey, J. Netw. Comput. Appl., vol. 97, pp. 35–47, 2017.

[19] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, 'Big data', Hadoop and cloud computing in genomics, J. Biomed. Inform., vol. 46, no. 5, pp. 774–781, 2013.

[20] S. Karimian-Aliabadi, D. Ardagna, R. Entezari-Maleki, E. Gianniti, and A. Movaghar, Analytical composite performance models for Big Data applications, J. Netw. Comput. Appl., vol. 142, pp. 63–75, 2019.