Toward Advanced Query Processing in Dataspaces

Christoph Quix[1,2]

[1]Hochschule Niederrhein University of Applied Sciences, Krefeld, Germany
[2]Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

Abstract

Dataspaces aim at enabling inter-organizational data exchange, emphasizing interoperability and data sovereignty of data assets. While current implementations focus on providing a foundational framework to enable secure, standards-based data sharing and sovereignty, they lack the robust query processing features needed to address emerging demands in distributed and federated data ecosystems. We present a vision for advancing dataspace technology by incorporating sophisticated query processing mechanisms and integrating features that ensure data sovereignty within traditional data management platforms such as data lakes.

Keywords

dataspaces, data integration, federated query processing, data sovereignty

1. The Need for Advanced Query Processing in Dataspaces

The original idea of dataspaces as envisioned by Franklin et al. [1] emphasized lightweight data integration and incremental development of an integrated, linked *personal* dataspace. Dataspaces should provide basic data access and interoperability between heterogeneous data sources while progressively enhancing integration through user-driven refinement and automated techniques. This approach features flexibility and usability, allowing users to interact with partially integrated data while supporting iterative improvements in data organization and querying capabilities.

In 2015, Fraunhofer in Germany started the Industrial Dataspace (IDS) initiative [2] which created a new view on dataspaces. Dataspaces were envisioned as a multi-sided platform for secure and trusted data exchange, guaranteeing data sovereignty with a decentralized architecture [3]. The development is governed by an institutionalized alliance of diverse stakeholders, i.e., the International Data Spaces Association (IDSA)[1]. First ideas of the IDS outline a dataspace as platform or market for data and services, in which data is described semantically in (central) metadata repositories. Data consumers can search the metadata for relevant datasets, invoke data services to integrate, transform or enrich data as desired, and finally use the data according to defined usage policies [4]. However, the work in the IDS project focused on the deployment of a trusted and secure connector framework [5].

Gaia-X has evolved from the IDS concept by extending its focus on data sharing and sovereignty into a broader framework that integrates cloud services, edge computing, and data ecosystems through standardized frameworks and governance mechanisms [6]. Gaia-X is a European initiative aimed at creating a secure, federated, and interoperable data infrastructure. It builds on IDS principles, such as trust and compliance, while expanding the ecosystem to include decentralized, federated infrastructures and a strong emphasis on transparency, openness, and digital sovereignty.

The priority for trust and data sovereignty is a significant strength, it also imposes limitations on the ability to support data processing across a distributed data ecosystem [7]. These limitations become particularly evident in use cases requiring:

(a) Federated Query Processing: The capability to process queries across multiple, independently managed datasets without compromising performance or accuracy.

(b) Semantic Enrichment: Leveraging metadata and domain-specific ontologies to enable more precise and meaningful query results.

(c) Granular Data Sovereignty: Enforcing fine-grained access control policies that align with legal and organizational requirements.

A lack of these features constrains the practical utility of dataspaces in scenarios where data-driven decision-making depends on seamless and secure integration of data in a distributed data ecosystem.

2. Integrating Dataspace Features into Modern Data Architectures

The evolution from data lakes to data meshes and data fabrics reflects a significant transformation in how organizations approach data management to address issues of scalability, governance, and accessibility [8]. Data lakes, originally designed to store large volumes of structured and unstructured data in a centralized repository [9], often faced challenges related to governance and usability. Without robust management and accessibility frameworks, many data lakes devolved into 'data swamps', where finding meaningful insights became increasingly difficult [10].

Data meshes emerged to solve these issues by decentralizing data governance. This paradigm treats data as a product, where responsibility for the quality, usability, and governance of data lies with domain-specific teams. This domain-driven ownership model ensures scalability while addressing the shortcomings of centralized approaches, such as those found in traditional data lakes.

In parallel, data fabrics focus on creating an interconnected layer that integrates metadata across disparate systems. By employing technologies such as AI, automation, and knowledge graphs, data fabrics enable seamless data discovery, improved lineage tracking, and enhanced integration across an organization's diverse data landscape. This approach prioritizes metadata-driven governance and

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

christoph.quix@hs-niederrhein.de (C. Quix)
https://www.hs-niederrhein.de/elektrotechnik-informatik/personen/quix (C. Quix)
0000-0002-1698-4345 (C. Quix)

[1]https://internationaldataspaces.org/

context-aware connectivity, providing a more dynamic and agile data ecosystem that supports advanced analytics and decision-making processes.

The concept of modern dataspaces aligns with these paradigms. Similar to the 'data as a product' philosophy in data meshes, dataspaces emphasize treating data assets as shared, governed resources designed for collaboration and interoperability. In both cases, the focus is on ensuring the quality, contextual relevance, and accessibility of data for specific stakeholders or applications. This shared emphasis underlines the mutual goal of enabling robust, domain-aware data collaboration across organizational boundaries.

However, sharing data products in a dataspace does not imply centralized governance or uniform data quality control. Similar to data meshes, data governance should be organized in a decentralized manner. Therefore, each participant should manage their own policies through *self-governing data products*. This approach aligns with the dataspace architecture, where data owners define and enforce usage policies for their data assets [4].

Additionally, dataspaces as well as data fabrics rely on semantic models to support semantic interoperability. In data fabrics, knowledge graphs serve as a foundational tool for modeling relationships and enriching metadata, allowing for enhanced data discovery, integration, and query capabilities. Similarly, dataspaces employ semantic models to achieve interoperability among heterogeneous data sources and domains (e.g., the IDS information model [11]). These models provide a shared understanding of data structures and relationships, which is essential for enabling meaningful cross-domain analytics and collaboration.

The interplay between these paradigms suggests a path toward convergence, where dataspaces could incorporate the principles of both data meshes and data fabrics. By blending the domain-centric ownership and product-oriented data management of data meshes with the semantic and automation-driven integration of data fabrics, dataspaces could emerge as a comprehensive framework for addressing modern data challenges. This evolution reflects a growing recognition of the need for distributed, interoperable, and semantically enriched data ecosystems capable of supporting diverse organizational and cross-domain needs.

The challenge lies in finding the optimal balance between unified semantic models for describing data assets and decentralized governance. In many dataspace projects, we have observed that a centralized approach to defining the core information model significantly slows down the bootstrapping process. A decentralized approach, as envisioned in data meshes, could accelerate this process but comes with the risk of diverging semantics. To mitigate these issues, collaborative ontology engineering methodologies need to be applied [12].

Enhancing data lakes with dataspace-inspired features can bridge the gap between centralized data repositories and the decentralized nature of dataspaces. Specifically, integrating features for data sovereignty and advanced query processing can yield transformative capabilities. By incorporating mechanisms like usage policies [13], a data lake can enforce access control, data provenance, and compliance policies. Databricks has proposed *Delta Sharing*[2], a protocol for sharing datasets between data lakes, or even between organizations. Although the protocol supports fine-grained access control, usage policies to support data sovereignty as in dataspaces are not yet covered.

On the other hand, enhancing dataspace frameworks, such as the Eclipse Dataspace Components[3], with more sophisticated federated query processing for heterogeneous datasets could offer a better usability of dataspaces. Data scientists require easy solutions for creating a Pandas data frame over heterogeneous data: an API as provided in Apache Spark, combined with Delta Sharing, and enriched with usage policies, could facilitate a true sovereign data science framework that integrates heterogeneous data access, data integration, and machine learning. Although it might be still challenging to merge all these features into one platform, we can leverage large-language models to support users in executing their tasks in such a platform [14, 15]. SEDAR, as an open-source data lake platform, offers a concrete foundation for these integrations [16]. Enhancing SEDAR with dataspace features could demonstrate the feasibility of such extensions and provide insights into performance trade-offs and usability.

3. Future Research Directions

We advocate for a paradigm shift in dataspace technology by prioritizing advanced query processing and seamless integration with traditional data management platforms. Leveraging existing innovations, such as the Delta Sharing Protocol, and extending platforms like SEDAR, can help realize the vision of a unified, sovereignty-aware data management ecosystem. However, several research challenges must still be addressed to fully enable this vision.

Policy-aware query execution requires embedding data sovereignty rules directly into query execution plans. Queries should be executed in compliance with access restrictions, data-sharing agreements, and computational constraints defined by data owners. *User-centric interfaces* should allow non-expert users to interact effectively with the dataspace. Many existing dataspaces suffer from poor usability, limiting their adoption and accessibility. Furthermore, *usage policies* must be extended beyond basic access control to include restrictions on query processing itself. Finally, *data quality management* remains a significant challenge in many dataspaces. A decentralized data quality framework, incorporating objective and standardized quality metrics, could help assess and improve data reliability while allowing participants to retain autonomy over their data assets.

Acknowledgments

This work has been sponsored by the German Federal Ministry of Education and Research in the funding program "Forschung an Fachhochschulen", project I²DACH (grant no. 13FH557KX0) and in the funding program "KI-Anwendungshub Kunststoffverpackungen – nachhaltige Kreislaufwirtschaft durch Künstliche Intelligenz", project KIOptiPack (grant no. 033KI111).

AI Disclosure Statement During the preparation of this work, the author used ChatGPT 4o in order to improve writing style, check grammar, and spelling. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

[2] https://github.com/delta-io/delta-sharing/

[3] https://projects.eclipse.org/projects/technology.edc

References

[1] M. J. Franklin, A. Y. Halevy, D. Maier, From databases to dataspaces: a new abstraction for information management, SIGMOD Rec. 34 (2005) 27–33. URL: https://doi.org/10.1145/1107499.1107502. doi:10.1145/1107499.1107502.

[2] B. Otto, Jürjens, J. Schon, S. Auer, N. Menz, S. Wenzel, J. Cirullies, Industrial Data Space - Digital Sovereignity over Data, Whitepaper, Fraunhofer-Gesellschaft, 2016. URL: https://www.fraunhofer.de/content/dam/zv/de/Forschungsfelder/industrial-data-space/Industrial-Data-Space_whitepaper.pdf.

[3] B. Otto, M. Jarke, Designing a multi-sided data platform: findings from the international data spaces case, Electron. Mark. 29 (2019) 561–580. URL: https://doi.org/10.1007/s12525-019-00362-x. doi:10.1007/S12525-019-00362-X.

[4] C. Quix, A. Chakrabarti, S. Kleff, J. Pullmann, Business process modelling for a data exchange platform, in: Proceedings of the Forum and Doctoral Consortium Papers Presented at the 29th International Conference on Advanced Information Systems Engineering, CAiSE 2017, Essen, Germany, June 12-16, 2017, volume 1848 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 153–160. URL: https://ceur-ws.org/Vol-1848/CAiSE2017_Forum_Paper20.pdf.

[5] H. Pettenpohl, M. Spiekermann, J. R. Both, International data spaces in a nutshell, in: Designing Data Spaces: The Ecosystem Approach to Competitive Advantage, Springer, 2022, pp. 29–40. URL: https://doi.org/10.1007/978-3-030-93975-5_3. doi:10.1007/978-3-030-93975-5_3.

[6] H. Tardieu, Role of Gaia-X in the european data space ecosystem, in: Designing Data Spaces: The Ecosystem Approach to Competitive Advantage, Springer, 2022, pp. 41–59. URL: https://doi.org/10.1007/978-3-030-93975-5_4. doi:10.1007/978-3-030-93975-5_4.

[7] S. Geisler, M. Vidal, C. Cappiello, B. F. Lóscio, A. Gal, M. Jarke, M. Lenzerini, P. Missier, B. Otto, E. Paja, B. Pernici, J. Rehof, Knowledge-driven data ecosystems toward data transparency, ACM J. Data Inf. Qual. 14 (2022) 3:1–3:12. URL: https://doi.org/10.1145/3467022. doi:10.1145/3467022.

[8] I. Blohm, F. Wortmann, C. Legner, F. Köbler, Data products, data mesh, and data fabric, Bus. Inf. Syst. Eng. 66 (2024) 643–652. URL: https://doi.org/10.1007/s12599-024-00876-5. doi:10.1007/S12599-024-00876-5.

[9] R. Hai, C. Koutras, C. Quix, M. Jarke, Data lakes: A survey of functions and systems, IEEE Trans. Knowl. Data Eng. 35 (2023) 12571–12590. URL: https://doi.org/10.1109/TKDE.2023.3270101. doi:10.1109/TKDE.2023.3270101.

[10] R. Hai, S. Geisler, C. Quix, Constance: An intelligent data lake system, in: Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016, ACM, 2016, pp. 2097–2100. URL: https://doi.org/10.1145/2882903.2899389. doi:10.1145/2882903.2899389.

[11] S. R. Bader, J. Pullmann, C. Mader, S. Tramp, C. Quix, A. W. Müller, H. Akyürek, M. Böckmann, B. T. Imbusch, J. Lipp, S. Geisler, C. Lange, The international data spaces information model - an ontology for sovereign exchange of digital content, in: The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II, volume 12507 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 176–192. URL: https://doi.org/10.1007/978-3-030-62466-8_12. doi:10.1007/978-3-030-62466-8_12.

[12] K. I. Kotis, G. A. Vouros, D. Spiliotopoulos, Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations, Knowl. Eng. Rev. 35 (2020) e4. URL: https://doi.org/10.1017/S0269888920000065. doi:10.1017/S0269888920000065.

[13] D. M. Mustafa, A. Nadgeri, D. Collarana, B. T. Arnold, C. Quix, C. Lange, S. Decker, From instructions to ODRL usage policies: An ontology guided approach, in: Proceedings of Workshops at the 50th International Conference on Very Large Data Bases, VLDB 2024, Guangzhou, China, August 26-30, 2024, VLDB.org, 2024. URL: https://vldb.org/workshops/2024/proceedings/LLM+KG/LLM+KG-15.pdf.

[14] S. Hoseini, A. Burgdorf, A. Paulus, T. Meisen, C. Quix, A. Pomp, Towards llm-augmented creation of semantic models for dataspaces, in: Proceedings of the Second International Workshop on Semantics in Dataspaces (SDS 2024) co-located with the 21st Extended Semantic Web Conference (ESWC 2024), Hersonissos, Greece, May 26, 2024, volume 3705 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3705/paper03.pdf.

[15] S. Hoseini, M. Ibbels, C. Quix, Enhancing machine learning capabilities in data lakes with automl and llms, in: Advances in Databases and Information Systems - 28th European Conference, ADBIS 2024, Bayonne, France, August 28-31, 2024, Proceedings, volume 14918 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 184–198. URL: https://doi.org/10.1007/978-3-031-70626-4_13. doi:10.1007/978-3-031-70626-4_13.

[16] S. Hoseini, A. Ali, H. Shaker, C. Quix, SEDAR: A semantic data reservoir for heterogeneous datasets, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023, ACM, 2023, pp. 5056–5060. URL: https://doi.org/10.1145/3583780.3614753. doi:10.1145/3583780.3614753.