The American Journal of Engineering and Technology
ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 159-165
DOI 10.37547/tajet/Volume07Issue09-10

Check for updates

Methodological Foundations for Merging Structured and Unstructured Sources in ML Pipelines

OPEN ACCESS

SUBMITTED 19 August 2025
ACCEPTED 02 September 2025
PUBLISHED 13 September 2025
VOLUME Vol.07 Issue 09 2025

CITATION
Olusesan Ogundulu. (2025). Methodological Foundations for Merging Structured and Unstructured Sources in ML Pipelines. The American Journal of Engineering and Technology, 7(09), 159–165. https://doi.org/10.37547/tajet/Volume07Issue09-10

Olusesan Ogundulu
Data Engineer at Alvarez & Marsal Holdings Tampa, FL, United States

Abstract: The article presents a theoretical and applied analysis of the methodological foundations for merging structured and unstructured data sources within machine learning systems. The study is based on an interdisciplinary approach that integrates architectural design of ML pipelines, data representation theory, and practices of heterogeneous format integration. Particular attention is paid to the analysis of recent scientific publications highlighting the application of Retrieve–Merge–Predict architectures, agent-based discovery systems, and multimodal frameworks involving large language models. Four stable strategies for data merging are identified, ranging from static unification to end-to-end processing within a unified training loop. The importance of selecting matching metrics and adaptation schemes when dealing with unstable data streams is demonstrated using experiments from Cappuzzo and Eltabakh. Special emphasis is placed on the methodological limitations of universal solutions, including the generalization paradox, sensitivity to structural evolution, and the lack of formalized testing scenarios in agent-oriented pipelines. It is shown that sustainable development of ML architectures requires a shift from linear ETL pipelines to coherent, iterative systems with internal adaptation and feedback from the model to the data. The article will be of interest to researchers in data preparation automation, developers of multimodal ML systems, data engineering specialists, and digital platform architects working with multi-format sources.

Keywords: data integration, multimodal pipelines, structured sources, unstructured data, machine learning, language models, agent systems, matching metrics, adaptive architectures, data lakes.

Introduction

Contemporary machine-learning practice in academic and industrial settings is characterised by the rapid growth of data volumes originating from heterogeneous sources. Alongside traditional tabular structures, unstructured information—text corpora, multimedia content, event logs and free-form annotations—is being used with increasing frequency [5]. This broadening of data types requires a reconsideration of architectural approaches to ML pipelines and introduces new methodological challenges linked to the coherent integration of heterogeneous sources.

The difficulty lies in establishing technical connectivity across formats while accommodating profound differences in structure, semantics and scale of data representation. Because most learning models presuppose harmonised feature spaces at input, the fusion of structured and unstructured sources becomes not merely a non-trivial engineering task but an independent domain of theoretical inquiry [2]. Against the backdrop of the growing influence of large language models, multimodal pipelines and agent-based systems, interest is mounting in formalising principles and criteria for such fusion that go beyond ad-hoc technical solutions.

Integrating disparate data sources demands well-grounded procedures for alignment, matching and joint utilisation within a single computational process [1]. These procedures must accommodate variations in information representation and account for the specifics of generation, context, trustworthiness and reproducibility. Theoretical reflection on these principles is particularly important for constructing transparent, adaptive and scalable ML systems capable of efficiently processing multi-format data streams.

Research objective to analyse the methodological foundations for unifying structured and unstructured sources within machine-learning workflows, identify prevailing approaches, underlying principles and current constraints, and outline avenues for their development in the context of coherent ML-pipeline architectures.

Materials and Methods

The methodological foundation of this study lies at the intersection of ML-system engineering architectures, data-representation theory and automation practices for processing heterogeneous sources, reflecting the interdisciplinary nature of unifying structured and unstructured formats. The primary tool for theoretical analysis is a review of scholarly publications addressing methods for synchronising, merging and integrating heterogeneous data within machine learning.

The investigation draws on sources covering both theoretical and applied aspects of data unification. Particular attention is given to Cappuzzo [1], which proposed the Retrieve–Merge–Predict architecture focusing on automatic table population leveraging data lakes. Carlson [2] was instrumental in establishing the conceptual basis by detailing principles for reliable and scalable inference on unstructured sources. D'Alessandro's study [3] presents a modular multimodal architecture designed to process structured and unstructured data within a single pipeline.

The review article by Dritsas [4] played a key role in outlining a generalised typology of intersections between machine learning and big-data–oriented infrastructures. Concepts of automated detection and matching of data across different modalities were examined based on the approach described by Eltabakh [5]. As an illustration of end-to-end pipelines with differentiated source integration, the system proposed by Hilprecht [6] was considered. The application of multimodal analysis in domain-specific fields such as medicine was analysed through Jandoubi [7], which emphasises the importance of comprehensively accounting for source characteristics when constructing diagnostic models.

Additional attention in the methodological review was devoted to Li [8], which explores the use of large language models as a bridging component between structured and unstructured sources. The systematic investigation by Sedlakova [9] describes best practices for handling unstructured healthcare data, with a focus on enrichment and normalization challenges. The collection of publications concludes with Wu [10], demonstrating the application of BERTopic topic modelling in the context of merging clinical records with tabular demographic data.

Thus, the methodological strategy is based on a comparative analysis of architectures, metrics, compatibility criteria and data-merging scenarios

presented across various ML domains. This approach has enabled the development of a theoretically grounded framework for analysing the principles and constraints of integrating structured and unstructured sources in contemporary ML pipelines.

Results

Based on the theoretical analysis of the literature, four methodological strategies have been identified, each representing a distinct paradigm for processing heterogeneous data within machine learning. These strategies should not be seen as hierarchically ordered stages or generations; rather, they correspond to different architectural and conceptual preferences in the design of analytical systems.

One common approach entails the preliminary harmonization of disparate sources into a unified format outside the core processing pipeline. Under this strategy, static integration is performed: feature extraction, data cleaning and transformation occur before data are fed into the model [4]. The tabular representations produced via external preparation remove the influence of the data's unstructured nature on the model architecture but at the expense of reducing the solution's flexibility and generalization capacity. This approach offers a high degree of control and reproducibility yet adapts poorly to highly variable data streams.

An alternative strategy relies on the dynamic inclusion of sources during processing, leveraging application programming interfaces and intermediate representations. Such integration is built on the ability to interact directly with external storages and services at runtime. Employing intermediate formats—such as indexed tables, unified semantic graphs or mapping dictionaries—enables reconciliation of entities across types and management of their relationships during execution [7]. Architectures embodying this approach demonstrate strong adaptability but demand significant effort to ensure real-time consistency and data verification.

A methodologically distinct direction comprises systems founded on language models and agent-based logic. Cappuzzo [1] presents a schema in which tabular data are automatically enriched with fragments from an external repository in response to model-generated queries. Carlson [2] emphasizes the design of a universal inference mechanism resilient to inputs of unpredictable structure. Eltabakh [5] illustrates agent coordination capabilities, where autonomous components discover and reconcile correspondences between entities. Li [8] amplifies this vector by treating the language model as an intermediary that facilitates transitions between formats and serves as a unifying interface between structured and unstructured information.

Finally, architectures that implement end-to-end processing of multiple data types within a single training loop play a prominent role. D'Alessandro [3] describes a system in which text, tabular and visual data are encoded in parallel and then projected into a unified feature space at the integration stage. This approach ensures resilience to modality differences and supports a cohesive interpretation within the model. Here, source integration is embedded as a structural element of the analytical framework rather than as an external function. A comparative analysis of these strategies indicates that each embodies its own paradigm of source handling: from fixed schemas to adaptive streaming processing, from manual mapping to semantically driven interaction, and from modular aggregation to architectural cohesion.

Analysis of methods for combining structured and unstructured data sources requires empirical measurement of indicators that reflect the degree of successful matching between a target table and external sources. To evaluate the performance of these methods, this study employed the dataset from Cappuzzo [1], which conducted a quantitative analysis of the accuracy of joining a base table with multiple additional sources of diverse nature. Table 1 presents five integration variants, ranging from a simple binary join to connections with US open government data [1].

Table 1 – Integration metrics for various sources and tables (Source: [1])

Base Table	Binary	YADL Base	YADL 10k	YADL 50k	Open Data US
Company Employees	0.20	0.33	0.37	0.25	0.26
Housing Prices	0.34	0.57	0.54	0.54	0.50

Schools	NA	NA	NA	NA	1.00
2021 US Accidents	0.26	0.44	0.44	0.43	0.31
US County Population	0.93	0.84	0.95	0.85	NA
US Elections	0.44	0.55	0.52	0.52	0.59

The metrics are expressed as the proportion of successful matches between the base table and each external source. The most robust results emerge when integrating with the "US County Population" table (up to 0.95 in the YADL 50k configuration), whereas the weakest performance is observed for the "Company Employees" source (ranging from 0.20 to 0.37 across configurations). Missing values (e.g., for "Schools" in several variants) indicate either a lack of shared entities or non-matching keys. These data demonstrate that integration effectiveness varies by domain and the quality of initial keys. Notably, an increase in sample size (for instance, moving from YADL Base to YADL 50k) does not always correlate with metric improvement—some sources reach a performance plateau.

Alongside general integration indicators, a critical aspect of the unification methodology is the selection of metrics that determine the closeness between entities from different sources. A comparative analysis of such metrics was conducted using the experiments reported by Eltabakh [5], which include containment, numeric, semantic and ensemble measures. Table 2 summarises the number of successfully answered queries, coverage rates and profiling times.

Table 2 – Comparison of individual metrics: containment, numeric, semantic, and ensemble models (Source: [5])

Benchmark	Metric	name	containment	numeric	semantic	CMDL ensemble
3A	RR	0.82	0.63	0.34	0.62	0.83
	Queries answered	99%	99%	87%	100%	100%
3B	RR	0.44	0.65	0.04	0.73	0.79
	Queries answered	75%	100%	20%	100%	100%

As shown in Table 2, the highest accuracy is achieved using numeric metrics (up to 99% on RR-3A) and ensemble schemes. At the same time, execution time increases substantially, especially for scenarios involving deep semantic matching. This underscores the need to balance accuracy against computational cost when designing data-integration systems.

Discussion

The analysis of existing solutions for unifying structured and unstructured sources has revealed several methodological contradictions that limit their universality and applicability in subject-specific contexts. Despite the rapid evolution of architectures aimed at generalized integration, many approaches demonstrate insufficient robustness when applied to scenarios requiring contextual sensitivity and high precision. One key limitation is the so-called paradox of universality. Systems built around monolithic data-processing pipelines are designed for scalability, repeatability and formal independence from domain specifics. However, this universality is achieved by abstracting away context-dependent data characteristics. Consequently, solutions that perform well on generalized test suites often prove unsuitable in specialized settings—such as clinical diagnostics or legal

document workflows. Jandoubi [7] emphasizes that the successful deployment of multimodal architectures in healthcare demands extensive tailoring to particular data types, formats and error categories. A similar conclusion emerges from the systematic review by Sedlakova [9], which highlights the necessity of manual enrichment, standardization and expert validation of unstructured medical records prior to their integration with tabular datasets.

Another constraint concerns the low resilience of integration solutions to structural changes in data sources. A system calibrated for a specific schema frequently fails when new attributes appear, key fields are modified or relationships between entities evolve. Eltabakh [5] demonstrates that even minor deviations in table or text formats can cause a sharp decline in matching accuracy, particularly in systems that rely on fixed rules or template-based mappings. The absence of mechanisms for automatic adaptation and contextual re-evaluation of structural relationships hinders the repeated and scalable deployment of these solutions. Furthermore, architectures lacking support for schema versioning and the tracking of data-structure evolution are unable to maintain the integrity of the integration process over extended time horizons.

Analysis of the methodological foundation for integrating structured and unstructured data sources in machine-learning systems highlights several challenges that affect both existing solutions and emerging research directions. These challenges concern the resilience of architectures to change, the manageability of training, the need to formalize evaluation practices, and the shift toward adaptive systems with intrinsic coherence across all analysis stages. One systemic difficulty is the high dynamism of data sources—particularly when data flow from data lakes, corporate repositories or external APIs. Architectures designed for static schemas often become unstable when formats, key fields, structural relationships or underlying semantics evolve. Cappuzzo [1] emphasizes that automatic table augmentation from external sources demands highly accurate matching metrics and adaptation mechanisms for volatile streams. The PipeWeaver approach described in that study demonstrates an attempt to accommodate variability in both structural and content parameters, signaling the need for further development of architectures resilient to source evolution.

Equally significant is the integration of feedback from the model back into data-preparation stages. Carlson [2] explores the concept of quality-aware training—an iterative process in which model performance informs subsequent strategies for data selection and transformation. Such a cyclical schema requires tight alignment between data ingestion, feature selection, training and validation, implying a move away from linear pipelines toward systems with controlled internal adaptation. Embedding large-scale language models into the evaluation loop opens opportunities for iterative refinement of sources based on training outcomes.

The creation of representative and reproducible testing scenarios remains an open problem for systems in which data preparation is handled by autonomous agents. Eltabakh [5] calls for the establishment of benchmark scenarios—control cases for assessing solution quality within agent-based architectures. The absence of such standards complicates comparative evaluation and limits the cumulative advancement of source-integration tools. Table 3 provides an overview of the key applied domains in which machine-learning methods are already employed to analyse complex, heterogeneous and partially unstructured data.

Table 3 – Overview of key studies across various applied domains of ML (Compiled by the author based on source: [4])

Topic	Description
Healthcare	Predictive analytics, medical imaging, drug discovery, data management
Finance	Risk management, fraud detection, algorithmic trading, CRM
Transportation & Smart Cities	Autonomous vehicles, traffic management, smart infrastructure

Manufacturing & Industry 4.0	Predictive maintenance, automation, quality control
Energy & Utilities	Energy optimization, smart grids, renewable integration
Education	Intelligent tutoring, student success analytics, automated grading
Legal & Compliance	Contract analysis, legal research, compliance systems

The data in Table 3 confirm the broad demand for integration approaches across sectors, with each domain imposing unique requirements on format, accuracy, reproducibility and interpretability of the merged data. This underscores the necessity of adaptive architectures in which data preparation, model training and evaluation are unified into a coherent, dynamically tunable system. The analysis of these challenges suggests that sustainable development of ETL/ELT processes will require rethinking the interconnections between processing layers and transitioning from fragmented solutions to holistic computational systems.

Conclusion

The conducted study systematised the key methodological approaches to unifying structured and unstructured data sources in machine learning and identified enduring paradigms that underpin the architecture of modern ML pipelines. The analysis revealed that, despite the diversity of technical solutions and tools, heterogeneous-data integration remains both an engineering and a conceptual challenge requiring a rigorous methodological foundation.

The strategies identified—from preliminary unification to end-to-end multimodal processing—map out the typical scenarios of interaction among data formats, merging logic and architectural designs. Special attention was given to models employing agent coordination and language-model interfaces, reflecting the emerging trend toward self-organising data-processing systems. At the same time, these solutions exhibit limitations related to contextual sensitivity, the evolution of source structures and the lack of formal adaptation mechanisms.

Empirical analysis confirmed that integration quality depends heavily on the characteristics of the original keys, the modalities involved and domain specificity. It was determined that the effectiveness of merging methods varies according to the metrics applied and the system's capacity to adapt to unstable, evolving data streams. This finding underscores the necessity of reconceptualising data preparation as part of a cyclical, iterative architecture rather than as an isolated stage.

It was demonstrated that contemporary ML-system design must move away from universal, static pipelines and toward coherent, modular and reproducible configurations that support feedback and retraining based on performance outcomes. Such designs become especially critical as ML applications expand—from healthcare and education to smart cities and energy—where model interpretability, accuracy and robustness depend directly on the quality of data integration.

Thus, the unification of structured and unstructured sources in ML workflows emerges not merely as a technical step but as the methodological framework for the entire analytical system. Future research should focus on developing architectures resilient to source dynamics, formalising matching metrics and embedding internal adaptation mechanisms, thereby transitioning from fragmented practices to holistic computational ecosystems.

References

1. Cappuzzo, R., Coelho, A., Lefebvre, F., Papotti, P., & Varoquaux, G. (2025). Retrieve, merge, predict: Augmenting tables with data lakes (arXiv:2402.06282). arXiv. https://doi.org/10.48550/arXiv.2402.06282
2. Carlson, J., & Dell, M. (2025). A unifying framework for robust and efficient inference with unstructured data (arXiv:2505.00282). arXiv. https://doi.org/10.48550/arXiv.2505.00282
3. D'Alessandro, M., Calabrés, E., & Elkano, M. (2024). A modular end-to-end multimodal learning method for structured and unstructured data (arXiv:2403.04866). arXiv. https://doi.org/10.48550/arXiv.2403.04866
4. Dritsas, E., & Trigka, M. (2025). Exploring the intersection of machine learning and big data: A survey. Machine Learning and Knowledge

Extraction, 7(1), 13. https://doi.org/10.3390/make7010013

5. Eltabakh, M. Y., Kunjir, M., Elmagarmid, A., & Ahmad, M. S. (2023). Cross modal data discovery over structured and unstructured data lakes (arXiv:2306.00932). arXiv. https://doi.org/10.48550/arXiv.2306.00932

6. Hilprecht, B., Hammacher, C., Reis, E., Abdelaal, M., & Binnig, C. (2022). DiffML: End-to-end differentiable ML pipelines (arXiv:2207.01269). arXiv. https://doi.org/10.48550/arXiv.2207.01269

7. Jandoubi, B., & Akhloufi, M. A. (2025). Multimodal artificial intelligence in medical diagnostics. Information, 16(7), 591. https://doi.org/10.3390/info16070591

8. Li, B., Jiang, G., Li, N., & Song, C. (2024). Research on large-scale structured and unstructured data processing based on a large language model. Preprints. https://doi.org/10.20944/preprints202407.1364.v1

9. Sedlakova, J., Daniore, P., Horn Wintsch, A., Wolf, M., Stanikic, M., Haag, C., Sieber, C., Schneider, G., Staub, K., Alois Ettlin, D., Grübner, O., Rinaldi, F., von Wyl, V., & University of Zurich Digital Society Initiative (UZH-DSI) Health Community. (2023). Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. PLOS Digital Health, 2(10), e0000347. https://doi.org/10.1371/journal.pdig.0000347

10. Wu, S.-W., Li, C.-C., Chien, T.-N., & Chu, C.-M. (2024). Integrating structured and unstructured data with BERTopic and machine learning: A comprehensive predictive model for mortality in ICU heart failure patients. Applied Sciences, 14(17), 7546. https://doi.org/10.3390/app14177546