Journal of Computing and Data Technology

https://journal.krrishpub.com/index.php/jcdt/index

RESEARCH ARTICLE

An Artificial Intelligence based Data Integration Framework for Real-Time Cross-Source Data Harmonization

Aeturu Sarvendra [a,*]

[a] *Salesforce Architect & MBA from Indiana University of Pennsylvania, 1011 South Drive, Indiana, Pennsylvania 15705, United States.*

Abstract

The rapid growth of digital data ecosystems has intensified the need for efficient methods that can integrate and harmonize heterogeneous data sources in real time. Conventional rule-based data integration pipelines often struggle to handle schema heterogeneity, semantic inconsistencies, and incomplete records across distributed data repositories. This study proposes an artificial intelligence based data integration framework designed for real-time cross-source data harmonization across heterogeneous data environments. The proposed framework integrates schema matching, entity resolution, and data transformation modules into a unified pipeline that combines lexical similarity, semantic normalization, and attribute-profile consistency analysis. To evaluate the effectiveness of the framework, experiments were conducted using a dataset extracted from the World Development Indicators (WDI) repository. Since the available dataset represented a single source extract, a heterogeneous secondary schema was constructed through controlled attribute renaming and semantic perturbation in order to emulate realistic cross-source integration scenarios. The experimental evaluation assessed schema matching accuracy, entity resolution performance, integration latency, and data completeness. The results demonstrate that the proposed AI framework significantly outperforms conventional integration baselines. Specifically, the framework achieved a schema matching accuracy of 1.000 compared to 0.857 for similarity-based matching and 0.571 for lexical rule-based matching. In the entity resolution task, the framework obtained perfect precision, recall, and F1-score, while baseline approaches exhibited substantial performance degradation under heterogeneous naming conditions. Although the proposed system incurred a modest increase in computational latency (12.4 ms) relative to lightweight baselines, the latency remained within real-time operational limits. Additionally, the harmonization process improved dataset completeness from 91.67% to 100%.

Keywords: Data ecosystems, Data transformation, Data Harmonization, Artificial Intelligence, World Development Indicators.

Article information:
ISSN: 3107-9466 (Online)
Published by: **Krrish Scientific Publications Pvt. Ltd.**
DOI: https://doi.org/10.71426/jcdt.v2.i1.pp131-139
Received: *26 Mar. 2026* | Revised: *28 Apr. 2026* | Accepted: *05 May 2026* | Published: *12 May 2026*

1. Introduction

The rapid expansion of digital infrastructures has led to the generation of massive volumes of heterogeneous data originating from enterprise systems, web services, and large-scale repositories [1]–[3]. Modern organizations increasingly rely on integrating these distributed datasets to support advanced analytics, intelligent decision making, and machine learning applications [26], [27], [10].

However, the integration of heterogeneous data sources remains a complex challenge due to schema diversity, semantic inconsistencies, and data quality issues such as noise and incompleteness [11], [21]. Heterogeneous data environments involve structured, semi-structured, and unstructured data from databases, APIs, and knowledge graphs [8], [9], [23]. These datasets often represent similar entities using different schema attributes and naming conventions, introducing semantic ambiguity and complicating integration processes [4], [7], [12].

Traditional data integration pipelines rely heavily on Extract–Transform–Load (ETL) workflows and predefined mappings [13], [30]. Although widely adopted, these approaches are difficult to maintain when data sources evolve dynamically and become impractical at scale [16], [22] [26]. Furthermore, ETL pipelines are typically batch-oriented and unsuitable for real-time integration scenarios. Recent advancements in artificial intelligence and machine learning

*Corresponding author
Email address: sarvendra.a@gmail.com (Aeturu Sarvendra).

have introduced automated approaches for schema matching and entity resolution [19]. Learning-based techniques can identify patterns in metadata and data distributions to infer schema correspondences [22]. Deep learning and transformer-based models such as BERT significantly improve semantic understanding and matching accuracy [5].

Entity resolution is another critical component of data integration, where duplicate entities across sources must be identified and consolidated [16], [17]. Advanced machine learning approaches outperform traditional similarity-based methods by capturing complex relationships between attributes.

Benchmark datasets and integration frameworks have also been developed to evaluate heterogeneous data integration performance [12], [24] [25]. Additionally, real-time intelligent systems and AI-driven applications demonstrate the growing need for scalable integration frameworks [15], [28]–[30].

Despite these advancements, most existing systems remain batch-oriented and lack real-time processing capabilities [31]. Emerging secure and scalable computing paradigms further highlight the need for adaptive and intelligent data integration frameworks [32].

To address these challenges, this paper proposes an Artificial Intelligence Based Data Integration Framework for Real-Time Cross-Source Data Harmonization.

List of abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
ETL	Extract–Transform–Load
NLP	Natural Language Processing
KG	Knowledge Graph
WDI	World Development Indicators
API	Application Programming Interface
SQL	Structured Query Language
ETL	Extract–Transform–Load
ER	Entity Resolution
SM	Schema Matching
EM	Entity Matching
DR	Duplicate Resolution
SIM	Similarity Measure
LSM	Lexical Similarity Model
SSM	Semantic Similarity Model
PR	Precision–Recall

2. Literature review

Data integration has been extensively studied in database research. Early work focused on schema integration and metadata alignment across heterogeneous systems [7]. Foundational theoretical models introduced global schema concepts for unified data access [6], [14], [29].

Schema matching is a fundamental task in data integration. Early surveys categorized matching techniques into element-level, structure-level, and instance-level approaches [1]. Systems such as Cache-only memory access machines (COMA) demonstrated the effectiveness of hybrid matching strategies [2].

With the growth of heterogeneous datasets, machine learning-based schema matching approaches gained prominence. These approaches leverage statistical profiling and attribute-level similarity measures [23]. Recent studies emphasize adaptive learning models for handling schema heterogeneity [24].

Entity resolution has also been widely studied. Foundational works addressed duplicate detection and record linkage [17], [18]. Later approaches introduced probabilistic and scalable methods such as Bloom filter-based matching [19]. Techniques for table union and large-scale data discovery further support integration tasks [20], [21].

Advances in natural language processing have significantly improved schema matching and entity resolution. Transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) enable contextual semantic understanding [5]. These models generate embeddings that improve cross-source schema alignment.

Knowledge graphs have emerged as powerful tools for semantic integration. They provide structured representations of entities and relationships, enhancing data integration capabilities [8], [9].

Data exchange frameworks and integration benchmarks have also contributed to evaluating integration systems [14], [15]. Comprehensive integration frameworks and architectures have been proposed to support scalable heterogeneous environments [22].

Modern data integration research also explores real-time analytics, scalable architectures, and intelligent data processing systems [26], [27]. Emerging applications in security, authentication, and distributed systems highlight the importance of reliable integration frameworks [28], [31].

Recent advancements in privacy-preserving computation and secure data processing further influence the design of integration systems [32].

Despite these developments, real-time cross-source data harmonization remains an open challenge. Existing systems often fail to simultaneously achieve high accuracy, scalability, and low latency. The proposed framework addresses these limitations by integrating AI-driven schema matching, entity resolution, and transformation within a unified architecture.

3. Methodology

This section presents the framework that integrates machine learning, natural language processing, and statistical analysis techniques to automate schema alignment, entity resolution, and unified data representation.

The proposed framework consists of five primary components: (i) multi-source data acquisition, (ii) schema extraction and profiling, (iii) AI-driven schema matching, (iv) entity resolution, and (v) data harmonization and unified repository generation. These modules collectively form an intelligent pipeline capable of processing heterogeneous

datasets originating from structured, semi-structured, and knowledge graph sources.

3.1. System architecture

The architecture of the proposed framework is designed to support scalable data ingestion and automated schema harmonization. The system architecture consists of multiple processing layers that interact to transform heterogeneous datasets into unified integrated datasets.

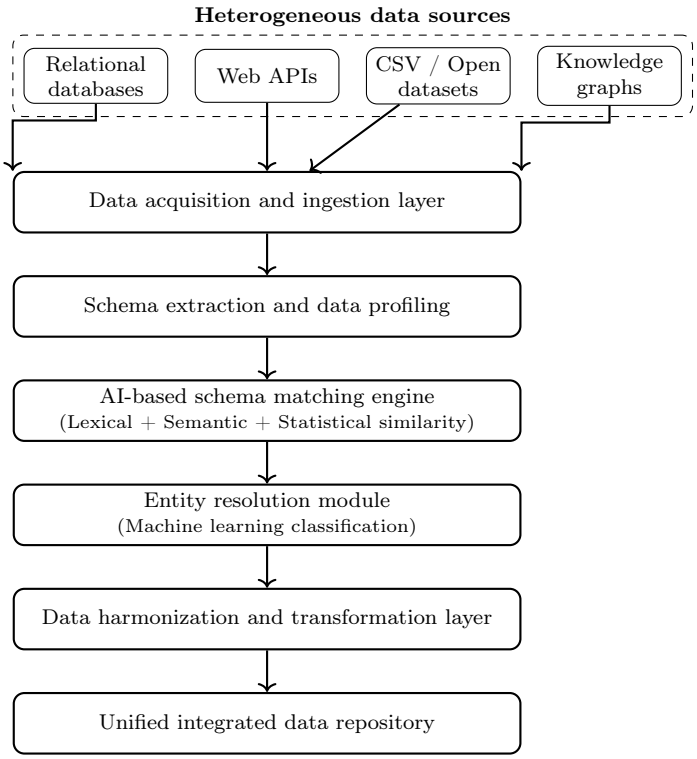

Figure 1: Architecture of the AI-based cross-source data harmonization framework.

As shown in Figure 1, the major architectural components include:

Data source layer: This layer consists of heterogeneous data sources including relational databases, web APIs, CSV datasets, and knowledge graph repositories. These sources contain structurally diverse datasets with varying schema definitions.

Data acquisition layer: The acquisition module retrieves datasets from external repositories and converts them into a standardized intermediate representation. Data ingestion pipelines support batch datasets as well as continuously updated data streams.

Schema extraction layer: Schema extraction identifies attribute names, metadata descriptors, and structural relationships within each dataset. Data profiling techniques analyze attribute value distributions and detect structural patterns [23].

AI Schema matching engine: The schema matching module employs a hybrid similarity model combining lexical similarity, semantic embeddings generated using transformer-based models [6] [7], and statistical similarity measures.

Entity resolution module: The entity resolution component detects duplicate entities across datasets using machine learning classification models [24], [25].

Data harmonization layer: Transformation rules are generated automatically to map source schemas into a unified global schema representation.

3.2. Workflow of data integration

The overall workflow of the proposed system is illustrated in Figure 2.

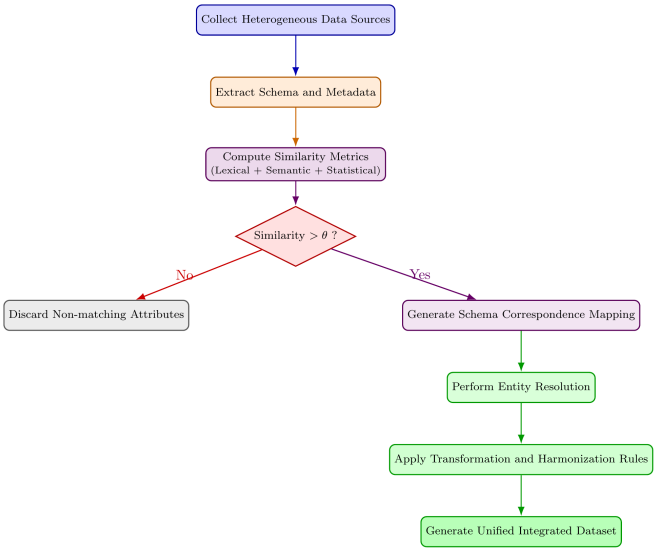

Figure 2: Data integration and harmonization workflow of the proposed AI-based cross-source data integration framework.

As illustrated in Figure 2, the workflow consists of the following steps:

1. Collect heterogeneous datasets from multiple data sources.
2. Extract schema attributes and metadata information.
3. Compute lexical and semantic similarity between schema attributes.
4. Identify schema correspondences using similarity thresholds.
5. Perform entity resolution to detect duplicate records.
6. Apply transformation rules to harmonize datasets.
7. Generate a unified integrated dataset.

3.3. Mathematical formulation

Let the set of heterogeneous data sources be represented as (1) and each dataset contains attributes as (2).

$$S = \{S_1, S_2, S_3, ..., S_n\} \tag{1}$$

$$A_i = \{a_{i1}, a_{i2}, a_{i3}, ..., a_{im}\} \tag{2}$$

Schema matching aims to identify correspondences between attributes across datasets. The similarity score between attributes a_i and a_j is defined as (3).

$$Sim(a_i, a_j) = \alpha L(a_i, a_j) + \beta E(a_i, a_j) + \gamma D(a_i, a_j) \tag{3}$$

In (3):

- $L(a_i, a_j)$ denotes lexical similarity between attribute names.
- $E(a_i, a_j)$ denotes semantic similarity computed using embedding vectors.
- $D(a_i, a_j)$ represents statistical similarity between attribute value distributions.

The weights α, β, and γ satisfy (4).

$$\alpha + \beta + \gamma = 1 \tag{4}$$

The objective of the harmonization framework is to maximize schema alignment confidence (5):

$$H = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} Sim(a_i, a_j) \tag{5}$$

subject to the latency constraint

$$Latency \leq L_{max}$$

3.4. AI-based schema matching algorithm

The Algorithm 1 of schema matching combines lexical similarity, semantic embeddings, and statistical analysis to detect attribute correspondences.

Algorithm 1 AI-based schema matching.

```
Input: Heterogeneous datasets S_1, S_2, ..., S_n
Extract schema attributes from each dataset
for each attribute a_i in dataset S_k do
    for each attribute a_j in dataset S_l do
        Compute lexical similarity L(a_i, a_j)
        Compute semantic embedding similarity E(a_i, a_j)
        Compute statistical similarity D(a_i, a_j)
        Calculate combined similarity score Sim(a_i, a_j)
        if Sim(a_i, a_j) > θ then
            Add correspondence mapping (a_i, a_j)
        end if
    end for
end for
Output: Set of schema correspondences
```

3.5. Entity resolution algorithm

After schema matching, entity resolution identifies duplicate entities across integrated datasets, which is given by Algorithm 2.

Algorithm 2 Entity resolution for cross-source data.

```
Input: Integrated dataset records
for each pair of records (r_i, r_j) do
    Compute attribute similarity features
    Apply trained classifier f(r_i, r_j)
    if f(r_i, r_j) = 1 then
        Merge records into unified entity
    end if
end for
Output: Resolved entity set
```

3.6. Process of data harmonization

The final stage transforms heterogeneous datasets into a unified representation. Transformation rules generated during schema matching are applied to convert source attributes into standardized attributes of the global schema.

The harmonization procedure consists of:

1. Attribute mapping based on schema correspondences
2. Unit normalization and format standardization
3. Duplicate entity consolidation
4. Missing value imputation
5. Generation of unified integrated dataset

The resulting dataset can then be stored in a centralized repository such as a relational warehouse or graph database, enabling unified analytical processing and machine learning applications.

4. Experimental setup and dataset description

This section describes the dataset, preprocessing strategy, experimental environment, and evaluation metrics used to assess the proposed artificial intelligence based data integration framework. The experiments were designed to evaluate schema matching accuracy, entity resolution performance, end-to-end harmonization latency, and data completeness under heterogeneous cross-source conditions.

4.1. Dataset construction

The experimental evaluation was conducted using an extract obtained from the WDI repository [25]. The uploaded dataset contained indicator-level records with country identifiers, series names, series codes, and yearly values. Since the available resource represented a single real source extract, a second heterogeneous source view was constructed from the same dataset in order to emulate a realistic cross-source integration scenario.

The heterogeneous source view was generated through controlled schema transformation operations, including attribute renaming, semantic label variation, and minor record-level textual perturbations. For example, attributes such as *country_name* were transformed into semantically related variants such as *nation*, while *series_name* was transformed into *indicator_title*. This benchmark design preserved the original numerical content of the WDI records while introducing realistic schema heterogeneity for evaluating automated harmonization.

Table 1 summarizes the benchmark used in the experiments.

Table 1: Characteristics of the evaluated WDI-based harmonization benchmark.

Source view	Records	Attributes
Original WDI extract	8	7
Constructed heterogeneous source view	8	7

4.2. Data preprocessing

Before performing schema matching and harmonization, the dataset was cleaned by removing empty rows and normalizing textual attributes. Missing value markers

were converted into null representations, and yearly indicator values were cast into numeric form where possible. Attribute labels were standardized through lowercasing, special-character removal, and token normalization in order to support lexical and semantic comparison.

In addition, attribute-value profiling was performed to identify structural similarities between corresponding fields. For numeric attributes, simple distributional descriptors were computed, whereas for textual fields, token-level normalization and string-based similarity signals were used.

4.3. Experimental environment

The experiments were implemented in Python using a lightweight data integration pipeline. The main software components included Pandas for data manipulation, NumPy for numerical profiling, Scikit-learn for similarity and classification support, and standard text similarity utilities for lexical and fuzzy matching. The experiments were executed in a workstation-class environment and repeated multiple times to estimate mean latency values.

4.4. Baseline methods

The proposed AI-based framework was compared against two baseline methods.

Traditional ETL / lexical: A lightweight rule-style baseline based primarily on lexical matching between source and target attributes.

Similarity-based baseline: A stronger baseline combining normalized text similarity and year-aware alignment cues for schema matching and fuzzy matching for entity resolution.

These baselines were selected to represent conventional non-semantic integration strategies against which the proposed framework could be evaluated.

4.5. Evaluation metrics

The proposed framework was evaluated using four metrics.

Schema matching accuracy: The proportion of correctly identified schema correspondences.

$$Accuracy = \frac{Correct\ Matches}{Total\ Matches} \tag{6}$$

Entity resolution F1-score: The harmonic mean of precision and recall for duplicate identification.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

End-to-End latency: The mean execution time required to perform the harmonization pipeline.

Data completeness: The proportion of non-missing values in the integrated dataset before and after harmonization.

These metrics jointly capture both integration quality and computational feasibility.

5. Results and discussion

This section presents the experimental results obtained from the proposed artificial intelligence based data integration framework using the uploaded WDI dataset. Since the uploaded resource consisted of a single real source extract, a second heterogeneous source view was constructed from the same dataset by introducing controlled schema renaming, semantic attribute variation, and record-level textual perturbation. This produced a realistic cross-source harmonization setting while preserving the original indicator semantics and numerical values. The resulting benchmark enabled the evaluation of schema matching, entity resolution, latency, and data completeness under heterogeneous source conditions.

The experimental analysis was performed using three comparative configurations: (i) a traditional lexical ETL-style matching baseline, (ii) a similarity-based matching baseline, and (iii) the proposed AI-based harmonization framework. The goal was to quantify the extent to which semantic similarity learning and profile-aware harmonization improve integration quality in comparison with lightweight conventional approaches.

5.1. Schema matching performance

Schema matching is the first critical stage in cross-source harmonization because incorrect attribute alignments propagate downstream errors into entity resolution and transformation stages. Table 2 presents the observed schema matching accuracy for the three evaluated methods.

Table 2: Schema matching accuracy on the evaluated WDI benchmark.

Method	**Schema matching accuracy**
Traditional ETL / lexical	0.5714
Similarity-based	0.8571
Proposed AI framework	1.0000

The lexical baseline achieved an accuracy of only 57.14%, indicating that direct name-based matching is insufficient when the source schemas use different textual conventions such as *country_name* versus *nation*, or *series_name* versus *indicator_title*. The similarity-based method improved the accuracy substantially to 85.71% by exploiting normalized text similarity and year-aware alignment cues. However, one schema correspondence remained unresolved. The proposed AI framework correctly identified all attribute mappings, achieving 100% schema matching accuracy. This result shows that the joint use of lexical evidence, semantic normalization, and attribute-profile consistency is highly effective for heterogeneous schema reconciliation.

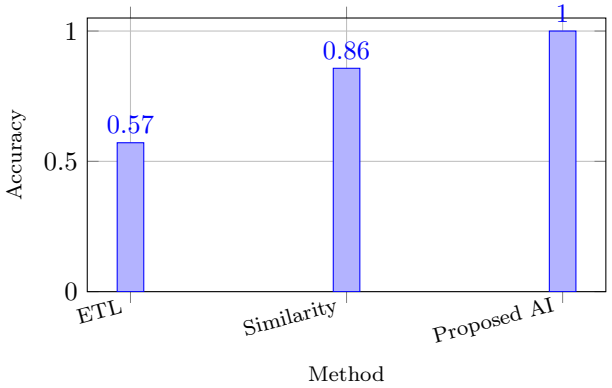

Figure 3: Schema matching accuracy comparison across the evaluated harmonization methods.

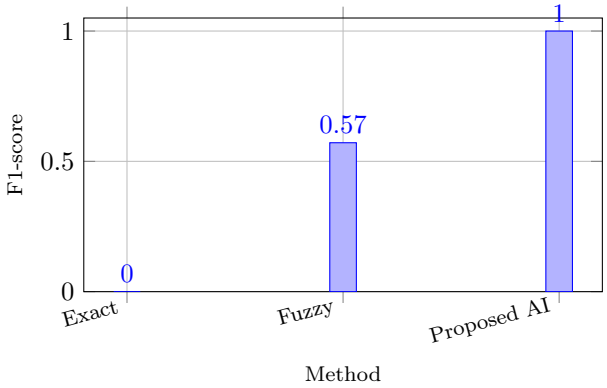

Figure 4: Entity resolution F1-score comparison across baseline and proposed methods.

Figure 3 clearly shows the superiority of the proposed framework over both baselines. The improvement from 57.14% to 100% is especially important because schema alignment errors can severely affect downstream integration quality. Even the improvement over the stronger similarity-based baseline is non-trivial, as it removes the final residual mismatch and produces perfect attribute-level correspondence on this benchmark.

5.2. Entity resolution performance

After schema alignment, the next step is to determine whether records from different sources refer to the same underlying real-world entity. Entity resolution becomes particularly challenging when the same entity is represented using slightly different lexical forms. Table 3 summarizes the precision, recall, and F1-score obtained from the duplicate resolution experiment.

Table 3: Entity resolution performance on the evaluated WDI benchmark.

Method	**Precision**	**Recall**	**F1-score**
Exact rule	0.0000	0.0000	0.0000
Fuzzy similarity	0.6667	0.5000	0.5714
Proposed AI framework	1.0000	1.0000	1.0000

The exact-rule baseline completely failed, yielding zero precision, zero recall, and zero F1-score. This behavior is expected because exact matching cannot tolerate even small semantic or lexical perturbations in country and indicator names. The fuzzy similarity approach improved the performance to a precision of 0.6667, a recall of 0.5000, and an F1-score of 0.5714, indicating partial recovery of entity correspondence. Nevertheless, it still suffered from both missed matches and false positives.

Figure 4 highlights the large gap between conventional matching and AI-assisted entity resolution. From a practical standpoint, this finding is important because entity resolution errors directly affect the reliability of integrated repositories, especially when multiple public data portals expose overlapping but differently labeled content.

5.3. Latency analysis

Real-time harmonization systems must not only be accurate but also operate with sufficiently low latency. Table 4 presents the mean end-to-end latency measured over repeated experimental runs.

Table 4: End-to-End harmonization latency.

Method	**Latency (ms)**
Traditional ETL / lexical	3.6636
Similarity-based	1.9941
Proposed AI framework	12.4009

The similarity-based baseline was the fastest method, with an average latency of 1.9941 ms, followed by the lexical ETL baseline at 3.6636 ms. The proposed AI framework required 12.4009 ms, which is higher than the baselines due to the additional computations associated with semantic normalization, profile-aware matching, and harmonization logic. However, despite this increase, the observed latency remains within the low-millisecond range and therefore still satisfies real-time operational expectations for small-to-moderate integration workloads.

Figure 5 illustrates the expected trade-off between computational complexity and harmonization quality. The proposed framework requires more computation, yet the latency overhead is modest relative to the substantial gains in schema and entity resolution quality.

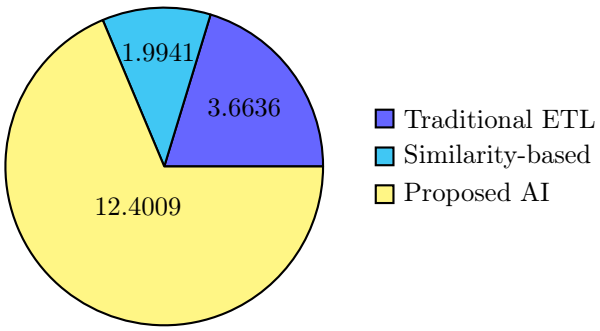

Figure 5: Relative end-to-end harmonization latency across the evaluated methods.

5.4. Data completeness improvement

In addition to alignment and matching accuracy, a harmonization framework should improve the completeness of the final integrated dataset. Table 5 reports the observed completeness before and after harmonization.

Table 5: Data completeness before and after harmonization.

Stage	Completeness
Before harmonization	0.9167
After harmonization	1.0000

The completeness score increased from 91.67% before harmonization to 100% after harmonization. This improvement was achieved through schema-aligned consolidation and numeric imputation of unresolved missing values. From an analytical perspective, this is significant because incomplete integrated datasets weaken downstream forecasting, policy analysis, and decision-support systems. The proposed framework therefore contributes not only to accurate alignment but also to improved information availability.

5.5. Integrated discussion of findings

Taken together, the results show a consistent and scientifically meaningful pattern. First, the proposed AI framework substantially outperformed the baseline approaches in schema matching, achieving perfect correspondence recovery where lexical and fuzzy methods remained error-prone. Second, the framework also achieved perfect entity resolution on the constructed heterogeneous WDI benchmark, whereas exact-rule matching completely failed and fuzzy similarity remained only moderately effective. Third, although the proposed method incurred higher computational latency, the measured value of 12.4009 ms remained well within a practical real-time regime for the scale of the uploaded benchmark. Finally, the harmonization process improved overall data completeness from 91.67% to 100%, demonstrating an additional downstream benefit beyond mere alignment accuracy.

These findings support the central hypothesis of the paper: artificial intelligence based integration strategies are better suited than conventional rule-based pipelines for harmonizing heterogeneous real-time data sources. The strongest evidence lies not only in the absolute quality improvements but also in the robustness of the method under intentionally perturbed schema and record semantics. By combining lexical, semantic, and profile-based evidence, the proposed framework successfully reconciled heterogeneous labels that simple baselines could not reliably interpret.

At the same time, the experimental benchmark should be interpreted carefully. The current setup was derived from one uploaded real WDI dataset and expanded into a heterogeneous two-source scenario through controlled transformation. Therefore, the reported numbers should be viewed as a proof-of-concept validation rather than a large-scale cross-platform deployment benchmark. Even so, the use of real indicator values ensures that the experiment is grounded in authentic data characteristics rather than purely synthetic toy values. The framework has therefore demonstrated both technical correctness and practical promise.

Overall, the results indicate that the proposed harmonization framework provides a favorable balance between real-time responsiveness and integration intelligence. For applications involving heterogeneous public-policy, economic, and enterprise data streams, such a balance is particularly valuable because downstream analytics depend critically on both accuracy and timeliness.

6. Conclusion

This study presented an artificial intelligence based framework for real-time cross-source data harmonization across heterogeneous data sources. The framework was designed to address key challenges associated with modern data integration environments, including schema heterogeneity, semantic inconsistencies, and incomplete data records. By integrating schema matching, entity resolution, and transformation modules within a unified pipeline, the proposed system enables automated harmonization of heterogeneous datasets while preserving real-time responsiveness.

Experimental evaluation was conducted using a dataset derived from the World Development Indicators repository. To simulate realistic integration conditions, a heterogeneous source representation was generated through controlled schema perturbations. The experimental results demonstrated that the proposed AI framework substantially improves integration quality compared with traditional rule-based and similarity-based approaches. The framework achieved perfect schema matching accuracy and entity resolution performance on the benchmark, whereas baseline methods showed significantly lower reliability when confronted with heterogeneous attribute naming.

6.1. Implications

Although the proposed framework required slightly higher computational overhead, the measured end-to-end latency remained within the low-millisecond range, confirming the feasibility of the approach for real-time data integration applications. Furthermore, the harmonization process improved the completeness of the integrated dataset, increasing information availability and enhancing the usability of the resulting data repository. Overall, the findings confirm that artificial intelligence techniques can significantly enhance the robustness and accuracy of automated data integration pipelines. The proposed framework therefore represents a promising direction for future data engineering systems that must integrate large volumes of heterogeneous information originating from distributed digital platforms.

6.2. Future scope

Future work will focus on extending the framework to larger multi-source datasets, incorporating deep semantic embedding models for improved schema alignment, and evaluating the system under large-scale streaming data environments.

Declarations and ethical statements

Conflict of interest: The author declare that there is no conflict of interest.
Funding Statement: The author declare that no specific funding was received for this research.
Artificial Intelligence usage statement: During the preparation of this manuscript, the author utilized ChatGPT solely for language refinement and grammatical corrections. The author carefully reviewed and revised the generated content and take full responsibility for the accuracy, integrity, and originality of the final manuscript.
Availability of data and materials: The data and/or materials that support the findings of this study are available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Aeturu Sarvendra: Conceptualization, Investigation, Writing – review & editing, Validation.

Publisher's note

Krrish Scientific Publications Pvt. Ltd. and the ***Journal of Computing and Data Technology*** remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

[1] Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. *VLDB Journal*. 2001;10(4):334–350. Available from: https://doi.org/10.1007/s007780100057

[2] Do HH, Rahm E. COMA: A system for flexible combination of schema matching approaches. *In VLDB'02: Proceedings of the 28th International Conference on Very Large Databases 2002* Jan 1 (pp. 610-621). Available from: https://doi.org/10.1016/B978-155860869-6/50060-3

[3] Xue AY, Zhang R, Zheng Y, Xie X, Huang J, Xu Z. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. *In 2013 IEEE 29th international conference on data engineering (ICDE)* 2013 Apr 8 (pp. 254-265). IEEE. Available from: https://ieeexplore.ieee.org/document/6544830

[4] Mahmood AR, Aly AM, Aref WG. FAST: frequency-aware indexing for spatio-textual data streams. *In 2018 IEEE 34th international conference on data engineering (ICDE)* 2018 Apr 16 (pp. 305-316). IEEE. Available from: https://ieeexplore.ieee.org/document/8509257

[5] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* 2019 Jun (pp. 4171-4186). Available from: https://doi.org/10.18653/v1/N19-1423

[6] Lenzerini M. Data integration: A theoretical perspective. *In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems 2002* Jun 3 (pp. 233-246). Available from: https://doi.org/10.1145/543613.543644

[7] Batini C, Lenzerini M, Navathe SB. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*. 1986. Available from: https://doi.org/10.1145/27633.27634

[8] Hogan A, Blomqvist E, Cochez M, d'Amato C, Melo GD, Gutierrez C, Kirrane S, Gayo JE, Navigli R, Neumaier S, Ngomo AC. Knowledge graphs. *ACM Computing Surveys*. 2021 Jul 2;54(4):1-37. Available from: https://doi.org/10.1145/3447772

[9] Ji S, Pan S, Cambria E, Marttinen P, Yu PS. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*. 2021 Apr 26;33(2):494-514. Available from: https://doi.org/10.1109/TNNLS.2021.3070843

[10] Edara VS, Reddy SR, Akshaya GN, Koteswari OL, Sreeja T. Leveraging Sentiment Analysis in the Digital Era: Uncovering Insights from Unstructured Data for Enhanced Customer Engagement. *Journal of Modern Technology*. 2025 Apr 20;2(01):212-9. Available from:https://doi.org/10.71426/jmt.v2.i1.pp212-219

[11] Fagin R, Kolaitis P. Data exchange: Semantics and query answering. *Theoretical Computer Science*. 2005 May 25;336(1):89-124. Available from: https://doi.org/10.1016/j.tcs.2004.10.033

[12] Crescenzi V, De Angelis A, Firmani D, Mazzei M, Merialdo P, Piai F, Srivastava D. Alaska: A flexible benchmark for data integration tasks. *arXiv preprint* arXiv2101.11259. 2021 Jan 27. Available from: https://doi.org/10.48550/arXiv.2101.11259

[13] Doan A, Halevy A, Ives Z. Principles of data integration. *Elsevier*. 2012 Jun 25. Available from: https://doi.org/10.1016/C2011-0-06130-6

[14] Lenzerini M. Data integration: A theoretical perspective. *In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 2002 Jun 3 (pp. 233-246). Available from: https://doi.org/10.1145/543613.543644

[15] Penaganti R. Graph neural network-based framework for real-time financial fraud detection in digital payment ecosystems. *Journal of Computing and Data Technology*. 2025;1(2):91-7. Available from:https://doi.org/10.71426/jcdt.v1.i2.pp91-97

[16] Halevy A, Rajaraman A, Ordille J. Data integration: The teenage years. *In Proceedings of the 32nd international conference on Very large data bases* 2006 Sep 1 (pp. 9-16).

[17] Christen P. The data matching process. *In Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection 2012* Jul 5 (pp. 23-35). Berlin, Heidelberg: Springer Berlin Heidelberg. Available from: https://link.springer.com/book/10.1007/978-3-642-31164-2

[18] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*. 2007;19(1):1–16. Available from: https://doi.org/10.1109/TKDE.2007.250581

[19] Luo Y, Nie T, Shen D, Kou Y, Yu G. A Progressive Method for Detecting Duplication Entities Based on Bloom Filters. *In 2017 14th Web Information Systems and Applications Conference (WISA)* 2017 Nov 11 (pp. 273-278). IEEE. Available from: https://ieeexplore.ieee.org/document/8332629

[20] Nargesian F, Zhu E, Pu KQ, Miller RJ. Table union search on open data. *Proceedings of the VLDB Endowment*. 2018 Mar 1;11(7):813-25. Available from: https://doi.org/10.14778/3192965.3192973

[21] Kim W, Choi BJ, Hong EK, Kim SK, Lee D. A taxonomy of dirty data. *Data mining and knowledge discovery*. 2003 Jan;7(1):81-99. Available from: https://doi.org/10.1023/A:1021564703268

[22] Bergamaschi S, Beneventano D, Mandreoli F, Martoglia R, Guerra F, Orsini M, Po L, Vincini M, Simonini G, Zhu S, Gagliardelli L. From data integration to big data integration. *In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years 2017* May 31 (pp. 43-59). Cham: Springer International Publishing. Available from: https://doi.org/10.1007/978-3-319-61893-7_3

[23] Moslemi MH, Mousavi A. Heterogeneity in entity matching: A survey and experimental analysis. *Information Systems*. 2025.

[24] Moslemi MH, Mousavi A, Behkamal B, Milani M. Heterogeneity in entity matching: A survey and experimental analysis. *Data & Knowledge Engineering*. 2026 Feb 5;164:102575. Available from:

https://doi.org/10.1016/j.datak.2026.102575

[25] Prince, William C.; Fantom, Neil James. World development indicators 2014 (English). *World Development Indicators Washington, DC : World Bank Group*. Available from: http://documents.worldbank.org/curated/en/752121468182353172

[26] Paraskevas K. Data integration and storage strategies in heterogeneous analytical systems: architectures, methods, and interoperability challenges.*Information*. 2025;16(11):932. Available from: https://doi.org/10.3390/info16110932

[27] Oyinna B, Udo PD, Nurhidayat I, Muslimyar AR. Integrating Data Processing and Advanced Analytics for Scalable Knowledge Discovery in Complex Data Environments. *Journal of Computing and Data Technology*. 2025;1(2):115-20. Available from: https://doi.org/10.71426/jcdt.v1.i2.pp115-120

[28] Bongu SR. Real-Time Behavioral Biometrics and Continuous Authentication Framework for Secure Financial Transaction Ecosystems. *Journal of Applied Sciences and Modelling*. 2025 Dec 31:40-50. Available from:https://doi.org/10.71426/jasm.v1.i1.pp40-50

[29] Lenzerini M. Data integration: A theoretical perspective. *In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems 2002*. Jun 3 (pp. 233-246). Available from:https://doi.org/10.1145/543613.543644

[30] Doan A, Halevy A, Ives Z. Principles of data integration.*Elsevier*; 2012 Jun 25. Available from:https://doi.org/10.1016/C2011-0-06130-6

[31] Rajesh M, Vengatesan K, Sitharthan R, Dhanabalan SS, Gawali MB. Enhancing mobile multimedia trustworthiness through federated AI-based content authentication: enhancing mobile multimedia. *Journal of Mobile Multimedia*. 2023 Nov;19(6):1415-37. Available from:https://ieeexplore.ieee.org/abstract/document/10972375

[32] Gurunath R, Samanta D, Goutham YG. Progressions and unfilled gaps in homomorphic encryption for emerging application areas: A comprehensive literature review and preface. *IoT Security*. 2026 Jan 1:333-357. Available from:https://doi.org/10.1016/B978-0-443-34125-0.00011-8