Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026252258

The Integration of Artificial Intelligence and Big Data: A Review of Technologies, Applications, and Challenges

Yifan Shi*

Faculty of Science, University of Ottawa, Ottawa, K1N 6N5, Canada

**Corresponding author: Yifan Shi.*

Abstract

Artificial intelligence (AI) and big data are now symbiotic pillars underpinning next-generation intelligent infrastructures. Big data serves as a rich, dynamic, and heterogeneous information reservoir—characterized by scale, variety, and velocity—while AI acts as the cognitive engine that extracts meaning from this deluge through self-directed learning, contextual pattern detection, and iterative decision refinement. Yet current scholarly work on their convergence tends to prioritize narrow technical enhancements or domain-specific deployments, falling short of offering a cohesive theoretical lens to articulate *how* these two paradigms co-evolve and synergize in practice. To bridge this conceptual and practical void, this study proposes an integrative analytical architecture grounded in three interlocking layers: (1) enabling technological foundations, (2) intelligent sense-making processes, and (3) context-sensitive deployment pathways. By unifying conceptual rigor with empirical grounding, this work advances a holistic, implementation-ready framework for building intelligent systems that are not only high-performing and scalable, but also ethically grounded, transparent, and attuned to real-world socio-technical contexts.

Keywords

artificial intelligence, big data, AI–big data integration

1. Introduction

Although the integration of artificial intelligence and big data has attracted extensive academic attention and industrial exploration, most existing studies remain relatively fragmented. Scholars generally focus on local model optimization or isolated application cases in single fields, while few form a systematic framework to elaborate the complete integration mechanism of the two technologies. Current research lacks in-depth discussion on interactive operation logic, practical implementation rules and realistic constraints affecting collaborative development.

To remedy this research gap, this paper adopts a cross-domain comprehensive analysis method to conduct overall research. It establishes a complete analytical framework covering basic integration principles, practical application modes and practical development barriers. The research sorts out the mutual promotion relationship between AI and big data, summarizes typical application effects in multiple fields, and analyzes key difficulties restricting large-scale popularization.

This work bridges theory and practice through two interrelated contributions. From a conceptual standpoint, it proposes a unified analytical framework for "data-driven intelligent analysis," delineating its underlying knowledge assumptions, functional mechanisms, and historical evolution across successive generations of AI and data infrastructure. From an applied perspective, it provides empirically substantiated, implementation-ready guidance to help organizations navigate the complexities of integrating AI with large-scale data systems—from strategic design and operational rollout to ongoing oversight and adaptation.

2. Big Data: Core Data Support for Artificial Intelligence Model Training

Big data serves as a foundational enabler for AI model training, which can be analyzed through three key aspects: volume, variety, and veracity of data.

From the perspective of data volume, deep learning–driven AI models consistently demonstrate pronounced data dependency. Empirical evidence indicates that limited training data substantially increases the risk of overfitting—where models inadvertently capture spurious correlations and dataset-specific noise rather than underlying patterns—thereby severely compromising their ability to generalize to new, unseen inputs. Big data's sheer volume furnishes rich, statistically meaningful samples, enabling neural networks to learn resilient, generalizable representations even within highly complex, high-dimensional parameter spaces. This principle is vividly illustrated in natural language processing: the transformative capabilities of GPT-family models stem largely from exposure to terabyte-scale textual corpora meticulously curated for relevance, coherence, and cleanliness. Moreover, data scale functions not merely as a quantity metric but as a critical determinant of reuse sustainability; as Yan et al. observe, "larger datasets can be repeated more times before the marginal benefit vanishes," [1] underscoring that scale intrinsically governs the longevity and utility ceiling of data assets in model development.

Concerning data diversity, real-world big data fundamentally consists of three major modal types: structured data (e.g., relational databases), semi-structured data (e.g., JSON or XML-formatted logs), and unstructured data (e.g., photographic images, time-synchronized video clips, raw audio signals, and human-language text). Such cross-modal variation compels AI systems to master unified representation learning across disparate input formats, while simultaneously generating more discriminative and interdependent supervision cues that accelerate convergence and improve out-of-distribution generalization. A quintessential use case is autonomous driving, where vehicles ingest precisely time-stamped sensory inputs from complementary sources—namely, high-resolution color camera feeds, dense 3D LiDAR scans, decimeter-level GNSS-aided inertial navigation paths, and calibrated millimeter-wave radar intensity profiles—facilitating holistic multimodal fusion that consistently elevates perception performance in terms of fault tolerance, measurement fidelity, and contextual responsiveness under dynamic real-world conditions. Zhang et al. point out that "data heterogeneity emerges from processing multimodal data that differ significantly in volume and processing cost, e.g., a high-resolution image is much larger and requires more resources for preprocessing than a line of text" [2], which may lead to delayed training progress. Their research highlights that directly tackling this variability is crucial for achieving efficient multimodal model training.

Regarding data quality, it is crucial to recognize that raw large-scale datasets collected from real-world environments usually undergo substantial preprocessing and rigorous validation prior to their application in training AI models. A wealth of empirical studies highlights widespread data quality issues—including incomplete or absent records, inconsistencies in structure across diverse data formats, and unreliable annotations arising from subjective judgments or inadequately standardized labeling protocols—all of which may significantly impair model performance, hinder training stability, and compromise practical deployment outcomes. In their systematic taxonomy of data quality defects for machine learning, Hagn et al. [3] point out that real-world datasets commonly exhibit various quality defects, which can significantly impair the performance and validity of ML models. They further demonstrate that inaccurate annotations and incomplete datasets are primary contributors to AI system failures—exemplified by the 2017 autonomous vehicle accident in Florida, in which a computer vision algorithm mistakenly classified a stationary truck trailer as part of the sky. This error was directly linked to insufficient diversity in training scenes and incorrect labeling within the training data. Consequently, robust data governance—including rigorous data cleaning, adaptive outlier identification, and consistent, transparent, and traceable annotation protocols—is essential for developing reliable AI systems. This principle is reinforced by real-world high-stakes applications: for instance, large-

scale automated utility billing platforms attain outstanding precision (error rates under 0.1%) only when underpinned by data that is not only highly accurate and temporally aligned but also rigorously validated for contextual relevance and integrity.

In summary, big data constitutes the essential input—commonly metaphorized as the "fuel"—for AI systems; yet, analogous to crude oil, it must be systematically processed, validated, and quality-controlled before it can effectively energize intelligent computation and deliver reliable model outcomes.

3. Artificial Intelligence: The Core Analytical Engine for Extracting Value from Big Data

If big data is analogous to fuel, then artificial intelligence functions as the engine that converts this fuel into actionable insight and operational power. The central role of artificial intelligence in unlocking the value of big data manifests in three interrelated capabilities: automated feature engineering, detection of complex, high-dimensional patterns, and robust optimization of predictive modeling and decision-making processes.

3.1 Automated Representation Learning

Traditional analytical methods rely heavily on domain experts to manually design features—such as statistical metrics or handcrafted data transformations—based on accumulated subject-matter knowledge. However, as datasets grow increasingly high-dimensional (often spanning thousands of features) and evolve from single-modality inputs to complex, multimodal data streams, this expert-dependent approach becomes increasingly limited in its flexibility, coverage, and consistency across applications. In contrast, deep learning—a core paradigm in artificial intelligence—overcomes this limitation by learning hierarchical, interpretable representations directly from raw input data, using end-to-end, gradient-driven optimization without requiring explicit feature engineering. As Shen et al. point out, "Excessive preference for local complementary features over cross-modal shared semantics adversely affects generalization performance" [4]. To address this limitation, their proposed MS2Fusion framework employs a dual-path parametric interaction mechanism based on state space models, enabling automatic cross-modal alignment. Experimental results demonstrate that this approach achieves state-of-the-art performance across multiple multispectral perception tasks without requiring task-specific modifications (Shen et al. [5]), providing empirical evidence that learning representations directly from unprocessed multimodal data—bypassing handcrafted rules and task-specific adaptations—diminishes reliance on domain expertise and enhances the model's capacity to generalize across varied modality pairings and previously unencountered situations.

3.2 Complex pattern recognition

Large-scale datasets often contain complex, nonlinear relationships that arise from high-dimensional feature spaces and evolving spatiotemporal interactions—patterns that traditional statistical methods like linear regression or principal component analysis (PCA) are inherently ill-suited to capture. By contrast, Graph Neural Networks (GNNs) specialize in learning representations from graph-structured data by propagating and refining information across node neighborhoods via learned aggregation functions. Unlike conventional neural architectures, GNNs explicitly encode relational inductive biases—capturing not only immediate adjacency relationships but also higher-order structural motifs and global graph-level invariants. As a result, they support effective knowledge distillation in domains where entities derive meaning and function from their networked associations: for instance, user interaction patterns in online platforms, logical entailments in ontological databases, and physicochemical properties encoded in atomic connectivity graphs. As Chen et al. state, "Graph Neural Networks (GNNs) have gained momentum in graph representation learning and boosted the state of the art in a variety of areas, such as data mining (e.g., social network analysis and recommender systems), computer vision (e.g., object detection and point cloud learning), and natural language processing (e.g., relation extraction and sequence learning)" [5]. Graph Neural Networks (GNNs) autonomously learn hierarchical, high-level features directly from large-scale graph-structured data, eliminating the need for handcrafted feature design. This represents a fundamental transformation in pattern analysis: whereas conventional statistical approaches struggle to model intricate, long-range interactions across nodes, GNNs harness AI-driven mechanisms to reveal nuanced structural and relational patterns—even among vertices that are topologically distant.

3.3 Predictive Decision-making Optimization

Reinforcement learning facilitates dynamic policy adaptation through trial-and-error interaction with real-world environments, whereas operations research delivers formal optimization methodologies for balancing competing objectives under hard constraints. As a result, AI systems operate as self-regulating decision entities—deriving adaptive, context-aware behaviors from longitudinal datasets and live telemetry. A case in point is AI-enhanced predictive maintenance in smart factories: by modeling temporal patterns in multi-sensor time-series data, it estimates remaining equipment lifespan and initiates maintenance workflows ahead of failure—thereby transforming maintenance from a reactive, schedule-based practice into a responsive, data-informed, and continuously refined operational discipline.

As Alginahi et al. state, "Reinforcement Learning(RL) as a transformative paradigm for adaptive control, intelligent optimization, and autonomous decision-making in smart factories" [6]. For example, within intelligent manufacturing ecosystems, AI-driven predictive maintenance frameworks analyze continuous, high-resolution sensor data from production machinery to estimate remaining useful life and associated uncertainty—automatically initiating targeted maintenance protocols prior to performance degradation or breakdown. This evolution—from corrective interventions to anticipatory asset management—embodies adaptive, feedback-guided decision-making: AI agents continuously update their operational strategies based on real-time system responses, enhancing sustained productivity and obviating the need for fixed, rule-based maintenance logic.

4. Typical application cases of the integration of AI and big data

4.1 Smart Government: LLM-Powered E-Government Architecture

George et al. [7] explored a modular architecture based on large language models (LLMs) for e-government services. Conventional e-government systems face significant challenges in information access efficiency, primarily stemming from data heterogeneity and intricately layered administrative service workflows. Based on retrieval-augmented generation (RAG) technology, the study proposed a multi-agent generative AI virtual assistant architecture using the Haystack framework, achieving scalability, transparency, and ethical compliance. Case validation showed significant improvement in public service efficiency, enabling natural language interaction for policy consultation and business guidance. This example illustrates how large language models are driving a transformation in e-government delivery—enabling more intelligent, self-executing service paradigms.

4.2 Smart City: Big Data-Assisted Government Hotline Document Classification

Zhang et al. [8] proposed a comprehensive system. Big Data-assisted Government Hotline classification System (BDCGHS) for government hotline business document classification. Government call centers handle vast volumes of daily textual records—rich in citizen sentiment and feedback—yet reliance on manual analysis proves inadequate for scaling to such high-throughput, unstructured data, ultimately compromising responsiveness and service effectiveness. The system integrates information entropy with TF-IDF weights to mine new words and constructs a nested balanced binary tree that accelerates new word mining to nearly five times the speed of traditional Trie trees. In two Chinese cities, BDCGHS has operated stably for over a year, achieving nearly 3% improvement in work order dispatch accuracy over the BERT model. This example highlights how artificial intelligence and large-scale data analytics contribute concretely to more effective, responsive, and evidence-based urban governance.

4.3 Comparative Analysis

Taken together, these two implementations illustrate that the architecture of AI–big data integration is determined less by industry sector and more by the intrinsic functional logic of the task. Despite both handling textual data in unstructured form, their technical implementations reflect fundamentally different priorities: the digital government service leverages verified, semantically rich policy texts and employs generative LLMs enhanced with external knowledge retrieval to deliver personalized, forward-looking guidance; conversely, the urban citizen hotline system manages large-scale, linguistically irregular inputs—marked by informal language, spelling variations, and fragmented expressions—using fine-tuned discriminative transformers,

supported by comprehensive data cleansing, standardization, and feature engineering workflows, to ensure timely, policy-aligned categorization and assignment of service requests. Ultimately, this architectural differentiation arises not from contextual conventions but from three core design determinants: the intended inference mode (generative synthesis versus discriminative judgment), the origin and integrity of input data (institutionally vetted versus crowd-sourced and uncurated), and the temporal orientation of the service logic (anticipatory engagement versus responsive intervention).

5. Current Challenges and Future Trends Based on Big Data and AI

5.1 Major Challenges

5.1.1 Computational Cost

Developing high-accuracy AI models demands substantial computational investment—particularly in specialized hardware like GPUs and TPUs—making it financially prohibitive for many small- and medium-sized enterprises as well as academic research groups. Wei et al. [9] point out that as model parameters continue to expand, training costs and carbon footprint issues are becoming increasingly prominent, necessitating the development of model compression and lightweighting technologies.

5.1.2 Data Quality and Privacy Security

Large-scale datasets frequently exhibit challenges such as incomplete entries, measurement noise, and skewed or inconsistent labeling—factors that undermine both the robustness and equitable performance of AI systems. Moreover, data throughout its lifecycle—from collection and storage to processing and sharing—often contains sensitive personal information and proprietary business intelligence, raising critical privacy and confidentiality concerns. Wei et al. [9] propose combining homomorphic encryption, Laplacian noise, and consortium blockchain consensus protocols to protect data privacy while effectively defending against poisoning attacks. However, their study also notes that federated learning still faces challenges in balancing communication efficiency, model performance, and security.

5.1.3 Model Interpretability

Deep neural networks are frequently characterized as opaque or “black-box” systems, owing to their complex, hierarchical architectures—which hinder transparent, interpretable explanations of their internal reasoning for stakeholders without technical expertise. Wei et al. [9] demonstrate that this limitation is particularly critical in high-stakes scenarios such as financial risk control and medical diagnosis, where even accurate predictions may struggle to gain regulatory compliance approval without interpretability.

5.2 Future Trends

The synergistic evolution of AI and big data is anticipated to unfold through three complementary pathways: On-Device Intelligence and Model Efficiency — transitioning inference capabilities from cloud servers to resource-constrained edge devices (e.g., smart sensors and embedded systems) via highly compressed, hardware-aware models to ensure real-time responsiveness, minimize network dependency, and conserve energy.

Ethically Grounded AI and Proactive Data Stewardship — embedding privacy safeguards, bias mitigation strategies, and transparent decision logic directly into the model development lifecycle, rather than as post-hoc add-ons; Foundation-Model-Driven Domain Generalization — exploiting the transferable semantic and structural knowledge of large-scale pretrained models to achieve high performance in niche domains using limited, domain-relevant data—thereby accelerating deployment, reducing annotation and computational overhead, and broadening AI accessibility across sectors.

6. Conclusion

It is essential to recognize that big data and artificial intelligence are not merely coexisting technologies but fundamentally interdependent enablers—each amplifying the capabilities of the other and jointly constituting the twin pillars underpinning modern intelligent systems. This research offers a comprehensive analysis of the

integration between artificial intelligence (AI) and big data, approached from three interrelated perspectives: (i) foundational technological drivers, (ii) real-world applications across distinct domains, and (iii) persistent challenges affecting operational viability. Big data serves as a critical input—supplying large-scale, heterogeneous datasets essential for training and refining AI models—while AI functions as an analytical engine, converting unstructured or high-volume data into meaningful insights, including real-time perception, anomaly detection, situational awareness, and evidence-informed decision support. Concrete applications in areas like digital governance and intelligent energy systems illustrate tangible improvements in administrative agility, citizen service quality, and overall infrastructure efficiency. Nevertheless, barriers such as infrastructure limitations in scaling AI solutions, tightening data protection requirements, and the "black-box" nature of complex AI algorithms hinder broad, ethically grounded deployment. To move beyond rudimentary automation toward resilient, user-centric, and environmentally adaptive AI systems, priority should be given to three mutually reinforcing strategies: lightweight, resource-aware model designs compatible with edge computing and diverse hardware platforms; AI frameworks explicitly engineered for traceability, equitable outcomes, and regulatory compliance; and modular, standards-based architectures that facilitate cross-sectoral knowledge transfer and continuous, context-sensitive learning.

References

[1] Yan, T., Wen, H., Li, B., Luo, K., Chen, W., & Lyu, K. (2026). Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression (Version v2). arXiv. https://doi.org/10.48550/arXiv.2511.13421

[2] Zhang, Z., Zhong, Y., Jiang, Y., Hu, H., Sun, J., Ge, Z., Zhu, Y., Jiang, D., & Jin, X. (2025). DistTrain: Addressing model and data heterogeneity with disaggregated training for multimodal large language models (Version v3). arXiv. https://arxiv.org/abs/2408.04275

[3] Hagn, M., Heinrich, B., Krapf, T., & Schiller, A. (2025). Handling imperfection: A taxonomy for machine learning on data with data quality defects. Decision Support Systems, 196, 114493. https://doi.org/10.1016/j.dss.2025.114493

[4] Shen, J., Zhan, H., Dong, S., Zuo, X., Yang, W., & Ling, H. (2026). Multispectral state-space feature fusion: Bridging shared and cross-parametric interactions for object detection. Information Fusion, 127, 103895. https://doi.org/10.1016/j.inffus.2025.103895

[5] Chen, C., Wu, Y., Dai, Q., Zhou, H. Y., Xu, M., Yang, S., Han, X., & Yu, Y. (2024). A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 10297–10318. https://doi.org/10.1109/TPAMI.2024.3445463

[6] Alginahi, Y. M., Sabri, O., & Said, W. (2025). Reinforcement learning for industrial automation: A comprehensive review of adaptive control and decision-making in smart factories. Machines, 13(12), 1140. https://doi.org/10.3390/machines13121140

[7] George Papageorgiou, Vangelis Sarlis, Manolis Maragoudakis & Christos Tjortjis. (2024). Enhancing E-Government Services through State-of-the-Art, Modular, and Reproducible Architecture over Large Language Models. Applied Sciences, 14 (18), 8259-8259. https://doi.org/10.3390/APP14188259.

[8] Zicheng Zhang, Anguo Li, Li Wang, Wei Cao & Jianlin Yang. (2024). Big data-assisted urban governance: A comprehensive system for business documents classification of the government hotline. Engineering Applications of Artificial Intelligence, 132, 107997-. https://doi.org/10.1016/J.ENGAPPAI.2024.107997.

[9] Wei, C., Yang, W. S., & Liu, W. (2025). A privacy-preserving federated learning framework based on consortium blockchain. Journal of Zhengzhou University (Natural Science Edition), 57(4), 23-29. https://zzdz.cbpt.cnki.net/portal/journal/portal/client/paper/01c66d864505742fda62d6ee024afd4e

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.