Data in Brief 56 (2024) 110853

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Heterogeneous data integration: Challenges and opportunities

I Made Putrama[a,b,*], Péter Martinek[a]

[a] Department of Electronics Technology, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Budapest, Hungary
[b] Department of Informatics, Faculty of Engineering and Vocational, Universitas Pendidikan Ganesha, Singaraja, Indonesia

ARTICLE INFO

Article history:
Received 22 June 2024
Revised 21 July 2024
Accepted 13 August 2024
Available online 29 August 2024

Keywords:
Big data
Integration
Ontology
Heterogeneous
Data sources
Review

ABSTRACT

Integrating multiple data source technologies is essential for organizations to respond to highly dynamic market needs. Although physical data integration systems have been considered to have better query processing systems, they pose higher implementation and maintenance costs. Meanwhile, virtual data integration has become an alternative topic that is increasingly attracting the attention of researchers in the current era of big data. Various data integration methodologies have been developed and used in various domains, processing heterogeneous data using various approaches. This review article aims to provide an overview of heterogeneous data integration research focusing on methodology and approaches. It surveys existing publications, highlighting key trends, challenges, and open research topics. The main findings are: (i) Research has been conducted in various domains. However, most focus on big data rather than specific study domains; (ii) researchers primarily focus on semantics challenges, and (iii) gaps still need to be addressed and related to integration issues involving semantics and unstructured data formats that must be thoroughly addressed. Furthermore, considering elements of cutting-edge technology, such as machine learning and data integration, about privacy concerns provides a chance for additional investigation. Finally,

* Corresponding author at: Department of Electronics Technology, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Budapest, Hungary.
E-mail addresses: putrama.imade@edu.bme.hu (I.M. Putrama), martinek.peter@vik.bme.hu (P. Martinek).
Social media: @i_putrama (I.M. Putrama)

https://doi.org/10.1016/j.dib.2024.110853

we provide insight into the potential for a broader review of integration challenges based on case studies.

1. Introduction

In today's big data era, the massive growth of information that can be obtained from various sources necessitates massive storage space and innovative methodologies and intelligent tools to understand the relationship between the properties of the data for extraction. This information will undoubtedly be essential for enterprises to continue providing innovative services and respond quickly to the needs of their customers by making decisions based on relevant data. For example, in the health sector, besides utilizing current evidence with limited information, health practitioners can make assessments based on historical data to provide patients with more effective and efficient solutions [1]. However, data has evolved from a limited amount of structured data to an abundance of various forms; thus, integrating it for analytical purposes becomes a challenging task. Because of the probability of information description and its diversity of features, conventional techniques such as entity mapping or relationship records via object references may no longer be efficient. Hence, new approaches and methods are required.

Data sources are often designed alongside specific applications, causing difficulties in collecting, integrating, reconciling, and extracting the data from these systems. It is even more problematic when there are redundant, duplicated, mixed, and incomplete data sets concerning particular business requirements [2]. Collecting and integrating such heterogeneous data has been regarded as a significant challenge, mainly due to the unstructured data and variety of data sources [3–5]. The heterogeneity of the data creates integration problems due to the dimensions of its volume, variety, and velocity.

The big data volume has quantities larger than conventional relational data, ranging from gigabytes to terabytes, petabytes, and even more. It has diversity in the form of structured, semi-structured, or unstructured formats from dispersed sources due to inconsistent data models and data content. In terms of velocity, it is generated at unprecedented speed, which must be dealt with promptly [6,4].

A generally challenging problem is when an enterprise has various information systems, including the diversity, autonomy, and distribution of data sources [7]. It is essential to integrate them into a comprehensive and unified system with a data set that is clean, accurate, and unduplicated so that information can be extracted easily for decision-making purposes. In the education sector, universities often maintain multiple separate information systems to manage different aspects of their operations. These systems are dedicated to specific functions such as tracking teacher and student activities, managing extracurricular events, conducting tracer studies, and handling administrative services. To facilitate decision-making, data from these systems must be well integrated, easy to extract, and ideally reflect real-time changes.

There are two types of data integration systems: *physical* and *virtual* data integration [8]. Physical data integration is generally more efficient in query processing because data from various sources will be copied into one unified data source, becoming the primary reference for data analysis. However, this system is expensive because it requires separate implementation and maintenance, and there is no guarantee that the data remains up-to-date. Additional processes are required to ensure the data is updated regularly. Examples of systems like this include data warehouses, data lakes, polystores, and data tamers. Meanwhile, the *virtual* data integration system is seen as more cost-effective and flexible because there is no need to create a separate data system or perform data transfer or replication. The data remains in the original source; the query will be processed through a query translation by a mediator, which will then execute it

according to query language specific to each data source as needed. However, the implementation becomes more complex, especially with various data heterogeneity issues that make it challenging to implement.

Many data integration techniques have been proposed over the years across various research focuses and domains using one or more combined techniques, including the use of Extract, Transform, and Load (ETL), Distributed or Federation databases, approaches based on ontology or semantic association, and graph-based solutions (see Table 1). These techniques mainly consist of a cooperative information system with a mediator and wrapper where the constructed mediator coordinates the data flow between the local source and the user application. In contrast, the wrapper will form a sort of mapping or directly interact with the local data source based on the queries built by the mediator [2]. In addition, information fusion is also one of the most common approaches used to integrate heterogeneous data sources. However, similar data from different sources can lead to semantic, structural, and syntactic heterogeneity issues. Meanwhile, because there is no common language, the system integration architecture must rely on custom queries to extract data from each of the relevant data sources [9]. To minimize this problem, researchers use an ontology-based data integration approach, which has been proven to have great potential to solve the heterogeneity problem among various data sources at the semantic level. This technique works in such a way that it can provide a joint knowledge base for different domain users. Naming and semantic conflicts that may be encountered are eliminated when the ontology is accepted as a common reference point by the system or other users who need it [10]. However, there is a fundamental problem with this approach. Building an ontology from scratch has been realized to be an expensive task. It requires a lot of domain expertise due to the semantic relationships in the data, which are generally not directly represented but are hidden due to assumptions that may be drawn and embedded when the system is being developed. This issue becomes even more difficult if schema changes also need to be considered, which require incorporation and proper handling of ontology evolution effectively and efficiently [10–12]. Additionally, integrating data with descriptions and various properties through traditional methods such as schema mapping, document association using entity references, and entity correspondence with unstructured data remains a significant challenge in the era of big data [5].

Despite the range of approaches discussed earlier, few studies have comprehensively reviewed existing solutions to extract the essence of issues, methodologies, and findings in a broader context. Existing reviews concentrate on specific aspects or domains, offering insights within a relatively limited scope (as shown in Table 2). For example, the review in [17] focuses on ontology-based integration in road asset management, [10] reviewed data integration with approaches for ontology/schema evolution, [18] reviewed the integration based on mapping relational databases using ontology and [19] specifically reviewed data integration related to manufacturing. Yet, many of these studies need a comprehensive perspective on the overall data integration landscape. Moreover, the abundance of current data necessitates different approaches, as traditional methods may have been outdated. Understanding the current research trends and approaches being actively investigated is crucial. Therefore, our study reviews previous research on integrating heterogeneous data in a general context with various aspects and domains in the form of a systematic review article. The aim is to fill existing gaps in the literature by providing a broader overview of data integration research, specifically focusing on the research domains undertaken, the approaches proposed by the authors, and potential areas for further investigation.

The rest of this review article is organized as follows: Section 2 explains the applied research methodology, Section 3 presents the review results, Section 4 discusses our findings, Section 5 provides additional analysis for the milestones and the most influential publications, and finally, Section 6 concludes the work with possible future research options included.

Table 1
Various data integration approaches in previous works.

Work	Approach	Contributions
Marek, M. [13]	Utilizes Extract, Transform, Load (ETL) processes to effectively integrate data from diverse sources.	• The study states that the ETL processes can significantly streamline the integration of heterogeneous data, improving data accessibility and usability. • The complexity and resource intensity of ETL processes, mainly when dealing with real-time data integration and large datasets, may require substantial computational power and time.
Ahamed, B. et al. [9]	It discusses techniques for data integration with a focus on integrating data at different levels by resolving inconsistencies and applying independent or unified queries over the available data.	• The study suggests improving the shortcomings and scope of data integration by enhancing the wrapper, optimizing the Query Optimization engine, and detecting and eliminating duplicates in the query process • It highlights challenges in incorporating heterogeneity and conflicts in data integration, the unavoidable delay associated with retrieving data from various data sources
Fusco, G. et al. [2]	Uses ontologies to represent both data sources and global schema	• The proposed approach allows unified access to heterogeneous data sources through a semi-automatic mediation system that creates a virtual view of the data. However, it does not support unstructured data • More research is needed to validate the quality of mapping views and optimize the mediation processes
Asfand-E-Yar, M. et al. [14]	Uses Ontology modeling algorithm, SPARQL query language, Query translation algorithm	• The main findings include integrating structural heterogeneous databases without retranslating databases or redesigning the existing structure, the advantages and disadvantages of the Data Translation Scheme, and the ease of integrating other databases using the proposed methodology. • Further research and improvement in the automation and integration process are still required, as well as a focus on integrating structured and semi-structured/non-structured databases. The study also suggests the potential benefit of integrating Machine Learning algorithms with Semantic Web Ontology to improve the integration process.
Nashipudimath, M. M. et al. [15]	Utilizes feature patterns and semantic analysis to improve the integration and indexing of large datasets.	• The proposed method enhances data retrieval efficiency and accuracy, offering improved indexing that supports better data integration and management in big data environments. • The index construction for query processing needs to be optimized to minimize the index column storage
Kolisetty, V. V. et al. [4]	Uses Probabilistic Semantic Association for data integration and classification for uncertain and unstructured data	• The proposed approach has been found effective in automating the semantic map creation and enhancing the accuracy of mapping for incoming data, which improves the data integration • The PSA employs a linear search algorithm, which tends to be slow for a higher number of records and, therefore, leads to a drop in data classification accuracy

(*continued on next page*)

Table 1 (*continued*)

Work	Approach	Contributions
Kolisetty, V. V. et al. [5]	Computes attribute conditional dependency and similarity index to enhance big data integration	• The study proposes an Attribute Conditional Dependency – Similarity Index (ACD-SI) method for integrating structured and unstructured collection datasets, which shows effectiveness compared to other classification methods and demonstrates better integration accuracy. • The approach may face challenges when scaling to huge datasets. Handling conditional dependencies between attributes may require extensive computational resources, which can adversely impact the accuracy of the integration results.
Vasiliev, D. A. et al. [16]	The study compares two data integration approaches, one using a mediator solution based on GraphQL and another using a unified GraphDB repository solution.	• The study suggests that GraphQL and GraphDB solutions enable clients to specify the data they need, but GraphQL is much more limited. However, it has the advantage of not requiring additional storage capacity. • GraphDB provides a more efficient solution as it is better for response time. Additionally, the Resource Description Framework (RDF) graph repository used in this solution allows semantic data interoperability. It can store vocabularies or ontologies that define the schema among various data sources.

Table 2
Literature reviews on data integration within narrower aspects/domains.

Work	Aspects/domains	Contributions
Kondylaksis, H. et al. [10]	Ontology-based data integration	• The study investigates the challenges of handling dynamic ontologies in data integration systems. • It outlines the characteristics of a data integration system capable of handling ontology evolution effectively.
Lei, X. al. [17]	Ontology-based data integration in road asset management	• The study investigates and highlights the road asset management where ontologies have been used. • It analyzes existing studies regarding ontology engineering in road asset management.
Haw, S. C. et al. [18]	Relational database mapping to ontology representation	• The study describes the general rules of transforming relational databases to ontology. • It discusses and compares various conversion tools such as RDBToOnto, Polfiet and Ichise, StdStrip+*K*, and BootOX.
Ferrer, B. R. et al. [19]	Ontology and database comparison in manufacturing	• The study develops a methodology to compare database and ontological models by seeing how data volume affects performance. • It analyzes how databases and ontologies complement each other when they exist together in a system.

2. Review methodology

This review article follows a structured methodology proposed by [20,21] to present a literature review. The method defines stages to conduct a review, including *plannining, conducting*, and *reporting* the review. Specifically, our literature review approach involves the following steps: (i) defining the scope and research questions, (ii) establishing inclusion and exclusion criteria, (iii) conducting the literature search using predefined methods, and (iv) reporting the survey results.

2.1. Scope and research questions

This review article summarizes the research on data integration and connecting technical solutions pursued by the researchers. The related questions that motivate this work include:

RQ1: What is the rising research focus related to data integration?
Given the problem of data extraction, integration, and analysis, where data itself accumulates very quickly in various forms and is supported by increasingly sophisticated technology, knowing the focus of the researchers' studies and general questions from the research carried out will help in exploring gaps that can be investigated in the future.

RQ2: What approaches are taken by researchers?
Identifying the various approaches that have been carried out will provide an overview of the latest methods and their limitations.

RQ3: What are the open questions regarding the heterogeneous data integration from studies carried out?
In addition to the knowledge obtained from answering the first question, the research results must have reached several significant findings. However, it is even more important to know what questions remain open according to the studies.

2.2. Search criteria

There are several categories of literature source search criteria used for inclusion. These are filtered later based on exclusion criteria to produce a list of relevant literature ready for analysis.

Inclusion Criteria We specifically include the following publication criteria for review:

- The data collected for review is searched and gathered electronically from Google Scholar, IEEE Xplore, ACM, ScienceDirect, Springer, and other relevant publication websites. We mainly selected publications with full text for this review article.
- The search categories include subjects such as Computer Engineering, Computer Science, and Information Science.
- The articles were published between 2012 and 2023 and explicitly referred to terms such as "integration" or "data integration" in their titles, abstracts, or main text.
- Only publications in English have been selected.

Exclusion Criteria Any publications that fall within the criteria below have been excluded from the review list:

- Publications about big data, the Internet of Things, or engineering that do not discuss data integration as one of the research focuses.
- Publications that emphasize application integration rather than data integration.

The literature review conducted on this study established inclusion and exclusion criteria to ensure the relevance and quality of the selected publications. Inclusion criteria focused on publications from reputable journals with full-text availability, ensuring access to comprehensive and credible information. Targeting fields such as Computer Engineering, Computer Science, and

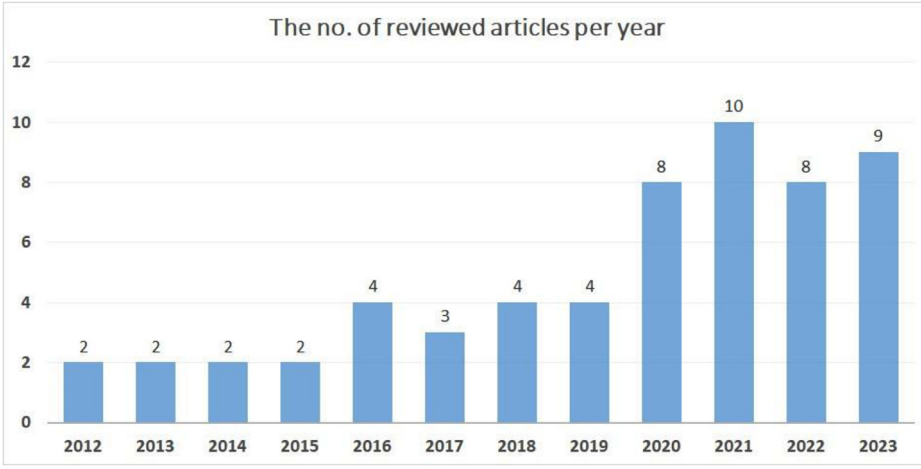

Fig. 1. The distribution of the collected articles.

Information Science, this review covers a broad but relevant scope of literature. Limiting the publication period to a fairly wide range ensures the incorporation of contemporary research and trends while limiting English language publications ensures accessibility and understanding. This study excluded articles on big data or IoT techniques that did not discuss data integration to prevent the review from straying into unimportant topics. Likewise, excluding publications that focus more on application integration than data integration ensures that the review remains focused on areas that only address data integration, even when related to big data. This selection process aims to increase the validity and applicability of the literature review findings.

2.3. Search method

Considering the citation analysis, particularly for STEM-related sources studied by [22,23], our initial review involved collecting 118 articles from pertinent publication databases by adhering to the inclusion criteria. Subsequently, we conducted a screening process to eliminate non-English publications and studies that did not specifically address topics as outlined in the exclusion criteria. As a result, 58 articles were included in our review analysis. To visualize the distribution of the publications collected, please refer to Fig. 1. The results are subsequently discussed in the following section, providing an overview of studies on data integration and addressing the predefined research questions.

3. Review results

Based on the collected publications, three research questions will be answered in this section. Firstly, we discuss the increasing research focus on data integration, addressed in Section 3.1 under research question RQ1. Subsequently, we cover the approaches used by researchers (RQ2) in Section 3.2 and summarize the open research questions discussed by researchers (RQ3) in Section 3.3.

3.1. Heterogeneous data integration research focus (RQ1)

The selected articles show their distribution based on the problem domains (Fig. 2) and the form of the processed integrated data (Fig. 3). Most of the domains are related to enter-

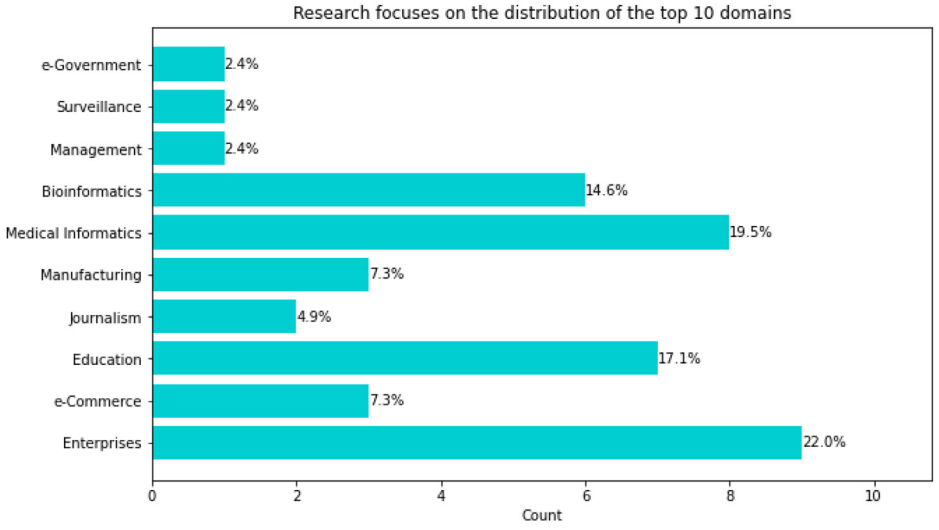

Fig. 2. The ten most significant problem domains studied in data integration.

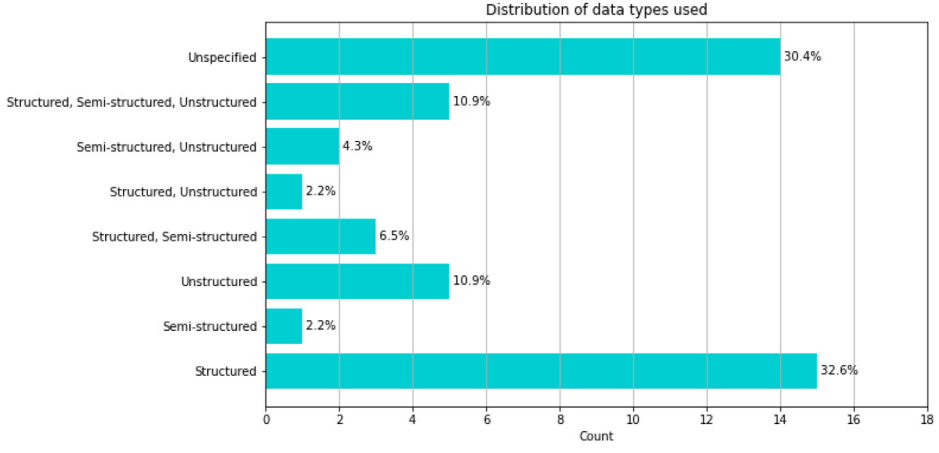

Fig. 3. The type of input data under examination in the study of data integration.

prises [24–27], medical informatics [2,28–31], education [14,32–35], e-commerce [12,18,36,37], and manufacturing [13,19,38]. When viewed from the distribution of data format processed during integration, as much as 32.6 % of data integration is still focused on structured data integration. This is based on the fact that many systems are still running using relational databases. Only some studies process semi-structured, unstructured data, or a mixture of both, while as many as 30.4 % of studies need to explicitly state the form of the data used in the research. In terms of the research's focus, the literature is grouped into three categories, namely integration related to semantics (52.38 %), physical data integration, integration issues/challenges (30.95 %), and big data integration (16,67 %), as seen in Table 3.

In the first group study, the researchers focus on translating or structuring user queries into standard database queries so that users will perceive that they are accessing a single data source rather than multiple data sources. There are also studies with a narrower focus, such as overcoming the time-consuming process of manually creating ontologies using ontology learning

Table 3
The researchers' research focus.

Group	Research focus
Semantics-related integration	Ontology-related data integration [2,17,39], Semi-auto ontology creation using ontology learning [40], Relational data to ontology representation [14,18,41], Ontology-based big data integration [3,6,42], Ontology and database comparison [19], Graph-based data integration [16,43–46], Automated ontology mapping and merging concept [47,48], Ontology-based Datawarehouse design [36], Relational Database (RDB) to RDF transformation [49], Ontology learning for relational database transformation [12], Open data integration [50], Machine Learning-based data integration [27,30,51–56].
Physical data integration, issues/challenges	SOA-based data integration [33,57,58], Metadata-based data integration [59], Schema matching [24,32,38,60], Feature extraction [13,61], Integration of dynamically changing heterogeneous data sources [29,35,37], Data integration issues/challenges [28,34,62], Knowledge discovery issues from data processing [63].
Big data integration	Unstructured data integration [64], big data integration overview [65], big data integration and classification [4,5,15,31], Data integration in an intelligent environment [66], Big Data integration platform [8,26], big data integration for the decision-making process [25,67].

Table 4
Specific approaches/techniques used by the researchers.

Work	Approaches/Techniques
Data Integration with specific algorithms	Decision Model and Annotation [67], Divide and Conquer Algorithm [34], Attribute Conditional Dependency [5], Similarity Index [51], Schema Matching [60], Schema Merging [2], Semantic data relation learning [15], Data correlation/relationship [35,49,56,58], Data Curation [29], Data Classification [30,31,61], Probabilistic Semantic Association [4], Data Harmonization [28,55], Transfer Learning [54].
Data Integration using ETL/Wrappers	Wrapper [37], Wrapper (RESTful services) [33,57], Rule-based Wrapper [59,66], Extract-Transform-Load (ETL) [62,65], NoSQL Databases [64], Virtual Data Integration [8,25,26], Schema Mapping [24,63], Rules [27], Data Warehouse [13], Dimensionality Reduction [53]
Data Integration with Ontology	Graph [46], Graph + Ontology [3,6,14,16,32,36,38,39,42,44,45,50,52], Ontology Mapping [12,17–19,40,41,47,48], Graph + Crowd Sourcing [43]

techniques. The rest of the studies in this group focus on how to work with schema changes when mapping to ontologies by adjusting data source modifications to keep the mapping from failing/crashing.

In the second group, researchers are more concerned with the data integration involving physical schema integration, schema matching, and integration challenges/issues, such as [28], who studied the challenges of integrating nonstandard data sources (from legacy systems) and highlighted issues such as syntax, content, and format heterogeneity.

In the last group, conducted between 2018 and 2023, several studies particularly explored big data integration in terms of integration architecture and technologies that may be used.

3.2. Integration approaches or techniques used (RQ2)

This section summarizes the integration approaches or techniques grouped into categories with specific algorithms, wrappers, and ontologies related, as shown in Table 4. The grouping is not clear-cut because several studies use specific algorithms, but they also use ontologies in their discussion. Conversely, some studies use ontology as the primary approach but also use wrappers. For example, in research conducted by [6], although in his research using ETL/wrapper in the proposed approach, the discussion of transformation from data sources to ontologies is more dominant, so we include them in the data integration groups related to ontologies.

The grouping will help to identify research that uses similar techniques, the situations in which these techniques are used, the nature of the problem, and how researchers use them. As we can see, most studies use graph and ontology techniques, followed by mediator/wrapper, and few use specific algorithms.

3.2.1. Data integration with specific algorithms

Articles in this group were published between 2012 and 2022. A study by [60] proposed a schema-matching solution based on a linguistic matcher to integrate various applications into enterprise systems. It compares the similarity of the schema contexts as entity nodes to perform schema integration. It also considers neighboring nodes when determining entity similarity. The algorithm considers the nodes that are most relevant to the similarity of a given entity. The authors evaluated the algorithm's performance with other schema matching methods such as NTA, Similarity Flooding (SF), and WordNet. They found that the proposed solution outperformed the others.

A study by [34] discusses heterogeneous databases' operability, integration issues, and impact on data analysis. The authors used a database connection instead of an interface connection to perform database integration. The proposed solution is heterogeneous distributed database systems (AHDDI), which uses a bottom-up technique and employs a Divide-Conquer algorithm to split a system instance into several small sub-instances of the original system. Although it employs such an algorithm, it is mainly an integration architecture. It does not discuss the mechanism of how data is integrated from each data source.

A big data integration solution was proposed by [67], which is based on a decision-making process. The proposed model is called the DMP-BDE model. It consists of four phases: *intelligence, design, choice*, and *implementation*. In the intelligence phase, decision requirements are specified. These requirements included the data and the knowledge to be extracted for decision-making. Then, in the second phase, possible courses of action are invented, developed, and analyzed to handle the required decision. An analytical model is built in this phase, and then the process continues with several proposed alternatives. The choice phase evaluates the alternatives, and finally, the selected alternatives are implemented in the last phase. Unfortunately, the proposed model in this study is still just a theory and needs to be thoroughly tested and evaluated. A similar approach was proposed by [28], who studied data integration from legacy systems related to the health domain. According to the study, data from legacy systems can be beneficial because they promote reusability and better clarity in clinical research. However, performing harmonization and combining large amounts of data takes much work. The study proposed an integration flow called data harmonization based on a case study to align unstandardized medical data from different sources. The approach simplified the process by only extracting relevant variables to answer specific, predefined research questions about treatment effects and patient characteristics. However, the study's weakness includes manual processes, which are still involved during the variables matching identification across different data sources. Several issues are highlighted, such as syntactical, content, and format heterogeneities encountered, as well as some loss of information and selection biases due to differences in granularity or data capture.

According to the studies in this group, the main problems of big data integration that they focus on relate to unstructured data, the complexity of variable data, inconsistent data models, and data content. An approach to classifying and integrating data is to use probabilistic semantics and association (PSA) and generate knowledge patterns for more extensive data sources [4]. Using a possible association, it can include potential matches and assign probabilistic values to the data in different locations to map to the respective classes. The PSA algorithm utilizes the Hadoop Framework data schema and modifies the Naive Bayes (NB) algorithm to automatically classify anonymous data and map them to their matching schema. The authors also used existing machine learning and data extraction approaches for the data integration. Initially, large data sources were integrated into a structured database, and then the classification and forecasting of structured data were performed using the PSA. Comparisons with other classification algorithms were also made, such as against the Naive Bayes, several Ensemble Classifiers ("AdaBoost," "Logic

Boost," and "Voting (VOT)"), as well as against the Split and Combine Linear Discriminate Analysis (SC-LDA).

Several studies pointed out that existing data integration approaches using ETL pose challenges related to data heterogeneity issues at the schema and instance levels, where different data sources may represent the same real-world entity in different ways [5,15]. Meanwhile, research on attribute extraction and selection supporting semantic approaches has shown promising results. A study by [5] proposed a method to improve the big data integration mechanism using the conditional dependence of attributes and similarity index. The study used a bibliography dataset from various sources, and it calculated the attributed conditional dependence (ACD) and the data's similarity index (SI). ACD predicts attribute relationships and SI, derived from the latent-semantic analysis (LSA) algorithm used to perform the integration. Then, the application users can perform queries against the integrated data. However, search performance, which uses indexing and querying, becomes a highly challenging task regarding data growth in real-time.

To perform search queries efficiently, [2,32] proposed a data integration query approach using Source wrapping, Schema matching, Schema merging, and Query Reformulation. The study proposed a framework called Data Integration Framework, a generalization of both Global as View (GaV) and Local as View (LaV). GaV provides a mapping approach corresponding to a view expressed in terms of global schema and a view expressed in terms of local schema. Meanwhile, in LaV, there is a correspondence between a view expressed in terms of local schema and a view expressed in terms of global schema. Similar studies were done by [15,50]. While [50] primarily utilizes web semantics technology, [15] suggested a mechanism for unsupervised feature selection by using a probabilistic feature pattern approach (PFP) through semantic classification with Naive Bayes. The approach implemented a feature transformation and selection to accurately map the data by computing the principal component analysis for multi-value data. The data was then indexed using a feature-based semantic aptitude analysis (F-LSA) method, which learns the semantic association relation with the categorized data to identify patterns.

For case studies, the authors conducted an experimental evaluation using the Hadoop Framework and measured two criteria: Purity and Normalized Mutual Information (NMI). A theoretical perspective of the theories is also discussed in detail by [68], who models the base of the data integration framework and how the query processing and reasoning can be done in both LaV and GaV. The source wrapping used a wrapper that outputs a local ontology describing the data source. From the local ontology, schema matching was done to build the respective global ontology of the data source. The next step was schema merging, which fuses the local and global schemas to generate a new virtual view. Query from global ontology into local ontologies was performed using query rewriting. Then, a query translation was performed from the local ontology into the corresponding data source.

3.2.2. Data integration using ETL/wrappers

Although it has been stated in the previous discussion of the first group of existing studies that the use of ETL in data integration has posed challenges, we still find many researchers studying the effectiveness of using the ETL/wrappers technique to perform data integration. The articles in this group were published in the period 2013–2023.

The first study was done by [59], who performed retrieval of heterogeneous data from websites. According to the authors, the principle of the existing web crawling technology has not been accurate enough, so it still produces a lot of junk information. The study proposed a combined solution of the metadata-based retrieval model by establishing an abstraction of semantic attributes with a rule-based web wrapper. The wrapper extracts structured data from semi-structured web sources using special extraction rules. Similar studies by [13,37,57,64,65,69] explore data integration challenges to integrate information and functionality of diverse IT systems across organizations using various communication methods and specific integrator solutions. In particular, the authors in [13] offered an ETL-based integration architecture with an extra method, such as employing virtual views called global schema and a set of mappings between the global schema and the source schemas. The ETL generates the global schema, which

is then processed by a business logic layer, which executes user queries, combines results, and returns them to the users.

ETL plays a great role in certain architecture, especially when the data can be historical and gathered at a particular time, such as once a day, a week, or a month. However, a study by [62] highlighted that it can be a problem if online data processing is required. In addition, a common architecture of ETL data pulls can result in network redundancy and unnecessarily increase the system load of data sources. To overcome these issues, the study suggests utilizing a native query for each data source, the Fast Access Method (FAM), with an asynchronous messaging system to handle distributed transaction requests. A study by [33] argued that a data integration system should allow users to specify what information is required without providing detailed instructions or even specifying its location. Conversely, it emphasizes that application-based integration necessitates efforts from all parties involved and is only appropriate for a limited number of applications. Furthermore, it stated that integration involving a middleware layer would require copying data from the sources into a different source system, commonly implemented as a data warehouse, which would disadvantage the approach. The authors emphasize virtual integration, which they believe provides several benefits, such as propagating data from sources to the integration system with almost zero latency. However, in the solution, which they called Resource Oriented Heterogeneous Data Integration Platform (ROHDIP), they still incorporated Service Oriented Architecture (SOA) with RESTful services to communicate between the mediated virtual schema and the data sources.

Other studies that have proposed approaches involving ETL/wrappers include those conducted by [2,8,25,42,58]. One of the most comprehensive integration platforms in this category was proposed by [8]. The solution's architecture in the study, Quarry, facilitates the complete data integration lifecycle, which supports collaborative, cross-organization analytics for the big data integration life cycle. Quarry contains four functional modules for the data preparation pipeline: Quarry Manager, Data Quality Manager, Flow Manager, and Integration Management. The authors also discussed virtual data integration systems, ideal for data exploration but unfeasible for intensive tasks such as predictive analysis. The authors performed a use case of building data integration and evaluation using Quarry modules.

3.2.3. Data integration with ontology

The articles in this group were published between 2012 and 2023, and they were mainly about data integration using graph and ontology approaches.

The first study by [36] discussed data warehouse systems integration based on ontology, assuming they share the same ontology. The authors discussed several aspects of integration such as ontology and taxonomy, formal definition, ontology-based database, querying ontology using Onto-QL query language, multidimensional paradigm in data warehouse, the definition of DW ontology, how to do reasoning, annotating in DW ontology, and finally present validation of the proposed method.

Studies by [39,40,47,48] focus on utilizing ontology and semantic approaches for data integration. One of the studies uses semantic search to replace conventional search engines, arguing that information searching using semantics and ontology allows users to develop new facts and use their keywords in different environments. However, the query facility in an ontology management system is not as robust and reliable as in a relational database. Therefore, they proposed a solution in the form of OWL-Map to map RDF/OWL to relational databases. A system architecture for the OWL-Map is presented and evaluated. A similar study was done by [6,9,14,18,38,41], discussing the data integration approach using a schema represented as Web Ontology Language (OWL) ontology with big data sources. However, as highlighted, building ontology manually is time-consuming and error-prone. One of the study approaches is first copying all data sources into a common NoSQL document-oriented database called MongoDB using an ETL wrapper. The next step is to generate an ontology corresponding to each data source using the OWL language transformation rules. Then, the resulting ontologies are merged into a global one. A tool such as M2Onto has been developed to convert MongoDB databases into OWL ontologies, which performs the automatic process.

A study by [43] explored graph-based data integration and took a different approach. According to them, existing studies on graph integration, which use automated integration of XML schemas, usually find identical elements between schemas but do not match individual instances or records between graphs. To achieve the most accurate graph matching at the lowest cost, they proposed an algorithm that interactively improves the graph matching by first asking about the most informative nodes. The authors first propose a random walk model for computing the similarity between every pair of nodes appearing in a different graph. Using the random walk model, authors derive an EM algorithm to estimate the similarities using semi-supervised learning. The next suggestion was an active selection algorithm, which discovers the most informative node to use as a query in crowdsourcing systems.

Other researchers proposed using ontology learning instead of deriving the ontology automatically from the data, such as [12], who proposed an ontology learning life cycle to perform relational databases (RDB) integration. A deeper study was done by [3], who stated that data integration, especially in big data ecosystems, should accommodate the flexibility of data change so that analysts can constantly adapt their analysis to such changes. Providing an overview of such evolving changes and heterogeneous set of data sources is a challenging problem. One approach to tackle this is to leverage Semantic Web technologies into so-called Ontology-based data access (OBDA). It achieves this by conceptualizing domain interest and allowing users to perform ontology-mediated queries (OMQs). However, OBDA has no means of adapting to resource changes. OBDA is built upon generic reasoning using Description Logic (DL) to represent schema mapping using the global-as-view (GAV) approach. Therefore, source changes could invalidate the mapping, thus causing queries to crash. In contrast, LAV could accommodate a dynamic environment, but it will be a computationally expensive task that might require reasoning. With this problem in mind, the study proposed a model approach and integrated evolving data from multiple providers using two levels of ontology based on RDF called Global and on Source Graph to accommodate schema evolution.

Similar approaches were also studied by [16,45,46], who integrated heterogeneous data sources into a unique graph, such as using Graph-QL middleware over heterogeneous data sources into a unified Graph-DB based on RDF standard to enable users to specify the exact data they require. In this group, various studies use machine learning techniques to do integration or queries, particularly about knowledge extraction procedures, matching estimation probabilities of the graph nodes of data sources, as well as data classification/clustering such as done by [27,42–44,52].

Recent studies by [17,19] under this group have also reviewed ontology-based information integration specifically related to road asset management and manufacturing domains. The review discussed research published between 2006 and 2019, mainly about ontology engineering, such as the modeling approach, the tool, data representation, and querying.

As discussed in the summary above, research development has been quite significant from the initial studies published from 2012 to 2023. Initially, the research focused on integrating structured data using ETL/wrappers techniques with limited semantic information extraction, such as RDF/OWL. Still, in the last few years, the research has been done by working on semi-structured and unstructured data integration. In addition, the latest research also utilizes various machine learning algorithms and techniques combined with graphs and ontologies.

3.3. Open research issues based on existing studies (RQ3)

Based on the analysis and previous findings, several studies explicitly mention gaps that can be studied further. These gaps are divided into three groups: *Data integration with specific algorithms, Data integration using ETL/Wrappers*, and *Data integration with ontology*. Each group encompasses various research gaps along with references to relevant studies. For example, the first group includes research gaps such as the need for experimentation with theoretical models and algorithms, validation of source wrapping techniques, and performance testing of integration algorithms. Similarly, the other groups highlight gaps related to the development of data virtu-

alization servers, evaluation of knowledge embedding models, and the use of machine learning algorithms for ontology integration. Each identified gap is detailed in Table 5, offering valuable insights into the current challenges and opportunities for further exploration within data integration.

4. Discussion

In this section, we discuss the results of the review we obtained while revealing some of our findings as a reference for further research. Our study is initiated by collecting and analyzing existing publications, and results are presented according to the respective research questions. Analyzing the distribution of selected publications from 2012 to 2023 (Fig. 1), we observed a significant increase in publications related to data integration that peaked in 2021. This indicates that active research will start in 2020. Although there was a slight downward turn in 2022, the next upward trend shows that research in this field is still relevant and actively researched by scientists.

To answer RQ1, the literature is separated into sections to determine how much research falls into a particular domain, what data format is processed, and the research's focus. Data integration is carried out in various domains, as explicitly stated in studies related to educational data, such as university libraries and national library systems, medical, government sector, e-commerce, and manufacturing, as shown in Fig. 2, but most of them do not clearly state specific domains (including those regarding big data integration) especially when they focus on integration based on certain algorithms where in some cases synthetic data is deemed sufficient. For example, a study by [41] discusses the transformation of relational databases into ontologies. It uses examples of a relational database with tables such as person, job, and students containing sample records to illustrate semantic representations where the solution is not meant for just one problem domain but for relational data in general. Regarding the data format used, as shown in Fig. 3, only a few studies deal with data other than structured data, and most others do not explicitly describe the data format being processed. In this case, most studies do not focus too much on the data format but emphasize other aspects, such as how the data is formed into ontologies and how to combine them [48]. Although our findings imply that most of them are related to big data integration problems, which ideally solve problems with various data formats, a few studies still deal with data other than structured data, thus leaving a gap for further exploration.

As shown in Table 3, the focus of research is dominated by topics related to semantic data integration, followed by studies that highlight issues/challenges in physical data integration, such as those studies with solutions using ETL and SOA approaches, as well as those researches focusing on schema matching. The rest of the studies focus on big data integration. Based on this disaggregation, we can conclude that virtual data integration topics have been explored for a considerable time and are still being actively researched. This is consistent with our finding, as presented in Table 5, which answers RQ3 that the explicitly mentioned research gaps are more related to ontology-based integration, leaving room for further exploration. For RQ2, it was discovered that, in general, virtual data integration techniques outnumber physical integration, as evidenced by [8,33]. Although the two approaches are complementary, virtual data integration is preferable because it offers greater flexibility and eliminates the need to copy data to an integrated system. However, this solution is more complex than the preceding physical data integration.

As shown in Table 4, although certain literature focuses on one form of integration and explains a specific method, the strategies employing ETL/wrapper and Graph/Ontology in some literature are connected because physical integration is necessary to build a complete integration solution, so physical and virtual integration approaches are both needed. One of the research topics that we believe is the most comprehensive was done by [8], which exhibits extraction using wrappers to disclose data from data sources into graph data representations.

Table 5
Research gaps.

Group	Research gaps
Data integration with specific algorithms	(1) Experiment to use the theoretical DMP-BDE model [67] (2) Perform experimentation to accurately validate source wrapping, schema matching, schema merging, query reformulation algorithm [2] (3) PSA can be used in the future for precise data mapping for various prediction tasks, allowing users to benefit from reliable prediction of heterogeneous data sources [4] (4) Perform integration performance testing for the ACD-SI algorithm under un-supervised method selection requirement and data indexing [5] (5) Addressing the challenges of data curation, standardizing formats and annotation for data release, and continued efforts to integrate multiple datasets [29]
Data integration using ETL /Wrappers	(6) Develop layer expansion in the direction of a data virtualization server, semi-automatic wrappers creation, or appropriate data source searching [13] (7) Apply parallel processing or parallel querying between the mediated schema RESTful service and the wrappers in the ROHDIP integration architecture [33] (8) Performing a comprehensive end-to-end assessment of the platform, with a focus on the critical functional modules and the usability of Quarry for non-technical users [8] (9) Training a knowledge embedding model to enrich the schema matching, evaluating the results by comparing it with existing baseline methods, and verifying the feasibility of the method in a real-world scenario by conducting a case study [63] (10) Making the data integration process completely automatic, optimizing the resulting schemes by eliminating redundant columns, and defining a strategy for a one-shot integration of n sources [27] (11) Adding methods of provenance information management to understand data origins and the process of analytical query results [25]
Data integration with ontology	(12) Formalize the methods for composing ontology modules and the mechanisms for updating the ontology in response to data source changes [5] (13) An integrated wrapper can be provided for all data sources instead of multiple wrappers. A wrapper generator can dynamically generate clauses depending on required data sources for dynamic data sources. Conflicting attribute identification can be done using machine learning approaches. Lastly, user input using a mechanism such as the Analytic Hierarchy Process (AHP) can be used to identify the most critical data from the conflicting attributes retrieved by a query [9] (14) Examine how improved query semantics may be used in query optimization, particularly in the context of distributed query processing [38] (15) Extending ontology with richer constructs to adjust to schema changes semi-automatically [3] (16) The ontology integration is not fully automatic; to overcome this, machine learning algorithms can be used in combination with Semantic Web Ontology [14] (17) By combining different Retrieval Information (RI) approaches with the agent's paradigm for creating dynamic and autonomous ontologies, we may improve the performance of the similarity identification module [48] (18) Use machine learning to automatically and efficiently populate a large number of ontology properties [44] (19) Providing security for data integration in healthcare is a significant challenge. It's also crucial to use semantic analytics for IoT smart data categorization and clustering [52] (20) Further data integration approach can adopt solutions based on the RDF graph repository to enable achieving semantic data interoperability [16] (21) To work on different structured sources such as Excel Spreadsheets, comma-separated value (CSV) as well as unstructured data sources for constructing ontologies [12] (22) Knowledge graphs can be (continuously) physically constructed and stored to reduce loading times (materialization). A graph database would be a suitable choice for storing knowledge graphs in ontologies [45] (23) Building a domain-specific application for smart tourist destinations using this semantic-based big data integration framework (STD) [42] (24) Incorporating more data from other learning management systems, updating the e-LION ontology with new relevant attributes, and aligning ontologies in different domains [32]

Most of the studies, especially those published in the last few years, describe the forms of virtual data integration using graph data or the semantics/ontologies method. In addition, we have highlighted a few expressed directly by the authors concerning the research gaps in Table 5 for RQ3. Gaps (1), (2), and (4) contain suggestions for future experiments and integration testing against the algorithms in the research group that use specific algorithms in data integration (which is also stated in Table 4). In contrast, gap (3) advises employing the algorithm for various class prediction tasks. As indicated by gaps (6), (8), and (9) in the wrapper/ETL-based data integration group, an extension as well as an assessment against the proposed platforms can be studied further. More gaps exist in the ontology-based data integration group, which each author has clearly explained. However, the most notable gap on the list is the lack of a virtual data integration solution that permits automatic semantic data integration, as in (13), (15), and (16), especially when considering unexpected or regular changes that occur in data sources. Machine learning techniques, which have become increasingly popular in other research areas, may also be explored. The discussion by [52] is an example of a current topic that can be referred to. It explores integrating data originating from IoT devices in the medical field using machine learning techniques such as unsupervised algorithms that deal with clustering and dimensionality reduction problems. A study may broaden its focus on integrating structured, semi-structured, and unstructured data. A recent study shows that data integration could be further explored for processing different structured data formats such as Excel sheet, CSV, as well as other unstructured data formats as indicated by gap (21), which is consistent with the findings answering the RQ1 that most of the studies we found still focused on structured data. However, sensitive data, such as personal information or medical records, cannot always be public, and we have yet to come across any data integration literature focusing on data security and privacy. As a result, as the gap (19) suggests, how to apply security while performing data integration is also an area that may be further investigated.

Given the research challenges identified through our findings in the publications reviewed, there is an urgent need for advancements in data integration. Addressing the issues of diverse and non-standard data formats in the enterprise environment is critical. Challenges in data mapping, redundancy, and integration processes that still need to be fully automated persist. This is proven by the same challenges that still exist in 2022–2023. Solving these obstacles is essential for improving integration performance and model accuracy during data utilization. Moreover, physical data integration seems no longer feasible in the era of big data with diverse sources. Instead, there is a growing demand for automatic or dynamic virtual data integration approaches capable of adapting to frequent schema changes in source databases. Exploring techniques such as machine learning, graph learning, and semantically-aware approaches can further enhance the efficiency of data integration processes. Additionally, ensuring the security of sensitive data during integration and sharing is crucial, necessitating the exploration of appropriate security measures for such scenarios. Furthermore, given the increasing use of semi-structured and unstructured big data, there is a need to investigate solutions that incorporate parallel/distributed mechanisms to address performance issues effectively. Overall, advancing data integration methodologies to address these challenges is imperative for harnessing the full potential of diverse and voluminous datasets.

5. Milestones and influential articles

This section provides additional analysis of the milestones and the most influential articles. Fig. 4 depicts the ten most frequently cited articles in the field, ranked by number of citations. These articles are further detailed in Table 6, where most articles come from leading publishers, with Elsevier being the top-ranked source. The highest-cited study, with 98 citations, focuses on ontology-based data warehouses [36]. This is followed by a similar study on ontology development for data integration [3] and [6], which garnered 77 and 59 citations, respectively. The next most cited articles worked on various topics, such as [61], which proposed an abstract and generic approach to data integration and received 51 citations. [52] proposed data integration

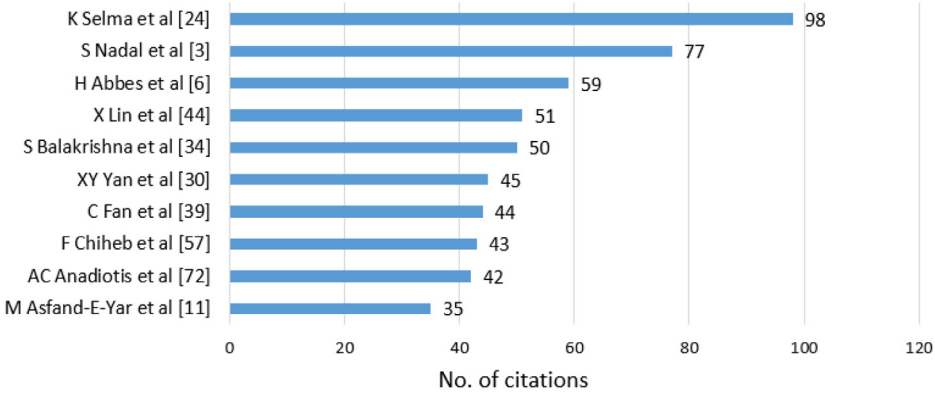

Fig. 4. The top ten most cited articles according to Google Scholar (as of July 2024).

related to IoT sensors, which received 50 citations. The next three studies [51,54,67] proposed using a machine learning approach for data integration, which received 45, 44, and 43 citations, respectively. The last two most cited articles [14,46] proposed data integration with solving semantics, syntactic, and structural issues, where they received 42 and 35 citations, respectively. The collection of studies highlights significant contributions to ontology-based and big data integration, addressing various challenges from IoT data integration to fault detection in transfer learning, reflecting the field's diversity and depth of research.

Following the significant contributions of studies across the publications, we compile a collection highlighting the most distinctive research and their yearly impact from 2012 to 2023, as shown in Table 7. In 2012, researchers started to adopt ontology to solve semantic issues during data integration. However, the architecture still used a materialized data warehouse fed by ontology-based databases. In 2013, researchers explored the use of machine learning, metadata, and rule-based approaches for data integration. Interestingly, in 2014, we found research advocating an ETL approach to address challenges related to integrating heterogeneous data that focused more on addressing semantic issues.

Over the span of the next five years (from 2015 to 2020), researchers delved into diverse data integration approaches. These included real-time data integration, leveraging NoSQL databases for big data integration, using crowdsourcing for precise data integration through annotation assistance, applying ontology-based methods for big data integration, harnessing machine learning techniques, and constructing knowledge graphs.

Research conducted between 2021 and 2022 started exploring data integration using graph methods. Concurrently, there was a rise in studies combining cloud-based solutions with deep learning techniques. By 2023, the focus shifted to semi-automated data integration methods, particularly those addressing data preprocessing to resolve syntactic and semantic problems, ultimately enhancing schema matching.

In summary, the analysis of data integration techniques from the past decade reveals notable advancements, transitioning from physical or materialized approaches to more practical and efficient virtual data integration solutions. However, as data complexity increases and problems such as redundancy cause syntax, structural, and semantic issues, both semi-automatic and fully automatic approaches continue to present challenging areas for further investigation.

Table 6
The research topics of the top ten most cited articles.

Ref	Title	Source	Research topics
24	"Ontology-based structured web data warehouses for sustainable interoperability: Requirement modeling, design methodology and tool"	Computers in Industry (Elsevier)	Building an ontology-based Data warehousing used to solve conflicts between heterogeneous data sources
3	"An integration-oriented ontology to govern evolution in Big Data ecosystems"	Information System (Elsevier)	Ontology development for modeling and integrating evolving data from multiple providers.
6	"Big Data Integration: A MongoDB Database and Modular Ontologies based Approach"	Procedia Computer Science (Elsevier)	Propose an approach to build an ontology for Big Data integration.
44	"Heterogeneous data integration by tree-augmented naïve Bayes for protein-protein interactions prediction"	Proteomics (Willey)	Analyze and identify flaws in existing data integration solutions and propose an abstract and general approach to address the identified issues.
34	"IoT sensor data integration in healthcare using semantics and machine learning approaches"	A Handbook of Internet of Things in Biomedical and Cyber-Physical Systems (Springer)	Highlight challenges in integrating and analyzing IoT sensor data using RESTful CoAP protocol to improve access to data in remote locations.
30	"Prediction of drug-target interaction by integrating diverse heterogeneous information source with multiple kernel learning and clustering methods"	Computational Biology and Chemistry (Elsevier)	Develop a novel computational method incorporating kernel learning and clustering approaches for integrating heterogeneous data sources to predict new drug-target interactions (DTI).
39	"A novel image-based transfer learning framework for cross-domain HVAC fault diagnosis: From multi-source data integration to knowledge sharing strategies"	Energy and Buildings (Elsevier)	Proposes an integration approach using tabular data from different systems for transfer learning for fault detection and diagnosis tasks.
57	"A new model for integrating big data into phases of decision-making process"	Procedia Computer Science (Elsevier)	Develop a theoretical model that integrates big data to improve the decision-making process in the organization.
72	"Graph integration of structured, semistructured and unstructured data for data journalism"	Information Systems (Elsevier)	Integrates arbitrary heterogeneous datasets into a unique graph for query and answer models.
11	"Semantic integration of heterogeneous databases of same domain using ontology"	IEEE Access (IEEE)	Integrates heterogeneous databases within the same domain, specifically focusing on Semantic, Syntactic, and Structural Heterogeneity.

Table 7
The contribution of the milestone publications.

Ref	Title	Year	Main contribution
24	"Ontology-based structured web data warehouses for sustainable interoperability: Requirement modeling, design methodology and tool"	2012	This study leverages ontology-based databases (OBDB) to construct data warehouses (DWs), resulting in DW models that semantically integrate sources from diverse origins.
44	"Heterogeneous data integration by tree-augmented naïve Bayes for protein-protein interactions prediction"	2013	This study integrates heterogeneous data sources using a tree-augmented naïve Bayes (TAN) classifier, which achieves higher classification accuracy than a manually constructed Bayesian network classifier and naïve Bayes.
43	"Integration and collection of heterogeneous data based on metadata"	2013	This study introduces a metadata-based retrieval model and rule-based web wrapper to address the problem of heterogeneous data collection and integration across different industries.
9	"Integration of Data from Heterogeneous Sources Using ETL Technology"	2014	This study proposes an architecture to solve data integration problems by performing syntactic integration with ETL workflows and supporting semantic integration through connections between data elements.
46	"Integration of heterogeneous databases"	2015	This study integrates data from different databases under different platforms, query languages, data models, and dependencies, enabling real-time applications and providing a transparent environment for users.
6	"Big Data Integration: A MongoDB Database and Modular Ontologies based Approach"	2016	This study introduces an approach to integrate big data based on ontology using NoSQL databases.
49	"A Schema-Based Approach to Enable Data Integration on the Fly"	2017	This study proposes a novel metadata model and processing approach that enables on-the-fly data integration in loosely coupled, heterogeneous environments.
26	"Integration of graphs from different data sources using crowdsourcing"	2017	This study presents an efficient algorithm to integrate two graphs from different sources using crowdsourcing to perform annotations and find matching data for precise data integration.
28	"Automated modeling assistance by integrating heterogeneous information sources"	2018	This study introduces a framework that can assist in integrating and querying heterogeneous information sources, such as ontologies, meta-models, and XML schemas, to provide modeling and meta-modeling assistance.
3	"An integration-oriented ontology to govern evolution in Big Data ecosystems"	2019	This study presents an algorithm that semi-automatically adapts ontology-based Big Data Integration upon new releases of data sources, which handles schema evolution and ensures correct retrieval of data from the most recent schema version and historical queries.
30	"Prediction of drug-target interaction by integrating diverse heterogeneous information source with multiple kernel learning and clustering methods"	2019	This study presents a novel computational method based on a machine learning approach, namely multiple kernel learning and clustering methods, to generate similarity matrices for data integration.
33	"Knowledge Graph Construction for Payment Data Risk Control"	2020	The study integrates data sources by building a knowledge graph based on payment data, which solves the problem of large amounts of complex data, allows for dynamic data import, and provides a combined text and graphical visualization.

(*continued on next page*)

Table 7 (*continued*)

Ref	Title	Year	Main contribution
36	"Semantic-based Big Data integration framework using scalable distributed ontology matching strategy"	2021	This study proposes a semantic-based Big Data integration framework that offers a unified view of data, ensuring accuracy, runtime efficiency, performance, and scalability in large-scale environments through optimization mechanisms and parallelism.
13	"Evaluation of data integration plans based on graph data"	2021	This study leverages graph data to perform data integration for seamless data integration with better response time and achieve semantic data interoperability through vocabularies and ontologies.
72	"Graph integration of structured, semi-structured and unstructured data for data journalism"	2022	The study integrates graph data from various datasets by introducing a centralized graph data warehouse architecture that enables matching across heterogeneous data sources while preserving the provenance of each node, extending previous work to novel data models, which are easy to install with minimal user effort.
64	"A framework of integrating heterogeneous data sources for monthly streamflow prediction using a state-of-the-art deep learning model"	2022	The study proposes a cloud-based deep learning system for data integration to fully capture image data's spatial and temporal features for machine learning tasks.
67	"A cloud-based deep learning model in heterogeneous data integration system for lung cancer detection in medical industry 4.0"	2022	The study proposes a similar cloud-based deep learning system for the semi-automatic data integration process.
66	"A semi-automatic data integration process of heterogeneous databases"	2023	The study introduces a semi-automatic data integration technique that preprocesses input data sources to normalize and standardize them, conducts syntactic and semantic analyses to identify similar tables, and finally performs data cleansing and merging of the identified tables.
56	"A semi-automated hybrid schema matching framework for vegetation data integration"	2023	The study introduces a framework that combines schema-level and instance-level matching techniques, resulting in a more effective and significantly improved schema-matching process compared to existing state-of-the-art methods.

6. Conclusion

A data integration mechanism is needed to extract data from various sources for specific purposes. However, heterogeneous data integration in the big data era is increasingly challenging, mainly due to the complexity of data in terms of format, semantics, and data structure issues.

To comprehend the development of numerous existing data integration studies, we conducted a literature review of research from various sources published from 2012 to 2023. We collected, classified, sorted, and processed 58 articles for review. The review was done based on the research questions, methods/techniques, and the open research gaps that the researcher explicitly stated.

From these results, we found that a few explicitly defined the focus of the research domain, and the rest focused on big data without mentioning a particular domain, especially research that focused on improvising the algorithms used in data integration. Likewise, regarding the form of data that is processed, only a small number of publications mentioned that they processed structured data; most of them did not specifically mention the form of the data, and because most of the discussion was related to big data, we concluded that there are still opportunities for unstructured or semi-structured data integration for further research. Regarding

the technique used, researchers tend to discuss virtual data integration related to data semantic problems. In line with this, several studies highlight the open challenges related to the automatic creation of ontologies from existing data and the challenges of adapting to schema changes. The use of techniques that have been successful in other fields, such as machine learning and methods to manage data privacy, is also room for future exploration. Finally, we have yet to find reports specifically based on case studies. Case studies help demonstrate many vital elements or factors, especially in the architecture of a complete data integration solution. Thus, for future work literature studies, analyzing solutions in industrial sectors with the latest heterogeneous data integration requirements can be investigated, too.

CRediT Author Statement

I Made Putrama: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Writing original draft/editing. **Peter Martinek:** Resources; Supervision; Validation; Providing review/feedback.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R.H. Hariri, E.M. Fredericks, K.M. Bowers, Uncertainty in big data analytics: survey, opportunities, and challenges, J. Big Data 6 (1) (2019).
[2] G. Fusco, L. Aversano, An approach for semantic integration of heterogeneous data sources, PeerJ Comput. Sci. 2020 (3) (2020) 1–30.
[3] S. Nadal, O. Romero, A. Abelló, P. Vassiliadis, S. Vansummeren, An integration-oriented ontology to govern evolution in Big Data ecosystems, Inf. Syst. 79 (2019) 3–19.
[4] V. VandanaKolisetty, D.S. Rajput, Integration and classification approach based on probabilistic semantic association for big data, Complex Intell. Syst. 9 (2021) 3681–3694.
[5] V.V. Kolisetty, D.S. Rajput, Big data integration enhancement based on attributes conditional dependency and similarity index method, Math. Biosci. Eng. 18 (6) (2021) 8661–8682.
[6] H. Abbes, F. Gargouri, Big data integration: a MongoDB database and modular ontologies based approach, Proc. Comput. Sci. 96 (2016) 446–455.
[7] M. Matyoqubov, A. Saidov, O. Kazakov, O. Rustamova, in: Enterprise Systems Data Integration," 2020 International Conference on Information Science and Communication Technology ICISCT 2020, 2020, pp. 2020–2022.
[8] P. Jovanovic, S. Nadal, O. Romero, A. Abelló, B. Bilalli, Quarry: a user-centered big data integration platform, Inf. Syst. Front. 23 (1) (2021) 9–33.
[9] B. Ahamed, T. Ramkumar, Data integration - challenges, techniques, and future directions: a comprehensive study, Indian J. Sci. Technol. 9 (44) (2016).
[10] H. Kondylakis, G. Flouris, D. Plexousakis, Ontology and schema evolution in data integration: review and assessment, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2009, pp. 932–947. vol. 5871 LNCS, no. PART 2.
[11] V. Ventrone, S. Heiler, Semantic heterogeneity as a result of domain evolution, ACM Sigmod Rec. 20 (4) (1991) 16–20.
[12] B. Ben Mahria, I. Chaker, A. Zahi, A novel approach for learning ontology from a relational database: from the construction to the evaluation, J. Big Data 8 (1) (2021).
[13] M. Marek, Integration of data from heterogeneous sources using ETL technology, Comput. Sci. 15 (2) (2014) 109.
[14] M. Asfand-E-Yar, R. Ali, Semantic integration of heterogeneous databases of same domain using ontology, IEEE Access 8 (2020) 77903–77919.
[15] M.M. Nashipudimath, S.K. Shinde, J. Jain, An efficient integration and indexing method based on feature patterns and semantic analysis for big data, Array 7 (April) (2020) 100033.
[16] D.A. Vasiliev, A.M. Ghiran, R.A. Buchmann, Evaluation of data integration plans based on graph data, Proc. Comput. Sci. 192 (2021) 1041–1050.
[17] X. Lei, P. Wu, J. Zhu, J. Wang, Ontology-based information integration: a state-of-the-art review in road asset management, Arch. Comput. Methods Eng. (2021).
[18] S.C. Haw, J.W. May, S. Subramaniam, Mapping relational databases to ontology representation: a review, in: ACM International Conference Proceeding Series, Part F1312, 2017, pp. 54–58.

[19] B. Ramis Ferrer, W.M. Mohammed, M. Ahmad, S. Iarovyi, J. Zhang, R. Harrison, J.L. Martinez Lastra, Comparing ontologies and databases: a critical review of lifecycle engineering models in manufacturing, Knowl. Inf. Syst. 63 (6) (2021) 1271–1304.
[20] T. Carrion, P. Vicente, G. Gonzalez, S. Aciar, R. Morales, Germania, Methodology for systematic literature review applied to engineering and education, in: 2018 IEEE Global Engineering Education Conference, 2018, pp. 1364–1373.
[21] B. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering PhD thesis, Keele University and Durham University Joint Report, Keele, U.K., 2007.
[22] R. Tomaszewski, A study of citations to STEM databases: ACM digital library, engineering village, IEEE Xplore, and MathSciNet, Scientometrics 126 (2) (2021) 1797–1811.
[23] A. Martín-Martín, E. Orduna-Malea, M. Thelwall, E. Delgado López-Cózar, Google scholar, web of science, and scopus: a systematic comparison of citations in 252 subject categories, J. Informetr. 12 (4) (2018) 1160–1177.
[24] M. Asif-Ur-Rahman, B.A. Hossain, M. Bewong, M.Z. Islam, Y. Zhao, J. Groves, R. Judith, A semi-automated hybrid schema matching framework for vegetation data integration, Expert Syst. Appl. 229 (PA) (2023) 120405.
[25] B. Silva, J. Moreira, R.L.d.C. Costa, Logical big data integration and near real-time data analytics, Data Knowl. Eng. 146 (2023) 102185.
[26] T. Cadden, J. Weerawardena, G. Cao, Y. Duan, R. McIvor, Examining the role of big data and marketing analytics in SMEs innovation and competitive advantage: a knowledge integration perspective, J. Bus. Res. 168 (2023) 114225.
[27] M. Barbella, G. Tortora, A semi-automatic data integration process of heterogeneous databases, Pattern Recognit. Lett. 166 (2023) 134–142.
[28] H.Le Sueur, I.N. Bruce, N. Geifman, N. Geifman, The challenges in data integration - Heterogeneity and complexity in clinical trials and patient registries of Systemic Lupus Erythematosus, BMC Med. Res. Methodol. 20 (1) (2020) 1–5.
[29] C. Pastrello, M. Abovsky, R. Lu, Z. Ahmed, M. Kotlyar, C. Veillette, I. Jurisica, Osteoarthritis Data Integration Portal (OsteoDIP): a web-based gene and non-coding RNA expression database, *Osteoarthr. Cart. Open* 4 (1) (2022) 100237.
[30] C. Gu, C. Dai, X. Shi, Z. Wu, C. Chen, A cloud-based deep learning model in heterogeneous data integration system for lung cancer detection in medical industry 4.0, J. Ind. Inf. Integr. 30 (2022) 100386.
[31] X. Yang, K.A. Hoadley, J. Hannig, J.S. Marron, Jackstraw inference for AJIVE data integration, Comput. Stat. Data Anal. 180 (2023) 107649.
[32] M. Paneque, M.d.M. Roldán-García, J. García-Nieto, e-LION: data integration semantic model to enhance predictive analytics in e-Learning, Expert Syst. Appl. 213 (PA) (2023) 118892.
[33] W. Shehab, S. M, E. Sallam, ROHDIP: resource oriented heterogeneous data integration platform, Int. J. Adv. Comput. Sci. Appl. 7 (9) (2016) 104–109.
[34] M.O. Hegazi, D.K. Saini, K. Zia, Moving from heterogeneous data sources to big data: interoperability and integration issues, Int. J. Adv. Comput. Sci. Appl. 9 (10) (2018) 207–212.
[35] L. Chen, M. Baza, H. Alshahrani, Data integration method of multi-source feedback evaluation for remote teaching quality, Mob. Netw. Appl. (2023).
[36] K. Selma, B. Ilyès, B. Ladjel, S. Eric, J. Stéphane, B. Michael, Ontology-based structured web data warehouses for sustainable interoperability: requirement modeling, design methodology and tool, Comput. Ind. 63 (8) (2012) 799–812.
[37] B. Garg, K. Kaur, Integration of heterogeneous databases, in: 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA 2015), 2015, pp. 1033–1038.
[38] D. Nicklas, T. Schwarz, B. Mitschang, A schema-based approach to enable data integration on the fly, Int. J. Coop. Inf. Syst. 26 (1) (2017) 1–41.
[39] G.M. Santipantakis, A. Glenis, K. Patroumpas, A. Vlachou, C. Doulkeridis, G.A. Vouros, N. Pelekis, Y. Theodoridis, SPARTAN: semantic integration of big spatio-temporal data from streaming and archival sources, Futur. Gener. Comput. Syst. 110 (2020) 540–555.
[40] A.G. Salguero, P. Delatorre, J. Medina, M. Espinilla, A.J. Tomeu, Ontology-based framework for the automatic recognition of activities of daily living using class expression learning techniques, Sci. Program. 2019 (iii) (2019).
[41] P. Mayadewi, B. Sitohang, F.N. Azizah, Study relational database transformation to ontology, in: 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing SNP*D 2018*, 2018, pp. 358–362.
[42] I. Mountasser, B. Ouhbi, F. Hdioud, B. Frikh, Semantic-based big data integration framework using scalable distributed ontology matching strategy, Distrib. Parallel Databases 39 (2021) 891–937.
[43] Y. Kim, W. Jung, K. Shim, Integration of graphs from different data sources using crowdsourcing, Inf. Sci. (Ny). 385-386 (2017) 438–456.
[44] Z. Wang, M. Guo, Z. Li, M. Tang, J. Yu, Knowledge Graph Construction For Payment Data Risk Control, 675, Springer, Singapore, 2020.
[45] C. Blankenberg, B. Gebel-Sauer, P. Schubert, Using a graph database for the ontology-based information integration of business objects from heterogenous Business Information Systems, Proc. Comput. Sci. 196 (2021) 314–323.
[46] A.C. Anadiotis, O. Balalau, C. Conceição, H. Galhardas, M.Y. Haddad, I. Manolescu, T. Merabti, J. You, Graph integration of structured, semistructured and unstructured data for data journalism, Inf. Syst. 104 (2021) 101846.
[47] H. Afzal, M. Waqas, T. Naz, OWLMap: fully automatic mapping of ontology into relational database schema, *Int. J. Adv. Comput. Sci. Appl.* 7 (11) (2016) 7–15.
[48] J. Kachaoui, A. Belangour, Local ontology merging in data ponds, in: 2020 International Conference on Intelligent Systems and Computer Vision, 423, 2020, pp. 3–9.
[49] M.S. Ángel, J. de Lara, P. Neubauer, M. Wimmer, Automated modeling assistance by integrating heterogeneous information sources, Comput. Lang. Syst. Struct. 53 (2018) 90–120.
[50] R. Kebede, A. Moscati, H. Tan, P. Johansson, Integration of manufacturers' product data in BIM platforms using semantic web technologies, Autom. Constr. 144 (September) (2022) 104630.
[51] X.Y. Yan, S.W. Zhang, C.R. He, Prediction of drug-target interaction by integrating diverse heterogeneous information source with multiple kernel learning and clustering methods, Comput. Biol. Chem. 78 (2019) 460–467.

[52] S. Balakrishna, M. Thirumaran, V.K. Solanki, IoT Sensor Data Integration in Healthcare using Semantics and Machine Learning Approaches, Springer International Publishing, 2020 vol. 165.
[53] S. Ahmed, M. Gentili, D. Sierra-Sosa, A.S. Elmaghraby, Multi-layer data integration technique for combining heterogeneous crime data, Inf. Process. Manag. 59 (3) (2022) 102879.
[54] C. Fan, W. He, Y. Liu, P. Xue, Y. Zhao, A novel image-based transfer learning framework for cross-domain HVAC fault diagnosis: from multi-source data integration to knowledge sharing strategies, Energy Build. 262 (2022) 111995.
[55] E. Negussie, O. González-Recio, M. Battagin, A.R. Bayat, T. Boland, Y. de Haas, A. Garcia-Rodriguez, P.C. Garnsworthy, N. Gengler, M. Kreuzer, B. Kuhla, J. Lassen, N. Peiren, M. Pszczola, A. Schwarm, H. Soyeurt, A. Vanlierde, T. Yan, F. Biscarini, Integrating heterogeneous across-country data for proxy-based random forest prediction of enteric methane in dairy cattle, J. Dairy Sci. 105 (6) (2022) 5124–5140.
[56] W. Xu, J. Chen, X.J. Zhang, L. Xiong, H. Chen, A framework of integrating heterogeneous data sources for monthly streamflow prediction using a state-of-the-art deep learning model, J. Hydrol. 614 (PB) (2022) 128599.
[57] A.S. Trunov, L.I. Voronova, V.I. Voronov, D.I. Sukhachev, V.G. Strelnikov, Legacy applications model integration to support scientific experiment, in: 2018 Systems of Signals Generating and Processing in the Field of on Board Communications, 2018, pp. 1–7.
[58] C. Varadharajan, V.C. Hendrix, D.S. Christianson, M. Burrus, C. Wong, S.S. Hubbard, D.A. Agarwal, BASIN-3D: a brokering framework to integrate diverse environmental data, Comput. Geosci. 159 (2022) 105024.
[59] L. Zhang, Integration and collection of heterogeneous data based on metedata, in: Proceedings of the 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering ICIII 2013, 1, 2013, pp. 205–208.
[60] B. Villanyi, P. Martinek, Towards a novel approach of structural schema matching, in: CINTI 2012 - 13th IEEE International Symposium on Computational Intelligence and Informatics, Proceedings, 2012, pp. 103–107.
[61] X. Lin, X.W. Chen, Heterogeneous data integration by tree-augmented naïve Bayes for protein-protein interactions prediction, Proteomics 13 (2) (2013) 261–268.
[62] M. Chromiak, M. Grabowiecki, Heterogeneous data integration architecture-challenging integration issues, Ann. Univ. Mar. Curie-Sklodowska. Sectio AI, Inform. 15 (1) (2015) 7.
[63] C. Ma, B. Molnar, A. Tarcsi, A. Benczur, Knowledge enriched schema matching framework for heterogeneous data integration, in: 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS 2022), 2022, pp. 183–188.
[64] J.V. Kumar, S.A. Moeed, C.M. Kumar, G. Ashmitha, Integration patterns of MongoDB GridFS for advanced data science and big data processing, *Mater. Today Proc.* (2021) no. xxxx.
[65] S. Vetova, Big heterogeneous data integration and analysis, AIP Conf. Proc. 2333 (2021).
[66] I.M. Putrama, P. Martinek, A hybrid architecture for secure big-data integration and sharing in smart manufacturing, Proc. Int. Spring Sem. Electron. Technol. 2023 (3) (2023).
[67] F. Chiheb, F. Boumahdi, H. Bouarfa, A new model for integrating big data into phases of decision-making process, Proc. Comput. Sci. 151 (2018) (2019) 636–642.
[68] M. Lenzerini, Data integration: a theoretical perspective, in: Proceedings of the Eighteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2002, pp. 233–246.
[69] A. Kadadi, R. Agrawal, C. Nyamful, R. Atiq, Challenges of data integration and interoperability in big data, in: Proceedings of the- 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, 2014, pp. 38–40.