Journal of Computer Science

Review Article

A Systematic Literature Review on Data Integration: Issues, Data Types, Architecture

Mimi Haryani Tasani and Mohd Kamir Yusof

Faculty of Informatics and Computing, University Sultan Zainal Abidin, Besut, Terengganu, Malaysia

Article history
Received: 04-01-2025
Revised: 20-05-2025
Accepted: 23-05-2025

Corresponding Author:
Mohd Kamir Yusof
Faculty of Informatics and Computing, University Sultan Zainal Abidin, Besut, Terengganu, Malaysia
Email: support@thescipub.com

Abstract: Data integration is an important process that enables organizations to merge data from various sources, creating a consolidated viewpoint that supports better decision-making, increased effectiveness, and deeper insights. Although data integration has advantages, it also comes with difficulties like maintaining data quality, adhering to compliance regulations, resolving technical issues, and merging with older systems. Effective management of data quality, robust data governance practices, and the use of advanced integration tools like data filtering are necessary to successfully overcome these challenges and ensure the harmonization and standardization of data. Systematic literature review explains by taking a strategic approach, companies can transform data integration into a valuable resource, establishing a base for improved analytics, customer understanding, and more efficient operations that bolster long-term growth and competitiveness. This research explores into the various complex obstacles of merging data in the realm of big data, focusing on structured, semi-structured, and unstructured data categories. With the progress of the healthcare sector, IoT, and big data analytics, there is a growing need for strong integration methods to manage the increasing volume and variety of data for better accessibility and usability. The research uncovers primary concerns in data integration including data quality and consistency, diverse data formats and standards, scalability and performance, and governance and compliance. It stresses the significance of recognizing the distinct traits and obstacles of various data forms to enable efficient integration. Moreover, the article examines different data integration models pointing out their advantages and disadvantages in meeting current data management requirements. The research seeks to improve data processing in today's data-driven world by sharing a detailed analysis of current obstacles and solutions in data integration.

Keywords: Data Integration, Big Data, Data Integration Models, Systematic Literature Review

Introduction

Because of the large amount and easy accessibility of multidimensional data, technological innovation has the significant possibility to greatly influence our daily lives in various fields, particularly in the healthcare industry, IoT and big data analysis. The quick expansion and utilization of data will give rise to a massive concept called big data (Krishnamoorthy et al., 2023; Rehman et al., 2022). However, big data techniques and applications pose numerous challenges and issues, such as big data intelligence, energy management, data security and privacy, scalability of computing infrastructure, data management, data interpretation, and real-time data processing. Since big data often involves several hazards, security and privacy are two of the most important issues. Different kinds of personal information that are sensitive, such as age, addresses, personal preferences, banking details, etc., have prompted research into methods to protect data confidentiality and private information (Deepa et al., 2022; Shafqat et al., 2020; Borges do Nascimento et al., 2021).

Due to the increasing amounts of different data types that being produced by businesses, conventional data storage systems frequently face challenges in meeting the requirements for scalability, flexibility, and cost-efficiency. A crucial process, data integration combines information from multiple sources to produce a coherent viewpoint that enables businesses to make wise decisions.

These many forms of data, including unstructured, semi-structured, and structured data, provide unique difficulties for the procedure (Piippola, 2024; Cauteruccioe et al., 2020; Rozony et al., 2024). Comprehension of these data categories is crucial for successful integration, as they determine the techniques and instruments employed to combine data smoothly.

Structured data is data organized in a specific and predetermined way, typically in tables with rows and columns. Each column in relational databases, Excel spreadsheets, and CSV files stands for a distinct attribute, and each row represents a distinct data point. Usually found such as tables containing specific columns like name, age, and date. It can be accessed using tools such as SQL and stored in relational databases or data warehouses. (Chinthapatla, 2024; Haleem et al., 2022). In essence, structured data is the most structured type of data, with clearly established boundaries and a straightforward layout, making it ideal for tasks that require precise searching and analysis. Simpler to examine and control because of its structured form. Backed by advanced tools and techniques for handling data. Nevertheless, structured data poses certain challenges. Limited in its flexibility, as it struggles with complex or varied data formats like unstructured data (e.g., images or social media posts). Requires set schemas, hindering its adaptability to dynamic or rapidly changing data sets (Erickson et al., 2020; Saura, 2021). Structured data is essential in areas such as business analytics, financial systems, and traditional databases, where consistency, accuracy, and effective processing are key.

Semi-structured data blends elements of structured and unstructured data, falling in between the two. It possesses certain organizational features (such as tags or markers) but does not have a fixed schema, enabling adaptability in both structure and content. Semi-structured data provides increased storage flexibility over fully structured data, while also preserving some hierarchy and relationships between data points. Frequently shown in forms such as XML, JSON, or YAML (Jun-Hee, 2021; Schäfer, 2023). Metadata gives information about the data, making it simpler to understand. Instances of data include JSON or XML files employed in APIs, HTML documents, sensor data featuring different attributes, and emails with specified headers like subject and sender, alongside unstructured body content (Liu et al., 2023). Needs specialized tools and platforms (such as MongoDB, Hadoop) to store and analyze data effectively. Preprocessing may be required to be compatible with structured data models for specific applications. Semi-structured data is commonly employed in contemporary applications such as web data exchanges, IoT systems, and data lakes, connecting the rigid structure with the flexibility of unstructured data (Sreepathy et al., 2024; Dabbèchi et al., 2021).Adapting to diverse and constantly changing data sources, unstructured data is more versatile than structured data. More convenient to handle and search than unorganized data with tools such as NoSQL databases.

Unstructured data is typically not as structured as structured data, making it more challenging to analyze due to its loose format and lack of definitive internal organization. Unstructured data does not fit into a set format or organized structure, which can complicate storage, analysis, and management. It includes a broad range of data types, which are frequently full of information but do not have a built-in structure (Rawat and Yadav, 2021; Ehsanul Majid et al., 2024). Information is not arranged in a structured manner with rows and columns. Incorporates text, pictures, clips, sound, and posts from social networks. analyses data, including emails and chat logs, using cutting-edge technologies like machine learning and Natural Language Processing (NLP). Posts and comments on social media platforms. Visuals, clips, and sound files. Content on websites, blogs, and articles (Li et al., 2022; Silverman et al., 2021). There are also difficulties associated with unstructured data types. Substantial processing and specific tools are required to reveal the intended significance. Difficult to store and manage in traditional relational databases. Rising costs of performing analysis. Unstructured data is essential for analyzing customer sentiment, recognizing images, and conducting big data analytics, offering new possibilities for valuable insights in adaptable formats (Zhang et al., 2022; Tayefi et al., 2021; Zhao and Chen, 2022). It comes with full of valuable data, providing profound understanding when examined thoroughly and captures complexity found in real-life situations, making it applicable for a variety of uses.

Data integration involves an architecture that incorporates framework and design principles to integrate information from several sources into a cohesive viewpoint. It lays the groundwork for data flow management, interoperability assurance, and efficient processing and analysis support (Ribeiro and Braghetto, 2021). Challenges in data integration architecture are frequently encountered. Data heterogeneity refers to differences in formats, schemas, and standards among sources. Scalability involves managing increasing data volumes and meeting real-time needs. Interoperability involves ensuring seamless communication among systems. Ensuring performance is maintained while integrating and accessing data. Data governance is the efficient management of quality, compliance, and security (Sadeghi et al., 2024; Javed et al., 2020). Data integration has conducted thorough research on data integration, which is a fundamental process in the data pre-processing phase of machine learning pipelines. Responsible data science involves acknowledging worries about data quality and bias when utilizing integrated data for analysis and model training (Sandhu, 2022; Hossain et al., 2022).

As data volume and diversity increase, existing architectures face challenges in supporting smooth integration, scalability, and interoperability, hindering the full potential of advanced analytics and applications.

As data management challenges continue to change, various new solutions are being implemented as outlined in SLR. Data kinds, including structured, semi-structured, and unstructured data, can be managed differently by different architectural models. Different methods of data integration are used depending on the needs of the system, the types of data, and the particular use cases. These new technologies are designed to tackle the increasing challenges of handling data on a large scale and in different settings, enabling data integration to be smoother, instant, and better suited to the needs of contemporary applications.

Methods

Review Methods

This study utilizes the Systematic Literature Review (SLR) method. An SLR involves identifying, evaluating, and examining relevant research data to address specific research inquiries (Van Dinter et al., 2021). This study aims to provide a thorough analysis of every step in the process. This work may provide as a springboard for further primary research in this area. To collect all pertinent primary studies, SLR is conducted. Kitchenham established rules that will be followed during this study. The planning, carrying out, and reporting stages are the three main tasks listed in the guideline, as seen in Fig. 1 (Van Dinter et al., 2021; Paul et al., 2021). References impacted this section's format, review process, and some of its figures.

Research Questions

To keep the review focused, the Research Questions (RQ) were well-defined. The research problems and factors motivating this literature review are presented in Table 1.

Search Strategy

Choosing digital libraries, coming up with a search word, performing a preliminary search, modifying the search term, and gathering a collection of important studies from digital libraries that support the search term are all steps in the searching process (Paul et al., 2021). The published research paper from the renowned database journal was located for this study using a certain search string.

Utilize electronic databases and keywords to locate the study materials. Furthermore, it also removes keywords that are still too broad, will be narrowed down, and even revises the keywords. The process of conducting title screening is also performed at this point.

The purpose of the search string was to automatically retrieve articles from the databases of Science Direct, ACM Digital Library, IEEE Xplore, and Web of Science while conducting a search. Based on the PIO paradigm, a list of phrases was developed to aid in defining the search string (Thumburu, 2021).

The search phrase that was finally entered was ((“architecture” OR “data types” OR “application” OR “data filtering”) AND (“data integration”). The results of the search are shown in Table 2.

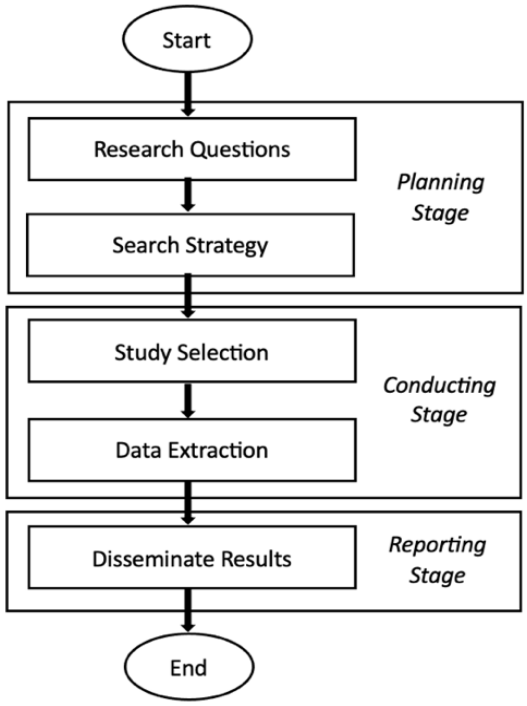

Fig. 1: Systematic Literature Review Steps

Table 1: Research Question

ID	Research Question	Motivation
RQ1	What are the issues in data integration?	To identify which technologies are used in the integration of data
RQ2	What are the data types in data integration?	To identify data types in data integration.
RQ3	What are the components of data integration architecture?	To identify the data integration architecture's constituent parts.

Table 2: Search Result of Each Database Journal

No	Database Journal	Number of Articles
1	ScienceDirect	692
2	ACM Digital Library	23
3	IEEE eXplore	271
4	Springer	2012
5	Web of Science (WoS)	26
Total		3024

Study Selection

Based on the inclusion and exclusion criteria, the primary research was chosen. In Table 3, the conditions are displayed.

After locating articles from multiple databases, review papers identified through specific keywords. Quality assessments are carried out using the 'research hierarchy' to determine whether papers that are being reviewed should be classified as research submissions or not. Conduct documentation review in a tabular format for both reviewed and unreviewed papers.

Figure. 2 shows the example of literature review process that have been conducted through this SLR. The original number of papers was decreased using criteria E2 during the first round of filtering. A total of 580 unique papers were found. Next, we found 162 potential papers following the application of criteria I4 and E1. Afterward, we utilize abstracts to find pertinent articles and subsequently employ I1 and I2 filtering. A total of 46 articles were found following this stage. In the last phase, we assess for complete text availability using I3 and E3 standards. 23 papers have been deemed acceptable.

Table 3: Inclusion and Exclusion Criteria

Inclusion Criteria	(I1) This review's main requirements are either regulatory compliance or data integration. (I2) Articles are written in English (I3) Full-text access is available for articles (I4) Papers published between 2020 and 2024
Exclusion Criteria	(E1) Extended abstracts, posters, short papers, surveys, magazines, books, chapters, editorials, notes, reviews, reference work, and reference work entries are not included. (E2) Similar articles from multiple databases and journals are not included. (E3) The study discusses data integration, however it is unrelated to architecture and applications.

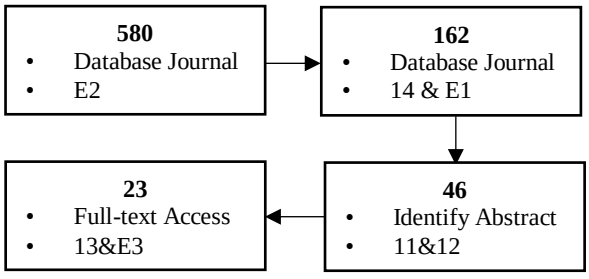

Fig. 2: Finding and Choosing Primary Studies

We recognise that the inclusion and exclusion criteria as well as the search query's key terms are responsible for this result. The implementation of data integration in different ways, for example, has been the subject of numerous research. One potential explanation is that the term application and architecture is unclear. It can be used to describe both a method and a reusable coding instrument.

Data Extraction

The selected primary studies provide information that is pertinent to addressing the research questions this review addresses. A list of all completed research papers by Scimago index may be found in Table 4. The Scimago Journal Ranking findings are displayed in Fig. 3.

Even though this work made a contribution, we found certain validity threats. Based on the 252 chosen articles, the analysis was conducted. This number of articles may jeopardise the findings of a secondary study. SLR is a relatively new field of study, which could account for the volume of research. Consideration should also be given to the search string's usage of the term SLR and its variants. Numerous publications that are beyond of scope could be found by searching with related terms like virtual learning environments, SLR experiences, and others. Finding works that at least address RQ1 was one of the primary motivations for maintaining the focus of our investigation. In order to guarantee the quality of this study, the research methodology was also founded on systematic literature review guidelines (Thumburu, 2022).

There is a risk to the validity of the technique because 2772 out of 3024 papers failed to validate it. These papers were retained in order to gather as many studies as possible and to provide a more comprehensive picture of the field.

Furthermore, they are research that have been examined by peers in the scientific community and published in journals and conferences, despite not meeting certain established quality requirements.

Table 4: Article and Number Post-Filtering Procedure

ID	Publication	Number of Articles
1	Journal Q1	56
2	Journal Q2	70
3	Journal Q3	72
4	Journal Q4	53
5	Conference	1
Total		252

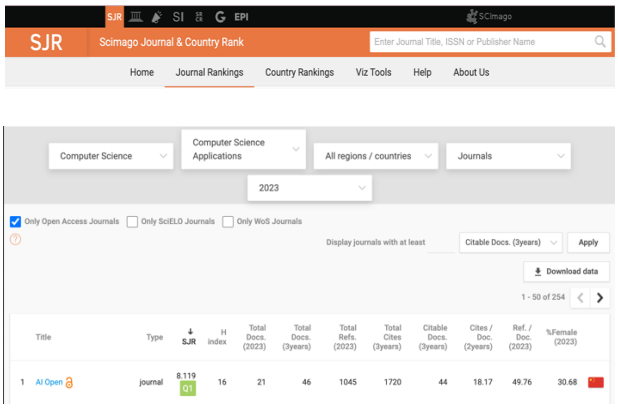

Fig. 3: Scimago Journal Ranking database

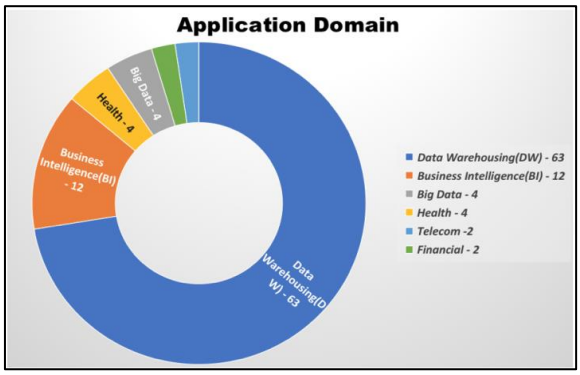

Fig. 3: ETL application domain (Sivudu, 2023)

Results

Issues in Data Integration

Numerous challenges in data integration. Table 5 shows the articles that related to issues in data integration.

According to Table V result, issues in data integration occur when attempting to merge data from various sources or systems. These matters tend to be intricate and differ based on the sources, technologies, and intentions in question. There are a few typical issues faced when dealing with data integration.

Inaccurate reporting can result from redundant records stored in various systems. Diverse formats or meanings of data across sources can lead to discrepancies. Inadequate data fields in certain systems may cause gaps in the combined dataset.

Ensuring that data is both complete and accurate is crucial for making decisions and training models. Continuous monitoring and honing of methods are necessary for eliminating bias during data cleaning (Mami, 2021; Zhang and Long, 2021). Using pre-trained models and expert knowledge can bring in potential bias risks from external sources.

Data Quality and Consistency

Different Data Formats and Standards

Systems may need different data formats such as relational databases or NoSQL, which may necessitate transformations. Certain data sources might produce information in formats that cannot be directly processed by others, such as XML compared to JSON, necessitating intermediary modifications (Zhang and Long, 2021).

Data collection includes collecting a variety of data types, such as genomic, clinical, and imaging data, from different sources. Integrating these datasets presents a difficulty because of variations in formats and standards (Raparthi et al., 2021; Pansara, 2023). Gathering and merging these datasets present major obstacles because of variations in formats, standards, and privacy rules. Data integration is the process of blending information from various origins to form a unified and standardized dataset for examination (George, 2022; Nargesian et al., 2022).

Scalability and Performance

Integration with large amounts of data can affect system performance in real-time integrations if not handled properly. Integrating data from various technologies and formats can be problematic due to delays (Mami, 2021). Following certain legal requirements, like GDPR, increases the level of complexity. Operational obstacles can also stem from the use of resources and possible bottlenecks. Businesses experience varying expenses because of utilizing numerous cloud services (Zhang and Long, 2021; Stojanovic et al., 2022). Effective cost optimization requires a dependable network infrastructure. Ensuring top-notch staff is essential for reducing operational costs (Stojanovic et al., 2022).

The growing amount of data and business needs has caused an increase in complexity in modern information systems. Contemporary businesses frequently use various data sources, such as relational and nonrelational databases, which lacks in terms of performance and the inflexibility of data model (Pansara, 2023; Ekundayo et al., 2023).

Data Governance and Compliance

Determining the ownership of different data pieces can be difficult, particularly when merging data from various departments or organizations. Regulations requiring cautious handling of personal data, like the CCPA or GDPR, make integration attempts more difficult (Mami, 2021, Martínez-García and Hernández-Lemus, 2022).

Table 5: Issues in Data Integration

No	Issues	Number of Articles	Reference
1	Data Quality and Consistency	5	(Mami, 2021; Zhang and Long, 2021; Raparthi et al., 2021; George, 2022; Stojanovic et al., 2022)
2	Different Data Formats and Standards	5	(Mami, 2021; Raparthi et al., 2021; George, 2022; Pansara, 2023; Nargesian et al., 2022)
3	Scalability and Performance	5	(Mami, 2021; Zhang and Long, 2021; Stojanovic et al., 2022; Pansara, 2023; Ekundayo et al., 2023)
4	Data Governance and Compliance	7	(Mami, 2021; Raparthi et al., 2021; Stojanovic et al., 2022; Martínez-García et al., 2022; Ekundayo et al., 2023, Nargesian et al., 2022; Anadiotis et al., 2022)
	Total		22

There is a growing interest in data mining within the medical field, especially in efficiently extracting information from Electronic Health Records (EHR). Analyzing EHR data is made challenging by its diverse mix of quantitative, qualitative, and transactional data, especially when dealing with multiple patients across institutions and centers. Initiatives in this field, such as extensive global partnerships, are essential for realizing the value of EHRs as a valuable data resource in healthcare (Zhang and Long, 2021; Stojanovic et al., 2022; Ekundayo et al., 2023).

Intelligence and information are essential for firms to maximise their decision-making process. Initiatives for data governance are clearly necessary to guarantee the availability and quality of the appropriate data for creativity and innovation. The storage of data must adhere to a sound framework based on principles to maintain data quality, especially as the volume of data continues to increase. Ultimately, the ability to maintain data quality remains a constant objective amidst the evolving frameworks for data storage Nargesian, et al., 2022; Anadiotis et al., 2022).

Various data types are essential for defining, arranging, and handling data from different sources in data integration. Effectively managing these different data types guarantees seamless integration and compatibility among (Raparthi et al., 2021). Each data type presents unique challenges for integration, especially when used in combination with others.

Data Types

Data integration involves three different categories of data. Table 6 displays the articles containing data types in data integration.

Table 6, indicates that integrating data from structured, semi-structured, and unstructured sources comes with distinct difficulties because of the varying ways in which these types of data are structured, stored, and managed (Abraham et al., 2019; Seenivasan, 2025). We discuss the issues to integrate based on data types below.

Structured Data

Structured data is very well-organized and easily searchable, usually kept in relational databases with a predetermined schema such as SQL databases (Abraham et al., 2019). Some issues while integrating structured data are (Malik et al., 2020; Jiang et al., 2022; Seenivasan, 2025):

(i) *Schema Mismatch:* Variances in schema between databases may necessitate in-depth mapping to align fields accurately
(ii) *Data Type Discrepancies:* Differences in data types (e.g., integers vs. strings) can result in mistakes during integration and analysis
(iii) *Limited Flexibility:* Structured data is strictly defined, making it challenging to adapt to new data types or business requirements without modifying the schema
(iv) *Data Quality Issues:* Even though structured data is usually more organized, problems such as missing values or duplicates may still arise, affecting integration methodologies

Semi-Structured Data

There are no strict schemas that apply to semi-structured data, yet it does include markers or tags to distinguish data elements (e.g., XML, JSON). The text below explains the challenges of data integration for semi-structured data (Abraham et al., 2019; Jiang and Zhao, 2022; Zhao and Chen, 2022; Weitzenboeck et al., 2022):

(i) *Parsing Complexity*: The extraction procedure and converting semi-structured data may be more challenging than structured data because of its flexible format, necessitating the use of specific tools for parsing
(ii) *Inconsistent Structures:* Varying structures can make it difficult to integrate data, like different JSON formats, leading to challenges in creating a cohesive perspective
(iii) *Schema Evolution:* Alterations to the data format as time passes may cause compatibility problems, requiring continuous modifications to integration procedures
(iv) *Data Quality Management:* Managing data quality can be challenging, as semi-structured data lacks the validation typically found in structured data

Table 6: Data Types

No	Data Types	Number of Articles	Reference
1	Structured Data	4	(Abraham et al., 2019; Malik et al., 2020; Jiang and Zhao, 2022; Seenivasan, 2025)
2	Semi-structured Data	4	(Abraham et al., 2019; Jiang and Zhao, 2022; Zhao and Chen; 2022; Seenivasan, 2025)
3	Unstructured Data	5	(Abraham et al., 2019; Jiang and Zhao, 2022; Weitzenboeck et al., 2022; Seenivasan, 2025; Gupta et al., 2022)
Total		13	

Unstructured Data

Unstructured data presents a difficulty in integration due to its absence of predefined format or structure, such as text documents, images, and videos. Data integration for unstructured data is facing several challenges (Abraham et al., 2019; Jiang and Zhao, 2022, 2022; Weitzenboeck et al., 2022; Seenivasan, 2025):

(i) *Lack of Standardization:* Unstructured data can present in various formats, hindering the development of consistent integration standards
(ii) *Data Processing Complexity:* Extracting useful insights from unstructured data adds complexity to data processing, frequently requiring the application of sophisticated techniques like machine learning or Natural Language Processing (NLP), which can be resource intensive
(iii) *Storage and Retrieval Challenges:* Challenges in storing and retrieving unstructured data may call for diverse storage solutions, such as data lakes, and retrieval techniques, making integration with structured and semi-structured data more complex
(iv) *Data Quality and Governance:* Ensuring the quality and adherence of unstructured data can be difficult because of its varied nature and absence of built-in structure

By grasping the unique difficulties related to combining structured, semi-structured, and unstructured data, businesses can enact focused tactics to enhance their data integration procedures and extract deeper insights from their wide-ranging data resources.

Data Integration Architecture

Different architectures are used in the integration of data. Table 7 shows various architectures in data integration.

Hybrid Integration ETL/ELT Architecture

Hybrid of Extract, Transform, Load (ETL) and Extract, Load, Transform (ELT) is a practical option for companies moving to modern data structures while still utilizing older systems. It enables the combination of the advantages of both worlds for effective, expandable, and compliant data integration (Rambabu et al, 2023). Before proceeding, there are detailed explanations of ETL and ELT that need to be provided.

Data is extracted from source systems, formatted for analysis, and then loaded into a destination data warehouse using traditional ETL techniques (Nwokeji and Matovu, 2021). This approach's systematic procedure, which guarantees that data is cleaned and altered before being stored, has led to its widespread use. Pre-processing can improve data consistency and quality, but because of the lengthy transformation phase, it may also cause latency (Yu et al., 2022).

Figure 3 (Sivudu, 2023), which identifies 6 domains using ETL solutions, shows the breadth of coverage in ETL research by application domain, geographic location, and publication frequency.

63 publications discuss data warehouses, 12 mention business intelligence, 4 mention big data, and 4 mention health. There were 3 papers that addressed transportation, 2 papers that addressed communications, and 2 publications that highlighted banking.

An open-source ETL tool based on research was developed to help with data conversion from the Observational Medical Outcomes Partnership (OMOP) CDM to the National Patient-Centered Clinical Research Network (PCORnet) Common Data Models (CDM). A dataset of 1000 randomly chosen patients from the Mayo Clinic's PCORnet CDM was used to assess the ETL tool. The effectiveness of the ETL tool was evaluated using gap analysis, data mapping accuracy, and information loss techniques (Zecchini, 2024).

The database mapping quality evaluation map is displayed in Fig. 4, and the needed fields are denoted by "*" (Zecchini, 2024). Three different transformation strategies value transformation, rule-based transformation, and concept code mapping were identified through the assessment of the ETL tool's data transformation and standardisation performance. Although the data from concept mapping presented difficulties, the source data in the value and rule-based transformation procedures was effectively transformed into the OMOP CDM.

Table 7: Data Integration Model

No	Architecture	Number of Articles	Reference
1	Hybrid Integration ETL/ELT	10	(Rambabu et al., 2023; Nwokeji and Matovu; 2021; Yu et al., 2022; Sivudu, 2023; Zecchini, 2024; Simitsis et al., 2023; Peng et al., 2024; Baunsgaard et al., 2021; Jarke and Quix, 2022; Stergiou et al., 2022)
2	Federated Based (FB)	8	(Qi et al., 2024; Gadekallu et al., 2026; Yin et al., 2020; Gu et al., 2024; Sahara and Aamer, 2022; Ngo et al., 2020; Rachakatla et al., 2021)
3	Data Warehousing (DW)	8	(Peng et al., 2024; Antunes et al., 2022; Prasetyo, 2022; Oliveira e Sá et al., 2024; Bogdanov et al., 2020; Rozony et al., 2024; Seenivasan, 2025; Ngo et al., 2020)
5	Data Virtualization (DV)	4	(Sandhu, 2022)
	Total		30

DOI: 10.3844/jcssp.2026.487.503

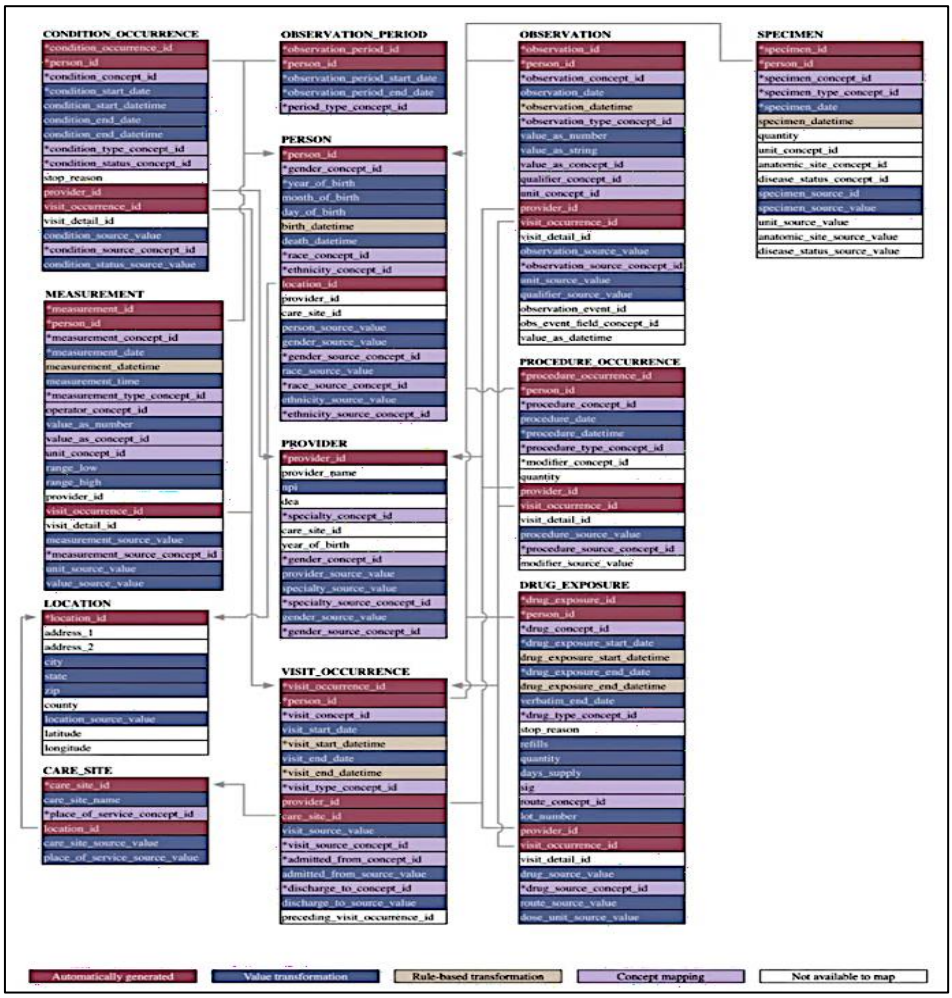

Fig. 4: Database mapping quality evaluation map (Zecchini et al., 2024)

Analysing the information loss during the concept code mapping process revealed challenges, especially with the manual, regular-expression, and vocabulary-based approaches (Fig. 5). loss of information throughout the ETL procedure. A. Information loss in concept mapping. B. Information loss in string mapping. C. Information loss via manual mapping. The figure's table name abbreviation indicates that we did not extract pertinent information from our initial PCORnet CDM-based database at the Mayo Clinic. This study offers insightful information for future improvements to our ETL procedure (Zecchini et al., 2024).

In conclusion, traditional approaches to ETL development often suffer from inefficiencies, resulting in lengthy development cycles and limited scalability.

To address these challenges, it is crucial to embrace modern ETL tools and methodologies, which offer improved efficiency, shorter development cycles, and enhanced scalability. By utilizing these advanced approaches, businesses can streamline their data integration processes, reduce operational costs, and obtain an advantage in the ever-changing digital market (Simitsis et al., 2023).

Data extraction from source systems comes first in ELT, then data loading into the target data warehouse, and finally transformation activities within the warehouse environment. This approach efficiently manages large-scale changes by leveraging the capabilities of contemporary data warehouses, particularly cloud-based platforms. Better scalability and lower data latency are inherent benefits of ELT, which can be tuned for best results by executing transformations as needed (Nwokeji and Matovu, 2021).

The emergence of ELT as a standard in data processing signifies a significant shift in the industry. The factors driving this transition include increased processing power, the rise of cloud computing, and the support of big data technologies like Hadoop and Spark. These developments have made ELT a compelling choice for data transformation, leveraging the computational power and scalability of modern data warehouses.

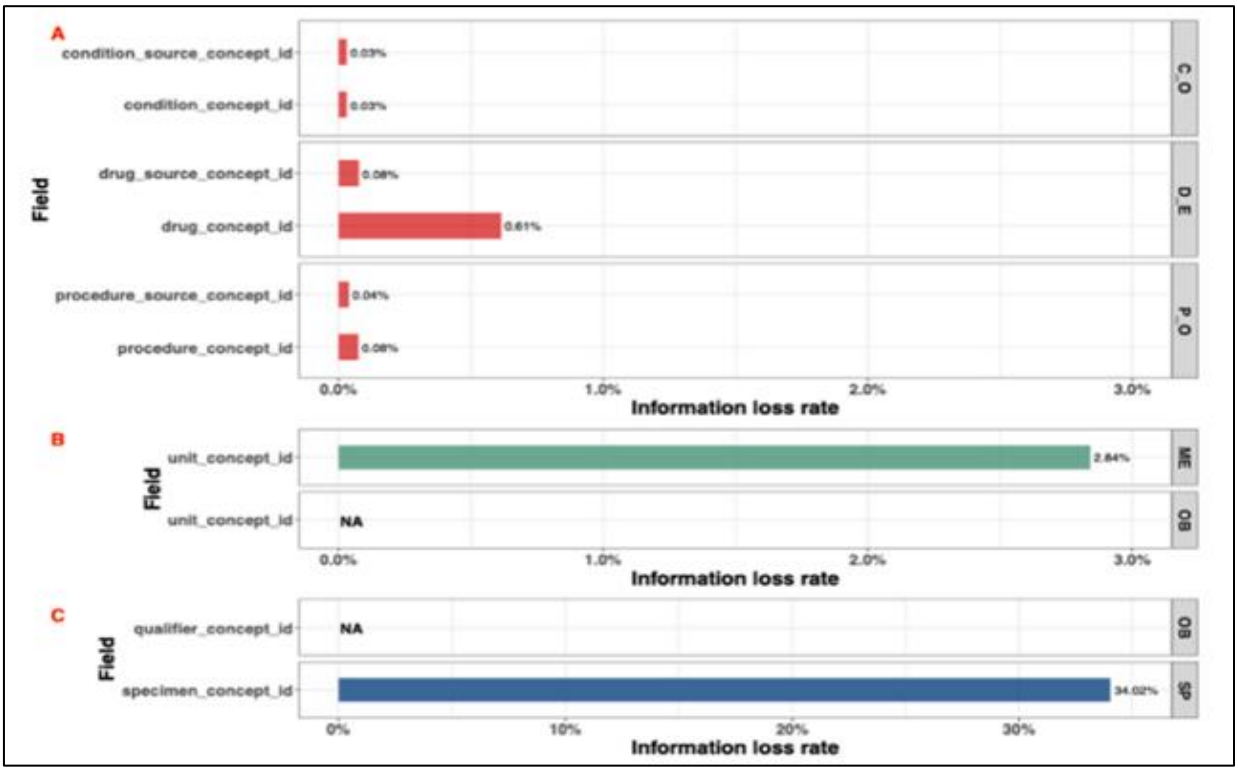

Fig. 5: Analysis of information loss (Zecchini et al., 2024)

Table 8: Comparative Analysis of ETL and ELT (Simitsis et al., 2023)

Criteria	ETL	ELT
Performance	Slower due to data transformations before loading	Faster due to the direct loading of raw data
Scalability	Limited by transformation engine capacity	Scales with data warehouse capabilities
Cost	Higher initial setup and maintenance costs	Lower initial setup costs, potentially higher operational costs

Table 8 shows a result, ELT has become a worthy substitute for ETL, offering speed, flexibility, and robust support for the analysis of large volumes of raw data. Table 1 shows the comparison of ETL and ELT (Peng et al., 2024).

ETL process involves three main steps: extraction, transformation, and loading (Baunsgaard et al., 2021). Data is imported into a data warehouse after being extracted from several sources and formatted accordingly. The explosion of big data presents the challenge of dealing with diverse data sources, including relational and NoSQL databases, open data from public administrations, and data from sensors and IoT devices. This diversity requires dedicated procedures for data management, and the ELT paradigm allows practitioners to transform and integrate relevant portions for business intelligence operations or efficient querying in relational database management systems.

There are three phases involved in Hybrid ETL/ELT processes. The initial stage is the ETL phase, used for immediate data transformation needs prior to loading (e.g., for compliance, standardization, or legacy system compatibility). Change takes place in a staging zone or middle server (Rambabu et al., 2023).

Perfect for systems that have restricted abilities to transform after loading. The next stage is the ELT phase which information is uploaded to a cloud-based data platform or a contemporary data warehouse such as Snowflake, Big Query, or Redshift. Next, processing tasks are managed within the destination system utilizing its computing capabilities and built-in tools. The third stage involves transitioning between ETL and ELT, as decided dynamically by the model on whether to transform data before or after loading, considering data attributes, performance requirements, and system capabilities (Jarke and Quix, 2022; Stergiou et al., 2022).

Hybrid ETL/ELT is being utilized by a technology company. A tech company employed a combination of ETL and ELT processes. While ETL was used for initial data cleaning and merging, ELT was utilized for more advanced analytics in the data warehouse. The hybrid ETL/ELT model poses challenges due to the bi-directional processes being more difficult, but offers

advantages such as increased efficiency, precise operations, and financial effectiveness (Rambabu et al., 2023; Jarke and Quix, 2022; Stergiou et al., 2022).

Data Federation Architecture

Federated learning has the potential to be widely used in various applications in practice. By automatically creating federated execution plans, this makes it possible to reuse and implement a wide variety of machine learning algorithms, data preparation methods, and model debugging techniques in federated contexts (Baunsgaard et al., 2021). A federated model for data integration allows querying and interacting with data from various sources without consolidating it into one central repository. The focus is to offer a consolidated, digital representation of information, while keeping the original data in its respective systems. This involves organizing and connecting data sources, establishing trust among participants, and providing data ownership services (Qi et al., 2024; Ngo et al., 2020).

The basic responsibilities in a data space and how they interact with one another are depicted in Fig. 6. A data space is a distributed data integration notion, as was previously discussed. As a result, there is no central data repository or data vault where data providers can deposit their data and where data consumers can access and retrieve it.

Based on a hypothetical situation that seeks to create a federated cooperative system, cooperative CSPs are essential for load-balancing data management and transmission, user authentication, and other tasks. As illustrated in Fig. 7, academic users can seamlessly access academic data through authorized CSPs, facilitating faster data management on the CSP Cloud platform (Yin et al., 2020).

Similar to hundreds or even thousands of devices in an IoT-centered federated system, a typical federated system has many clients. While aggregating models, many clients must transmit their local updates to the network simultaneously, leading to potential communication congestion issues caused by bandwidth limitations (Gu et al., 2024).

Federated data integration has important characteristics. Initially, data virtualization generates a virtual layer to access and interrogate data from different systems instantly without the need to move or duplicate the data. Some popular tools are Denodo, Red Hat Data Virtualization, and Dremio (Ngo et al., 2020; Rachakatla et al., 2021). Additionally, Query Federation enables queries to cover various sources and merge the outcomes into a cohesive result. SQL-based interfaces frequently aid in communication between systems.

Thirdly, heterogeneous source compatibility enables support for a variety of data sources such as relational databases, NoSQL databases, APIs, files, and cloud systems. Metadata Management involves maintaining information about data sources to increase query processing efficiency and ensure accurate data mapping. Additionally, Security and Access Control aim to retain the security features of source systems and facilitate federated access control policies.

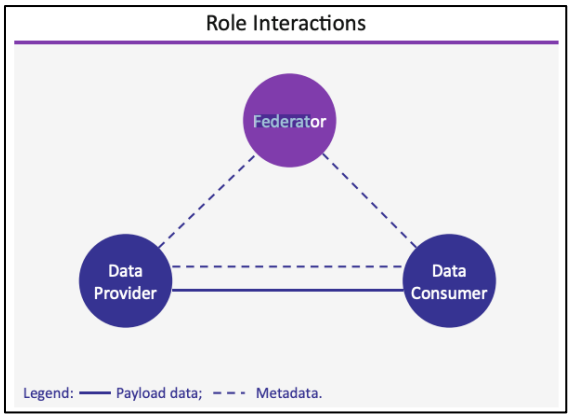

Fig. 6: Data Space Roles (Gadekallu et al., 2026)

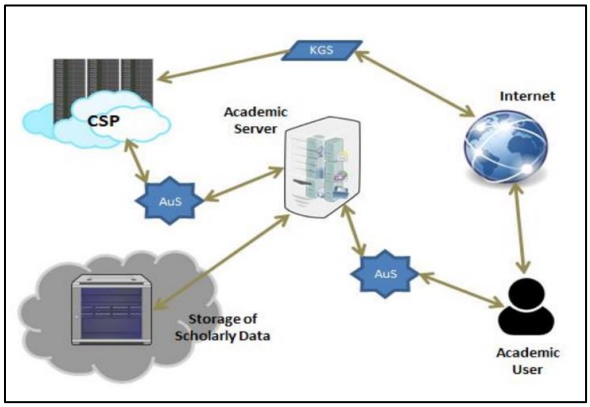

Fig. 7: System Architecture (Yin et al., 2020)

IoT's growth has led to many big data services and applications. Traditional AI/ML for big data faces critical issues like data privacy, variety, communication efficiency, and scalability. Federated Language is a breakthrough in addressing these challenges. This paper provides a comprehensive survey on Federated Language's use in big data. It highlights the potential of Federated Language and proposes future exploration directions (Sahara and Aamer, 2022; Rachakatla et al., 2021).

Data Warehouse Architecture

Data warehousing is a fundamental component for successful predictive modelling as it combines large quantities of financial data, both historical and real-time, from various sources into one central storage system. By utilizing this unified data, AI-based predictive models, including machine learning algorithms, deep learning structures, and other advanced statistical methods, can

provide detailed insights and predictions that are accurate and actionable (Antunes et al., 2022; Oliveira e Sá et al., 2024).

Data warehousing began by gathering data from operational databases into centralized warehouses to provide business leaders with analytical insights for decision support and business intelligence (BI). Information in these storage facilities would be recorded using the schema-on-write method, guaranteeing that the data structure was designed for efficient downstream BI utilization. This is commonly known as the initial generation of data analytics platforms (Prasetyo, 2022; Rozony et al., 2024).

Figure 8 (Oliveira e Sá et al., 2024) shows data warehousing concept. The goal of data warehousing, a specialized approach to data management, is to make it easier to combine, store, and analyze vast volumes of data from several sources. Fundamentally, a data warehouse is a central storage space that combines information from various operational systems, allowing organizations to conduct thorough data analysis and business intelligence tasks. The main goal of data warehousing is to aid in the decision-making process by offering a consolidated data view that allows for querying and analysis across different dimensions (Oliveira e Sá et al., 2024).

As seen in Fig. 9, the four separate modules that make up a typical DW system's design are Raw Data, Extraction Transformation Loading (ETL), Integrated Information, and Data Mining. In this instance, the Raw Data (source data) module is first stored in spreadsheets, flat files, and SQL, among other storage methods. It is often necessary to clean raw data, fix noise and outliers, and deal with missing numbers. Then, using an Extract, Transform, Load (ETL) module, the data must be combined and reinforced before being moved into a data warehouse (Bogdanov et al., 2020).

Data warehousing is a centralized storage unit that integrates information from various sources, making it easier to gain accurate insights for analysis. Data warehousing mainly stores structured data, structured into schemas like star or snowflake schemas, and includes historical data. Preserves past data to facilitate the analysis of patterns over time. It is specifically optimized for queries, catering to read-heavy workloads that involve intricate queries and aggregations. Utilizes indexing, partitioning, and caching to improve performance. The Data Integration model utilizes procedures such as ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform) to sanitize and organize data. Peng et al., 2024; Bogdanov et al., 2020; Rozony et al., 2024)

Different types of data warehouses include Enterprise Data Warehouse (EDW), which is a centralized platform for analytics across the entire organization. Operational Data Store (ODS), employed for close to real-time data analysis, frequently serves as a temporary storage location prior to data being transferred to the EDW (Seenivasan, 2025).

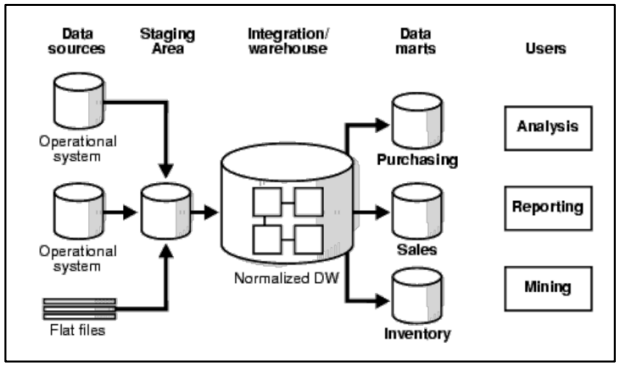

Fig. 8: Data Warehousing Concept (Oliveira e Sá et al., 2024)

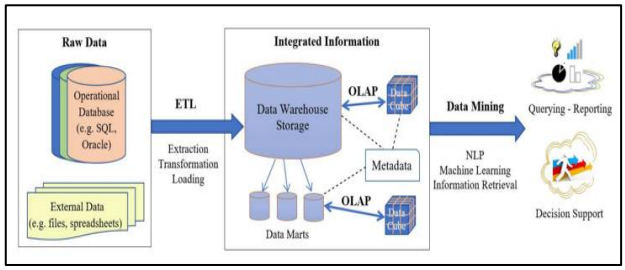

Fig. 9: Agricultural Data Warehouse Architecture (Bogdanov et al., 2020)

A few instances of cloud data warehouses are Azure Synapse, Amazon Redshift, Google Big Query, and Snowflake. Allow for expansion, flexibility, and reduced costs for facilities and equipment (Ngo et al., 2020).

The guarantee of data consistency, which ensures constant definitions and a trustworthy source of truth, is one benefit of data warehousing in data integration. Scalability is perfect for historical analysis as it's designed for handling and examining extensive amounts of historical information. Quality of data is important for ensuring that data during integration is clean and accurate. The Data Integration Model has limitations in real-time capabilities, causing challenges for traditional data warehouses in comparison to event-driven models. Moving data around and physically consolidating it can require a lot of resources and be expensive. Traditional data warehouses primarily support structured data, while modern requirements frequently involve semi-structured or unstructured data, resulting in limited support for unstructured data (Oliveira et al., 2024; Bogdanov et al., 2020; Rozony et al., 2024).

Modern extensions like Snowflake and Big Query are cloud-based warehousing solutions that are well-suited for integrating with real-time and unstructured data. Architecture of lake houses that merges data lakes and warehouses to accommodate various data types and integration patterns (Ngo et al., 2020). To conclude, data warehousing is a traditional data integration model

designed for centralized, structured, and historical data, but it is not as suitable for real-time and decentralized needs when compared to more recent models.

Data Virtualization (DV) Architecture

Data virtualization is a method of data integration that allows users to retrieve and analyze data from various sources across different locations without the need to transfer or duplicate the data. It hides the specifics of the data sources and offers an up-to-date, simulated display of the data (Sandhu, 2022). There are multiple features of data virtualization. Unified data access combines data from different sources such as databases, APIs, and data lakes to form a cohesive virtual layer. Users interact with data as if it were all in one place. Real-time data integration allows for real-time querying and retrieval, removing the need for ETL or data duplication.

Having diverse sources of support enables the processing of different kinds of data from both on-site and cloud environments. Managing metadata involves using metadata to link data sources, enabling schema modifications and enhancing query performance. Security and governance involve upholding strict control over access and compliance with rules, while respecting the policies of the initial system.

Figure 10 shows hybrid cloud frameworks or data virtualization technologies which provide a solution for handling data format and protocol inconsistencies through the creation of a virtual layer that displays a cohesive view of data from various systems. Data virtualization allows organizations to retrieve and analyze data from various sources without the need to physically transfer or replicate it. This method makes data integration easier by offering a uniform interface for querying and analyzing data, no matter its initial format or location. Organizations can enhance integration processes and boost data access in hybrid cloud environments by using data virtualization.

To prove that data virtualization's importance in current data integration is depicted in Fig. 11.

Figure 11 indicates that 31% of participants' companies consider decreasing latency in ETL routines a top priority for enhancing data pipelines and preparation. As the amount and variety of data grow, along with an increase in user demand for up-to-date data, the difficulties of ETL latency tend to rise. Using data virtualization can be a substitute for displaying data without physically moving it; 14% aim to enhance their systems by establishing a data virtualization layer.

Advantages of using data virtualization includes quicker access to insights is achieved by eliminating the delays common in traditional ETL processes through direct data access. Save on storage costs by avoiding duplicate data. Able to adjust easily to alterations in data sources or new integration requirements. Reduce duplication, maintains accuracy by accessing real-time data instead of depending on duplicates. Simplified data management, offers a single interface to access distributed data, decreasing complexity.

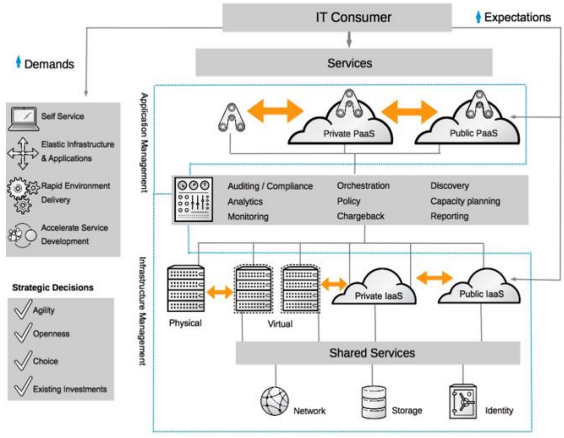

Fig. 10: Hybrid cloud frameworks (data virtualization)

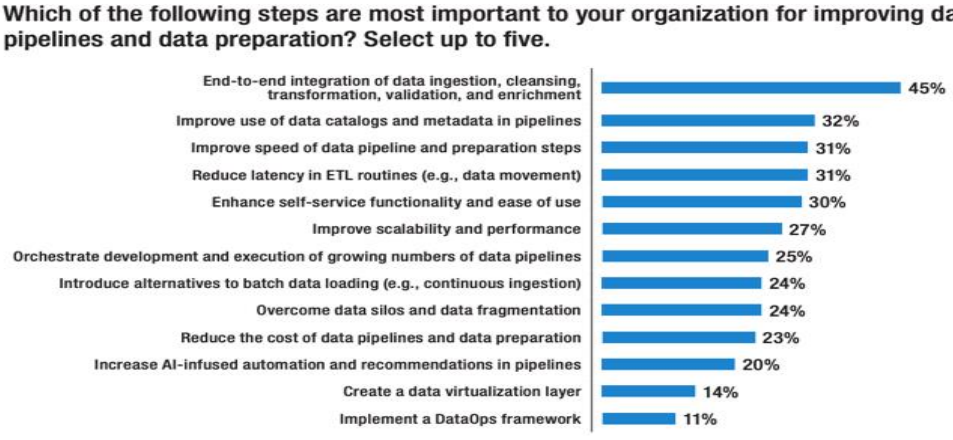

Fig. 11: Based on answers from 364 respondents' satisfaction

Data virtualization is becoming popular as a contemporary, flexible method for data integration, particularly for real-time analytics, hybrid cloud setups, and scenarios where little data transfer is important.

Discussion and Conclusion

According to the study, the Table 9 presents a comparison of data integration architectures based on previous section.

Table 9: Comparisons of Data Integration Architectures

Architecture / Aspects	Hybrid Integration ETL/ELT	Data Federation	Data Warehouse	Data Virtualization
Data Movement	*Centralized + Virtual*	*Virtual, On-Demand*	*Centralized, Batch*	*Virtual, Real-Time*
Real-Time Support	*High*	*Medium*	*Limited*	*High*
Performance	*Flexible, Use-Case Driven*	*Medium for Ad Hoc Queries*	*High for Analytics*	*Dependent on Source*
Scalability	*High*	*Low*	*High*	*Medium*
Complexity	*High*	*Low*	*Medium-High*	*Low-Medium*
Governance	*Flexible*	*Source-Based*	*Centralized*	*Source-Based*
Cost	*High*	*Low-Medium*	*High*	*Medium*
Use Cases	*Mixed Workloads*	*Ad Hoc Queries*	*Historical Analysis*	*Real-Time Reporting*

Data integration is essential for combining data from various sources to improve decision-making, efficiency, and insights. Challenges consist of upholding data accuracy, dealing with regulations, handling technical differences, and incorporating with legacy systems. Meeting these requirements involves robust data management practices, effective data quality assurance, and advanced integration solutions. A strategic method turns data integration into an asset for enhanced analytics, customer insight, and operations efficiency.

The paper presents a comprehensive review of data integration, emphasizing its challenges, types of data, and architectural aspects. Data integration brings together data from different sources to improve decision-making, boost analytics, and increase operational efficiency. Crucial for industries such as healthcare, IoT, and big data analytics. Nevertheless, data integration poses challenges such as deficiencies in data, including missing, redundant, or inconsistent information. Harmonization is needed for various schemas, formats, and standards that are different from each other. Growing amounts of data and the requirement for instant processing are putting pressure on systems. GDPR necessitates strong data governance and security measures. Various types of data were also discussed including Structured Data, Semi-structured Data, and Unstructured Data.

According to the information presented in Table 7, it shows the comparisons among the data integration architectures. Hybrid ETL/ELT merges conventional ETL (Extract, Transform, Load) with contemporary ELT (Extract, Load, Transform) methods to enhance effectiveness. Federated systems allow for querying multiple sources without needing to consolidate data centrally. Data warehousing is a centralized storage solution for structured data, designed for analytics rather than real-time requirements. Data virtualization allows for real-time access to data without the need for duplicating the data.

The SLR employs a methodical approach, criteria for including or excluding studies, and assessments of quality to pinpoint relevant studies. Concentrating on articles released between 2020 and 2024, encompassing prominent journals and databases. Sophisticated integration techniques such as hybrid ETL/ELT and data virtualization provide flexibility and scalability. Conventional approaches like data warehousing struggle to meet the needs of real-time and unstructured data.

Nevertheless, there is also a requirement for frameworks that can handle dynamic and diverse data while improving methods of governance and inventive tools to guarantee adherence and safety. Data integration is essential for contemporary analytics; however, it involves addressing obstacles such as quality, diversity, and adherence. Merging classic designs with contemporary advancements can facilitate smooth incorporation between various data types and sources.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Both the authors have equally contributed to this manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, *49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Anadiotis, A. C., Balalau, O., Conceição, C., Galhardas, H., Haddad, M. Y., Manolescu, I., Merabti, T., & You, J. (2022). Graph integration of structured, semistructured and unstructured data for data journalism. *Information Systems*, *104*, 101846. https://doi.org/10.1016/j.is.2021.101846

Antunes, A. L., Cardoso, E., & Barateiro, J. (2022). Incorporation of ontologies in data warehouse/business intelligence systems - a systematic literature reviewof Ontologies in Data Warehouse/Business Intelligence Systems - A Systematic Literature Review. *International Journal of Information Management Data Insights*, *2*(2), 100131. https://doi.org/10.1016/j.jjimei.2022.100131

Baunsgaard, S., Boehm, M., Chaudhary, A., Derakhshan, B., Geißelsöder, S., Grulich, P. M., Hildebrand, M., Innerebner, K., Markl, V., Neubauer, C., Osterburg, S., Ovcharenko, O., Redyuk, S., Rieger, T., Rezaei Mahdiraji, A., Wrede, S. B., & Zeuch, S. (2021). ExDRa: Exploratory Data Science on Federated Raw Data. *Proceedings of the 2021 International Conference on Management of Data*, 2450–2463. https://doi.org/10.1145/3448016.3457549

Bogdanov, A., Degtyarev, A., Shchegoleva, N., Khvatov, V., & Korkhov, V. (2020). Big Data Virtualization: Why and How? *Proceedings of the 22nd Conference on Information Technologies: Algorithms, Models, Systems*, 11–21.

Borges do Nascimento, I. J., Marcolino, M. S., Abdulazeem, H. M., Weerasekara, I., Azzopardi-Muscat, N., Gonçalves, M. A., & Novillo-Ortiz, D. (2021). Impact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future Studies. *Journal of Medical Internet Research*, *23*(4), e27275. https://doi.org/10.2196/27275

Chinthapatla, S. (2024). Data Engineering Excellence in the Cloud: An In-Depth Exploration. *International Journal of Engineering, Science & Mathematics*, *13*(3), 11–19.

Cauteruccio, F., Giudice, P. L., Musarella, L., Terracina, G., Ursino, D., & Virgili, L. (2020). A Lightweight Approach to Extract Interschema Properties from Structured, Semi-Structured and Unstructured Sources in a Big Data Scenario. *International Journal of Information Technology & Decision Making*, *19*(03), 849–889. https://doi.org/10.1142/s0219622020500182

Dabbèchi, H., Haddar, N. Z., Elghazel, H., & Haddar, K. (2021). NoSQL Data Lake: A Big Data Source from Social Media. *Proceedings of the 20th International Conference on Hybrid Intelligent Systems (HIS 2020)*, *169*, 93–102. https://doi.org/10.1007/978-3-030-73050-5_10

Deepa, N., Pham, Q.-V., Nguyen, D. C., Bhattacharya, S., Prabadevi, B., Gadekallu, T. R., Maddikunta, P. K. R., Fang, F., & Pathirana, P. N. (2022). A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems*, *131*, 209–226. https://doi.org/10.1016/j.future.2022.01.017

Ehsanul Majid, M., Marinova, D., Hossain, A., E. H. Chowdhury, M., & Rummani, F. (2024). Use of Conventional Business Intelligence (BI) Systems as the Future of Big Data Analysis. *American Journal of Information Systems*, *9*(1), 1–10. https://doi.org/10.12691/ajis-9-1-1

Ekundayo, T., Bhaumik, A., & Chinoperekweyi, J. (2023). Identifying the core data governance framework principle: a framework comparative analysis. *Organization Leadership and Development Quarterly*, *5*(1), 30–53.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). *Autogluon-tabular: Robust and accurate automl for structured data*. https://doi.org/https://doi.org/10.48550/arXiv.2003.06505

Gadekallu, T. R., Pham, Q.-V., Huynh-The, T., Feng, H., Fang, K., Pandya, S., Liyanage, M., Wang, W., & Nguyen, T. T. (2026). Federated learning for big data: A survey on opportunities, applications, and future directions. *Engineering Applications of Artificial Intelligence*, *166*, 113614. https://doi.org/10.1016/j.engappai.2025.113614

George, J. (2022). Optimizing hybrid and multi-cloud architectures for real-time data streaming and analytics: Strategies for scalability and integration. *World Journal of Advanced Engineering Technology and Sciences*, *7*(1), 174–185. https://doi.org/10.30574/wjaets.2022.7.1.0087

Gu, Z., Corcoglioniti, F., Lanti, D., Mosca, A., Xiao, G., Xiong, J., & Calvanese, D. (2024). A systematic overview of data federation systems. *Semantic Web*, *15*(1), 107–165. https://doi.org/10.3233/sw-223201

Gupta, A., Selvaraj, P., Singh, R. K., Vaidya, H., & Nayani, A. R. (2022). The Role of Managed ETL Platforms in Reducing Data Integration Time and Improving User Satisfaction. *Journal for Research in Applied Sciences and Biotechnology*, *1*(1), 83–92. https://doi.org/10.55544/jrasb.1.1.12

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *2*(4), 100089. https://doi.org/10.1016/j.tbench.2023.100089

Hossain, M. R., Akhter, F., & Sultana, M. M. (2022). SMEs in Covid-19 Crisis and Combating Strategies: A Systematic Literature Review (SLR) and A Case from Emerging Economy. *Operations Research Perspectives*, *9*, 100222. https://doi.org/10.1016/j.orp.2022.100222

Jarke, M., & Quix, C. (2022). *Federated Data Integration in Data Spaces*. https://doi.org/10.1007/978-3-030-93975-5_11

Javed, A., Malhi, A., Kinnunen, T., & Framling, K. (2020). Scalable IoT Platform for Heterogeneous Devices in Smart Environments. *IEEE Access*, *8*, 211973–211985. https://doi.org/10.1109/access.2020.3039368

Jiang, L., & Zhao, Z. (2022). JSONSki: streaming semi-structured data with bit-parallel fast-forwarding. *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 200–211. https://doi.org/10.1145/3503222.3507719

Jun-Hee, L. (2021). *Space-efficient Representation of Semi-structured Document Formats Utilizing Succinct Data Structures*.

Krishnamoorthy, S., Dua, A., & Gupta, S. (2023). Role of emerging technologies in future IoT-driven Healthcare 4.0 technologies: a survey, current challenges and future directions. *Journal of Ambient Intelligence and Humanized Computing*, *14*(1), 361–407. https://doi.org/10.1007/s12652-021-03302-w

Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., & Radev, D. (2022). Neural Natural Language Processing for unstructured data in electronic health records: A review. *Computer Science Review*, *46*, 100511. https://doi.org/10.1016/j.cosrev.2022.100511

Liu, X. (2023). *Service-oriented tools for checking the identity of XML documents*.

Malik, A., Burney, A., & Ahmed, F. (2020). A Comparative Study of Unstructured Data with SQL and NO-SQL Database Management Systems. *Journal of Computer and Communications*, *08*(04), 59–71. https://doi.org/10.4236/jcc.2020.84005

Mami, M. N. (2021). *Strategies for a semantified uniform access to large and heterogeneous data sources*.

Martínez-García, M., & Hernández-Lemus, E. (2022). Data Integration Challenges for Machine Learning in Precision Medicine. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.784455

Nargesian, F., Asudeh, A., & Jagadish, H. V. (2022). Responsible Data Integration: Next-generation Challenges. *Proceedings of the 2022 International Conference on Management of Data*, 2458–2464. https://doi.org/10.1145/3514221.3522567

Ngo, V. M., Khac, N. A. L., & Kechadi, M. T. (2020). Data warehouse and decision support on integrated crop big data. *International Journal of Business Process Integration and Management*, *10*(1), 17. https://doi.org/10.1504/ijbpim.2020.113115

Nwokeji, J. C., & Matovu, R. (2021). A Systematic Literature Review on Big Data Extraction, Transformation and Loading (ETL). *Intelligent Computing*, 308–324. https://doi.org/10.1007/978-3-030-80126-7_24

Oliveira e Sá, J., Gonçalves, R., & Kaldeich, C. (2024). Benchmark of Market Cloud Data Warehouse Technologies. *Procedia Computer Science*, *239*, 1212–1219. https://doi.org/10.1016/j.procs.2024.06.289

Pansara, R. (2023). Unraveling the Complexities of Data Governance with Strategies, Challenges, and Future Directions. *Transactions on Latest Trends in IoT*, *6*(6), 46–56.

Paul, J., Lim, W. M., O'Cass, A., Hao, A. W., & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies*, *45*(4), O1-16. https://doi.org/10.1111/ijcs.12695

Peng, Y., Bathelt, F., Gebler, R., Gött, R., Heidenreich, A., Henke, E., Kadioglu, D., Lorenz, S., Vengadeswaran, A., & Sedlmayr, M. (2024). Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review. *JMIR Medical Informatics*, *12*, e52967. https://doi.org/10.2196/52967

Piippola, T.-J. (2024). *Data Strategy Handbook as Guide Towards Data-Driven Organization*.

Prasetyo, A. P. (2022). *Investment Analysis of Enterprise Data Warehouse Implementation with IT Balanced Scorecard and Information Economics*.

Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G., & Piccialli, F. (2024). Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems*, *150*, 272–293. https://doi.org/10.1016/j.future.2023.09.008

Rachakatla, S. K., Ravichandran, P., & Machireddy, J. R. (2021). The Role of Machine Learning in Data Warehousing: Enhancing Data Integration and Query Optimization. *Journal of Bioinformatics and Artificial Intelligence*, *1*(1), 82–104.

Rambabu, V. P., Althati, C., & Selvaraj, A. (2023). ETL vs. ELT: Optimizing Data Integration for Retail and Insurance Analytics. *Journal of Computational Intelligence and Robotics*, *3*(1), 37–84.

Raparthi, M., Gayam, S. R., Kasaraneni, B. P., Kondapaka, Kiran Kumar, Pattyam, S. P., Putha, S., & Thuniki, P. (2021). AI-Driven Decision Support Systems for Precision Medicine: Examining the Development and Implementation of AI-Driven Decision Support Systems in Precision Medicine. *Journal of Artificial Intelligence Research*, *1*(1), 11–20.

Rawat, R., & Yadav, R. (2021). Big Data: Big Data Analysis, Issues and Challenges and Technologies. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012014. https://doi.org/10.1088/1757-899x/1022/1/012014

Rehman, A., Naz, S., & Razzak, I. (2022). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems*, *28*(4), 1339–1371. https://doi.org/10.1007/s00530-020-00736-8

Ribeiro, M. B., & Braghetto, K. R. (2021). A data integration architecture for smart cities. *Anais*. 36th Brazilian Symposium on Databases (SBBD 2021), Rio de Janeiro, Brazil (Online event).

Rozony, f. z., Aktar, m. n. a., Ashrafuzzaman, m., & islam, a. (2024). a systematic review of big data integration challenges and solutions for heterogeneous data sources. *academic journal on business administration, innovation &sustainability*, *4*(4), 1–18.

Sadeghi, M., Carenini, A., Corcho, O., Rossi, M., Santoro, R., & Vogelsang, A. (2024). Interoperability of heterogeneous Systems of Systems: from requirements to a reference architecture. *The Journal of Supercomputing*, *80*(7), 8954–8987. https://doi.org/10.1007/s11227-023-05774-3

Sahara, C. R., & Aamer, A. M. (2022). Real-time data integration of an internet-of-things-based smart warehouse: a case study. *International Journal of Pervasive Computing and Communications*, *18*(5), 622–644. https://doi.org/10.1108/ijpcc-08-2020-0113

Sandhu, A. K. (2022). Big data with cloud computing: Discussions and challenges. *Big Data Mining and Analytics*, *5*(1), 32–40. https://doi.org/10.26599/bdma.2021.9020016

Saura, J. R. (2021). Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics. *Journal of Innovation & Knowledge*, *6*(2), 92–102. https://doi.org/10.1016/j.jik.2020.08.001

Schäfer, N. (2023). *On Enabling Efficient and Scalable Processing of Semi-Structured Data*.

Seenivasan, D. (2025). ETL vs ELT: Choosing the right approach for your data warehouse. *SSRN Electronic Journal*, *7*(2), 110–122. https://doi.org/10.2139/ssrn.5148194

Shafqat, S., Kishwer, S., Rasool, R. U., Qadir, J., Amjad, T., & Ahmad, H. F. (2020). Big data analytics enhanced healthcare systems: a review. *The Journal of Supercomputing*, *76*(3), 1754–1799. https://doi.org/10.1007/s11227-017-2222-4

Silverman, G. M., Sahoo, H. S., Ingraham, N. E., Lupei, M., Puskarich, M. A., Usher, M., Dries, J., Finzel, R. L., Murray, E., Sartori, J., Simon, G., Zhang, R., Melton, G. B., Tignanelli, C. J., & Pakhomov, S. V. (2021). NLP Methods for Extraction of Symptoms from Unstructured Data for Use in Prognostic COVID-19 Analytic Models. *Journal of Artificial Intelligence Research*, *72*, 429–474. https://doi.org/10.1613/jair.1.12631

Simitsis, A., Skiadopoulos, S., & Vassiliadis, P. (2023). The History, Present, and Future of ETL Technology. *Proceedings of the 25th International Workshop on Data Warehousing and OLAP*, 3–12.

Sivudu, P. (2023). ACCELERATING DATA Integration: Harnessing The Power of A Model-Driven Framework For Etl Process Development. *American Journal of Philological Sciences*, *3*(05), 52–58.

Sreepathy, H. V., Dinesh Rao, B., Kumar Jaysubramanian, M., & Deepak Rao, B. (2024). Data Ingestions as a Service (DIaaS): A Unified Interface for Heterogeneous Data Ingestion, Transformation, and Metadata Management for Data Lake. *IEEE Access*, *12*, 156131–156145. https://doi.org/10.1109/access.2024.3479736

Stergiou, C. L., Psannis, K. E., & Gupta, B. B. (2022). InFeMo: Flexible Big Data Management Through a Federated Cloud System. *ACM Transactions on Internet Technology*, *22*(2), 1–22. https://doi.org/10.1145/3426972

Stojanovic, A., Horvat, M., & Kovacevic, Z. (2022). An overview of data integration principles for heterogeneous databases. *Proceeding of the Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, 1111–1116. https://doi.org/10.23919/mipro55190.2022.9803579

Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., & Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Computational Statistics*, *13*(6), 136–152. https://doi.org/10.1002/wics.1549

Thumburu, S. K. R. (2022). Data Integration Strategies in Hybrid Cloud Environments. *Innovative Computer Sciences Journal*, *8*(1).

Van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, *136*, 106589. https://doi.org/10.1016/j.infsof.2021.106589

Weitzenboeck, E. M., Lison, P., Cyndecka, M., & Langford, M. (2022). The GDPR and unstructured data: is anonymization possible? *International Data Privacy Law*, *12*(3), 184–206. https://doi.org/10.1093/idpl/ipac008

Yin, F., Lin, Z., Kong, Q., Xu, Y., Li, D., Theodoridis, S., & Cui, S. R. (2020). FedLoc: Federated Learning Framework for Data-Driven Cooperative Localization and Location Data Processing. *IEEE Open Journal of Signal Processing*, *1*, 187–215. https://doi.org/10.1109/ojsp.2020.3036276

Yu, Y., Zong, N., Wen, A., Liu, S., Stone, D. J., Knaack, D., Chamberlain, A. M., Pfaff, E., Gabriel, D., Chute, C. G., Shah, N., & Jiang, G. (2022). Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration. *Journal of Biomedical Informatics*, *127*, 104002. https://doi.org/10.1016/j.jbi.2022.104002

Zecchini, L. (2024). *Integrazione di dati on-demand*.

Zhang, L., Qi, Z., & Meng, F. (2022). A Review on the Construction of Business Intelligence System Based on Unstructured Image Data. *Procedia Computer Science*, *199*, 392–398. https://doi.org/10.1016/j.procs.2022.01.048

Zhang, Y., & Long, Q. (2021). Assessing fairness in the presence of missing data. *Advances in Neural Information Processing Systems*, 16007–16019.

Zhao, Y., & Chen, J. (2022). A Survey on Differential Privacy for Unstructured Data Content. *ACM Computing Surveys*, *54*(10s), 1–28. https://doi.org/10.1145/3490237