Data Analytics and Artificial Intelligence in the New Scenario of Data Spaces

Juan Trujillo[1,*,†], Alejandro Reina-Reina[1,†] and Gustavo Candela[1,†]

[1] Lucentia Research Group (DLSI), University of Alicante, San Vicent del Raspeig, Spain

Abstract

Data spaces present significant opportunities for organizations to collaborate and leverage decentralized, interoperable, and secure data for advanced analytics and Federated Machine Learning. However, challenges such as ensuring data quality, managing privacy, and integrating heterogeneous data formats and semantics remain critical. Addressing these challenges requires robust data governance, real-time quality assurance, and the adoption of AI-data integration tools to improving decision-making and generating value in dynamic business ecosystems.

Keywords

Data Spaces, Collaborative Data Management, Data Governance, Federated Machine Learning

1. Data Spaces and Collaborative Models: Challenges and Opportunities in Data Management

Today, the efficient use of data is a key factor in the success of an organization. The advancement of organizations toward more collaborative and interconnected models can foster the emergence of new ways to manage and leverage data [1].

In this regard, data spaces present opportunities to enhance advanced analytics and foster collaboration between companies, representing a crucial evolution that could shift the big data paradigm, currently based on data warehouses and data lakes. However, this evolution requires a high level of maturity, with big data management and AI development being key. Organizations must adapt to a dynamic and complex data infrastructure, considering aspects such as decentralization, interoperability, and data quality, which pose essential technical and organizational challenges to optimize the use of data in strategic decision-making and value creation [2].

2. Data spaces in business ecosystems

To date, organizations rely primarily on traditional big data architectures such as data warehouses and data lakes. These are designed to consolidate structured data in a centralized environment. However, with the advent of data spaces, there is the challenge of managing data in a distributed and decentralized manner while simultaneously adhering to the principles of interoperability, security, privacy, and data governance [2]. Furthermore, data spaces must address the challenge of fostering trust-based collaboration among participating organizations [3].

Despite the challenges, data spaces enable organizations to collaborate with each other and share information without the need to move data from their original sources. The advantages of data spaces are invaluable, as they facilitate new opportunities for the metrics necessary for KPI monitoring through distributed analytics models [4]. Moreover, data spaces can support Federated Machine Learning [5], allowing companies to learn collaboratively while ensuring that personal data remain confidential and private, without leaving the organization's private environment. This also helps to comply with privacy regulations.

3. Evaluating and improving the quality of data-driven services

An inherent challenge that users will face in data spaces is the evaluation of the quality and reliability of the data they consume [6]. It is true that data spaces offer a rich ecosystem of information sources, but their heterogeneity raises critical questions: How can we ensure that the shared data is accurate, up-to-date, and free from biases? What mechanisms can guarantee that data-driven services are consistent and reliable enough to be integrated into my business process?

Furthermore, in a context where companies can consume third-party information to feed their business models, it is crucial to have metrics that allow the evaluation of data quality aspects [7], such as completeness, accuracy, or consistency. This is because if the data are of low quality, the results obtained, as well as AI-based decisions, may be erroneous [8].

To address this issue, organizations must implement robust data governance processes, combined with technological tools that enable the auditing and certification of data quality, even in real time. Furthermore, the use of quality labels or certifications could become a standard in these data spaces, similar to how products or services are evaluated in other sectors.

4. Use of Heterogeneous Data with Different Formats and Semantics

Data heterogeneity has been a recurring obstacle in traditional information management systems, but it becomes even more pronounced in distributed and decentralized

DOLAP 2025: 27th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT 2025, March 25, 2025, Barcelona, Spain

[*] Corresponding author.

[†] These authors contributed equally.

jtrujillo@dlsi.ua.es (J. Trujillo); alejandro.reina@ua.es (A. Reina-Reina); gcandela@dlsi.ua.es (G. Candela)

0000-0003-0139-6724 (J. Trujillo); 0000-0001-7116-9338 (A. Reina-Reina); 0000-0001-6122-0777 (G. Candela)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

ecosystems [9]. In this regard, organizations face another key challenge, which is the ability to work with data that does not necessarily share the same format, structure, or semantics as their own data [10]. For example, a transport service provider may receive traffic data from one source and weather data from another, but analytical models must be able to integrate these datasets to extract valuable insights that can support decision-making.

The solution to this challenge is not trivial; however, it necessarily involves the use of standards such as ontologies, as well as advanced technologies, including integration and automatic transformation tools. In this context, approaches based on embeddings [11] or AI [12] may facilitate the alignment of heterogeneous data, enabling its effective integration and utilization.

Acknowledgments

This work has been co-funded by the AETHER-UA project (PID2020-112540RB-C43) funded by the Spanish Ministry of Science and Innovation; the ENIA Chair of Artificial Intelligence from the University of Alicante (TSI-100927-2023-6) funded by the Recovery, Transformation and Resilience Plan from the European Union Next Generation through the Ministry for Digital Transformation and the Civil Service; and the BALLADEER (PROMETEO/2021/088) project funded by the Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital (Generalitat Valenciana), and Spanish Ministry of Digital Transformation (TSI-100121-2024-10).

References

[1] L. M. Camarinha-Matos, H. Afsarmanesh, N. Galeano, A. Molina, Collaborative networked organizations – concepts and practice in manufacturing enterprises, Computers & Industrial Engineering 57 (2009) 46–60. doi:10.1016/j.cie.2008.11.024.

[2] F. Möller, I. Jussen, V. Springer, A. Gieß, J. C. Schweihoff, J. Gelhaar, T. Guggenberger, B. Otto, Industrial data ecosystems and data spaces, Electronic Markets 34 (2024) 41. doi:10.1007/s12525-024-00724-0.

[3] M. Huber, S. Wessel, G. Brost, N. Menz, Building Trust in Data Spaces, Springer International Publishing, 2022, pp. 147–164. doi:10.1007/978-3-030-93975-5_9.

[4] R. Kalmar, B. Rauch, J. Dörr, P. Liggesmeyer, Agricultural Data Space, Springer International Publishing, 2022, pp. 279–290. doi:10.1007/978-3-030-93975-5_17.

[5] B. Farahani, A. K. Monsefi, Smart and collaborative industrial iot: A federated learning and data space approach, Digital Communications and Networks 9 (2023) 436–447. doi:10.1016/j.dcan.2023.01.022.

[6] S. R. Carroll, I. Garba, O. L. Figueroa-Rodríguez, J. Holbrook, R. Lovett, S. Materechera, M. A. Parsons, K. Raseroka, D. Rodriguez-Lonebear, R. Rowe, R. Sara, J. D. Walker, J. Anderson, M. Hudson, The CARE principles for indigenous data governance, Data Sci. J. 19 (2020) 43. URL: https://doi.org/10.5334/dsj-2020-043. doi:10.5334/DSJ-2020-043.

[7] I. ISO, Iec 25012. software engineering–software product quality requirements and evaluation (square)–data quality model, International Organization for Standardization (2000).

[8] Y. Alkatheeri, A. Ameen, O. Isaac, A. Al-Shibami, M. Nusari, The mediation effect of management information systems on the relationship between big data quality and decision making quality, Test Engineering and Management 82 (2020) 12065–74.

[9] M. Naeem, T. Jamal, J. Diaz-Martinez, S. A. Butt, N. Montesano, M. I. Tariq, E. D. la Hoz-Franco, E. De-La-Hoz-Valdiris, Trends and future perspective challenges in big data, in: Advances in Intelligent Data Analysis and Applications, volume 253, Springer, Singapore, 2022, pp. 309–325. doi:10.1007/978-981-16-5036-9_30.

[10] W. Fan, H. Lu, S. E. Madnick, D. Cheung, Discovering and reconciling value conflicts for numerical data integration, Information Systems 26 (2001) 635–656. doi:10.1016/S0306-4379(01)00043-6.

[11] C. Koutras, G. Siachamis, A. Ionescu, K. Psarakis, J. Brons, M. Fragkoulis, C. Lofi, A. Bonifati, A. Katsifodimos, Valentine: Evaluating matching techniques for dataset discovery, 2021. URL: https://arxiv.org/abs/2010.07386. arXiv:2010.07386.

[12] J. García-Carrasco, A. Reina, A. Lavalle, A. Maté, J. Trujillo, A conceptual model-based approach for exploiting large language model embeddings in automatic data integration, 2024. doi:10.2139/ssrn.5024901.