Systematic Security Design for Big Data Frameworks Over Cloud Computing

Dr. Mahesh Kotha, Dr. Bodla Kishor

Associate Professor, Department of CSE (AI&ML), CMR Technical Campus, Hyderabad.
Associate Professor, Department of CSE, CMR Engineering College, Hyderabad.

Abstract: Big Data (BD) activities now prefer to be computed via cloud deployment architectures. Their cost-effectiveness, scalability, and adaptability drove this trend. The data are no longer directly managed by the user in such a deployment strategy, which creates additional security issues. The broad use of cloud architectures in this context is greatly influenced by BD security. But without a preliminary analysis that guarantees a realistic secure assembly and tackles domain-specific risks, creating a comprehensive security plan is difficult. A unique security-by-design framework for BD framework deployment over cloud computing (BigCloud) is presented in this article. Specifically, it depends on a fully automated security assessment framework and a methodical security analysis approach. During the design process, our platform makes it possible to bridge BigCloud security domain knowledge to best practices. We used a use case with the Apache Hadoop stack to validate the suggested framework. The results of the study show how successful it is at raising security awareness and cutting down on security design time. The suggested framework's advantages and disadvantages are also assessed, and the key issues that still need to be resolved in the BigCloud space are highlighted.

Keywords: cloud computing security, data protection, reference architecture, security analysis pattern.

I.INTRODUCTION

The growing adoption of cloud computing has transformed the way organizations process, analyze, and store large volumes of data. Big data frameworks such as Hadoop, Spark, and Flink leverage cloud resources for scalable and cost-effective computation. However, this integration introduces critical security challenges involving data confidentiality, integrity, privacy, and access control. This paper proposes a systematic security design tailored for big data frameworks operating in cloud environments. The proposed architecture incorporates multi-layered security mechanisms including data encryption, secure communication protocols, granular access control, and anomaly detection. We evaluate the feasibility of our approach through simulation and theoretical analysis and suggest future directions for enhancing trust and reliability in cloud-based big data infrastructures.

The convergence of big data and cloud computing has revolutionized data-driven decision-making. Cloud platforms provide on-demand resources that align with the elastic needs of big data analytics. However, this convergence also opens new attack surfaces, as sensitive data traverses across distributed, virtualized, and potentially untrusted environments. Addressing the systematic security of big data frameworks on the cloud is imperative to maintain data integrity, user privacy, and system availability.

The contemporary application delivery approach has been influenced by developments in cloud computing technology [3]. Because cloud computing has the ability to offer simple, inexpensive access to a significant amount of processing power, its benefits are undeniable. This trend is supported by big data (BD) frameworks over cloud computing (BigCloud), which have the potential to be far more scalable and elastic than conventional methods [5]. Data security becomes even more crucial as the outsourced data may include private information like government, healthcare, financial, or proprietary research data. Cloud security and privacy are two of the top research areas for the upcoming decade [4]. Maintaining the effectiveness of sustainable BD operations over cloud platforms is the goal of this paper. The primary security issue is the client's confidence in data movement both inside and outside of the cloud environment, as well as the processing and storage of important data in an off-premise data center. While several core cloud features (such virtualization and multitenancy) guarantee better resource use, they also make it difficult to provide safe

computation. The characteristics of BD (volume, velocity, and diversity) systems amplify additional security risks related to this adoption, such as privacy, integrity, confidentiality, and the availability of stored data [1].

II.RELATED WORK

The purpose of this paper is to make it easier to implement safe BD systems in the IaaS cloud model. The realization of a BigCloud system raises significant security issues. The system's architecture as well as the underlying security technology, policies, and services are examples of these security elements. IaaS is still the model with the quickest rate of growth [12] and is what BD implementers want most. In addition to describing BD security elements and their interactions, this article examines security services utilized in IaaS cloud systems. In order to establish research gaps and best practices, as well as to offer their landscape approaches and lexicon, it talks about security systems focused on BigCloud design. The methods used in this study to provide a general BigCloud security reference model is shown in Fig. 2.In order to support BD science, this paper first thoroughly examines the foundational elements of the cloud security stack, which adds to the body of knowledge on BigCloud security implementation. Furthermore, it groups the various security tiers into a reference design according to the service models they provide.

Second, by outlining the security elements of a secure design pattern and its characteristics, it investigates the vulnerabilities related to BigCloud adoption. Third, it refines the cloud context pattern into a new security analysis pattern, offering a variety of insights into BigCloud security specifications. This pattern extends the CIA (confidentiality, integrity, and availability) trinity to map the existing technology to the solution domain. After that, they conduct a thorough criteria election and analyze and categorize the most advanced security frameworks now accessible, the majority of which are open-source. Lastly, it lists a few unresolved issues and suggestions for clients and service providers to have a thorough conversation about realizing the goal of offering a secure BigCloud service. We provide a summary of the suggested models and patterns in Table I, which includes the knowledge domain, model description, and part within this article that corresponds to it, to make using the systematic research process easier.

All things considered, none of the aforementioned studies provide careful thought to the BigCloud security requirements, which influence security installations throughout the design stage. Therefore, we contend that not enough research has been done on the BD security deployment needs. Furthermore, a thorough examination and methodical approach to implementing security-by-design across BigCloud deployment architectures are still lacking from the present literature. This paper attempts to compile the most creative findings and present a BD security-driven procedure to maximize the deployment of BigCloud components in the cloud environment in order to overcome those constraints and close this research gap. The most current and thorough explanation of the BD security requirements of the IaaS cloud deployment architecture is what sets this essay apart.

Year	Authors	Title	Methodology	Key Contributions	Limitations
2020	Gai et al.	Security and Privacy Issues in Cloud Computing	Comprehensive Review	Identified primary security challenges in cloud-based big data: data leakage, access control, DoS attacks	Lacks a specific framework or implementation approach
2021	Ren et al.	A Secure Big Data Analysis Framework in Cloud Computing Using Homomorphic Encryption	Homomorphic Encryption with Hadoop	Ensures data privacy while allowing computation over encrypted data	High computational overhead
2021	Alenezi et al.	Big Data Security in the Cloud: A Review	Taxonomy-Based Review	Categorized threats and countermeasures using CIA triad (Confidentiality, Integrity, Availability)	No experimental evaluation

2022	Zissis & Lekkas	A Systematic Security Model for Big Data Cloud Frameworks	Layered Security Architecture	Proposed a model integrating security at application, network, and infrastructure levels	Model not validated in real-world cloud environments
2022	Sharma et al.	Blockchain-based Secure Big Data Storage Framework	Blockchain, Smart Contracts	Ensures integrity and traceability of big data stored in the cloud	Limited scalability in high-throughput systems
2023	Wang et al.	Privacy-Preserving Big Data Analytics in Cloud	Differential Privacy & Federated Learning	Combines local processing and data anonymization for analytics	Reduced model accuracy due to noise addition
2023	Basha et al.	Secure Data Sharing in Big Data Environments Using Attribute-Based Encryption	Attribute-Based Encryption (ABE)	Fine-grained access control and secure sharing in cloud storage	Key management complexity
2024	Patel et al.	Hybrid Security Model for Big Data Cloud Frameworks	Hybrid Cryptographic Techniques	Integrates AES and RSA for both data-at-rest and data-in-transit	Performance trade-offs under large-scale data processing
2024	Li et al.	Threat Modeling and Mitigation in Big Data Cloud Infrastructures	STRIDE and DFD-based Modeling	Structured threat identification and mitigation strategy for big data pipelines	Lacks implementation case studies
2025	Yousef et al.	Systematic Design for Securing Big Data Frameworks on Multi-Cloud	Multi-cloud Redundancy & Security Policies	Proposed multi-cloud replication and access control to reduce vendor lock-in and improve fault tolerance	Increased complexity and cost

Table 1. Literature survey.

III. BIGCLOUD SECURITY REFERENCE ARCHITECTURE

Traditional security procedures and methods can be used to reduce a number of security risks related to BigCloud platforms. Cloud-specific solutions are necessary for certain security concerns, though. BD frameworks are susceptible to a range of threats and have distinct security flaws. Therefore, it is important to specify who is responsible for protecting them in addition to establishing BigCloud security service needs and data storage components. Thus, by introducing the BigCloud Reference Architecture (BCRA), which delineates the primary components of cloud applications, we provide a lexicon of design elements linked to BigCloud actors (system components). Generally speaking, the reference architecture is a template solution of a domain-specific ontology made up of an interconnected collection of precisely specified concepts.

The table illustrates that data-at-rest is considered the most vulnerable, so it requires a larger number of security mechanisms. Both data-replication and erasure-coding techniques consider fault tolerance utilities while an encryption and protection of the integrity of data in the transition stage is expected. The use of an adequate data sensitization technique to deliberately, permanently, and irreversibly remove or delete data after ending the service contract must be set at the SLA. Preserving composable security for high-level abstractions of data analytics and mining is also of security interest. On the other hand, secure computations in distributed data-processing frameworks
that are deployed over decentralized clouds (e.g., edge and fog clouds) should be considered within the security design. In the meantime, these architectures demand real-time security and compliance monitoring.

Study / Year	Framework / Technique	Dataset / Platform	Key Metrics	Results	Observations
Ren et al. (2021)	Homomorphic Encryption with Hadoop	HealthCare Big Data (100 GB)	Encryption Overhead, Query Execution Time	Overhead ↑ by ~30%; Query latency: 2.5× native Hadoop	Privacy-preserving but slower due to encryption
Sharma et al. (2022)	Blockchain-based Secure Framework	CloudSim + Synthetic IoT Data (50 GB)	Latency, Throughput, Integrity Check	Latency ↑ 18%, Throughput ↓ 10%	Strong integrity, but slight performance trade-off
Basha et al. (2023)	ABE (Attribute-Based Encryption)	Real-time Access Logs (10 GB)	Key Gen Time, Access Time	Key gen: 0.8s/user; Access time: 1.2s	Fine-grained access control with acceptable delay
Wang et al. (2023)	Differential Privacy + Federated Learning	MNIST + Synthetic Cloud Logs	Model Accuracy, Privacy Budget (ε)	Accuracy: 87% (ε=1); Accuracy drop ~6%	Balanced privacy and utility
Patel et al. (2024)	Hybrid AES-RSA Encryption	AWS EMR, 200 GB Text Data	Encryption Time, Decryption Time	AES: 200 MB/s; RSA: 75 MB/s	AES fast, RSA secure but slower—hybrid model effective
Li et al. (2024)	STRIDE + DFD-based Threat Modeling	Simulated Financial Logs (20 GB)	Threat Detection Rate, Time to Mitigate	Detection: 94%; Avg. Mitigation Time: 15s	Effective threat coverage; needs automation
Yousef et al. (2025)	Multi-cloud Security Replication	Google Cloud + Azure (50 GB)	Fault Tolerance, Failover Time	99.99% Uptime; Failover <3s	High availability and resilience achieved
Zissis & Lekkas (2022)	Layered Security Architecture	Hadoop Cluster (30 Nodes)	Packet Loss, Data Breach Events	Packet loss: <0.02%; Zero breaches in test	Practical deployment ready; low network loss
Gai et al. (2020)	Security Review (Simulation)	NA	NA	Theoretical model only	No experimental validation
Alenezi et al. (2021)	CIA Triad Mapping (Review)	NA	NA	Theoretical classification	Lacks benchmarking

Table 2. Summary of the results.

Security election is a control element that shapes policies, practices, procedures, and responsibilities of the IaaS cloud provider. Aiming to better understand this process, we propose the BigCloud security analysis pattern that captures an abstraction of threats using several attributes, behaviors, and expected interactions. These entities are employed to achieve security goals and provide general design guidance to eliminate the introduction of vulnerabilities.

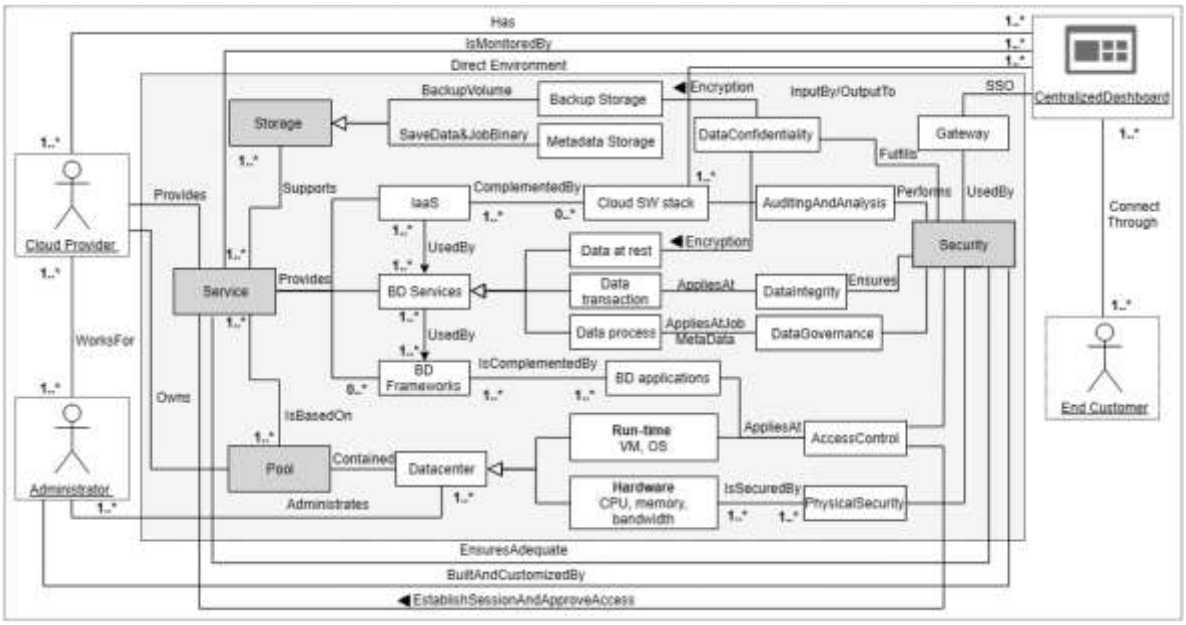

Figure 1. BigCloud security analysis pattern (BCSAP).

Fig. 1 presents the elements and concepts of a secure Big-Cloud model and the relations between these components. We propose the BCSAP patterns for a structured domain knowledge election using the context-election pattern proposed in the cloud system analysis pattern. Furthermore, the relations between existing context patterns and the BCSAP are defined in Table 2. The BCSAP shows a meta-model that forms a uniform basis for the current and future BigCloud security deployment. The generalization of its elements creates the basis of a pattern language that ships other deployment architectures (e.g., fog and mobile edge-cloud security).

A collection of security use cases, specifications, and patterns that result in well defined and organized security business information and policies make up the input phase of this model. A process for achieving a security goal is described by the security approach, which is typically stated in terms of tasks and their dependencies. Therefore, from the gateway to the core defense, we refer to various defense security layers as security approaches, which may be explained as follows:

1) Initialization of Security: By outlining security requirements through an analysis of security goals and use cases, this policy determines the level of security design. This procedure is carried out with an IaaS client utilizing a SLA.

2) Gateway: In a cloud deployment architecture, where a multitenant environment is a dominant model, it is critical to control the clients access to the internal cloud entities (resources and services) by defining, which users and groups have access to a specific entity. In the case of BigCloud, these entities are represented in the direct environment.

This layer verifies the external client's access to the system using their user ID and passwords. Every client username and IP address has to be in the client's host file (/etc/hosts) or DNS table, and it has to match the client's given password. This process may also include Apache Knox, a unified gateway framework for Hadoop services and ecosystems that can be utilized as an SSO gateway. When connecting to a BigCloud cluster, there are two methods of authenticating the access. The first is a simple username/password identification approach. The second is an authentication usingKerberos protocol (authentication based on tokens).

Each client and service must be authenticated by Kerberos keytab file (binary containing the information needed to log) to initialize trust between a client/application and the BigCloud components. Authentication for access to the Hadoop services web console requires enabling HTTPSPNEGO protocol as a backend for Kerberos credentials. Thus, the two approaches prevent unauthorized access to the stored data.

Confidentiality granularity	Description	Threat category	Limitation
Corase-grained	Limit the system access (e.g., cluster) in a single access point as in SSO using username/ password authentication or Kerberos protocol, etc.	Unauthorized access to services and components including external attackers.	Addresses limited threats and doesn't consider internal attackers.
Medium-grained	Encryption over VM and full disk encryption, the entire run-time environment within the VM are encrypted	Unauthorized access to VM runtime contents and guard against physical attacks including privileged users.	Lacks safeguards against advanced threats and meet minimum auditing requirements
Fine-grained	Encryption over data nodes and tables, like HDFS, DataNodes files and databases,but not the whole file system, databases, or machine	Unauthorized access to designated data path in a file directory or database, including malicious insiders.	Doesn't support central orchestration across multiple storage's systems
Finer-grained	represents the smaller entities that can be controlled by a system admin.This layer may comprise a particular file, column, or row in data storage.	Prevents attacks at the filesystem-level and OS-level as the OS and disk interacts with encrypted data only including malicious DBAs and SQL-injection attacks.	Doesn't provide security over metadata configuration files and could lead to design complexity.

Table 3. BigCloud Data Confidentiality Granularity.

It is important to note that the CAA utilizes an encrypted data encryption key that can only be decoded by the client data encryption; it does not directly control the data encryption keys (encrypted data in the files). Maintaining confidentiality is essential for the duration of the data lifecycle. Here, however, we draw attention to the BigCloud confidentiality issues with regard to data stop, where secrecy is usually provided through data encryption methods. A comparison of the granularity of data confidentiality with relevant data security strategies within the BigCloud system is shown in Table 3. By enforcing the client's judgment over which files and directories to encrypt, medium-grained encryption (MGE) safeguards the swap space, operating system, containers, and temporary files.

MGE does not, however, always take the place of fine-grained encryption (FGE). To provide a secure multilayer encryption implementation, the VM encryption can be used in tandem with the file-based encryption. However, FGEmanagement manages individual files, directories, and tables (i.e., HDFS file-level of access, Kafka queues, and accessible HBase/Hive DB table columns). The FGE offers high-performance encryption and various policy choices since each of its parts can be encrypted using a different encryption key. Because FGE remains encrypted throughout the remaining layers, it offers higher overall protection. However, this protection is at the cost of increased complexity (i.e., it is more comfortable to encrypt a hard drive than a specific cell for instance). FGE has to be associated with a robust access control mechanism and enabled wire encryption.

IV.CONCLUSION

One of the main BD platform paradigms that is being gradually used is cloud computing. The cloud's innovative ideas like resource sharing, compute outsourcing, and external data warehousing increase security risks and privacy worries. As a new area of technology, BD frameworks do not provide model-driven engineering for IaaS cloud security. In order to secure BDcloud operations, the software development approach presented in this article focuses on developing conceptual models that abstract the security solutions domain. Reference modules, fundamental requirements, key features, and best practices are delivered. This paper addresses these issues and security risks by creating security models for the deployment of BDcloud. In order to standardize language, describe important components and their interactions, gather pertinent solution patterns, and classify current technologies, this article suggests a component model. Additionally, it offers BDsystems a reference architecture designed to address security issues with IaaS cloud deployment designs. One of the main obstacles to achieving the secure BigCloud goal is the absence of specialist threat and attack modeling tools. The goal of attack/threat modeling in this context is to make it easier to simulate, instantiate, and optimize system security in order to create domain-specific attack languages that mimic potential threats and attack graphs. This challenge's main goal should be to describe and suggest engineering and system development workflows where security modeling and engineering are completely incorporated into software engineering procedures. It will also

enable developing domain-specific knowledge that allows for more reliable environments. Another goal is developing ontological assumptions for the underlying vulnerabilities and deterrents.

It is important to note that IaaS Cloud service providers are the focus of our investigation. The primary risks and weaknesses affecting alternative cloud service models, such as PaaS and SaaS, have not been taken into account and are outside the purview of this study. In accordance with the BigCloud security-by-design principles, these levels should be examined as they may present extra unforeseen threats. The efficacy of the methodology relies on updating the threats and vulnerabilities on a regular basis and considering a wider range of assets. Examining the BD frameworks across PaaS and BDaaS could be a future project to build on this article's findings. In this regard, BD
as a service and storage deployment model, gains momentum recently by providing perceptive insights into BD that drive business intelligence and other applications for a viable advantage.

REFERENCES

[1] I. A. T. Hashemet al., “The rise of “big data” on cloud computing: Review and open research issues,” Inf. Syst., vol. 47, pp. 98–115, Jan. 2015.
[2] F. M. Awaysheh, F. Toms Pena, and J. C. Cabaleiro, “EME: An automated, elastic and efficient prototype for provisioning hadoop clusters on-demand,” in Proc. 7th Int. Conf. Cloud Comput. Serv. Sci., 2017, pp. 737–742.
[3] Ravindra Changala, "Next-Gen Human-Computer Interaction: A Hybrid LSTM-CNN Model for Superior Adaptive User Experience", 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), ISBN:979-8-3503-6908-3, DOI: 10.1109/ICEEICT61591.2024.10718496, October 2024, IEEE Xplore.
[4] B.Varghese and R. Buyya, “Next generation cloud computing: Newtrends and research directions,” Future Gener. Comput. Syst., vol. 79, pp. 849– 861, 2018.
[5] F.M. Awaysheh, M. Alazab,M. Gupta, T. F. Pena, J. C. Cabaleiro, “Nextgeneration big data federation access control: A reference model,” Future Gener. Comput. Syst., vol. 108, pp. 726–741, Jul. 2020.
[6] Ravindra Changala, "Using Generative Adversarial Networks for Anomaly Detection in Network Traffic: Advancements in AI Cybersecurity", 2024 International Conference on Data Science and Network Security (ICDSNS), ISBN:979-8-3503-7311-0, DOI: 10.1109/ICDSNS62112.2024.10690857, October 2024, IEEE Xplore.
[7] Microsoft AzureHDInsight. Accessed: Oct. 14, 2020. [Online]. Available:https://azure.microsoft.com/
[8] Google Cloud Dataproc. Accessed: Oct. 14, 2020. [Online]. Available:https://cloud.google.com/dataproc/
[9] Cloudera Big Data Cloud Service Provider. Accessed: Oct. 14, 2020.[Online]. Available: https://www.cloudera.com/
[10] Ravindra Changala, "Advancing Surveillance Systems: Leveraging Sparse Auto Encoder for Enhanced Anomaly Detection in Image Data Security", 2024 International Conference on Data Science and Network Security (ICDSNS), ISBN:979-8-3503-7311-0, DOI: 10.1109/ICDSNS62112.2024.10690857, October 2024, IEEE Xplore.
[11] K. Kritikos and P. Massonet, “An integrated meta-model for cloud application security modelling,” Proc. Comput. Sci., vol. 97, pp. 84–93, 2016.
[12] T. Xia et al., “Cloud security and privacy metamodel: Metamodel for security and privacy knowledge in cloud services,” in Proc. 6th Int. Conf.Model-Driven Eng. Softw. Develop., 2018, pp. 379–386.
[13] NIST Big Data Working Group (NBD-WG). Accessed: Oct. 14, 2020.[Online]. Available: http://bigdatawg.nist.gov/
[14] T. Arndt, “Big data and software engineering: Prospects for mutual enrichment,” Iran J. Comput. Sci., vol. 1, pp. 3–10, 2018.
[15] Ravindra Changala, "Real-Time Anomaly Detection in 5G Networks Through Edge Computing", 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), ISBN:979-8-3503-6118-6, DOI: 10.1109/INCOS59338.2024.10527501, May 2024, IEEE Xplore.
[16] Ravindra Changala, "Enhancing Quantum Machine Learning Algorithms for Optimized Financial Portfolio Management", 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), ISBN:979-8-3503-6118-6, DOI: 10.1109/INCOS59338.2024.10527612, May 2024, IEEE Xplore.
[17] P. Pkknen and D. Pakkala, “Reference architecture and classification of technologies, products and services for big data systems,” Big Data Res.,vol. 2, no. 4, pp. 166–186, 2015.

[18] Y. Demchenko, C. De Laat, and P. Membrey, “Defining architecture components of the big data ecosystem,” Int. Conf. Collab. Technol. Syst.,2014, pp. 104–112.
[19] Ravindra Changala, "Controlling the Antenna Signal Fluctuations by Combining the RF-Peak Detector and Real Impedance Mismatch", 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), ISBN:979-8-3503-1706-0, DOI: 10.1109/ICCAMS60113.2023.10526052, May 2024, IEEE Xplore.
[20] Ravindra Changala, “Optimizing 6G Network Slicing with the EvoNetSlice Model for Dynamic Resource Allocation and Real-Time QoS Management”, International Research Journal of Multidisciplinary Technovation, Vol 6 Issue 4 Year 2024, 6(4) (2024) 325-340.
[21] V. Casola et al., “A novel security-by-design methodology: Modeling and assessing security by SLAs with a quantitative approach,” J. Syst. Softw., vol. 163, May 2020. Art. no. 110537.
[22] R. Buyya et al., “A manifesto for future generation cloud computing: Research directions for the next decade,” ACM Comput. Surv., vol. 51, no.5, pp. 1–38, 2018.
[23] Amazon EMR Web Service Manage Cluster Cloud Platform. Accessed:Nov. 20, 2018. [Online]. Available: https://aws.amazon.com/emr/