



Systematic literature review for Architectural Approaches in Big Data Systems: Integration of Processing, Security, and Advanced Analytics

Revisión sistemática de literatura para los Enfoques Arquitectónicos en Sistemas de Big Data: Integración de Procesamiento, Seguridad y Analítica Avanzada

Deivid Sebastian Armero Gualguan¹  

¹ Universidad de Nariño. Nariño, Colombia.

Submitted: 07-06-2026 | Revised: 14-08-2026 | Accepted: 19-09-2026 | Published: 20-09-2026

ABSTRACT

Modern Big Data architectures have moved away from monolithic models to adopt distributed approaches based on microservices and containers. In this context, tools such as Kubernetes and Docker allow applications to be deployed and managed with greater flexibility, enabling self-scaling and efficient orchestration that improves availability and resilience. In terms of data processing, the Lambda and Kappa architectures remain key references. Both seek to integrate batch analysis with real-time processing, either through differentiated layers or through continuous flows, achieving consistency without compromising performance. Data management has also evolved. Instead of large centralized repositories prone to degrade into "data swamps", approaches such as domain-oriented design (DDD) and federated governance are adopted. This distributes responsibility between teams and improves the quality and traceability of information. Security is integrated transversally through service mesh, secure protocols and, in some cases, blockchain, reinforcing communication and data integrity. Finally, MLOps and advanced analytics are incorporated through automated pipelines that operate between the cloud and edge computing, optimizing the life cycle of models in distributed environments.

Keywords: Big Data, software architecture, distributed systems, data processing, data lakes.

RESUMEN

Las arquitecturas modernas de Big Data han dejado atrás los modelos monolíticos para adoptar enfoques distribuidos basados en microservicios y contenedores. En este contexto, herramientas como Kubernetes y Docker permiten desplegar y gestionar aplicaciones con mayor flexibilidad, habilitando el autoescalado y una orquestación eficiente que mejora la disponibilidad y la resiliencia. En cuanto al procesamiento de datos, las arquitecturas Lambda y Kappa siguen siendo referencias clave. Ambas buscan integrar el análisis batch con el procesamiento en tiempo real, ya sea mediante capas diferenciadas o a través de flujos continuos, logrando consistencia sin comprometer el rendimiento. La gestión del dato también ha evolucionado. En lugar de grandes repositorios centralizados propensos a degradarse en "data swamps", se adoptan enfoques como el diseño orientado al dominio (DDD) y la gobernanza federada. Esto distribuye la responsabilidad entre equipos y mejora la calidad y trazabilidad de la información. La seguridad se integra de forma transversal mediante service mesh, protocolos seguros y, en algunos casos, blockchain, reforzando la comunicación y la integridad de los datos. Por último, MLOps y la analítica avanzada se incorporan mediante pipelines automatizados que operan entre la nube y el edge computing, optimizando el ciclo de vida de modelos en entornos distribuidos.

Palabras clave: Big Data, arquitectura de software, sistemas distribuidos, procesamiento de datos, data lakes.

INTRODUCTION

The growing generation of large-scale data has intensified the need for architectural approaches capable of efficiently managing its processing, storage, and analysis. In this context, Big Data systems are conceived as infrastructures designed to handle large volumes of heterogeneous data, generated at high velocity and originating from multiple sources, through distributed and scalable technologies. The integration of privacy-preserving techniques into Big Data analysis frameworks

Data is crucial for ensuring data security; the challenge lies in applying advanced machine learning (ML) algorithms to sensitive data without compromising privacy[1]. However, the increasing complexity of these environments poses new challenges for the coherent integration of their different components, especially in scenarios that require efficient processing, data protection, and advanced analytical capabilities.

Currently, the state of the art in Big Data architectures shows an evolution toward hybrid and integrated models that combine batch and real-time processing, incorporate advanced security and privacy mechanisms, and facilitate the adoption of advanced analytics and machine learning. In this context, approaches such as Lambda, Kappa, and microservices-based architectures, together with the use of cloud and edge technologies, have enabled the construction of more flexible, scalable, and adaptive ecosystems. Consequently, these architectures represent a response to the need to process data in a timely manner without sacrificing consistency or security.

Based on this overview, this work adopts a systematic literature review process aimed at analyzing the main architectural approaches in Big Data systems. In particular, questions are raised regarding the integration of processing paradigms, the incorporation of security and privacy mechanisms, and the inclusion of advanced analytics and machine learning components within these architectures. These questions enable the structuring of a critical and comparative analysis of various proposals, contributing to a deeper understanding of current solutions and their implications in real-world scenarios.

Finally, the document is organized into several sections that address the questions posed. Initially, architectural approaches for integrating batch and real-time processing are analyzed. Subsequently, data security and protection mechanisms are examined, and the incorporation of advanced analytics and machine learning in these environments is studied. Complementary sections are also included that address other relevant aspects identified during the review, and the general conclusions of the study are presented at the end.

METHODS

This systematic literature review was developed through a structured methodological approach aimed at identifying, selecting, and analyzing relevant studies in the field of Big Data systems and their architectural approaches. The objective was to consolidate scientific evidence related to

the integration of processing, security, and advanced analytics in distributed environments, including their relationship with cloud computing technologies. The process began with the definition of the research questions, which guided the search strategy and thematic delimitation. Subsequently, information was collected from 5 academic databases: DOAJ, IEEE Xplore, ResearchRabbit, Redalyc, and Scopus, yielding a total of 120 articles. These were filtered using inclusion and exclusion criteria based on their relevance, methodological quality, and correspondence with the study's axes.

The selection and analysis of the studies were carried out following guidelines for systematic reviews widely used in Software Engineering, which ensured the traceability and rigor of the process. In particular, methodological principles proposed in the literature were considered, such as those of Barbara Kitchenham and the PRISMA guidelines.

Finally, the selected articles were analyzed through a qualitative synthesis, which enabled the identification of significant trends, approaches, and contributions in the design of Big Data architectures.

Research Questions

The objective of this study is to analyze architectural approaches in Big Data systems, with emphasis on the integration of processing, security, and advanced analytics. To construct the state of the art, the following research questions were formulated:

—RQ0: What are the main trends, approaches, and challenges in the design of architectu-

res for Big Data systems oriented toward the integration of processing, security, and advanced analytics?

—RQ1: How are microservices-based architectures designed and evaluated to support Big Data systems?

—RQ2: What architectural practices and orchestration strategies are used in Big Data systems deployed in container-based environments?

—RQ3: How do architectural decisions influence data management, governance, and integration within Big Data and Data Lakes ecosystems?

—RQ4: What architectural approaches are employed to integrate batch and real-time processing in Big Data systems?

—RQ5: How are security, privacy, and data protection mechanisms incorporated into software architectures for Big Data systems?

—RQ6: How do modern software architectures integrate machine learning and advanced analytics components in Big Data ecosystems?

Table 1. Definition of general concepts with PIPOH. Adapted from [2].

Criterion	Description
Population	Big Data systems and software architectures oriented toward the processing of large data volumes
Intervention	Architectural approaches; integration of batch and real-time processing; security and privacy mechanisms; and use of analytics advanced and machine learning
Results	Scientific publications (articles and conferences) that describe Big Data architectures, including processing, security, and analytical
Professionals	Software engineering, computer science, data engineering
Context	Academic, industrial, and Big Data research environments

Search

Table 2. Term, synonyms, and filters to compose the search string. Adapted from [2].

Criterion	Description	Additional filters
Big Data architectures	Big Data systems, data architectures, and distributed data systems.	>= 2020
	Real-time processing, stream processing, hybrid processing.	
	Data protection, information security, privacy-preserving	
	Advanced analytics, AI in Big Data, and data analytics.	
	System design, distributed architecture	

Table 3. Main term and database. Adapted from [2].

Criterion	Description
Big Data architectures	IEEE Xplore
	DOAJ
	SCOPUS
	REDALYC

Table 4. Search String. Adapted from [2].

Main term	Search string
Big Data Architectures	("Big Data Architectures" OR "Data architectures" OR "Distributed data systems") AND ("batch processing" OR "stream processing" OR "real-time processing") AND ("security" OR "privacy" OR "machine learning" OR "analytics")

Selection

Table 5. Dates of consultation and file download. Adapted from [2].

Database	Search date
IEEE Xplore	04/03/2026
DOAJ	04/03/2026
Scopus	04/03/2026
REDALYC	04/03/2026

Design and evaluation of microservices-based architectures for Big Data systems.

In digital health systems, microservice architectures integrate distributed infrastructures, NoSQL databases, and advanced analytics mechanisms, enabling the application of machine learning algorithms to large datasets for pattern identification and decision support[3]. The adoption of microservices represents an evolution toward decentralized models; however, their systematic integration into Big Data ecosystems remains limited, necessitating reference architectures that articulate modern software engineering with large-scale analytics platforms[4].

Table 6. Quality assessment in search and selection processes. Adapted from [2].

Term Main	Results	Duplicates	Excluded	Relevant	Base of Data
Big Data Architectures	30	0	22	8	IEEE Xplore
	23	0	16	7	DOAJ
	84	0	64	20	Scopus
	16	1	10	6	REDALYC
Total	153	1	112	41	

Hybrid frameworks that combine microservices with deep learning improve scalability, modularity, and computational efficiency in bioinformatics applications, facilitating real-time processing and interoperability between distributed systems[5]. Reference architectures evaluated using the ATAM method have demonstrated improvements in performance, availability and modifiability, with response times of less than one second and greater capacity to evolve in the face of changes in requirements or volumes of data[4].

Event-driven architectures enable each domain to manage its own information while global consistency is maintained through asynchronous communication, favoring service autonomy and dynamic adaptation of the system[6]. Practical implementation requires container-based infrastructures and platforms such as Kubernetes, which automatically scale services according to demand and maintain the established quality-of-service levels and service-level objectives[7].

Architectural practices and orchestration strategies for Big Data systems in container-based environments.

Microservices deployed in containers managed through Kubernetes allow dynamic scaling of computational resources to maintain quality and service-level objectives, demonstrating that autoscaling optimizes infrastructure under variations in demand[7]. The orchestration of Big Data flows through persistent containers, prioritizing data locality, reduces latency and improves performance, showing that separating the control, data, and compute layers allows the construction of more efficient and scalable platforms[8].

Cloud-native architectures deploy containerized components managed with Docker and Kubernetes, automating deployment and resource management through configuration files that describe the operation of pods within the cluster[9]. Microservices-based collection and analysis systems allow components such as collectors, download modules, and analysis systems to scale independently within the cluster, improving resilience and the capacity to process large volumes of information[10].

Reactive IoT applications with Big Data analytics apply orchestration patterns such as multi-zone deployments, node affinity, and StatefulSets on Kubernetes, controlling container placement and guaranteeing availability and stable performance[11]. Deep learning frameworks for biological data implement microservices encapsulated in containers orchestrated with Kubernetes and Docker Swarm, executing analyses in parallel within distributed clusters with fault tolerance and elastic scalability[5].

Reference architectures for distributed Big Data employ domain-oriented design with microservices communicating through a service mesh on Kubernetes, using sidecar proxies for load balancing, fault control, and telemetry without modifying application code[4]. Adaptive Big Data platforms deploy autonomous microservices in containerized environments with independent scaling by domain, relying on a service mesh to manage communication and scaling policies that enable dynamic adaptation to workload changes[6].

Health telemonitoring systems structured with microservices in Docker containers managed by Kubernetes allow each component to be replicated and scaled independently, integrating analytics and machine learning services with high availability under increasing demand[3]. Architectures that integrate Big Data and machine learning in distributed environments leverage containerization to partition AI models into parallel tasks executed in pods managed by Kubernetes, reducing execution times through parallelization strategies controlled by the orchestrator[12].

Standard Big Data technologies integrated into containerized environments employ Kubernetes as a middleware platform to schedule Docker containers on cluster nodes, managing resource allocation and facilitating large-scale simulations within distributed infrastructures[13]. Big Data architectures oriented toward advanced services organize containerized microservices in cloud-virtualized infrastructures, enabling modules from different providers to integrate independently and scalably, favoring extensibility and the construction of open data-based ecosystems[14].

Influence of architectural decisions on data management, governance, and integration within Big Data and Data Lake ecosystems.

Domain-based decentralized architectures address critical governance problems through federated control models that distribute responsibility for data, preventing the formation of “data swamps” and facilitating interoperability between services in highly distributed environments[4]. Domain-oriented design with federated governance establishes distributed mechanisms for quality control and metadata management, where each domain retains ownership of its data while global coherence is ensured through asynchronous communication between services[6].

Functional layered architectures manage heterogeneous data through common models and semantic harmonization processes, with a central governance layer that coordinates integration, preprocessing, storage, and querying, eliminating silos and facilitating interoperability[14]. Modern cloud-based architectures decouple storage and compute, automating data pipelines and implementing flexible security and compliance mechanisms compared with traditional monolithic approaches[15].

An adequate Big Data architecture design enables the coherent collection, storage, and processing of information; however, the absence of quality controls in Data Lakes generates inconsistent data that hinders the use of analytical systems[16]. Data Lake architectures with integrated security and monitoring mechanisms strengthen governance by incorporating components that supervise data integrity and lifecycle, facilitating anomaly detection without relying on external solutions[17].

Multilevel governance architectures improve data quality and integration through dynamic rules and traceability mechanisms such as lineage analysis, ensuring consistency and reliability for complex analytical processes[18]. The use of ontology-oriented architectures solves integration problems in heterogeneous environments by providing a common mediated schema that facilitates semantic interoperability among multiple sources within Big Data ecosystems[19].

Layered Data Lake architectures efficiently manage the data lifecycle through cleaning and structuring processes prior to storage, avoiding disorganized environments and facilitating quality and real-time analysis[20]. The integration of blockchain into Big Data architectures strengthens

governance through anonymity, access control, and traceability mechanisms, where ETL rules regulate storage and guarantee the integrity of information throughout its lifecycle[21].

Decisions about distributed database architectures influence data availability, consistency, and interoperability, since models based on replication and fragmentation can generate integration difficulties and limit portability between providers[22]. The incorporation of the “mask” element within the mediator-wrapper model improves system flexibility and scalability by redistributing responsibilities among components, facilitating schema management and reducing complexity in integrating heterogeneous sources[23].

Layered architectures in IoT and Big Data environments manage integration through components specialized in acquisition, validation, aggregation, and analysis, guaranteeing data consistency before their use in analytical processes[24]. The relationship between data architecture and governance requires incorporating control, security, and modeling mechanisms from the system design, since the heterogeneity of devices and the distribution of data require structures that ensure the protection and quality of information[25].

Federated learning-based architectures integrate data from multiple entities without compromising privacy, through authentication and encryption mechanisms that guarantee access control and governance without creating dependence on centralized repositories[26]. Layered storage architectures manage data quality through cleaning, standardization, and transformation processes at each level, ensuring that the information is suitable for analysis in Big Data systems[25].

Hybrid architectures that combine distributed storage and blockchain improve governance in Data Lakes through scalable integrity verification mechanisms, where the separation between off-chain storage and on-chain anchoring reduces costs and enables real-time audits[27]. Architectures based on data profiles optimize integration in Data Lakes by automating the discovery of relationships between datasets, facilitating interoperability between heterogeneous schemas and reducing the computational complexity of traditional integration methods[28].

Security, privacy, and data protection mechanisms within software architectures for Big Data systems.

A hybrid architecture for Data Lakes integrates cryptographic mechanisms into the data pipeline through authenticated encryption and verification structures based on Merkle and blockchain, processing and validating each record before storage to guarantee integrity, authenticity, and traceability[29]. Modern Big Data architectures integrate security through Kerberos authentication, granular access control with Apache Ranger, and cloud identity management services, directly coupled to storage and processing systems[15].

A secure Data Lake architecture incorporates components dedicated to integrity monitoring as embedded services, integrating anomaly detection systems and checksum controllers that continuously monitor the processes of data loading, processing, and access[17]. A reference architecture incorporates security and privacy as cross-cutting concerns, integrating identity and access management together with policy enforcement mechanisms through sidecars that control access and governance in a federated manner in distributed systems[6].

A blockchain-based architecture for decentralized identity management uses verifiable identifiers to control access and data flow, guaranteeing privacy and traceability through smart contracts over the lifecycle of the information[35]. An architecture with a privacy-oriented analytics layer integrated throughout the entire data flow allows machine learning models to be executed on encrypted and anonymized information using techniques such as differential privacy and federated learning[1].

A Big Data architecture for IoT integrates security components oriented toward identity management, authentication, and authorization through standards such as OAuth2 and XACML, inserted directly into the interaction flow among devices, services, and data platforms[36]. A distributed architecture integrates security through service meshes and sidecar proxies that implement secure communication with mTLS and granular authorization policies, applying consistent access and privacy controls in distributed data environments[4].

An architecture for massive data collection integrates differential privacy directly into the ingestion pipeline, protecting the data from the moment of its capture before being stored or processed[37]. An architecture that integrates encryption, authentication, access control, and blockchain guarantees the security, integrity, and transparency of the data throughout the entire supply chain within a Big Data environment[38].

Integration of modern software architectures, machine learning components, and advanced analytics in Big Data ecosystems.

An Earth system digital twin architecture integrates AI and machine learning components as core modules within information and simulation subsystems, combining physical models with AI-based models according to precision constraints, computational resources, and response time requirements[39]. A Data Lake processing architecture integrates a machine learning model into a hybrid offline-online pipeline that analyzes massive data profiles to infer the quality of joins between attributes and optimizes the execution of large-scale analytical operations[28].

An MLOps-based architecture integrates machine learning components, such as automated pipelines deployed on Kubernetes, in which each stage of preprocessing, modeling, and inference runs as an independently orchestrated pod to support batch processing in Big Data environments[10]. A multilayer architecture encapsulates machine learning and artificial intelligence components by integrating tools such as Spark MLlib and TensorFlow to manage the complete model lifecycle, feeding batch and streaming flows[14].

An extension of Big Data reference architectures incorporates specific modules for the machine learning lifecycle, integrating model development, packaging, deployment, and distribution with AIOps and DevOps practices in distributed infrastructures[40]. An MLOps-oriented architecture integrates advanced analytics through distributed pipelines for model acquisition, processing, and execution, automatically deployed in heterogeneous edge, fog, and cloud environments through standardized containers and inference engines[41].

An architecture based on the Lambda paradigm integrates descriptive, predictive, and prescriptive analytics components at the logical level, combining batch processing for model training with stream processing for generating real-time predictions from industrial data streams[34]. A microservices-based architecture implements generative deep learning models as independent, modular services orchestrated within Big Data workflows to facilitate scalability, updating, and parallel execution of specialized analytical tasks[5].

A collaborative cloud-edge architecture functionally distributes machine learning components, executing inference on edge devices for real-time processing while the cloud manages storage and continuous retraining of models[42]. A Data Lake architecture integrates machine learning models as embedded services within monitoring components, where algorithms such as Isolation Forest, K-means, and convolutional networks perform anomaly detection, clustering, and real-time classification[17].

REFERENCES

1. A. Cuzzocrea y I. Belmerabet, «Enhanced Privacy-Preserving Cloud Big Data Analytics with Decentralized SOLID Personal Online Datastores», *Procedia Comput. Sci.*, vol. 270, pp. 1956-1967, 2025, doi: 10.1016/j.procs.2025.09.316.
2. J. A. Jiménez-Toledo, «Consideraciones en los procesos de enseñanza-aprendizaje para un primer curso de programación de computadores: una revisión sistemática de la literatura».
3. H. Calderón-Gómez et al., «Telemonitoring System for Infectious Disease Prediction in Elderly People Based on a Novel Microservice Architecture».
4. P. Ataei, «Cybermycelium: a reference architecture for domain-driven distributed big data systems», *Front. Big Data*, vol. 7, p. 1448481, nov. 2024, doi: 10.3389/fdata.2024.1448481.
5. A. F. El-Sayed, S. M. Darwish, N. M. Labib, y M. S. El-Hadidi, «Generative deep learning in bioinformatics: a microservices-driven paradigm for big data environment», *J. Big Data*, vol. 13, n.o 1, p. 7, dic. 2025, doi: 10.1186/s40537-025-01320-5.
6. P. Ataei y M. Atemkeng, «Terramycelium: a reference architecture for adaptive big data systems».
7. L. M. Ruiz, P. P. Pueyo, J. Mateo-Fornes, J. V. Mayoral, y F. S. Tehas, «Autoscaling Pods on an On-Premise Kubernetes Infrastructure QoS-Aware», *IEEE Access*, vol. 10, pp. 33083-33094, 2022, doi: 10.1109/ACCESS.2022.3158743.
8. A.-A. Corodescu et al., «Big Data Workflows: Locality-Aware Orchestration Using Software Containers», *Sensors*, vol. 21, n.o 24, p. 8212, dic. 2021, doi: 10.3390/s21248212.
9. Y. Ding, «Research on Management and Optimization of Big Data Computing Engine Based on Cloud Native Technology IT Architecture», *Procedia Comput. Sci.*, vol. 243, pp. 910-917, 2024, doi: 10.1016/j.procs.2024.09.109.
10. J. Pastor-Galindo, H.-Â. Sandlin, F. G. Mármol, G. Bovet, y G. M. Pérez, «A Big Data architecture for early identification and categorization of dark web sites», *Future Gener. Comput. Syst.*, vol. 157, pp. 67-81, ago. 2024, doi: 10.1016/j.future.2024.03.025.
11. L. Jurado Pérez y J. Salvachúa, «An Approach to Build e-Health IoT Reactive Multi-Services Based on Technologies around Cloud Computing for Elderly Care in Smart City Homes», *Appl. Sci.*, vol. 11, n.o 11, p. 5172, jun. 2021, doi: 10.3390/app11115172.
12. P. Pääkkönen, «Extending reference architecture of big data systems towards machine learning in edge computing environments», 2020.
13. R. Knapen et al., «Efficient and scalable crop growth simulations using standard big data and distributed computing technologies», *Comput. Electron. Agric.*, vol. 236, p. 110392, sep. 2025, doi: 10.1016/j.compag.2025.110392.
14. M. Pau, P. Kapsalis, Z. Pan, G. Korbakis, D. Pellegrino, y A. Monti,
15. «MATRYCS—A Big Data Architecture for Advanced Services in the Building Domain», *Energies*, vol. 15, n.o 7, p. 2568, abr. 2022, doi: 10.3390/en15072568.
16. W. Elouataoui y Y. Gahi, «Empirical Evaluation of Big Data Stacks: Performance and Design Analysis of Hadoop, Modern, and Cloud Architectures», *Big Data Cogn. Comput.*, vol. 10, n.o 1, p. 7, dic. 2025, doi: 10.3390/bdcc10010007.
17. T. Nguyen, «Data quality management in big data: Strategies, tools, and educational implications», *J. Parallel Distrib. Comput.*, 2025.
18. S. Y. Muratov y S. B. Muravyov, «Framework architecture of a secure big data lake», *Procedia Comput. Sci.*, vol. 229, pp. 39-46, 2023, doi: 10.1016/j.procs.2023.12.005.
19. F. Zhao, «Design and Implementation of Data Quality Optimization Strategy and Multi-Level Data Governance Platform based on Big Data Processing and Dynamic Rule Extension», *Procedia Comput. Sci.*, vol. 262, pp. 638-647, 2025, doi: 10.1016/j.procs.2025.05.095.
20. J. Peral, A. Ferrandez, D. Gil, R. Muñoz-Terol, y H. Mora, «An Ontology-Oriented Architecture for Dealing With Heterogeneous Data Applied to Telemedicine Systems», *IEEE Access*, vol. 6, pp. 41118-41138, 2018, doi: 10.1109/ACCESS.2018.2857499.
21. A. P. Thenata, «Data Pipeline Architecture with Near Real-Time Streaming Multiple Source Indonesian Online News Data Lake», *JISAJurnal Inform. Dan Sains*, vol. 3, n.o 1, jun. 2020, doi: 10.31326/jisa.v3i1.657.
22. A. Alwi, N. A. Sasongko, Suprpto, Y. Suryana, y H. Subagyo, «Blockchain and big data integration design for traceability and carbon footprint management in the fishery supply chain», *Egypt. Inform. J.*, vol. 26, p. 100481, jun. 2024, doi: 10.1016/j.eij.2024.100481.
23. W. Khan, T. Kumar, C. Zhang, K. Raj, A. M. Roy, y B. Luo, «SQL and NoSQL Database Software Architecture Performance Analysis and Assessments—A Systematic Literature Review», *Big Data Cogn. Comput.*, vol. 7, n.o 2, p. 97, may 2023, doi: 10.3390/bdcc7020097.
24. J. Dončević, K. Fertalj, M. Brčić, y A. Krajna, «Mask-Mediator-Wrapper: A Revised Mediator-Wrapper Architecture for Heterogeneous Data Source Integration», *Appl. Sci.*, vol. 13, n.o 4, p. 2471, feb. 2023, doi: 10.3390/app13042471.
25. D. Fawzy, S. M. Moussa, y N. L. Badr, «The Internet of Things and Architectures of Big Data Analytics: Challenges of Intersection at Different Domains».
26. M. J. Anwar, A. Q. Gill, F. K. Hussain, y M. Imran, «Secure big data ecosystem architecture: challenges and solutions», *EURASIP J. Wirel. Commun. Netw.*, vol. 2021, n.o 1, p. 130, dic. 2021, doi: 10.1186/s13638-021-01996-2.
27. X. Xin, X. Xia, L. Zhao, Z. Zhu, y X. Cui, «A Verifiable Federated Learning Mechanism for Digital Museum Collection Based on Homomorphic Proxy Re-Authentication Technology».
28. V. Saengthong, T. Mantharngkul, C. Wongsuwan, K. Ritthaisong, y S. Fugkaew,

29. «DHT-Backed Ancestor-Assisted Merkle Verification for Scalable and Real-Time Log Integrity in Cloud Data Lakes».
30. M. Maynou, S. Nadal, R. Panadero, J. Flores, O. Romero, y A. Queralt, «Freyja: Efficient Join Discovery in Data Lakes», *IEEE Trans. Knowl. Data Eng.*, vol. 38, n.o 4, pp. 2277-2288, abr. 2026, doi: 10.1109/TKDE.2026.3656786.
31. M. U. Demirezen y T. S. Navruz, «Performance Analysis of Lambda Architecture-Based Big-Data Systems on Air/Ground Surveillance Application with ADS-B Data», *Sensors*, vol. 23, n.o 17, p. 7580, ago. 2023, doi: 10.3390/s23177580.
32. A. Tembo Welo, H. Lubaki Kinzonzi, N. Bila Khonde, y E. Mbuyi Mukendi, «Big Data Architectures and Concepts», *J. Innov. Inf. Technol. Appl. JINITA*, vol. 5, n.o 2, pp. 97-104, dic. 2023, doi: 10.35970/jinita.v5i2.1876.
33. D. Martínez-Mosquera, R. Navarrete, y S. Luján-Mora, «Development and Evaluation of a Big Data Framework for Performance Management in Mobile Networks», *IEEE Access*, vol. 8, pp. 226380-226396, 2020, doi: 10.1109/ACCESS.2020.3045175.
34. B. Zhang, L. Yu, Y. Feng, L. Liu, y S. Zhao, «Application of Workflow Technology for Big Data Analysis Service», *Appl. Sci.*, vol. 8, n.o 4, p. 591, abr. 2018, doi: 10.3390/app8040591.
35. A. Bousdekis y G. Mentzas, «Enterprise Integration and Interoperability for Big Data-Driven Processes in the Frame of Industry 4.0», *Front. Big Data*, vol. 4, p. 644651, jun. 2021, doi: 10.3389/fdata.2021.644651.
36. S. Huang, Z. Xie, J. Wang, P. Lv, y W. Wang, «Design and implementation of big data processing system based on Hadoop», *Procedia Comput. Sci.*, vol. 259, pp. 1115-1122, 2025, doi: 10.1016/j.procs.2025.04.065.
37. Z. Yang et al., «BDIDA-IoT: A Blockchain-Based Decentralized Identity Architecture Enhances the Efficiency of IoT Data Flow», *Appl. Sci.*, vol. 14, n.o 5, p. 1807, feb. 2024, doi: 10.3390/app14051807.
38. D. M. Berbes Villalón, L. Sánchez Jiménez, M. De La Iglesia Campos, M. E. Díaz Aguirre, y T. Delgado Fernández, «An IoT architecture for smart cities based on the FIWARE platform», *Rev. Cienc. Tecnol.*, n.o 38, pp. 20-27, oct. 2022, doi: 10.36995/j.recyt.2022.38.003.
39. T. Gouasmi, S. Laswed, y A. H. Kacem, «Secure and Efficient Big Data Collection with Differential Confidentiality and Machine Learning», *Procedia Comput. Sci.*, vol. 270, pp. 3860-3869, 2025, doi: 10.1016/j.procs.2025.09.511.
40. H. Zhu, «E-Commerce Big Data Logistics Supply Chain Control System Based on Artificial Intelligence and Blockchain», *Procedia Comput. Sci.*, vol. 259, pp. 923-930, 2025, doi: 10.1016/j.procs.2025.04.045.
41. J. L. Moigne et al., «“What Now/What Next/What If” Or NASA Earth System Digital Twins», *IEEE Geosci. Remote Sens. Mag.*, vol. 14, n.o 1, pp. 362-382, feb. 2026, doi: 10.1109/MGRS.2025.3609241.
42. P. Pääkkönen y D. Pakkala, «Extending reference architecture of big data systems towards machine learning in edge computing environments», *J. Big Data*, vol. 7, n.o 1, p. 25, dic. 2020, doi: 10.1186/s40537-020-00303-y.
43. R. Miñón, J. Díaz-de-Arcaya, A. I. Torre-Bastida, y P. Hartlieb, «Pangea: An MLOps Tool for Automatically Generating Infrastructure and Deploying Analytic Pipelines in Edge, Fog and Cloud Layers», *Sensors*, vol. 22, n.o 12, p. 4425, jun. 2022, doi: 10.3390/s22124425.
44. M. Yang, S. Qian, y X. Wu, «Real-time fire and smoke detection with transfer learning based on cloud-edge collaborative architecture», *IET Image Process.*, vol. 18, n.o 12, pp. 3716-3728, oct. 2024, doi: 10.1049/ipr2.13187.

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Deivid Sebastian Armero Gualguan.
Writing–review and editing: Deivid Sebastian Armero Gualguan.