Available Online: 30 January 2026 DOI: 10.54254/2977-3903/2026.31591

Research status and application scenario expansion of unstructured big data analysis technology

Runhan Fan

College of Software, Xinjiang University, Urumqi, China

2053466702@qq.com

Abstract. In the context of the profound evolution of the big data era, unstructured big data has become the main body of data resources, and its value extraction capability is directly related to the intelligent transformation process of all industries. This paper focuses on the processing and analysis technologies of unstructured big data, systematically reviewing the research status and expansion directions of application scenarios in this field. Research findings indicate that while artificial intelligence and cloud computing technologies have driven a revolutionary leap in processing capabilities, this field still faces a core contradiction: a significant gap exists between the fragmented state of key technologies and the end-to-end solutions required by real business, compounded by the prevalent data quality shortcomings, which together constitude the main obstacles to value realization. Therefore, through the literature analysis method, this paper summarizes existing technologies and application practices, and proposes a four-layer integrated framework centered on scenario-driven and intelligent orchestration from the two dimensions of technical integration architecture and scenario integration. The objective is to provide theoretical references and practical guidance for building efficient and implementable intelligent solutions for unstructured data.

Keywords: big data, unstructured, technology integration, scenario-driven, value extraction

1. Introduction

Although the industry has reached a consensus that unstructured big data will be the core source of future value creation, the conversion of raw data into value insights remains constrained by application gaps [1]. Despite significant breakthroughs in technologies such as natural language processing and computer vision, these single-point breakthroughs often remain isolated when confronting complex real-world business scenarios. The fundamental contradiction lies in the absense of a systematic framework capable of organically integrating and intelligently orchestrating these fragmented technical capabilities to form end-to-end solutions that directly address business needs.

Rural finance audit optimization serves as an illustrative example. A typical business requirement involves the entire chain from the collection of unstructured data, multimodal storage, and intelligent analysis to structural visualization [2]. Although existing research has provided a wealth of single-point tools, it has not fully addressed a key question: how to dynamically select and assemble the most appropriate technical

components for specific business scenarios. This gap in the connection from technical points to solutions is precisely the core bottleneck that prevents technology from being transformed into real productivity.

This bottleneck is particularly prominent in specific business practices. The intelligent review of bills in supply chain finance is a typical case.. This scenario involves the extraction and cross-validation of structured information from bill images of multiple formats and sources. Existing research has provided Optical Character Recognition (OCR) technology based on deep learning for bill recognition, rule-based and machine learning-based review models for bill compliance checks, and graph neural network-based correlation analysis algorithms for anomaly detection [3]. However, these technologies are often developed and optimized independently. When faced with a review task involving various bill types such as Value-Added Tax (VAT) invoices, freight documents, and customs declarations, system designers are confronted with numerous practical choices: in the OCR stage, how to dynamically select or integrate the most appropriate recognition models for bills with varying clarity and formats to balance speed and accuracy; in the information verification stage, how to determine the calling sequence and conditions of rule engines and machine learning models to account for the historical credit status of different enterprises; when multiple bill information is interrelated, how to configure parameters of graph analysis algorithms to identify complex fraud patterns; more importantly, when the timeliness requirements of the review change, can the entire process be rapidly adjusted to prioritize the critical path? Existing research provides insufficient systematic guidance on these dynamic assembly and optimization decisions [4].

To address this gap, this study aims to bridge the divide between technical components and practical solutions. This paper shifts focus from optimizing individual technologies to exploring the construction of integrated, scenario-driven technical application solution. Specifically, this study aims to design and validate an end-to-end platform capable of intelligently perceiving business needs and dynamically orchestrating and scheduling underlying technical components. The goal is to realize a critical transformation from decentralized to collaborative and from experimental to production in the processing of unstructured big data. Ultimately, this study intends to provide a practical path for exploring the value of unstructured data.

2. Research status and core contradictions of unstructured big data analysis technology

The processing and analysis of unstructured data are currently in a critical stage characterized by tension and technical paradoxes. On one hand, along the technical chain from data preprocessing, intelligent storage to advanced analytical algorithms, various "breakthroughs at the point" continue to emerge, rendering the technical toolkit extremely prosperous. On the other hand, in core fields such as finance and healthcare, business practices generally encounter the dilemma of converting data into value. This study argues that the fundamental reason for this gap is not the absense of a single technical capability, but rather the lack of an effective systematic connection framework between highly fragmented technical supply and highly integrated business demands. Therefore, this section will first systematically review the current key technical status of unstructured data processing, and then deeply analyze the core contradiction between the technologies and end-to-end business scenarios. Based on this analysis, this paper ultimately proposes and elaborates on an integrated solution framework centered on scenario-driven and intelligent orchestration, thereby providing a new systematic perspective for exploring pathways to realize the value of unstructured data. This understanding constitutes the logical starting point and core issue of subsequent research.

2.1. Technical status: "point-like breakthrough" and "depth development" of single point capabilities

At the stages of data preprocessing and feature extraction, research focuses on transforming the unorganized raw data into analyzable and structured information. For text data, mainstream studies indicate that technical approaches have evolved from early rule-based methods, through statistical models, to the current phase dominated by deep learning and pre-trained models, excelling in tasks such as sentiment analysis and topic identification [3]. Meanwhile, for more general unstructured digital resources, metadata extraction is the first step in value unlocking. These two mainstream methods exhibit distinct applicability: rule-based methods demonstrate efficiency in format standardization scenarios, while statistical method are more effective in processing variable documents. Adopting a hybrid approach that integrates the advantages of both strategies to enhance accuracy and robustness has emerged as a key development trend in this field [4].

At the stage of data storage and management infrastructure, the transformation of the technical paradigm is particularly evident. Traditional relational databases encounter scalability bottlenecks when dealing with unstructured data due to their fixed table structure. By comparing NoSQL, the research clearly indicates that NoSQL technologies represented by document databases and graph databases [5], with their flexible data models, efficient horizontal scalability, and direct expression capabilities for complex relationships, have become the preferred technical approach for managing massive unstructured data. Notably, in processing graph-structured data, empirical evidence confirms that graph databases outperformed traditional databases in query efficiency and pattern scalability.

At the stage of core analysis algorithm and models, research is advancing towards higher accuracy and automation. In the high-dimensional and redundant challenges of unstructured big data classification, some studies have proposed an integrated solution that combines neural network cleaning, supervised dimension reduction, and improved decision tree algorithms [6]. Such research represents a typical direction in current research: through meticulously designed multi-stage pipelines, in-depth optimization of specific analysis tasks is conducted to achieve performance superior to that of traditional methods across core indicators such as accuracy and information gain.

Thus, spanning feature extraction, storage management and intelligent algorithms, the "technical toolbox" for processing unstructured data has become increasingly complete and efficient. Various technologies are constantly pursuing in-depth development in their specific problem domains, forming a vibrant yet fragmented technical ecosystem.

2.2. Core contradiction: the structural tension between fragmented technologies and end-to-end business requirements

Despite significant advancements in foundational technologies, their development exhibits a highly specialized and fragmented characteristic, creating a sharp contradiction with the continuity, systematicness, and value orientation demanded by real-world business scenarios. This contradiction is first manifested in the gap between component-based technology supply and integrated business demands. The current technological ecosystem resembles a collection of numerous sophisticated yet independent "technical islands", while typical business demands such as financial risk control auditing [2] require the seamless connection of a series of heterogeneous tasks such as document parsing, image recognition, graph association analysis, and report generation into a smooth automated pipeline. However, the absence of standardized integration interfaces and collaborative frameworks renders the selection and assembly of viable solutions from an extensive and complex technology landscape a highly demanding, highly customized, and complex project, seriously hindering the effective transformation of technology into productivity. Secondly, the contradiction between a

focus on local optimization orientation and overall process efficiency leads to inadequate coherence in data flow, state management, and exception handling between tasks, resulting in a complete data analysis process relying on a large amount of temporary manual bonding and maintenance, making the entire system fragile, rigid, and difficult to adapt to changes in requirements, and unable to form an agile closed loop from business feedback to model optimization. Moreover, the mismatch between technical measurement standards and business value perception poses additional challenges. The precise rate, recall rate, and other model indicators delivered by the technical team, and the final effects such as cost reduction and efficiency improvement, risk avoidance. This disconnect creates a "semantic gap" that hinders the accurate assessment and communication of technological progress value.

This contradiction is particularly profound in specific industry scenarios. In the medical field, for example, in the case of image-assisted diagnosis, clinical processes require seamless integration of lesion detection, segmentation, benign and malignant classification, and report generation, but the technology supply consists of independent single-point models. The absence of an intelligent framework capable of understanding clinical contexts and coordinating heterogeneous tasks compels hospitals to allocate substantial resources to customized integration during system deployment, resulting in fragile systems that are challenging to continuously optimize [1]. In the field of financial risk control, such as credit approval, the business end requires an end-to-end intelligent connection solution, while the technology side often independently optimizes single-point indicators such as OCR, anti-fraud, and credit assessment, ignoring the collaborative efficiency between models. This fundamental tension reveals that the bottleneck restricting the development of this field is no longer the lack of single-point technical capabilities, but the absence of a systematic framework capable of flexibly integrating fragmented technologies, connecting the data value chain, and closely aligning with business goals [7].

2.3. Core concept design: establish an end-to-end integrated framework oriented towards scenarios

The effective unlocking of the value of unstructured data hinges on bridging the gap between the fragmentation of aforementioned technologies and the integration of business needs. This requires research to move beyond the sole pursuit of technical performance and instead construct a systematic framework capable of achieving intelligent integration and collaboration across various components. The research aims to propose a new framework centered on scenario-driven and intelligent orchestration. This framework no longer treats technologies as a collection of discrete tools but reconfigures them into an organic ecosystem. Its core mission is to interpret business intentions and automatically convert these intentions into efficient and reliable data processing procedures, ultimately achieving seamless transformation from raw data to decision insights.

The design of this framework adheres to three core principles. The first is the scenario-driven principle, which serves as the ultimate goal of organizing and scheduling all technical components, to optimally meet the end-to-end requirements of specific business scenarios, whether it is post-loan monitoring in financial risk control or auxiliary diagnosis in healthcare. The second is the loosely coupled and modular principle, by standardizing interfaces to decouple data, algorithms, and capabilities, enabling each component to evolve independently and generating infinite possibilities through combination. Finally, it is the principle centered on intelligent orchestration, clearly defining the intelligent core of the framework as an engine capable of task planning, resource scheduling, and process execution. This engine undertakes the critical translation and command functions of connecting business language with technical implementation.

Based on the above principles, this research envisages the construction of a four-layer collaborative framework. The unified data resource layer constitutes the cornerstone of the entire framework. Its role

extends beyond mere data storage to the unified collection, management, and preprocessing of multi-source, multi-modal, and multi-format raw data. This layer must integrate the inclusiveness of a data lake with the standardization of a data warehouse, and may introduce graph databases to handle complex inter-entity relationship networks, thereby providing a high-quality and accessible data foundation for the upper-level analysis. To achieve efficient unified data governance, this layer adopts the "lake-and-database integration" architecture as the core. The data lake is responsible for cost-effectively ingesting massive volumes of heterogeneous source data in its original format while maintaining complete data lineage. This layer defines a unified data model and quality rules to build a logical data warehouse layer with high-performance query capabilities, converting the unstructured and semi-structured data in the "lake" into structured information that can be used for analysis. Above this lies the modular capability layer, which offers a systematic response to the current prosperous yet fragmented technological ecosystem. Various advanced processing and analysis technologies, whether based on rule-based document parsing [3], deep learning-based image recognition, or complex natural language understanding models, are encapsulated into standardized "capability units" with clear interfaces and explicit functions. These units resmble Lego blocks, their internal implementation can be a black box, but they provide stable service commitments externally, thereby encapsulating the complexity of the technology and releasing flexibility to the upper layer. Simultaneously, a "capability description file" is introduced. This file not only declares functions but also describes its applicable scenarios, data modalities processed, dependencies on upstream and downstream units, and adjustable parameters in a machine-readable metadata format, providing structured knowledge for the dynamic discovery and combination of the intelligent orchestration engine.

The core of the entire framework lies in the intelligent orchestration engine. Its essence is a closed-loop controller based on the "planning - scheduling - execution - monitoring" cycle. It receives requests from the business layer, which may be expressed in natural language or in structured templates. After receiving a task, the engine's primary function is to parse and divide the scenario, decomposing it into a series of ordered atomic operations. Then, based on the goals, constraints, and current system state of each atomic task, it dynamically selects one or more capability units from the module layer for combination to construct the optimal execution path. During this process, it must not only consider the matching degree of functions, but also balance performance, cost, and reliability. Finally, the engine is responsible for initiating the task pipeline, monitoring the entire execution process, managing data flow between stages, and handling possible exceptions to ensure the robustness and coherence of the process. Through this engine, the framework achieves a transition from a rigid, pre-defined workflow to a flexible, dynamically generated intelligent pipeline.

Ultimately, all underlying capabilities and the results of complex orchestration are presented to the end users through the contextual application layer. This layer provides an intuitive interface that enables business experts to define analytical scenarios, explore results, and iterate on requirements through configuration or natural interaction. It could be a risk dashboard embedded in an existing business system, or an independent intelligent analysis workbench. The existence of this layer ensures that the technical value can be perceived and utilized in a format that is familiar and relevant to business personnel, thus truly completing the "last mile" delivery from data to decision-making.

The integrated framework constructed by this research achieves fundamental innovation not through an original breakthrough in specific underlying technologies, but through systematic reorganization and top-level design that revolutionizes the paradigm of the existing technological ecosystem. It integrates scattered "technical islands" into a coherent whole that can directly respond to and execute business intentions by defining clear architectural logic and collaborative mechanisms. This framework decouples technical components through standardized interfaces and enables flexible combination through an intelligent engine as

the core driver for dynamic orchestration and full-process collaboration. Its ultimate objective is to shift unstructured data analysis from a highly dependent "customized skill" on expert manual integration to a standardized system engineering driven directly by business requirements, with efficient processes and stable, reliable results. This paradigm shift represents a key leap that consolidates dispersed technological potential into continuous business value, and provides a reusable systematic blueprint for unlocking the value of unstructured data.

3. Conclusions

This study reveals that the core bottleneck of unstructured data analysis lies in the systematic mismatch between the fragmentation of technologies and the integration requirements of business. The key to breaking through in the future lies in driving a fundamental shift in the research paradigm: shifting from pursuing the ultimate performance of individual technologies to building end-to-end solutions that can intelligently orchestrate technical components and precisely respond to the value demands of scenarios. This requires collaboration between academia and industry, as well as systematic construction at the levels of standards, platforms, and talents, ultimately achieving the leap from "tool innovation" to "value creation" in this field.

Naturally, this study also has certain limitations. Firstly, as an inductive study based on literature, the proposed integrated framework still needs to be verified and iteratively refined in real-world industrial environment. Secondly, this study does not deeply explore the data privacy and security issues in different industries. Future research can build upon this foundation to conduct in-depth case studies, develop platform prototypes, and continuously validate this technical paradigm in practice.

Looking ahead, the path to unlocking the value of unstructured data analysis will inevitably be deeply integrated with technological trends such as the native architecture of artificial intelligence, autonomous agents, low-code development, and natural interaction. Future research can build upon this framework to explore how the orchestration engine can integrate the complex task planning and reasoning capabilities of large language models, how to achieve cross-scenario and cross-domain capability migration and reuse, and how to construct a more intuitive "dialogue-based" data analysis interface. Ultimately, the processing of unstructured data will no longer be a sophisticated "craft", but will evolve into a basic service capability embedded in business processes and directly driven by demands, continuously providing core impetus for the digitalization and intelligence processes of various industries.

References

[1] Wang, Y. L. (2012). Comparative study of graph database NEO4J and relational database. *Modern Electronics Technique, 35*(20), 77–79. https: //doi.org/10.16652/j.issn.1004-373x.2012.20.045

[2] Yu, X. (2024). Research on optimization of unstructured data audit in rural cooperative banks in the context of big data. *Financial Forum*, (3), 90–92.

[3] Yi, X. Y., & Yi, M. Z. (2022). Unstructured data text classification algorithm based on SWOT analysis. *Science and Technology Innovation and Application, 12*(29), 25–28+33. https: //doi.org/10.19981/j.CN23-1581/G3.2022.29.006

[4] Zhang, X. Q. (2022). Research on the extraction methods of metadata for unstructured digital resources. *Jiangsu Science and Technology Information, 39*(27), 29–32+43.

[5] Yang, Q. H. (2023). Analysis of NoSQL database technology based on big data background. *Computer Knowledge and Technology, 19*(24), 67–69. https: //doi.org/10.14004/j.cnki.ckt.2023.1282

[6] Tang, K. L., & Zheng, H. (2024). An optimization method for unstructured big data classification based on improved ID3 algorithm. *Journal of Jilin University (Information Science Edition), 42*(5), 894–900. https: //doi.org/10.19292/j.cnki.jdxxp.2024.05.009

[7] Dai, P. J., Zhang, W. J., Li, X. Y., & Gao, Y. (2025). Exploration of the application and optimization path of unstructured data in human resources in the era of big data. *National Circulation Economy*, (11), 169–172. https: //doi.org/10.16834/j.cnki.issn1009-5292.2025.11.029