Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2023, 10(4):63-70

Research Article ISSN: 2394 - 658X

Cross-Domain Data Engineering: Challenges, Solutions, and Future Directions

Paraskumar Patel

Neal Analytics
Bellevue, WA, USA

ABSTRACT

In an era characterized by the exponential growth of big data, the discipline of cross-domain data engineering has become increasingly vital, serving as the backbone for innovation and strategic decision-making across various sectors. This paper comprehensively explores the multifaceted challenges and opportunities within cross-domain data engineering, highlighting its critical role in harnessing the power of disparate data sources for enhanced analytical insights. The research delves into the complexities of integrating, processing, and analyzing data across diverse domains, addressing key challenges such as data heterogeneity, quality and consistency issues, scalability, and privacy and security concerns. This study offers a roadmap for overcoming the inherent obstacles in cross-domain data analysis by examining advanced technological solutions, including semantic web technologies, data virtualization, and machine learning for data integration. Furthermore, the paper discusses the evolving landscape of privacy regulations and the importance of robust data governance frameworks. By leveraging case studies and outlining future directions, this research contributes to the ongoing discourse on cross-domain data engineering, providing valuable insights and guidance for researchers and practitioners aiming to navigate the complexities of this field. Ultimately, the paper underscores the significance of cross-domain data engineering in unlocking the full potential of big data, paving the way for innovative solutions and strategic advantages in the competitive global market.

Key words: Cross-Domain Data Engineering, Data Heterogeneity, Interoperability Standards, Cross-domain Data Marketplaces, Cross-Disciplinary Collaboration, Cross-Domain Data Integration Techniques

INTRODUCTION

In the era of big data, efficiently managing and analyzing information across diverse domains has become crucial for innovation, decision-making, and maintaining competitive advantages. Cross-Domain Data Engineering emerges as a pivotal field, addressing the complexities and challenges of integrating, processing, and analyzing data that spans multiple disciplines or areas of expertise. This paper delves into the intricacies of cross-domain data engineering, highlighting its pivotal role in leveraging the full potential of data collected from disparate sources.

The significance of cross-domain data engineering cannot be overstated. It underpins the success of multidisciplinary projects, from healthcare research combining clinical and genomic data to financial services integrating market data with social media trends. Bridging the gaps between different data realms enables a more comprehensive analysis, fostering insights that would remain obscured within siloed data environments.

However, this integration is fraught with challenges. Data heterogeneity, quality and consistency issues, scalability concerns, and privacy and security risks are just a few obstacles that professionals in this field must navigate. Moreover, the evolving regulatory and compliance requirements landscape adds another layer of complexity to cross-domain data engineering efforts.

This paper aims to provide a comprehensive overview of these challenges, shedding light on the current state of cross-domain data engineering practices and the pressing issues that must be addressed. By exploring technological solutions, case studies, and future directions, we offer a roadmap for researchers and practitioners alike, aiming to advance the field and overcome the barriers to effective cross-domain data analysis.

By examining the multifaceted challenges and exploring innovative solutions, this paper contributes to the ongoing discourse on cross-domain data engineering, offering insights and guidance for overcoming the obstacles inherent in this critical area of data science.

BACKGROUND AND RELATED WORK

Data engineering has evolved significantly from its nascent stages in database management to the sophisticated, scalable systems we see today. Initially, data engineering practices focused on efficiently storing, retrieving, and managing data within single-domain applications. This involved using relational databases, ETL (extract, transform, load) processes, and basic data warehousing techniques to support business intelligence and analytics within an organization.

As technology advanced, so did the complexity and volume of data, leading to more advanced data storage solutions like NoSQL databases, data lakes, and real-time processing frameworks. These technologies enabled organizations to handle big data, support more complex queries, and derive insights from unstructured data sources.

A. Evolution into Cross-Domain Contexts

The need for cross-domain data engineering emerged from the realization that valuable insights often require integrating data across various sources, domains, and formats. As businesses and research institutions began to understand the potential of leveraging diverse data sets, the focus shifted towards creating interoperable, scalable systems that could process and analyze data from multiple domains.

This shift has been driven by several factors, including the proliferation of IoT devices generating vast amounts of data, the rise of machine learning and AI requiring diverse data sets for training models, and the increasing importance of data privacy and governance necessitating sophisticated data management practices.

Cross-domain data engineering involves dealing with challenges such as data heterogeneity, where data from different domains may follow different schemas or formats; data integration, requiring sophisticated techniques to merge data from disparate sources; and data privacy and security, ensuring that cross-domain data usage complies with regulations and ethical guidelines.

B. Summary of Existing Literature on Challenges in Cross-Domain Data Engineering

The existing literature on cross-domain data engineering highlights several key challenges. One of the primary concerns is the issue of data interoperability, which involves ensuring that systems and applications can exchange and use information from different domains seamlessly. Researchers have proposed various approaches to address this, including using standardized data formats, ontologies, and data transformation techniques.

Another significant challenge is maintaining data quality and consistency across domains. This includes dealing with missing data, inconsistencies, and errors that may arise when integrating data from multiple sources. Techniques such as data cleansing, validation, and enrichment are crucial to ensuring that cross-domain data is accurate and reliable.

Data privacy and security are also significant concerns in cross-domain data engineering. Data from various sources makes ensuring compliance with data protection laws and ethical guidelines complex. Solutions involve implementing robust data governance frameworks, secure data-sharing protocols, and advanced encryption techniques.

Finally, scalability and performance issues arise when dealing with large volumes of cross-domain data. Researchers and practitioners have explored distributed computing, cloud-based solutions, and advanced data processing frameworks to address these challenges, ensuring that cross-domain data engineering systems can handle the scale and complexity of modern data landscapes.

In summary, cross-domain data engineering is a rapidly evolving field that addresses the challenges of integrating and analyzing data across different domains. The literature highlights the complexity of these challenges, including interoperability, data quality, privacy, and scalability, and proposes various solutions to overcome them. As technology advances, ongoing research and development in this area are crucial to unlocking the full potential of cross-domain data insights.

CROSS-DOMAIN DATA ENGINEERING CHALLENGES

Cross-domain data engineering encompasses integrating and analyzing data from multiple sources, presenting a unique set of challenges stemming from the diversity and complexity of data landscapes. These challenges can broadly be categorized into data heterogeneity, data quality and consistency, scalability and performance, privacy and security, and regulatory compliance.

A. Data Heterogeneity

One of the primary challenges in cross-domain data integration lies in the heterogeneity of data, a multifaceted issue that significantly complicates the process. This heterogeneity manifests in several forms, including structural, semantic, and system variations, each presenting unique obstacles. Structural heterogeneity is evident in the diverse data models and formats across different domains. For instance, the stark differences between relational databases and NoSQL databases, as well as structured versus unstructured data, exemplify this challenge. On the other hand, semantic heterogeneity arises from the differences in meaning, interpretation, or usage of similar data across various domains, leading to potential misinterpretations and errors during data analysis. Lastly, system heterogeneity highlights the divergences in the underlying systems, platforms, or technologies employed to store, process, and manage data, further complicating integration efforts.

These disparate data assets often exist in a fragmented and proprietary state, lacking a centralized or integrated approach that facilitates easy access and utilization. This fragmentation erects significant barriers for stakeholders to leverage these data assets for analysis and decision-making, underlining the pressing need for more open and structured data ecosystems [1]. Overcoming these forms of heterogeneity demands sophisticated integration techniques capable of reconciling these differences and offering a unified data view[2]. Such advancements are crucial for enabling accurate and efficient data analysis, ensuring stakeholders can fully exploit the value of integrated data across diverse domains.

B. Data Quality and Consistency

Ensuring data quality and consistency is paramount for reliable analysis and decision-making. Challenges in this domain include Inconsistent Data Formats, where variations in data formatting and representation can introduce inaccuracies; Data Errors and Incompleteness, which can significantly skew analysis results; and Temporal Consistency, the necessity to synchronize data from different sources in time for accurate trend analysis. Addressing these challenges involves implementing robust data validation and cleaning processes and adopting standardized data formats and protocols to ensure data integrity [3].

C. Scalability and Performance

As data volumes continue to grow exponentially, the importance of scalability and performance in data management systems cannot be overstated. These systems must possess scalable architectures that can dynamically adjust to handle increasing amounts of data without suffering from performance degradation. This means that as more data becomes available for processing, these systems should be able to scale up efficiently to manage the load, ensuring that data processing remains fast and efficient with minimal latency. Efficient data processing is crucial for analyzing large datasets effectively, and robust resource management strategies are needed to maintain an optimal balance between workload and system performance.

Moreover, real-time data analytics has heightened the need for systems that can process and analyze data instantly, particularly in scenarios across different domains. This requires the capability to process streaming data from diverse sources, necessitating systems with high-throughput and low-latency processing capabilities. Advanced technologies such as in-memory computing, stream processing frameworks, and real-time analytics platforms are essential to meet these challenges [4]. These technologies enable processing large volumes of data in real time, facilitating timely insights and decisions based on the most current data available. In summary, addressing the dual challenges of scalability and performance in the face of growing data volumes and the demand for real-time analytics is paramount for organizations aiming to leverage data effectively in their decision-making processes.

D. Privacy and Security

In the realm of cross-domain data engineering, the handling of sensitive or personal data introduces significant privacy and security challenges. The assurance of data privacy involves the ethical use of personal data. It requires adherence to stringent privacy regulations such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. On the other hand, data security is primarily concerned with safeguarding data against unauthorized access, breaches, and theft [5]. This is particularly complex regarding cross-border data flow, which introduces additional legal and regulatory challenges associated with transferring data across different geographic and jurisdictional boundaries.

To mitigate these risks, it is imperative to implement robust security protocols, including encryption and anonymization techniques, to ensure the privacy and security of data. When outsourcing work internationally, especially in data-sensitive sectors like healthcare, the challenges of cross-border data flow become even more pronounced. Companies must navigate the complexities of adhering to stringent regulations such as GDPR and HIPAA, not just within their borders but also in the international landscape. This necessitates a comprehensive understanding of global data protection laws and the development of solid data governance frameworks. These frameworks are crucial for managing cross-jurisdictional transfers effectively, ensuring data privacy and security are maintained across the entire data lifecycle [6].

E. Third-Party Data Management and Control

The management and control of third-party data are fraught with complexities that organizations must navigate to maintain data integrity, privacy, and compliance with contractual obligations. The dependency on external data providers necessitates carefully examining the terms and conditions imposed, which often include limitations on the amalgamation of datasets. Such restrictions can significantly hinder the ability to perform comprehensive analyses. Additionally, the requirement for temporary staff or contractors to access sensitive data introduces the need for robust data access policies and legal frameworks, including Non-Disclosure Agreements (NDAs), to protect confidential information.

Moreover, the reliance on third-party infrastructures, such as Application Programming Interfaces (APIs) for data retrieval, presents risks associated with data availability and integrity. Ensuring consistent and reliable access to these external data sources is paramount to avoid disruptions in data workflows and ensure the continuity of analytical processes. Challenges also arise in integrating third-party data, often necessitating additional engineering efforts to address data quality issues, such as gaps or inaccuracies. This compromises the data quality and places an increased burden on data engineering teams, requiring advanced expertise in data preprocessing and quality assurance.

F. Interoperability

Interoperability is a significant challenge, especially when integrating data and systems from different domains. It involves the technical aspects of making diverse systems work together and aligning business processes and objectives across domains. Achieving interoperability requires standards, protocols, and APIs that facilitate communication and data exchange between disparate systems and agreements on data governance and sharing models.

G. Diverse Technological Ecosystems

In cross-domain data engineering, professionals frequently engage with a broad spectrum of technologies and architectures, necessitating a versatile skill set to navigate effectively across varied technological ecosystems. This diversity stems from companies employing distinct technology stacks, thus complicating integration efforts and compelling data engineers to undertake continuous learning and adaptation. Moreover, disparate departments may adopt multiple storage technologies within a single organization, inadvertently fostering data duplication and inconsistencies. Integrating these diverse data sources into a unified repository presents considerable challenges. It requires deploying sophisticated data integration strategies and tools to achieve a cohesive and accurate data landscape. This scenario underscores the complexities of managing data across different technological ecosystems, highlighting the importance of flexibility, expertise, and innovative solutions in data engineering.

H. Data Integration and Governance

As data is integrated from multiple domains, establishing a comprehensive data governance framework becomes imperative. This includes defining policies for data access, quality, security, and usage. The challenge lies in creating a governance model that is flexible enough to accommodate the diverse nature of data and stringent enough to ensure compliance and best practices are followed. Effective data governance ensures the integrated data landscape is trustworthy, secure, and efficiently managed.

Cross-domain data engineering thus requires a multifaceted approach to address these challenges, involving technical, legal, and organizational strategies to harness the full potential of integrated data analysis.

TECHNOLOGICAL SOLUTIONS AND APPROACHES

A. Advanced Data Integration Techniques

Unified Data Access Layers: Implementing a unified data access layer involves creating a standardized interface that abstracts the complexities of underlying data sources[7]. By doing so, data consumers can access and query data from different domains without needing to understand the specific details of each source. This approach simplifies integrating heterogeneous data by providing a common framework for interaction. Furthermore, the utilization of generics in a layered architecture on the .NET platform showcases an efficient means for realizing a universal data access layer, enhancing reusability, and reducing time consumption in software projects.

Semantic Web Technologies: Leveraging semantic web technologies such as RDF and OWL involves creating a standard data model and vocabulary for describing data across domains [8]. By establishing a semantic layer, organizations can ensure that data from various sources is interpretable and usable in a unified context. This approach addresses semantic heterogeneity by enabling data integration based on shared meaning rather than just syntactical similarities.

Data Virtualization: Data virtualization technology offers a real-time integration approach by providing a consolidated virtual data view across multiple domains [9]. This abstraction layer hides the technical details of data stored in disparate sources, allowing for easier access and analysis without needing physical data movement. By virtualizing data, organizations can overcome the challenges of data silos and achieve a unified view of their data landscape.

B. Data Quality Management

Automated Data Cleansing Tools: Automated data cleansing tools help improve data quality by identifying and correcting errors, filling in missing values, and standardizing data formats. These tools use algorithms and rules to cleanse data at scale, ensuring consistency and accuracy across domains. Organizations can reduce the time and effort required to maintain high-quality data by automating the cleansing process.

Continuous Data Validation: Implementing continuous data validation processes involves checking data for accuracy and consistency at various stages, such as during data entry and throughout its lifecycle. By continuously monitoring data quality, organizations can detect anomalies and errors in real time, preventing the propagation of incorrect information. This proactive approach to data validation ensures ongoing data integrity and reliability for analysis.

C. Scalability and Performance Optimization

Cloud-native Architectures: Adopting cloud-native architectures allows organizations to leverage the scalability and flexibility of cloud computing for data processing. By designing applications and infrastructure that are native to the cloud, organizations can dynamically scale resources to handle varying data volumes and computational demands. This elasticity enables organizations to manage large-scale data processing tasks cost-effectively while maintaining optimal performance.

Distributed Data Processing Frameworks: Utilizing distributed data processing frameworks such as Apache Spark or Apache Flink enables efficient handling of large-scale data across domains. These frameworks support parallel processing and real-time analytics, allowing organizations to promptly process and analyze massive datasets. Organizations can achieve high throughput and low latency for data processing tasks by distributing workloads across multiple nodes or clusters.

D. Privacy and Security Enhancements

Advanced Encryption Techniques: Employing advanced encryption techniques such as homomorphic encryption ensures the security of sensitive information across domains. Homomorphic encryption allows for computations to be performed on encrypted data without decrypting it, preserving privacy while enabling analysis. Organizations can protect against unauthorized access and data breaches by encrypting data at rest and in transit.

Differential Privacy: Implementing differential privacy techniques in data analysis adds noise to the data or queries to prevent the identification of individuals from aggregated data. This approach allows organizations to derive insights from datasets while preserving the privacy of individual records. Organizations can balance the need for data-driven decision-making with individual privacy rights by incorporating differential privacy into data analysis workflows.

E. Interoperability and Data Governance

Standardized Data Formats and Protocols: Promoting standardized data formats and protocols facilitates interoperability between different systems and domains. Organizations can simplify data exchange and integration by adopting standard formats such as JSON or XML and protocols such as RESTful APIs. This standardization reduces the complexity of data integration efforts and ensures compatibility between disparate systems.

Comprehensive Data Governance Frameworks: Establishing comprehensive frameworks involves defining policies and procedures for managing data quality, security, privacy, and usage across domains. These frameworks provide guidelines for data management practices and ensure compliance with regulatory requirements. By implementing robust data governance practices, organizations can maintain control over their data assets while mitigating data integration and usage risks.

F. Leveraging Machine Learning for Data Integration

Machine Learning Models for Data Matching: Applying machine learning models to automate the matching and linking related data across different sources improves the efficiency and accuracy of data integration efforts. These models can learn patterns and relationships in the data, enabling organizations to automatically identify and reconcile duplicates or inconsistencies. Organizations can streamline the integration process and reduce manual effort by leveraging machine learning for data matching.

Natural Language Processing (NLP) for Semantic Integration: Utilizing NLP techniques to interpret and integrate data based on semantic meaning enhances the depth of integrated data analysis. NLP algorithms can extract contextual information from text data and map it to a common semantic framework, facilitating the integration of unstructured data with structured datasets. Organizations can gain deeper insights from diverse data sources and improve decision-making by incorporating NLP into data integration workflows.

G. Continuous Learning and Adaptation

In the rapidly evolving field of cross-domain data engineering, professionals must continuously learn and adapt to new technologies and methodologies. Staying abreast of emerging trends and best practices enables professionals to effectively address evolving challenges and capitalize on opportunities for innovation. By investing in continuous learning and professional development, organizations can build expertise within their teams and remain competitive in the dynamic landscape of data engineering.

FUTURE DIRECTIONS

The ongoing evolution of technologies, methodologies, and the increasing demand for sophisticated data analysis across various domains will likely shape future trends and directions in cross-domain data engineering. Based on the comprehensive overview provided in the article, the following future trends and directions can be anticipated:

Integration of Advanced AI and Machine Learning Techniques: As AI and machine learning technologies mature, their integration into cross-domain data engineering will become more prevalent. This will enhance the capabilities for data matching and semantic integration and enable the development of more intelligent, adaptive data management and analysis systems. Future systems could autonomously improve their data integration and analysis processes based on learning from previous experiences, leading to more efficient and accurate insights generation [10], [11], [12].

Expansion of Data Privacy Technologies: With privacy regulations becoming more stringent globally, the adoption of advanced privacy-preserving technologies such as homomorphic encryption and differential privacy is expected to grow. Furthermore, innovations in privacy-preserving data analysis techniques that enable data utilization without compromising individual privacy will become critical. This includes the development of more sophisticated data anonymization methods and privacy-by-design frameworks for cross-domain data engineering platforms [13].

Advancements in Real-time Data Processing and Analytics: The demand for real-time data analytics is increasing across sectors. Future developments in cross-domain data engineering will likely focus on enhancing real-time data processing and analysis capabilities. This involves improving the scalability and performance of data processing frameworks and adopting more efficient stream processing technologies. Such advancements will enable organizations to derive timely insights from vast data generated across domains [14].

Evolution of Data Governance and Interoperability Standards: As data becomes more integral to strategic decision-making, establishing comprehensive data governance frameworks and interoperability standards will be crucial. Future trends may include the development of universal data governance models that can be adapted across domains and more robust and flexible interoperability standards. This would facilitate smoother data integration and sharing, ensuring data quality, security, and regulation compliance [15].

Greater Emphasis on Decentralized Data Architectures: Decentralized data architectures like blockchain may play a more significant role in cross-domain data engineering. These architectures offer potential solutions for enhancing data security, integrity, and transparency in data transactions. As these technologies evolve, they could provide a foundation for creating more secure and trustable data ecosystems, particularly in environments where data sharing and collaboration are essential [16].

Development of Cross-domain Data Marketplaces: The concept of data marketplaces, where data from various domains can be exchanged or sold, may gain traction. These marketplaces would facilitate access to a broader range of data, enabling organizations to enrich their analyses and insights [1]. However, this would also necessitate advancements in data standardization, privacy protection, and intellectual property rights management to ensure fair and secure data exchange [17].

Focus on Human-Centric Data Engineering Approaches: As technology advances, there will be an increasing emphasis on designing cross-domain data engineering systems that are technically efficient, ethically responsible, and aligned with human values. This involves considering the societal impact of data engineering projects and ensuring inclusivity, fairness, and transparency in data-driven decision-making processes [18].
Adoption of Cloud-native and Edge Computing Technologies: The shift towards cloud-native architectures and the integration of edge computing technologies will continue to influence cross-domain data engineering. These technologies offer scalable, flexible data processing and storage solutions, enabling more efficient data management across distributed environments. Future trends will likely involve the seamless integration of cloud and edge computing resources to support distributed data processing and analytics at scale [19].

Increased Collaboration Across Disciplines: Finally, the future of cross-domain data engineering will depend on fostering stronger collaborations between data engineers, domain experts, policymakers, and other stakeholders. By working together, these groups can address the complex challenges of cross-domain data integration, ensuring that data engineering efforts lead to meaningful, impactful outcomes across various fields [20].

CONCLUSION

In big data's vast and intricate landscape, cross-domain data engineering has emerged as a critical field that transcends traditional boundaries, offering innovative solutions to the complex challenges of integrating and analyzing data across multiple domains. This article has systematically explored the multifaceted challenges inherent in cross-domain data engineering, including data heterogeneity, quality and consistency issues, scalability and performance demands, privacy and security concerns, and the evolving regulatory compliance landscape. Through a comprehensive examination of technological solutions, case studies, and future directions, we have underscored the significance of advanced data integration techniques, data quality management, scalability and performance optimization, privacy and security enhancements, and the necessity for interoperability and robust data governance frameworks.

The future of cross-domain data engineering is poised for significant evolution, driven by the integration of advanced AI and machine learning techniques, the expansion of data privacy technologies, advancements in real-time data processing and analytics, and the evolution of data governance and interoperability standards. Additionally, decentralized data architectures, the development of cross-domain data marketplaces, and a focus on human-centric data engineering approaches highlight the forward trajectory of the field. As we look ahead, adopting cloud-native and edge computing technologies and increased collaboration across disciplines will be crucial in addressing the complex challenges and unlocking the full potential of cross-domain data insights.

In conclusion, cross-domain data engineering represents a pivotal area of research and practice that holds the key to harnessing the transformative power of data across diverse domains. By addressing the challenges and embracing the technological advancements and collaborative efforts outlined in this article, researchers, practitioners, and organizations can unlock new avenues for innovation, decision-making, and maintaining competitive advantages in the era of big data.

REFERENCES

[1]. A. Mavrogiorgou et al., “A Cross-domain Data Marketplace for Data Sharing,” ACM International Conference Proceeding Series, pp. 72–79, Oct. 2022, doi: 10.1145/3571697.3571707.

[2]. H. Hu, H. Wang, and B. Zheng, “Challenges in Managing and Mining Large, Heterogeneous Data,” LNCS, vol. 6588, pp. 462–462, 2011, doi: 10.1007/978-3-642-20152-3_40.

[3]. S. McClean, B. Scotney, and M. Shapcott, “Using Domain Knowledge to Learn from Heterogeneous Distributed Databases,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3213, pp. 171–177, 2004, doi: 10.1007/978-3-540-30132-5_28/COVER.

[4]. A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq, “Challenges of data integration and interoperability in big data,” Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, pp. 38–40, 2014, doi: 10.1109/BIGDATA.2014.7004486.

[5]. P. Christen, D. Vatsalan, and V. S. Verykios, “Challenges for privacy preservation in data integration,” Journal of Data and Information Quality (JDIQ), vol. 5, no. 1–2, Sep. 2014, doi: 10.1145/2629604.

[6]. R. H. Khokhar, B. C. M. Fung, F. Iqbal, D. Alhadidi, and J. Bentahar, “Privacy-preserving data mashup model for trading person-specific information,” Electron Commer Res Appl, vol. 17, pp. 19–37, May 2016, doi: 10.1016/J.ELERAP.2016.02.004.

[7]. M. Sooriyabandara et al., “Unified Link Layer API: A generic and open API to manage wireless media access,” Comput Commun, vol. 31, no. 5, pp. 962–979, Mar. 2008, doi: 10.1016/J.COMCOM.2007.12.025.

[8]. D. Yang, L. Li, and L. Sun, “Layered Graph Data Model for dataspaces management,” 2011 IEEE 3rd International Conference on Communication Software and Networks, ICCSN 2011, pp. 234–237, 2011, doi: 10.1109/ICCSN.2011.6014430.

[9]. X. Wang, R. Feng, W. Dong, X. Zhu, and W. Wang, “Unified access layer with PostgreSQL FDW for heterogeneous databases,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10578 LNCS, pp. 131–135, 2017, doi: 10.1007/978-3-319-68210-5_14/FIGURES/5.

[10]. S. C. Y. Lu, “Machine learning approaches to knowledge synthesis and integration tasks for advanced engineering automation,” Comput Ind, vol. 15, no. 1–2, pp. 105–120, Jan. 1990, doi: 10.1016/0166-3615(90)90088-7.

[11]. M. Birgersson, G. Hansson, and U. Franke, “Data Integration Using Machine Learning,” Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOCW, vol. 2016-September, pp. 313–322, Sep. 2016, doi: 10.1109/EDOCW.2016.7584357.

[12]. S. Amershi et al., “Software Engineering for Machine Learning: A Case Study,” Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019, pp. 291–300, May 2019, doi: 10.1109/ICSE-SEIP.2019.00042.

[13]. X. L. Dong and T. Rekatsinas, “Data integration and machine learning,” Proceedings of the VLDB Endowment, vol. 11, no. 12, pp. 2094–2097, Aug. 2018, doi: 10.14778/3229863.3229876.

[14]. Y. Li and A. Ngom, “Data integration in machine learning,” Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015, pp. 1665–1671, Dec. 2015, doi: 10.1109/BIBM.2015.7359925.

[15]. C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, “Integrating Machine Learning with Human Knowledge,” iScience, vol. 23, no. 11, Nov. 2020, doi: 10.1016/J.ISCI.2020.101656.

[16]. M. Picard, M. P. Scott-Boyer, A. Bodein, O. Périn, and A. Droit, “Integration strategies of multi-omics data for machine learning analysis,” Comput Struct Biotechnol J, vol. 19, pp. 3735–3746, Jan. 2021, doi: 10.1016/J.CSBJ.2021.06.030.

[17]. A. Nazir, “SEAMLESS AUTOMATION AND INTEGRATION OF MACHINE LEARNING CAPABILITIES FOR BIG DATA ANALYTICS,” International Journal of Distributed and Parallel Systems (IJDPS, vol. 8, no. 3, 2017, doi: 10.5121/ijdps.2017.8301.

[18]. D. Wang et al., “Human-AI Collaboration in Data Science,” Proc ACM Hum Comput Interact, vol. 3, no. CSCW, Nov. 2019, doi: 10.1145/3359313.

[19]. Y. ting Zhuang, F. Wu, C. Chen, and Y. he Pan, “Challenges and opportunities: from big data to knowledge in AI 2.0,” Frontiers of Information Technology and Electronic Engineering, vol. 18, no. 1, pp. 3–14, Jan. 2017, doi: 10.1631/FITEE.1601883/FIGURES/2.

[20]. A. Y. Levy, “Combining Artificial Intelligence and Databases for Data Integration,” pp. 249–268, 1999, doi: 10.1007/3-540-48317-9_10.