








## REVIEW

# Overview on Data Ingestion and Schema Matching

## Visión general de la ingestión de datos y la correspondencia de esquemas

Oumaima El Haddadi<sup>1,2</sup>  , Max Chevalier<sup>1</sup> , Bernard Dousset<sup>1</sup>, Ahmad El Allaoui<sup>2,3</sup> , Anass El Haddadi<sup>2</sup> , Olivier Teste<sup>1</sup> 

<sup>1</sup>IRIT, SIG Team, Paul Sabatier University, Toulouse, France.

<sup>2</sup>LSA, SDIC Team, Abdelmalek Essaadi University, Tetouan, Morocco.

<sup>3</sup>STI, IDMS Team, FST, Moulay Ismail University of Meknes, Morocco.

Cite as: Haddadi OE, Chevalier M, Dousset B, Allaoui AE, Haddadi AE, Teste O. Overview on Data Ingestion and Schema Matching. Data and Metadata 2024;3:219. <https://doi.org/10.56294/dm2024219>.

Submitted: 01-11-2023

Revised: 28-12-2023

Accepted: 07-02-2024

Published: 08-02-2024

Editor: Prof. Dr. Javier González Argote 

### ABSTRACT

This overview traced the evolution of data management, transitioning from traditional ETL processes to addressing contemporary challenges in Big Data, with a particular emphasis on data ingestion and schema matching. It explored the classification of data ingestion into batch, real-time, and hybrid processing, underscoring the challenges associated with data quality and heterogeneity. Central to the discussion was the role of schema mapping in data alignment, proving indispensable for linking diverse data sources. Recent advancements, notably the adoption of machine learning techniques, were significantly reshaping the landscape. The paper also addressed current challenges, including the integration of new technologies and the necessity for effective schema matching solutions, highlighting the continuously evolving nature of schema matching in the context of Big Data.

**Keywords:** Data Management; Schema Matching; Data Ingestion; Heterogeneous Schema; Dynamic Environment.

### RESUMEN

Esta visión general rastreó la evolución de la gestión de datos, haciendo la transición desde procesos tradicionales de ETL para abordar desafíos contemporáneos en Big Data, con un énfasis particular en la ingestión de datos y la coincidencia de esquemas. Exploró la clasificación de la ingestión de datos en procesamiento por lotes, en tiempo real y híbrido, destacando los desafíos asociados con la calidad de datos y la heterogeneidad. En el centro de la discusión estaba el papel del mapeo de esquemas en la alineación de datos, demostrando ser indispensable para vincular diversas fuentes de datos. Los avances recientes, especialmente la adopción de técnicas de aprendizaje automático, estaban remodelando significativamente el panorama. El documento también abordó desafíos actuales, incluida la integración de nuevas tecnologías y la necesidad de soluciones efectivas de coincidencia de esquemas, resaltando la naturaleza en constante evolución de la coincidencia de esquemas en el contexto de Big Data.

**Palabras clave:** Gestión de Datos; Coincidencia de Esquemas; Ingestión de Datos; Esquema Heterogéneo; Entorno Dinámico.

### INTRODUCTION

In today's landscape, information and data play a central role in understanding scientific, economic, political, and social issues, triggering transformative change in working environments. The growing interest in

Big Data in science, management and finance underscores the critical need to effectively process diverse types of data. This imperative goes beyond simply transforming data into knowledge for decision-making.

Addressing these information needs; requires well-defined data management processes that cover data collection, storage, and exploitation. While initial reliance was on manual techniques, these have now been supplanted by advanced processes and tools to manage the complexities of big data. The ETL (Extract, Transform, Load) process has become the cornerstone of data integration, gaining in importance with the advent of data warehouses.<sup>(1)</sup> Later, users adopted the ELT (Extract, Load, Transform) solution, often associated with data lakes, to mitigate data latency.<sup>(2)</sup> Despite these developments, data management remains a critical area, with ongoing challenges and continual improvements.

Researchers are also focusing on data ingestion, recognizing its role as either a precursor to or an integral part of the integration process. Schema matching, a key technique in this area, facilitates the integration of diverse data types by identifying semantically and/or contextually linked objects.<sup>(3)</sup> This technique is crucial for data alignment, maintaining data in their initial states and establishing connections. While these methods adeptly handle data heterogeneity, there are still uncertainties regarding their effectiveness in addressing the dynamism of evolving data environments.

This comprehensive overview explores the current state and potential advancements in data ingestion in Section 2. Section 3 delves into data alignment. Both sections cover the exploration, current challenges, and future prospects.

### Exploring data ingestion

As an emerging area of research, contemporary researchers are directing their attention towards data ingestion in the realm of Big Data, a domain often interchangeably associated with data integration. In the literature, the term "data ingestion" is commonly interlinked with "data integration," denoting a process that encompasses the extraction, loading, and transformation of data, akin to the procedures observed in ETL (Extract, Transform, Load) processes.<sup>(4)</sup> As of today, we have identified a total of 372 documents employing the term "data ingestion" in Web of Science and 564 documents in Scopus. In comparison, back in 2022, the figures were 78 documents on Scopus and 53 on Web of Science, illustrating a noticeable increase in research activity. However, despite these results, we found only 54 documents on Scopus and 29 documents on Web of Science that specifically focused on research related to data management.

As depicted in Figure 1, "data ingestion" emerges as a multi-disciplinary research topic, demonstrating a robust presence in computer science and engineering across both Web of Science and Scopus. Its application extends across various scientific and technological domains, underscoring its crucial role in the management, processing, and analysis of diverse datasets. The broad distribution across disciplines underscores the interdisciplinary nature of research on data ingestion.

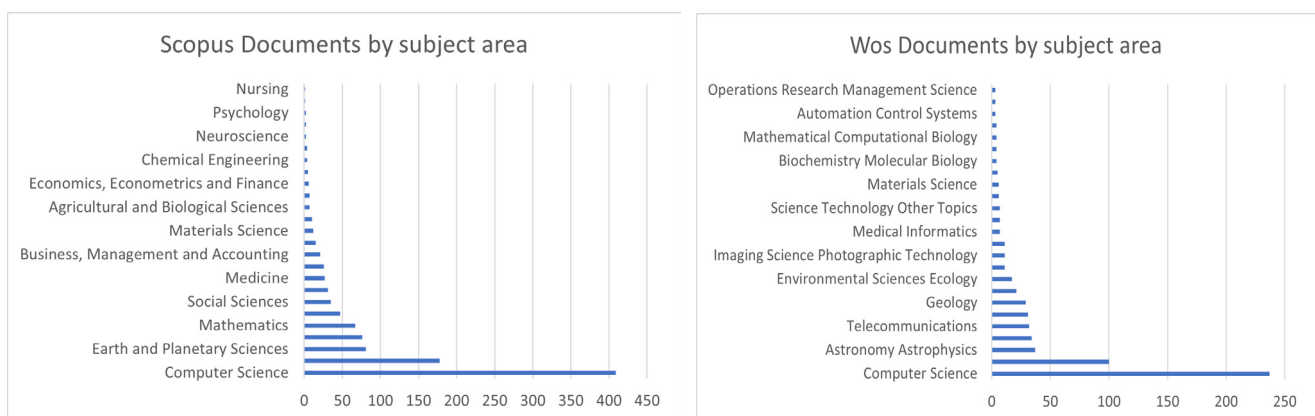


Figure 1. Scopus and Web of Science Documents by subject area

These findings signify a growing interest and recognition of the significance of data ingestion in research, with researchers and scholars from diverse fields actively contributing to the exploration and advancement of this crucial aspect of data management and analysis.

### Types and Classification

Data ingestion is the process of transferring data from various sources (databases, flat files, APIs, datasets, etc.) to a single storage location, where it can undergo further analysis.<sup>(5,2)</sup> This has been a longstanding issue addressed in previous research, covering technologies such as data integration, duplication, maintaining integrity constraints, and bulk data loading. Data ingestion processes encompass manual, semi-automatic, and

automatic methods, as indicated by <sup>(6)</sup>.

According to recent research by <sup>(7,8,9)</sup>, data ingestion can be classified into three types, predominantly dependent on the processing of big data:

- Batch processing: This model emerged to address the issue of Volume. It is used for processing stored and historical data, featuring scalability, distribution, parallelism, fault tolerance, high latency, and the utilization of large amounts of static data. It's important to note that in batch data ingestion, data is collected and transferred in batches at regular intervals.
- Real-time OR stream processing: This model is more geared towards resolving the issue of Velocity, storing only finalized results. It is characterized by low latency, continuous and unlimited data flow, distributed, parallel, and fault tolerance. Therefore, streaming data ingestion exclusively involves data collected in real-time (or near real-time), loaded almost immediately into the target location.
- Hybrid processing: This is a novel model implemented in 2014 to address both volume and speed concerns, providing a combination of batch processing and stream processing. It is primarily developed by two techniques: Lambda architecture <sup>(10)</sup> and Kappa Architecture.<sup>(11)</sup> This model is characterized by low latency, massive and streaming data, scalability, combining batch and real-time results.

The choice of the type of ingestion depends mainly on the type, structure, and format of the data collected, as shown in figure 2.

To date, there are few articles addressing with data ingestion issues. Typically, ETL is the conventional solution for data ingestion. However, to meet the emerging need for real-time data processing, <sup>(4)</sup> proposed a new layer adjacent to ETL called Data Collection. In this framework, the data collection layer employs the Kafka tool to ingest data, transferring it to the ETL layer for necessary transformations and processing before storing it in the data warehouse. Similarly, <sup>(12)</sup> used another tool called Apache Flume for data collection, representing the data ingestion phase. Another innovative framework presented by <sup>(13)</sup> uses two tools, Apache Kafka and Apache Flume, as an ingestion data layer to address the challenges of data heterogeneity and velocity.

The main challenges facing data processing architectures are manifold. With the appropriate data ingestion tool, one can instantly import, process and store data from various data sources.<sup>(12)</sup> A first category of tools comprises software that iteratively collects data through preconceived and industrialized tasks. Most of these tools are offered by the Apache Foundation, and can also be used to aggregate, convert and clean data before ingestion.<sup>(14)</sup> Many other data ingestion tools have been developed in recent years. Table 1 summarizes some tools cited in the work of.<sup>(15,16)</sup>

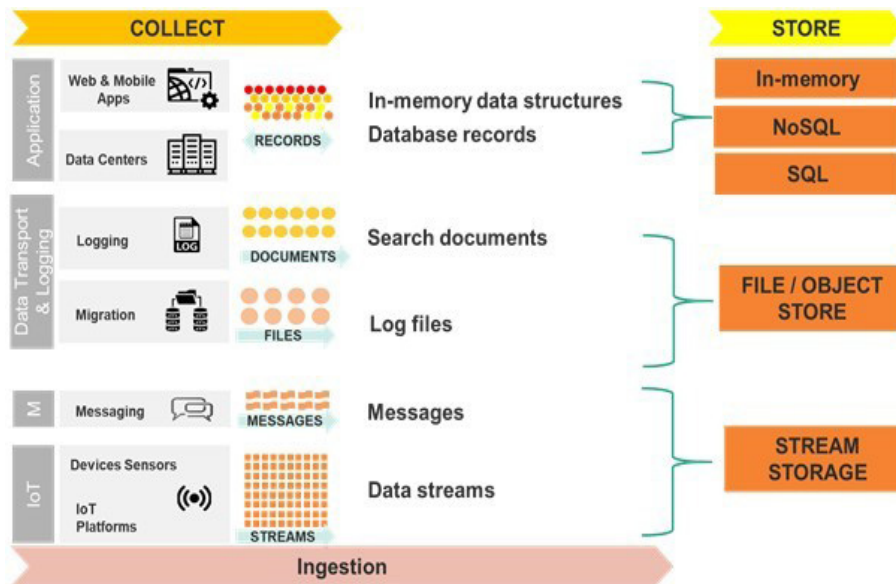


Figure 2. Ingestion types

Table 1. Data Ingestion Tools

Tool Name	Description
Apache Kafka	An open-source system for ingesting data from several sources in real-time; it can be used to collect high throughput parallel data for loading into Hadoop.
Apache Sqoop	An open-source system that can be used for transferring bulk data in both directions between relational databases, data warehouses, and HDFS or Hadoop data stores such as HBase or Hive.
Apache Flume	A distributed and reliable ingestion tool that can be used to collect, aggregate streaming data from many different sources and to push out the serialized data, using mechanisms called data sinks, to a centralized data store such as HDFS or HBase on Hadoop or Cassandra.

Astera Centerprise	A visual data management and integration tool to build bidirectional integrations, complex data mapping, and data validation tasks to streamline data ingestion.
Fluentd	A dedicated open-source platform for data ingestion. It unifies data in a data warehouse.
Elastic Logstash	An open-source data ingestion tool. It is a server-side data processing pipeline that ingests data from many sources, transforms it simultaneously, and then sends it to storage.
Apache Nifi	A data ingestion tool; it provides a powerful and reliable way to process and distribute information.

### Contemporary Challenges

As with all technologies and approaches, there are various issues associated with data ingestion <sup>(4,17)</sup> such as data cleansing (standard issues), data quality (find unnecessary data) and synchronization of data from multiple sources (whether to adopt synchronization of all data formats or integrated on a single storage medium). these issues can impact data ingestion and pipeline performance. Therefore, <sup>(18,19)</sup> have listed a set of challenges to be overcome in this area, which it is advisable to take into consideration and study as soon as possible:

- **Data quality:** The challenge arises when working with a multitude of variety of data sources. Problems are mainly related to the inconsistent data formats, data duplication and missing values, making the analysis unreliable. Therefore, data aggregation should be performed after proper data analysis and preparation.
- **Heterogeneous data:** Importing data involves combining data from different sources that were developed independently. Each source will have its own definitions, schemas, and data structure (tables, XML, unstructured text, etc.).
- **Sluggish processes:** With the increase of data volume and diversity. Writing codes for data ingestion and manual mapping for extracting and loading data has become tedious. The move towards the automation of data ingestion has become essential, especially since old methods of data ingestion are not fast enough to cope with the large volume and variety of data sources.
- **Cost:** the cost of the infrastructure as well as the cost of the proposed ingestion tools make the task of ingesting data very expensive.
- **Legacy data:** There is still important data that is stored in an old form, such as spreadsheets and ad-hoc structures. It is challenging to combine this data with modern data structure like XML.
- **Unreliability:** Incorrect data ingestion can lead to unpredictable connectivity issues, disrupting communication and leading to data loss.
- **Security:** Data security is a significant challenge when data is moved, as it is distributed in multiple phases throughout the ingestion process. Meeting security standards during the data ingestion process requires the introduction of advanced security techniques, such as data encryption and anonymous access to sensitive information.

In light of all this research, data ingestion emerges as a technology with a double-edged sword. Researchers often overlook data ingestion as a genuine research design, merely providing definitions and selecting tools for their pipelines based on their specific needs. Nonetheless, data ingestion plays a pivotal role in the collection and import of data, serving as the foundation for any data pipeline. To unearth potential issues stemming from such a data process, engaging in practical experimentation is necessary. This involves working with heterogeneous data of varying sizes using diverse tools, and perhaps even formulating new data ingestion processes.

Additionally, in the previous challenges set, researchers do not address the evolution of data as an issue to contend with. However, considering the evolution of ingested data as a solution for this problem can optimize the alignment process.

### Exploring data alignment

Among the advanced techniques explored by the scientific community to achieve data alignment, schema matching takes a central role. This approach aims to identify objects that are semantically and/or contextually linked, allowing for the establishment of correspondences between concepts from different data sources within a given data space, characterized by schema heterogeneity. In the specialized literature, researchers in this field have defined several objectives related to the use of schema matching, including data exploration, discovering joins between datasets, enriching the data space, and aligning data. <sup>(20,21,22,3)</sup>

Through research on Scopus and Web of Science using the keyword "Schema Matching" we found 1040 and 555 documents, respectively, on Scopus and Web of Science. The years 2009 and 2008 represent the peaks of research in this field. The primary objective of alignment is to preserve ingested data in their original states and establish connections between them. This implies avoiding the integration of all data into a centralized database (e.g., a warehouse) with a single schema. This approach proves particularly relevant given the dynamic nature of data spaces, where data schemas evolve. Consequently, a new challenge arises - keeping alignments updated as data schemas evolve.

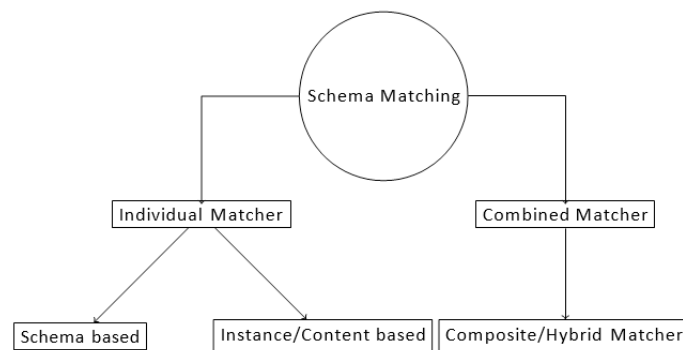
While each solution implements different algorithms, they share common foundations for the most part. Indeed, a set of criteria and information at our disposal will help us achieve automatic matchings, although exploiting this information often requires additional processing. Here is a brief overview of the techniques used to leverage the main information derived from the data that matching algorithms typically employ.

**Common Approaches to Schema Matching**

Alignment entails creating connections between diverse data schemas, often employed in data discovery and knowledge graph applications.<sup>(23)</sup> Additionally, the alignment principle can function as a preliminary step in data integration. In the realm of this approach, previous works include ontology alignment<sup>(20)</sup> and the development of data dictionaries and catalogs.<sup>(24,25)</sup> Schema matching stands out as the most prevalent technique for aligning data.

Schema matching is a method used to identify correspondences between concepts from different distributed and heterogeneous data sources. It is considered a fundamental operation for schema integration and data processing. Moreover, schema matching finds extensive applications in various domains such as databases, web applications, and more.<sup>(20,21)</sup>

There are various techniques for schema matching, such as linguistic matching, instance-based matching, structure-based matching, constraint-based matching, hybrid matching, and rule-based matching.



**Figure 3.** Hierarchical structure of schema matching approaches

According to the hierarchical structure depicted in figure 3, schema matching methods can be classified into two main types, as inspired by the original structure.<sup>(21)</sup> These categories are as follows:

- **Individual Matcher:** This category includes matchers that are either instance-based or schema-based. For instance, the Cupid algorithm<sup>(26)</sup> is a schema-based matcher that constructs tree structures to represent schema elements, using a weighted combination of linguistic and structural matching to determine the similarity between elements.
- **Combined Matcher:** This category comprises matchers that integrate both schema and instance information. Such matchers may use a composite or hybrid approach and can operate either automatically or with manual intervention to optimize matching accuracy.

Based on this classification, we have organized some schema matching solutions in table 2.

Paper	Instance Based	Schema based	Combined Matcher	Description
Cupid <sup>(26)</sup>		X		Cupid employs a schema-based approach, focusing on matching schema elements using linguistic and structural properties.
Coma++ <sup>(20)</sup>	X	X		Coma++ is a combined matcher using both instance-based and schema-based techniques for efficient schema matching.
SemProp <sup>(27)</sup>			X	SemProp use a combined approach with a focus on semantic properties for aligning schemas.
AlMatch <sup>(28)</sup>			X	AlMatch is centered around artificial intelligence methods, using a combined matcher approach for dynamic schema alignment.
SFIMatch <sup>(33)</sup>	X			SFIMatch a schema-free instance matching algorithm based on virtual document similarity, emphasizing data instance utility in schema matching.

### Recent Advances in Schema Matching

In the last few years, the field of schema matching has seen notable progress, driven by the infusion of deep learning and attention-based models into traditional matching processes. Researchers have begun using deep neural networks to adjust similarity matrices, thus refining the matching outcomes without human intervention.<sup>(29,30)</sup> The application of attention mechanisms, as demonstrated by<sup>(31)</sup>, provides a deep learning solution to automate schema matching, particularly beneficial in domains like healthcare where data privacy is paramount.

Furthermore, holistic matching approaches such as hMatcher<sup>(32)</sup> have shown promise in achieving high matching accuracy by employing context-based semantic similarity measures.

Another significant advancement has been the use of crowdsourcing for schema matching.<sup>(34)</sup> By leveraging collective knowledge, it has become possible to further improve the precision of schema matching by integrating analyses of datasets conducted by individuals capable of understanding them. In addition, major progress has also been achieved in the domain of matching semi-structured data, notably NoSQL, which now exhibit better quality despite the limited information that can be extracted from data structures.<sup>(35)</sup>

### Schema Matching Challenges

As schema matching technology progresses, it faces new challenges due to the huge amount of data and the evolution of data environments. These challenges are being addressed by recent studies, which explore the integration of machine learning and deep learning:

- **Integration of Advanced Learning Techniques:** With the rise of deep learning in schema matching, there are complexities such as the need for extensive training data and managing model intricacies. Recent research by<sup>(31)</sup> addresses these challenges by using state-of-the-art natural language processing techniques to obtain semantic mappings between source and target schemas. In addition,<sup>(36)</sup> have addressed these challenges by studying the problem of schema mapping from XML to RDF, suggesting a reconfigurable pipeline for semi-automatic schema matching (REPSASM) that offers an environment where users can experiment with their own schema-matching procedures.
- **Scalability and Efficiency:** The growth of databases has increased the demand for scalable schema matching. The Valentine system,<sup>(26)</sup> for instance, facilitates the organization and execution of large-scale matching experiments, which is vital in today's data-rich environment. On the other hand, a dynamic graph framework introduced in<sup>(37,40,41,42,43)</sup> models the interaction between natural language utterances and database schemas, potentially advancing the handling of evolving schemas. There is still a need for research to perform incremental schema matching on dynamic data. This is because schema matching can be particularly costly when dealing with multiple and complex data sources, requiring a global recalculation each time one of the input schemas is modified.
- **Handling Noisy and Incomplete Data:** Incomplete and noisy data present significant hurdles to accurate schema matching. Research such as that presented in<sup>(38)</sup> and<sup>(30)</sup> provides insights into addressing these data imperfections, thereby enhancing the interpretability and accuracy of schema matching.
- **Human's Role in Schema Matching:** The extent of human involvement in schema matching processes remains a topic of debate. A recent study by<sup>(39)</sup> discusses the changing roles of humans and machines in schema matching, highlighting the impact of big data and machine learning advancements.

Each of these challenges points towards an aspect of schema matching that is currently under active development. The integration of human cognitive models, as well as advanced machine learning and deep learning techniques, represents the forefront of research in this area. As schemas become more complex, voluminous and dynamic, these challenges will only become more critical to address.

### CONCLUSION

In conclusion, the evolving landscape of information and data has led to transformative changes across scientific, economic, political, and social domains. The increasing interest in Big Data emphasizes the crucial role of efficiently processing diverse data types. From the historical evolution of ETL processes to the contemporary challenges and improvements in data integration, the importance of well-defined processes for data collection, storage, and exploitation is evident.

The focus on data ingestion, particularly in the context of Big Data, has seen a surge in research activity, with multidisciplinary contributions spanning computer science and engineering. The classification of data ingestion into batch processing, real-time/stream processing, and hybrid processing reflects the evolution of techniques to address challenges such as data quality, heterogeneous data.

As far as data alignment is concerned, schema matching emerges as a central technique to create connections between semantically linked objects in diverse data sources. The classification of schema matching solutions into individual and combined matchers provides insights into the approaches employed. Recent advances, including the adoption of deep learning, incremental matching, crowdsourcing, and handling semi-structured

data, showcase the dynamic nature of this field.

Looking ahead, addressing the ever-evolving needs of schema matching involves the integration of advanced learning techniques, addressing domain-specific challenges, improving matching quality, and exploring scalability considerations and incremental schema matching to cope with dynamic data environments. Challenges such as data quality, heterogeneity, cost, legacy data, unreliability, and security persist and require careful consideration in ongoing research endeavors in data management and analysis. The integration of schema matching with instance matching continues to be a focus area, presenting opportunities for advancements in matching quality and speed. The upcoming evolution of schema matching is anticipated to revolve around efficiently managing the dynamism of evolving data schemas.

## REFERENCES

1. Souibgui M, Atigui F, Zammali S, Cherfi S, Yahia SB. Data quality in ETL process: A preliminary study. *Procedia Computer Science* [Internet]. 2019;159. Available from: <https://doi.org/10.1016/j.procs.2019.09.223>
2. Informatica [Internet]. [cited 2023 Oct 18]. What Is Data Ingestion? Available from: <https://www.informatica.com/resources/articles/what-is-data-ingestion.html>
3. Alserafi A. Dataset Proximity Mining for Supporting Schema Matching and Data Lake Governance [PhD Thesis]. Universitat Politècnica de Catalunya, BarcelonaTech; 2021.
4. Meehan J, Tatbul N, Aslantas C, Zdonik S. Data Ingestion for the Connected World. In: *CIDR'17*. 2017.
5. Hoseini S, Ali A, Shaker H, Quix C. SEDAR: A Semantic Data Reservoir for Heterogeneous Datasets. In: *32nd ACM International Conference on Information and Knowledge Management* [Internet]. ACM; 2023. p. 5056-60. Available from: <https://doi.org/10.1145/3583780.3614753>
6. Yihun AM, Stanislava S. Learning analytics for higher education: proposal of big data ingestion architecture. *SHS, Web of Conferences* [Internet]. 2021; Available from: <https://doi.org/10.1051/shsconf/20219202002>
7. Giebler C, Stach C, Schwarz H, Mitschang B. BRAID - A Hybrid Processing Architecture for Big Data. In: *7th International Conference on Data Science, Technology and Applications*. SCITEPRESS - Science and Technology Publications; 2018.
8. Miloslavskaya N, Tolstoy A. Big Data, Fast Data and Data Lake Concepts. *Procedia Computer Science*. 2016;88.
9. Pal G, Li G, Atkinson K. Big Data Ingestion and Lifelong Learning Architecture. In: *2018 IEEE International Conference on Big Data, Big Data 2018*. 2018.
10. Marz N, Warren J. *Big data: principles and best practices of scalable real-time data systems*. Shelter Island, NY: Manning; 2015.
11. Kreps J. Questioning the Lambda Architecture [Internet]. 2014. Available from: <https://www.oreilly.com/radar/questioning-the-lambda-architecture/>
12. Podhoranyi M. A comprehensive social media data processing and analytics architecture by using big data platforms: a case study of twitter flood-risk messages. *Earth Sci Inform*. 2021;14:913-29.
13. Pal G, Atkinson K, Li G. Managing Heterogeneous Data on a Big Data Platform: A Multi-criteria Decision Making Model for Data-Intensive Science. In: *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. 2020.
14. Sawadogo P, Darmont J. On data lake architectures and metadata management. *Journal of Intelligent Information Systems*. 2021;56(1).
15. Sharjeel A. What is Data Ingestion: Process, Tools, and Challenges Discussed [Internet]. 2020. Available from: <https://dataintegrationinfo.com/what-is-data-ingestion/>
16. Armoogum S, Li X. Big Data Analytics and Deep Learning in Bioinformatics With Hadoop. In: *Deep Learning*

and Parallel Computing Environment for Bioengineering Systems. Elsevier; 2019.

17. Ahmed H, Mun J, Park Y, Choi J. A schema generator for collected data from wearable devices for reliable data ingestion. In: ACM International Conference Proceeding Series. 2019.

18. Abdallah ZS, Du L, Webb GI. Data Preparation. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer US; 2017.

19. Naeem T. Data Ingestion - Definition, Challenges, and Best Practices [Internet]. 2020. Available from: <https://www.astera.com/type/blog/data-ingestion/>

20. Aumueller D, Do H, Massmann S, Rahm E. Schema and ontology matching with COMA++. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data. 2005.

21. Bernstein PA, Madhavan J, Rahm E. Generic schema matching, ten years later. Proceedings of the VLDB Endowment. 2011;4(11):695-701.

22. Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. The VLDB Journal. 2001;10(4):334-50.

23. Auza-Santiv  n JC, D  az JAC, Cruz OAV, Robles-Nina SM, Escalante CS, Huanca BA. Bibliometric Analysis of the Worldwide Scholarly Output on Artificial Intelligence in Scopus. Gamification and Augmented Reality 2023;1:11-11. <https://doi.org/10.56294/gr202311>.

24. Castillo JIR. Aumented reality im surgery: improving precision and reducing risk. Gamification and Augmented Reality 2023;1:15-15. <https://doi.org/10.56294/gr202315>.

25. Castillo-Gonzalez W, Lepez CO, Bonardi MC. Augmented reality and environmental education: strategy for greater awareness. Gamification and Augmented Reality 2023;1:10-10. <https://doi.org/10.56294/gr202310>.

26. Aveiro-R  balo TR, P  rez-Del-Vall  n V. Gamification for well-being: applications for health and fitness. Gamification and Augmented Reality 2023;1:16-16. <https://doi.org/10.56294/gr202316>.

27. Chaudhri V, Baru C, Chittar N, Dong X, Genesereth M, Hendler J, et al. Knowledge Graphs: Introduction, History and Perspectives. AI Magazine. 2022;43(1):17-29.

28. Ehrlinger L, Schrott J, Melichar M, Kirchmayr N, W  b W. Data Catalogs: A Systematic Literature Review and Guidelines to Implementation. In: DEXA 2021 Workshops, Communications in Computer and Information Science. Springer International Publishing, Cham; 2021. p. 148-58.

29. Diamantini C, Giudice PL, Musarella L, Potena D, Storti E, Ursino D. A New Metadata Model to Uniformly Handle Heterogeneous Data Lake Sources. In: New Trends in Databases and Information Systems. Springer International Publishing, Cham; 2021. p. 165-77.

30. Koutras C, Siachamis G, Ionescu A, Psarakis K, Brons J, Fragkoulis M, et al. Valentine: Evaluating Matching Techniques for Dataset Discovery. In: IEEE 37th International Conference on Data Engineering (ICDE). 2021.

31. Castro Fernandez R, Mansour E, Qahtan AA, Elmagarmid A, Ilyas I, Madden S, et al. Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE). IEEE; 2018.

32. H  ttasch B, Truong-Ngoc M, Schmidt A, Binnig C. It's AI Match: A Two-Step Approach for Schema Matching Using Embeddings [Internet]. 2022. Available from: <http://arxiv.org/abs/2203.04366>

33. Shraga R, Gal A, Roitman H. ADnEV: Cross-Domain Schema Matching using Deep Similarity Matrix Adjustment and Evaluation. Proceedings of the VLDB Endowment. 2020;13(9):1401-15.

34. Cappuzzo R, Papotti P, Thirumuruganathan S. Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks. In: 2020 ACM SIGMOD International Conference on Management of Data. 2020.



35. Zhang J, Shin B, Choi JD, Ho JC. SMAT: An Attention-Based Deep Learning Solution to the Automation of Schema Matching. In: *Advances in Databases and Information Systems*. Springer International Publishing; 2021. p. 260-74.
36. Yousfi A, Yazidi M, Zellou A. hMatcher: Matching Schemas Holistically. *International Journal of Intelligent Engineering and Systems*. 2020;13:490-501.
37. Amrouch S, Mostefai S. A Schema-Free Instance Matching Algorithm Based on Virtual Document Similarity. *The International Arab Journal of Information Technology*. 2022;19(3A).
38. Zhang CJ, Chen L, Jagadish HV, Zhang M, Tong Y. Reducing Uncertainty of Schema Matching via Crowdsourcing with Accuracy Rates [Internet]. 2018. Available from: <http://arxiv.org/abs/1809.04017>
39. Amghar S, Cherdal S, Mouline S. A Schema Integration Approach for Big Data Analysis. *Ingénierie Des Systèmes d'Information*. 2023;28(2).
40. Liao X, Bottelier J, Zhao Z. A Column Styled Composable Schema Matcher for Semantic Data-Types. *Data Sci J*. 2019;18(25).
41. Hui B, Geng R, Ren Q, Li B, Li Y, Sun J, et al. Dynamic Hybrid Relation Network for Cross-Domain Context-Dependent Semantic Parsing [Internet]. 2021. Available from: <https://arxiv.org/abs/2101.01686>
42. Zhou Q, Liu X, Wang Q. Interpretable duplicate question detection models based on attention mechanism. *Information Sciences*. 2021
43. Gal A, Shraga R. Human's Role in-the-Loop [Internet]. arXiv; 2022, Available from: <http://arxiv.org/abs/2204.14192>

#### **FINANCING**

No financing.

#### **CONFLICT OF INTEREST**

The authors declare that there is no conflict of interest.

#### **AUTHORSHIP CONTRIBUTION**

*Conceptualization:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Data curation:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Formal analysis:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Acquisition of funds:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Research:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Methodology:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Project management:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Resources:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Software:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Supervision:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Validation:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Display:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Drafting - original draft:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.

*Writing - proofreading and editing:* Oumaima El Haddadi, Max Chevalier, Bernard Dousset, Ahmad El Allaoui, Anass El Haddadi, Olivier Teste.