

Security-by-Design Distributed Big Data Architectures: A Bibliometric and Conceptual Analysis of Governance, Privacy, and Integrity Layers (2015–2026)

Abubakari Saddiq Adams

abubakrsiddiq10@gmail.com

Korle-Bu Teaching Hospital

Research Article

Keywords: Big data security, Security-by-Design, distributed systems, differential privacy, blockchain integrity, federated learning, data governance, bibliometric analysis

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10259282/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

The exponential growth of distributed big data systems has significantly intensified challenges related to security, privacy, and governance. While modern frameworks such as Apache Spark and Apache Flink enable scalable and efficient data processing, security mechanisms are frequently implemented as post-hoc add-ons rather than being integrated into core system design—a critical architectural deficiency that exposes organizations to systemic vulnerabilities. This paper presents a comprehensive bibliometric and conceptual analysis of 210 peer-reviewed works on secure distributed big data architectures published between 2015 and 2026, drawn from the Scopus database. Employing the Bibliometrix toolkit in R, our analysis maps the evolution of security-centric research through performance analysis, science mapping, and thematic clustering. The dataset reveals a high annual growth rate of 23.28%, a peak of 83 publications in 2021, and a fragmented research landscape where privacy-preserving techniques, blockchain-based integrity mechanisms, and data governance models are developed largely in isolation. To address this gap, we propose the Security-by-Design Distributed Big Data Architecture (SDBDA)—a six-layer framework that embeds governance, access control, integrity, privacy, processing, and secure data management directly into the architectural core. Grounded in empirical bibliometric findings and validated through conceptual analysis, the SDBDA provides a scalable, policy-aware foundation that balances analytical utility with the data minimization imperative, mitigating the “Panopticon effect” inherent in modern data collection systems.

I. Introduction

The rapid proliferation of big data technologies has fundamentally transformed the operational landscape of modern computing. Distributed architectures, characterized by the convergence of cloud computing, edge/fog nodes, and Internet of Things (IoT) ecosystems, now underpin critical infrastructure across healthcare, finance, smart cities, and national development programs [1], [2]. Foundational processing frameworks including Apache Hadoop, Apache Spark, and Apache Flink, have substantially advanced computational efficiency in data-intensive environments, enabling real-time analytics at petabyte scale. However, this rapid technical advancement has been accompanied by a persistent and consequential architectural deficiency: security is routinely treated as a peripheral concern, appended to systems after deployment rather than embedded into their foundational design [3].

This design paradigm yields systemic vulnerabilities. Traditional big data architectures were engineered with performance and horizontal scalability as primary objectives, frequently at the expense of security and privacy guarantees. The resulting systems expose sensitive data to unauthorized access, privilege escalation, and integrity attacks across distributed storage and computation layers. Contemporary developments in telecommunication data monetization [4] and smart infrastructure deployment further compound this challenge, as the velocity of innovation consistently outpaces the maturity of global privacy regulatory frameworks. The “Panopticon effect” wherein data collection becomes pervasive, asymmetric, and difficult to govern has emerged as a defining concern of the modern data economy.

A review of the extant literature reveals three principal research trajectories that have developed largely in isolation:

1. privacy-preserving computation, encompassing differential privacy and secure multi-party computation.
2. (ii) blockchain-based data integrity and provenance assurance;
3. (iii) access control and governance frameworks. This fragmentation impedes the development of holistic, production-ready security architectures.

Crucially, data governance encompassing policy enforcement, regulatory compliance, and data sovereignty remains largely absent from core architectural designs, particularly in developing-economy contexts where national data protection regimes are nascent [5].

This study addresses these gaps through two primary contributions. First, it presents a systematic bibliometric analysis of 210 Scopus-indexed works (2015–2026) that maps the intellectual structure, thematic evolution, and collaborative dynamics of security-focused distributed big data research. Second, it proposes the Security-by-Design Distributed Big Data Architecture (SDBDA), a six-layer framework that embeds security, privacy, and governance as foundational system primitives rather than optional extensions. The SDBDA is grounded in the empirical findings of the bibliometric analysis and draws upon the most impactful mechanisms identified therein.

The remainder of this paper is organized as follows. Section II reviews related work across three thematic domains. Section III details the bibliometric methodology. Section IV presents and analyzes the results. Section V introduces the proposed SDBDA framework. Section VI discusses implementation considerations. Section VII concludes and outlines future research directions.

II. Related Work

Research on distributed big data security has evolved substantially since 2015, transitioning from a peripheral concern to a central design imperative. This section surveys three interrelated research streams that form the empirical and conceptual foundation for the proposed SDBDA framework.

A. Privacy-Preserving Computation in Distributed Systems

Privacy-preserving techniques constitute the most active research domain in the corpus, with 45 documents indexed under the “big data” keyword and 22 under “federated learning.” The theoretical foundation of this domain rests on differential privacy (DP), which provides a rigorous mathematical framework for limiting individual record disclosure under statistical query models. Practical deployment of DP in federated recommendation systems has been advanced by mechanisms such as Adap DP-FR, which jointly optimizes model performance and privacy through adaptive sensitivity calibration [6]. The privacy-utility trade-off (PUT) inherent in DP mechanisms remains a central engineering challenge, with empirical studies reporting accuracy reductions of 1% to 2% under standard privacy budgets.

Federated learning (FL) has emerged as a dominant paradigm for privacy-preserving distributed analytics, enabling model training across decentralized data silos without raw data transmission. Qu et al. [1] the most cited work in the dataset with 308 citations demonstrated the viability of blockchained federated learning for cognitive computing in Industry 4.0 networks, establishing a foundational integration of FL and blockchain. This was extended by Chai et al. [7] (296 citations), who proposed a hierarchical blockchain-enabled FL algorithm for knowledge sharing in Internet of Vehicles (IoV) environments, demonstrating scalability under dynamic network topologies. For healthcare applications, Haripriya et al. [8] introduced a privacy-enhanced collaborative big data analysis framework combining FL with homomorphic encryption, while Ghayvat et al. [9] (126 citations) proposed CP-BDHCA, a blockchain-based confidentiality-preserving scheme specifically designed for healthcare clouds and electronic health record (EHR) management.

Secure Multi-Party Computation (MPC) provides an alternative paradigm for privacy-preserving analytics, enabling joint computation over private datasets without trusted third parties. Wu et al. [10] constructed a maliciously secure two-party mixed-protocol framework for data-driven computation tasks, utilizing committed oblivious linear evaluation (C-OLE) based on homomorphic commitments to provide security guarantees in the malicious model, a significantly stronger adversarial assumption than the semi-honest model prevalent in earlier MPC frameworks. For distributed storage and search, Liu et al. [11] (87 citations) proposed a privacy-preserving multi-keyword searchable encryption scheme optimized for distributed systems, addressing the dual challenge of privacy and search efficiency in cloud-hosted data environments.

Edge-cloud collaborative frameworks represent a third avenue for privacy-preserving distributed analytics, offloading sensitive computations to network edge nodes to reduce data transport risk. Firouzi et al. [12] (267 citations) provided a comprehensive analysis of edge-fog-cloud convergence dynamics in AI-driven IoT environments. The FogChain architecture [13] integrates fog computing with blockchain, delegating analytics to the network edge while maintaining a tamper-resistant audit trail. Privacy-Assured FogCS [14] extends this paradigm through chaotic compressive sensing for secure industrial big image data, demonstrating the applicability of fog-integrated architectures in industrial IoT settings.

B. Blockchain-Based Integrity and Trust Mechanisms

Blockchain technology is extensively explored as a mechanism for ensuring data integrity, provenance, and auditability in distributed systems. The dataset contains 29 documents indexed under “blockchain,” underscoring the centrality of this technology to the field. Three sub-domains have emerged: lightweight blockchain platforms for resource-constrained environments, hardware-assisted integrity schemes, and blockchain-enabled collaborative data sharing.

Lightweight blockchain platforms address the performance constraints of deploying distributed ledgers on IoT devices and edge nodes. The Tikiri platform [15] (77 citations) employs Apache Kafka as a high-throughput messaging layer to achieve transaction rates compatible with IoT workloads, while the Rahasak architecture [16] (59 citations) provides a scalable enterprise blockchain framework with

configurable consensus mechanisms. The convergence of IoT and Distributed Ledger Technologies (DLT) has been comprehensively surveyed by Farahani et al. [17] (194 citations), who identify interoperability, scalability, and regulatory alignment as the principal open challenges.

Hardware-assisted integrity mechanisms leverage Physical Unclonable Functions (PUFs) inherent manufacturing variations in semiconductor devices to bind data integrity guarantees to the physical substrate. PUFchain 3.0 [18] integrates PUF-based device authentication with a distributed ledger for robust identity assurance in healthcare cyber-physical systems, providing cryptographic binding between physical hardware identity and data provenance records. This approach is particularly relevant for supply chain applications and medical device ecosystems, where hardware-level trust anchors are essential.

Blockchain-enabled collaborative data sharing leverages decentralized ledger architectures to enable privacy-preserving data exchange across organizational boundaries. BeepTrace [19] (146 citations) demonstrated a blockchain-enabled privacy-preserving contact tracing system during the COVID-19 pandemic, balancing epidemiological utility with individual privacy rights. The CP-BDHCA scheme [9] further demonstrates blockchain's applicability to healthcare data marketplaces, enabling fine-grained access control over EHR data with cryptographic verifiability. Cha et al. [20] (102 citations) proposed a blockchain-empowered cloud architecture based on secret sharing for smart city applications, providing both confidentiality and integrity guarantees for heterogeneous urban data streams.

C. Security-by-Design and Governance Frameworks

The "Security-by-Design" principle mandating that security properties be specified, verified, and enforced at the architectural design stage rather than retrofitted post-deployment has gained significant traction since the promulgation of GDPR in 2018. The most directly relevant work in the corpus, by Awaysheh et al. [3] (113 citations), provides a comprehensive Security-by-Design framework for big data deployments over cloud computing, introducing a layered security taxonomy that distinguishes infrastructure, platform, and application security concerns. Their analysis of Apache Hadoop, Spark, and related frameworks reveals fundamental gaps in the native security primitives of each platform, motivating external governance overlays.

Privacy-by-Design, a precursor principle articulated by Cavoukian [21], has been operationalized for big data environments by several researchers. Salas and Domingo-Ferrer [22] (38 citations) analyzed fundamental anonymization techniques and their limitations under big data conditions, demonstrating that classical k-anonymity and l-diversity models are insufficient when combined with auxiliary information and re-identification attacks. Rajamäki and Simola [23] examined the application of Privacy-by-Design in open-source intelligence (OSINT) and big data analytics, highlighting the tension between investigative utility and privacy rights.

GDPR compliance frameworks for IoT and distributed systems have been proposed by Alamri et al. [24] (47 citations), who introduced a GDPR-compliant framework for IoT-based personal health records using blockchain, and by Rhahla et al. [25], who formulated actionable guidelines for GDPR compliance in big

data systems. The category-based access control framework proposed by Obrezkov et al. [28] provides a privacy-aware authorization model aligned with GDPR's purpose limitation and data minimization principles.

A notable gap identified across all three research streams is the absence of governance-aware architectural frameworks that integrate policy enforcement, privacy protection, and integrity assurance into a unified, deployable system design. Moreover, while the corpus contains several regionally specific applications in China, Australia, and Europe, governance frameworks tailored to Sub-Saharan African digital ecosystems, where data sovereignty concerns are acute and regulatory infrastructure is nascent, are conspicuously absent, representing an opportunity for future research.

III. Methodology

This study adopts a bibliometric research design to systematically evaluate the intellectual evolution, thematic structure, and collaborative patterns of security-focused distributed big data research. Bibliometrics provides a quantitative and reproducible methodology for analyzing large scholarly corpora, enabling the identification of emergent research fronts, highly influential works, and structural gaps in a field's knowledge base.

A. Data Source and Retrieval Strategy

Bibliographic data were retrieved from the Scopus database, selected for its comprehensive coverage of peer-reviewed engineering and computer science literature and its structured metadata format. The search was conducted on May 3, 2026, employing the following structured query targeting the intersection of distributed systems, big data processing, and security-by-design principles:

TITLE-ABS-KEY ("big data" AND "distributed" AND ("security" OR "privacy" OR "blockchain" OR "federated learning") AND ("architecture" OR "framework" OR "design"))

The search was restricted to English-language articles and reviews published between January 2015 and May 2026. After automated de-duplication and manual screening for topical relevance, the final dataset comprised 210 documents from 144 unique publication sources, exported in BibTeX format for computational processing.

B. Analytical Procedures

The raw bibliographic data were processed in the R statistical computing environment using the Bibliometrix package (v4.x), a validated toolchain for quantitative science mapping. The analytical workflow comprised two complementary phases.

Performance analysis quantified the productivity and impact of the research field through standard bibliometric indicators, including annual publication counts, citation distributions, h-index computations

for authors and venues, and country-level productivity metrics. Fractionalized authorship counting was employed to normalize contributions across multi-author collaborative works.

Science mapping techniques were applied to reveal the conceptual and intellectual structure of the field. Keyword co-occurrence analysis was performed on author-assigned keywords to identify thematic clusters, with co-occurrence networks constructed using cosine similarity normalization to mitigate the effect of highly frequent terms. Bibliographic coupling was applied at the document level to identify research fronts, and co-citation analysis was employed to map the intellectual foundations of the field. Thematic mapping employed the strategic diagram framework, positioning research themes by density and centrality to identify motor themes, basic themes, niche themes, and emerging/declining themes.

IV. Results and Analysis

A. Bibliometric Performance Overview

The dataset comprises 210 documents from 144 unique source venues, reflecting a broad and interdisciplinary research landscape. Table I summarizes the key bibliometric performance indicators derived from the Scopus export.

TABLE I

Bibliometric Performance Summary of the Scopus Dataset (2015–2026)

Metric	Value
Total Documents	210
Unique Sources (Journals/Venues)	144
Annual Growth Rate (CAGR)	23.28%
Peak Publication Year	2021 (83 articles)
Average Document Age	4.74 years
Average Citations per Document	21.93
Most Productive Country	China (59 articles)
Highest Impact Country	Australia (avg. 87.00 citations/article)
Most Productive Author	Zhang Y. (9 articles)
Most Cited Document	Qu et al. [1] – 308 citations

The field exhibits a high compound annual growth rate (CAGR) of 23.28%, indicating rapid and sustained expansion of research interest. Scientific production surged markedly in 2021, with 83 publications representing 39.5% of the total corpus in a single year. This concentration reflects the convergence of

three accelerating trends: widespread enterprise adoption of Hadoop/Spark architectures, the maturation of federated learning as a privacy-preserving paradigm, and heightened regulatory attention following landmark GDPR enforcement actions and the COVID-19 pandemic's acceleration of digital infrastructure deployment. The average document age of 4.74 years and average citation count of 21.93 per document indicate a field with a recent, high-impact body of work. Notably, publications in 2025 and 2026 (36 combined) signal continued momentum, with emerging themes in IoT privacy regulation [4] and Secure-by-Design approaches for cloud-native analytics platforms [5].

B. Geographical and Author Productivity Analysis

China leads in absolute publication volume with 59 articles, reflecting substantial national investment in data infrastructure and security research, concentrated at institutions including Harbin Institute of Technology, Beijing Institute of Technology, and Shandong Normal University. The United States (24 articles) and India (24 articles) represent co-equal second-tier contributors, followed by Italy (10), Australia (9), Germany (8), and Canada (7).

Research impact, measured by average citations per article, reveals a markedly different hierarchy. Australia demonstrates the highest research impact at 87.00 average citations per article, driven by high-impact contributions from researchers at the University of Wollongong including co-authors of the most cited works in the dataset. This citation asymmetry between volume leaders (China) and impact leaders (Australia) suggests distinct research cultures: China's large-volume output encompasses a broad range of system implementations, while Australia's smaller, more targeted output achieves higher international influence through foundational theoretical contributions.

At the author level, Zhang Y. leads in total article count (9), consistent with the broader observation of high Chinese author productivity. Na Na leads in fractionalized contributions (score: 7.00), reflecting consistent multi-author collaborative output. The most cited individual work is by Qu et al. [1], published in IEEE Transactions on Industrial Informatics (2021) with 308 citations, followed closely by Chai et al. [7] in IEEE Transactions on Intelligent Transportation Systems (296 citations) and Firouzi et al. [12] in Information Systems (267 citations).

C. Science Mapping and Thematic Cluster Analysis

Keyword co-occurrence analysis of 210 documents identified four primary thematic clusters, representing the dominant conceptual axes of the field. These clusters are described below in order of keyword frequency.

Cluster 1 Foundational Big Data Technologies (45 occurrences, "big data"): This cluster encompasses the core processing and storage frameworks (Hadoop, Spark, Kafka) that constitute the substrate of distributed big data systems. Works in this cluster are primarily concerned with performance optimization, fault tolerance, and horizontal scalability, with security treated as a secondary concern. The

prevalence of this cluster as the most frequent keyword underscores that security research must engage directly with the constraints and primitives of these established frameworks.

Cluster 2 Blockchain and Integrity (29 occurrences, “blockchain”): This cluster encompasses blockchain-based integrity, provenance, and trust mechanisms, including lightweight IoT platforms (Tikiri [15], Rahasak [16]), enterprise blockchain architectures, and DLT-IoT convergence frameworks [17]. The cluster’s centrality in the co-occurrence network reflects blockchain’s dual role as both an integrity primitive and a governance enabler.

Cluster 3 Federated Learning and Privacy (22 occurrences, “federated learning”; 19 occurrences, “privacy”): This cluster encompasses privacy-preserving machine learning, federated analytics, and differential privacy mechanisms. The tight co-occurrence between “federated learning” and “differential privacy” keywords reflects the prevalent integration of these two paradigms in the literature, as demonstrated by Qu et al. [1], Zhou et al. [26], and Adap DP-FR [6].

Cluster 4 Intelligent and Edge Security (19 occurrences, “cloud computing”; 14 occurrences, “security”; 13 occurrences, “internet of things”): This cluster spans edge computing security frameworks [27], IoT intrusion detection systems [28], and AI-driven security analytics for smart city and smart grid applications [13], [29]. The cluster reflects the expanding attack surface of distributed systems at the network edge, where resource constraints preclude conventional cryptographic approaches.

Thematic mapping analysis, positioning clusters by centrality (relevance to the field) and density (internal development), identifies Cluster 2 (blockchain integrity) and Cluster 3 (FL/privacy) as motor themes, well-developed and central to the field. Cluster 4 (intelligent edge security) presents as an emerging theme with high centrality but lower density, indicating rapid growth from a nascent base. Notably, data governance and regulatory compliance themes (GDPR, data sovereignty) appear as niche themes: internally developed but weakly connected to the dominant research clusters, confirming the governance gap motivating this work.

V. Proposed Security-by-Design Distributed Big Data Architecture (SDBDA)

A. Framework Motivation and Design Principles

The bibliometric analysis reveals a fundamental structural gap: the four thematic clusters identified in the corpus operate in relative isolation, with governance and compliance themes weakly connected to technical security research. Existing frameworks either address privacy-preserving computation without governance enforcement [1], [6], [7], or provide governance taxonomies without technical implementation guidance [3], [21]. The SDBDA is designed to bridge this gap through a vertically integrated, six-layer architecture that embeds security, privacy, and governance as foundational system primitives.

The framework is governed by five core design principles derived from the empirical findings of the bibliometric analysis. The principle of Security-by-Design mandates that each architectural layer exposes a formal security interface, preventing the ad hoc addition of security mechanisms post-deployment. The Governance-First principle requires that policy enforcement and regulatory compliance constraints be evaluated at the governance layer before any data operation is executed at lower layers. The Privacy Minimization principle, operationalizing the GDPR data minimization requirement and the AI-by-Corporate-Design governance model [30], mandates that only the minimum data necessary for a declared analytical purpose be admitted to the processing pipeline. The Verifiable Integrity principle requires that all data transformations be recorded in a tamper-resistant, cryptographically verifiable audit log. Finally, the Adaptive Utility principle acknowledges the privacy-utility trade-off and mandates that privacy mechanisms (e.g., differential privacy noise calibration) be dynamically adjusted to maintain analytical utility within declared privacy budgets.

B. Layer Specifications

The SDBDA comprises six vertically integrated layers, each with defined security responsibilities, interfaces, and implementation mechanisms. Table II provides a consolidated specification of all six layers.

TABLE II

SDBDA Layer Specification: Functions, Mechanisms, and Key References

Layer	Primary Function	Specific Mechanism	Key Reference
Governance	Policy enforcement & data sovereignty compliance	AI-by-Corporate-Design protocols; GDPR/data minimization enforcement	[26], [27]
Access Control	Identity management & authorization	Zero Trust Architecture; Category-Based Access Control (CBAC)	[28], [29]
Integrity	Data authenticity & provenance assurance	PUFchain 3.0; blockchain-based immutable audit logging	[11], [30]
Privacy	Sensitive data protection & anonymization	Adaptive Differential Privacy (Adap DP-FR); FPE for data at rest	[6], [31]
Processing	Secure distributed computation	Federated Synergy Learning (FSyL); MPC mixed-protocol framework	[32], [8]
Data Layer	Secure ingestion, storage & transport	Format-Preserving Encryption; Apache Kafka-based ingestion (Tikiri)	[9], [33]

The Governance Layer serves as the apex control plane of the SDBDA, responsible for policy specification, regulatory compliance enforcement, and data sovereignty management. This layer

implements the AI-by-Corporate-Design framework [30], which extends Cavoukian's Privacy-by-Design principles to encompass AI system governance, mandating that data processing purposes be declared, bounded, and auditable. The governance layer exposes a policy decision point (PDP) interface, evaluated before any data ingestion operation, and maintains a compliance registry mapping data flows to applicable regulatory frameworks (GDPR, CCPA, Ghana Data Protection Act). For smart metering and energy data applications, the governance layer incorporates multi-regime compliance rules as analyzed by Komnios [31].

The Access Control Layer implements a hybrid authorization model combining Zero Trust Architecture (ZTA) with Category-Based Access Control (CBAC). Zero Trust eliminates the implicit trust assumptions of perimeter-based models by requiring continuous authentication and authorization for every access request, regardless of network location [32]. CBAC extends role-based access control by introducing data sensitivity categories (e.g., personally identifiable, commercially sensitive, public) as a first-class authorization dimension, enabling purpose-limited access aligned with GDPR's purpose limitation principle [28]. The access control layer interfaces directly with the governance layer's PDP, ensuring that authorization decisions reflect current policy state.

The Integrity Layer provides cryptographic assurance of data authenticity and provenance throughout the processing pipeline. PUFchain 3.0 [18] is employed as the hardware-assisted integrity anchor, binding data provenance records to the physical identity of originating devices through PUF-based attestation. All data transformation events are recorded as immutable transactions on a permissioned blockchain (e.g., Hyperledger Fabric), providing a tamper-evident audit trail verifiable by authorized governance stakeholders. The integrity layer exposes a provenance query interface, enabling downstream consumers to verify the origin, transformation history, and custody chain of any dataset.

The Privacy Layer implements adaptive differential privacy through the Adap DP-FR mechanism [6], which dynamically calibrates noise injection to maintain a target privacy budget ϵ while minimizing accuracy loss. For data at rest, Format-Preserving Encryption (FPE) preserves the structural properties of sensitive fields (e.g., date formats, identifier structures) necessary for analytical processing, while providing cryptographic confidentiality [5]. The privacy layer also implements data subject rights management supporting right-to-erasure and right-to-access requests through integration with the governance layer's compliance registry.

The Processing Layer provides secure distributed computation through Federated Synergy Learning (FSyL) [33], an enhanced federated learning protocol that incorporates secure aggregation, model poisoning detection, and asynchronous update handling. For tasks requiring data fusion across organizational boundaries, the processing layer optionally employs the MPC mixed-protocol framework of Wu et al. [10], providing malicious-model security guarantees for joint computation over privately held datasets. The processing layer is designed to operate within the privacy budget enforced by the privacy layer, receiving calibrated noise parameters as a service dependency.

The Data Layer provides secure data ingestion, storage, and transport services. High-throughput ingestion is implemented through a Tikiri-compatible Apache Kafka pipeline [15], providing encrypted message transport and ordered delivery guarantees. Distributed storage employs encrypted data partitioning with cryptographic access tokens managed by the access control layer, ensuring that storage nodes cannot access plaintext data independently. Data transport between layers is secured through TLS 1.3 with mutual certificate authentication, preventing man-in-the-middle attacks across the distributed infrastructure.

C. Inter-Layer Security Dependencies

A defining characteristic of the SDBDA is the formal specification of security dependencies between adjacent layers. Unlike the ad hoc security additions characteristic of existing frameworks, each SDBDA layer consumes security guarantees from the layer below and provides security guarantees to the layer above, creating a vertically integrated trust chain. The governance layer's policy decisions constrain the access control layer's authorization logic; the access control layer's authorization tokens gate the integrity layer's write operations; the integrity layer's provenance records inform the privacy layer's sensitivity classifications; and the privacy layer's noise parameters constrain the processing layer's analytical operations. This dependency structure ensures that a security guarantee at any layer is transitively conditioned on the guarantees of all underlying layers, preventing partial deployment that leaves security gaps.

VI. Implementation Considerations

A. Privacy-Utility Trade-Off Management

Practical deployment of the SDBDA must navigate the Privacy-Utility Trade-off (PUT), the fundamental tension between the analytical accuracy of data processing operations and the strength of privacy guarantees provided. Differential privacy mechanisms introduce calibrated noise that degrades analytical accuracy; empirical evaluations of Adap DP-FR report a 1%–2% accuracy reduction under standard privacy budgets ($\epsilon \in [0.1, 1.0]$) compared to non-private baselines [6]. The SDBDA addresses the PUT through adaptive ϵ -budget management at the privacy layer, enabling operators to specify accuracy thresholds that are automatically translated into privacy budget allocations, subject to governance layer compliance constraints that enforce minimum privacy guarantees.

B. Blockchain Performance and Scalability

Blockchain-based integrity mechanisms introduce transaction latency and storage overhead that must be managed in high-throughput data environments. The Tikiri platform [15] achieves throughput of approximately 3,000 transactions per second using Apache Kafka as the messaging substrate, sufficient for most enterprise IoT deployments. For environments requiring higher throughput, the SDBDA integrity layer supports selective audit logging recording only governance-significant events (schema changes, access policy modifications, cross-boundary data transfers) rather than all individual data operations.

PUFchain 3.0's hardware-assisted attestation further reduces on-chain state by anchoring device identity verification in hardware, minimizing redundant re-authentication transactions [18].

C. Edge-Cloud Resource Allocation

Effective SDBDA deployment must leverage edge-cloud collaborative frameworks to balance computational load across the processing hierarchy. Computationally intensive cryptographic operations (homomorphic encryption, MPC protocols) are preferentially allocated to cloud infrastructure with dedicated cryptographic acceleration. Privacy-sensitive preprocessing operations (local differential privacy injection, data minimization filtering) are executed at edge nodes to minimize sensitive data transmission, consistent with the "compute to data" principle advocated by Firouzi et al. [12]. The governance layer's data sovereignty constraints additionally restrict cross-jurisdiction data flows, requiring that data subject to national data protection regulations (e.g., Ghana Data Protection Act) be processed within compliant jurisdictional boundaries.

D. Deployment Context: Emerging Digital Ecosystems

The SDBDA framework is designed with explicit consideration for deployment in emerging digital ecosystems, where data infrastructure maturity, regulatory capacity, and technical expertise may be constrained. Ghana's developing digital economy presents a paradigmatic deployment context, where rapid expansion of mobile financial services, health information systems, and smart infrastructure intersects with nascent data protection regulation. The governance layer's configurable compliance registry enables adaptation to jurisdiction-specific regulatory requirements, while the lightweight blockchain options (Tikiri, Rahasak) provide integrity guarantees compatible with resource-constrained network environments. Future empirical validation of the SDBDA will prioritize these deployment contexts, where the gap between data collection capabilities and security governance infrastructure is most acute.

VII. Conclusion and Future Work

This paper has presented a comprehensive bibliometric and conceptual analysis of 210 Scopus-indexed works on secure distributed big data architectures published between 2015 and 2026 and has proposed the Security-by-Design Distributed Big Data Architecture (SDBDA) as a response to the structural fragmentation identified therein. The bibliometric analysis reveals a field experiencing rapid growth (23.28% CAGR), dominated by four thematic clusters foundational big data technologies, blockchain integrity, federated learning and privacy, and intelligent edge security that operate with insufficient integration. Governance and regulatory compliance themes are identified as niche areas with high internal development but weak connectivity to dominant technical research clusters, representing the principal structural gap addressed by the SDBDA.

The SDBDA addresses this gap through a six-layer vertically integrated architecture that embeds governance, access control, integrity, privacy, processing, and secure data management as foundational

system primitives rather than post-hoc additions. The framework's formal inter-layer security dependencies create a vertically integrated trust chain, ensuring that security guarantees are compositional and that partial deployment cannot create exploitable security gaps. Implementation considerations address the privacy-utility trade-off, blockchain performance constraints, and edge-cloud resource allocation, with particular attention to deployment in emerging digital ecosystems where governance infrastructure is nascent.

Several directions are identified for future work. First, prototype implementation and empirical validation of the SDBDA in production-scale environments is required to validate the theoretical security guarantees under real-world adversarial conditions. Particular priority is accorded to deployments in the Ghanaian and broader West African digital ecosystem, where the framework's governance-first design addresses acute data sovereignty concerns. Second, formal security analysis of the SDBDA's inter-layer dependency structure using established security modeling frameworks (e.g., Attack-Defense Trees, STRIDE) would provide rigorous verification of the framework's compositional security properties. Third, the integration of AI-driven threat detection at the governance and access control layers leveraging the intelligent security models identified in Cluster 4 represents a promising extension that would enable adaptive, context-aware policy enforcement. Finally, longitudinal bibliometric monitoring of the field's evolution, with particular attention to the governance and compliance niche theme, will track whether the integration gaps identified in this analysis are progressively addressed by the research community.

Declarations

Ethics approval and consent to participate:

Not applicable. This study did not involve human participants, human data, or animal data; it is a bibliometric and conceptual analysis of publicly available, published scholarly literature.

Consent for publication:

Not applicable.

Competing interests:

The author declares that he has no competing interests.

Funding:

This research received no external funding from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

A.S.A. conceived the study, performed the literature search and bibliometric data extraction, conducted the analysis using the Bibliometrix toolkit in R, designed the proposed SDBDA framework, and wrote and reviewed the manuscript. A.S.A. is the sole author and takes full responsibility for the content of this work.

Acknowledgements:

Not applicable.

Use of generative AI:

Generative AI tools were not used to generate original research content, data, or conclusions in this manuscript.

Data Availability

The bibliometric dataset analyzed during this study, comprising 210 Scopus-indexed records retrieved on 3 May 2026 using the search strategy described in Section III.A, is available from the corresponding author upon reasonable request.

References

1. Qu Y, Pokhrel SR, Garg S, Gao L, Xiang Y. A Blockchain Federated Learning Framework for Cognitive Computing in Industry 4.0 Networks. *IEEE Trans Ind Informat.* 2021;17:2964–73. 10.1109/TII.2020.3007817.
2. Farahani B, Firouzi F, Luecking M. The convergence of IoT and distributed ledger technologies (DLT): Opportunities, challenges, and solutions. *J Netw Comput Appl.* 2021;177. 10.1016/j.jnca.2020.102936.
3. Awaysheh FM, Aladwan MN, Alazab M, Alawadi S, Cabaleiro JC, Pena TF. Security by Design for Big Data Frameworks Over Cloud Computing. *IEEE Trans Eng Manag.* 2022;69:3676–93. 10.1109/TEM.2020.3045661.
4. Cheng YW, Bilal A. Regulatory data privacy concerns in proliferating IoT era of telecommunication data monetization. *Computing*, 2025.
5. Lagos-Obando J, Aillapan G, Fenner-López J. A Secure-by-Design Approach to Big Data Analytics Using Databricks and Format-Preserving Encryption. *Appl Sci*, 2025.
6. Zheng G et al. Adap DP-FR: Adaptive Differential Privacy for Federated Recommendation. *Lect Notes Comput Sci*, 16305, 2026.

7. Chai H, Leng S, Chen Y, Zhang K. Blockchain-Enabled Federated Learning Algorithm for Knowledge Sharing in Internet of Vehicles. *IEEE Trans Intell Transp Syst.* 2021;22:3975–86. 10.1109/TITS.2020.3002712.
8. Haripriya R, Khare N, Pandey M, Bist P. A privacy-enhanced framework for collaborative Big Data analysis in healthcare using federated learning and homomorphic encryption. *J Big Data*, 12, 2025.
9. Ghayvat H, Pandya S, Bhattacharya P, Zuhair M, Rashid M, Hakak S, Chen W. Clouds and Applications. *IEEE J Biomed Health Informat.* 2022;26:1937–48. 10.1109/JBHI.2021.3097237. CP-BDHCA: Blockchain-Based Confidentiality-Privacy Preserving Big Data Scheme for Healthcare.
10. Wu Y, Wang X, Susilo W, Yang G, Jiang ZL, Wang H, Wu T. Efficient maliciously secure two-party mixed-protocol framework for data-driven computation tasks. *Comput Stand Interfaces.* 2022;80. 10.1016/j.csi.2021.103571.
11. Liu X, Yang G, Susilo W, Tonien J, Liu X, Shen J. Privacy-Preserving Multi-Keyword Searchable Encryption for Distributed Systems. *IEEE Trans Parallel Distrib Syst.* 2021;32:561–74. 10.1109/TPDS.2020.3027003.
12. Firouzi F, Farahani B, Marinšek A. The convergence and interplay of edge, fog, and cloud in the AI-driven Internet of Things (IoT). *Inf Syst.* 2022;107. 10.1016/j.is.2021.101840.
13. Mayer AH, Rodrigues VF, da Rosa Righi R, da Costa CA, Bez AR, Kong N. FogChain: A Fog Computing Architecture Integrating Blockchain and Internet of Things. *IEEE Internet Things J*, 8, 2021.
14. Zhang Y, Wang P, Huang H, Zhu J. Privacy-Assured FogCS: Chaotic Compressive Sensing for Secure Industrial Big Image Data in Fog Computing. *IEEE Trans Ind Informat*, 2021.
15. Bandara E, Tosh D, Foytik P, Shetty S, Ranasinghe N. and H. De Zoysa, Tikiri—Towards a lightweight blockchain for IoT. *Future Gener Comput Syst*, 119, 2021.
16. Bandara E, Liang X, Foytik P, Shetty S, Ranasinghe N, De Zoysa H. Rahasak—Scalable blockchain architecture for enterprise applications. *J Syst Archit*, 2021.
17. Farahani B, Firouzi F, Luecking M. The convergence of IoT and distributed ledger technologies (DLT): Opportunities, challenges, and solutions. *J Netw Comput Appl.* 2021;177. 10.1016/j.jnca.2020.102936.
18. Bathalapalli VKVV, Mohanty SP, Kougianos E, Panda BK. PUFchain 3.0: Hardware-Assisted Distributed Ledger for Robust Authentication in Healthcare Cyber-Physical Systems, *Sensors*, vol. 24, 2024.
19. Xu H, Zhang L, Onireti O, Fang Y, Buchanan WJ, Imran MA. BeepTrace: Blockchain-Enabled Privacy-Preserving Contact Tracing for COVID-19 Pandemic and Beyond. *IEEE Internet Things J.* 2021;8:3915–29. 10.1109/JIOT.2020.3025953.
20. Cha J, Singh SK, Kim TW, Park JH. Blockchain-empowered cloud architecture based on secret sharing for smart city. *J Inf Secur Appl.* 2021;57. 10.1016/j.jisa.2020.102686.
21. Cavoukian A, Chibba M. Start with Privacy by Design in All Big Data Applications. *Big Data Secur Priv Trust*, 2018.

22. Salas J, Domingo-Ferrer J. Some Basics on Privacy Techniques, Anonymization and their Big Data Challenges. *Math Comput Sci*, 2018.
23. Rajamäki J, Simola J. How to apply privacy by design in OSINT and big data analytics? *in Proc. ECCWS*, 2019.
24. Alamri B, Javed IT, Margaria T. A GDPR-Compliant Framework for IoT-Based Personal Health Records Using Blockchain, *in Proc. IEEE CCEM*, 2021.
25. Rhahla M, Allegue S, Abdellatif T. Guidelines for GDPR compliance in Big Data systems. *J Inf Secur Appl*, 2021.
26. Zhou P, Wang K, Guo L, Gong S, Zheng B. A Privacy-Preserving Distributed Contextual Federated Online Learning Framework with Big Data Support in Social Recommender Systems. *IEEE Trans Knowl Data Eng.* 2021;33:824–38. 10.1109/TKDE.2019.2936565.
27. Mehmood MY, Oad A, Abrar M, Munir HM, Hasan SF, Muqet HA. Edge Computing for IoT-Enabled Smart Grid, *Secur. Commun. Netw.*, vol. 2021, 2021. 10.1155/2021/5524025
28. Obrezkov D, Sohr K, Malaka R. A Category-Based Framework for Privacy-Aware Collaborative Access Control, *in Proc. IEEE Workshops*, 2021.
29. Chen J, Ramanathan L, Alazab M. Holistic big data integrated artificial intelligent modeling to improve privacy and security in data management of smart cities. *Microprocess Microsyst.* 2021;81. 10.1016/j.micpro.2020.103722.
30. Cicu GC. Bridging Traditional Corporate Governance and Technology: the 'AI Corporate Design' Framework. *Italian Law J*, 10, 2024.
31. Komnios K. Smart metering and multi-regime governance in EU energy law. *J World Energy Law Bus*, 19, 2026.
32. Preuveneers D, Joosen W. Security and privacy controls for streaming data in extended intelligent environments. *J Ambient Intell Smart Environ*, 8, 2016.
33. Fu S, Dong F, Shen D, Chen R et al. DESIGN: Online Device Selection and Edge Association for Federated Synergy Learning, *ACM Trans. Intell. Syst. Technol.*, 2024.