

Cross-Sector, Efficient, Trusted Data Sharing in Dataspaces

Soumya Kanti Datta
Digiotech, Estonia
Email: soumya@digiotech.com

Abstract—This paper presents a holistic architecture for enabling cross-sector, trusted data sharing in European Dataspaces. The proposed architecture overcomes the sector-specific Dataspace challenges by introducing novel extensions to state-of-the-art Dataspace connectors, specifically designed for handling vast amounts and streaming data scenarios. The architecture implements intelligent data placement and caching mechanisms that optimise data access patterns across sector boundaries, alongside secure connectivity for data sharing. With detailed workflow presentation, this paper demonstrates how the functional components enable efficient discovery and sharing of cross-sectoral data among Dataspace participants.

Keywords—*Dataspace; Data sharing; Discovery; Multi-Cloud.*

I. INTRODUCTION

A Dataspace is a decentralised data sharing system that enables controlled and secure exchange of data between organisations while preserving data sovereignty [1]. It involves a standardised framework where data owners (providers) maintain complete control over their data while making it accessible to authorised consumers through secure connectors. The key distinction from traditional data sharing approaches is that a Dataspace does not copy, store, or centralise data but instead creates a secure, standardised environment for controlled, trustworthy data sharing while leaving the data under the provider's control. In industrial settings, Dataspaces implement a provider-consumer architecture that maintains data ownership while enabling controlled sharing. For software systems, Dataspaces provide a unified solution through standardised connectors and data plane extensions that handle various data types. This standardisation dramatically simplifies system architecture while improving reliability.

To benefit from the Dataspace concept, emerging implementations focused on vertical specific industries, e.g., agriculture [2], mobility, healthcare, manufacturing [3], and energy [4]. However, such industry or sector specific Dataspaces often build around data silos, sometimes shared in a community, supply chain, or geography. Such practices lead to discovery and access challenges, let alone practically adapt to the needs of an organisational parameters to such data sources, and then easily configure to connected data consumer and provider. Moreover, current Dataspace connectors [5] are mostly concentrated on discreet data transfer within a specific sector while very few research have investigated decentralised data sharing [6]. There is a visible lack of research on two important aspects – (a) sharing and management of vast amounts and streaming data across industry verticals while maintaining security and performance and (b) intelligent data placement and caching strategies for efficient data sharing.

The EU-funded DS2 project aims to deliver a reference implementation which can truly address the above aspects enabling cross-sector, efficient, trusted data sharing across Dataspaces. The key novel aspect of DS2 is a component called Intersector Dataspace Toolkit (IDT) which is deployed at each Dataspace and network connected to any other IDT-enabled Dataspace. Once connected, DS2 will enable

discovery of data formats, contexts (including language, culture etc), privacy and security and other non-functional requirements to extract and map data together, controlling the data life cycle between the data consumer and provider.

This paper focuses on a core building block of DS2 – cross-sector, trusted, multi-cloud data sharing module and its architecture. The novel contributions of the paper include: (i) a holistic architectural framework that enables efficient transfer of vast amounts and streaming data between participants of Dataspaces from data stores distributed across multi-cloud storage infrastructure, (ii) extensions to the Dataspace connector data plane specifically designed for vast amount and streaming data sharing, and (iii) intelligent data placement and caching strategies that optimise data access across sector boundaries while maintaining security and performance. The paper also presents detailed workflow analyses, component descriptions emphasising cross-sector interactions, and comprehensive validation metrics with defined performance thresholds to validate the architecture's effectiveness in real-world scenarios. This research advances the state-of-the-art in industrial Dataspaces by providing a secure, efficient solution for cross-sector data sharing, particularly relevant for ongoing and emerging European Dataspace ecosystems.

II. CROSS-SECTOR TRUSTED DATA SHARING MODULE ARCHITECTURE

In compliance with the Dataspace principles, the module architecture has been designed to create the underlying soft infrastructure to allow for the efficient sharing of discreet, vast amounts, and streaming data from multi-cloud data stores. This includes novel solutions on intelligent data placement and caching of data along with establishment of secure connectivity between distributed data stores. The modular architecture is depicted in Figure 1 and is described below.

Catalogue module: It securely stores a description metadata of the Dataspaces and implements an interconnected search and retrieval system for a consumer participant to discover [7] data offer (s) and relevant provider participant(s) details.

Use case application: These are the high-level applications of users such as those in the use cases of DS2. Such applications can require data from multiple data stores, which are presumed to be distributed to multi-cloud storage. Also, the data sharing happens through multiple Dataspace provider participants. Each use case application can directly consume the obtained data and/or temporarily store them in a local storage for combining the data arriving in batches (in case of vast data transfer) or processing in future.

Temporary data store: A local data storage which are used by the use case applications to store data for (very) short term.

Dataspace connector (data plane): It facilitates secure and efficient transfer of data between participants in the DS2 ecosystem while ensuring compliance with agreed-upon data governance policies and handling data routing. In DS2, the data plane supports three types of data sharing – discrete data,

vast amounts of data, and streaming data. The connector implementation (e.g., Eclipse Dataspace connector) at present supports push/pull style discrete data. The connector is extended to support vast and streaming data sharing which are one of the main novel aspects and contributions of the presented work.

Dataspace consumer participant: It is composed of two sub-components.

- **Consumer Participant data offer discovery service:** This service facilitates cross-sector data offer discovery by implementing dual functionalities. First, it publishes metadata descriptions of data offers to the Catalogue module, enabling centralised discovery across different industrial sectors. Second, it enables consumer participants to discover data offers spanning multiple sectors within the DS2 ecosystem. The service implements metadata structures that accommodate diverse data types from multiple sectors, ensuring interoperability through standardised metadata schemas.
- **Data retrieval service:** This component implements sophisticated retrieval mechanisms optimised for cross-sector data access. It enables seamless data retrieval across sector boundaries while maintaining sector-specific security and compliance requirements. The service implements adaptive retrieval strategies that optimize performance based on the characteristics of cross-sector data access patterns. It supports both direct data consumption and temporary storage options, accommodating various cross-sector use cases with different data processing requirements.

Dataspace provider participant: It includes following sub-components.

- **Provider participant data offer discovery service:** It publishes a description of the data offer of each participant to the Tier 1 catalogue module for a centralised discovery by the consumer participant.
- **Intelligent data placement and caching:** This component implements cross-sector optimization strategies for data placement and caching across multi-cloud environments. It analyses access patterns from consumer participants across different industrial sectors, enabling predictive caching that accounts for sector-specific data access behaviours. The component employs sophisticated algorithms that consider cross-sector data relationships, optimising cache locations based on interdependencies between different industrial domains. The intelligent data placement strategy ensures optimal use of storage resources by distributing data based on access frequency, storage costs, and performance requirements. This sub-component is comprised of cache invalidation and consistency manager, data placement controller, intelligent data placement engine, and data caching.

Dataspace connector data plane: It supports three types of data transfer:

- **Push/pull data transfer:** Data can be delivered to consumer counterpart through this sub-component either via a push model (where the Dataspace sends data automatically at intervals or when triggered) or a pull

model (where consumer participant requests data when needed). While this is supported by default in the Dataspace connector, it is needed to accomplish discrete or tiny amounts of data transfer.

- **Vast data extension:** This novel component implements specialised mechanisms for handling large-scale data transfers across sector boundaries. The data partitioning and compression sub-component employs adaptive algorithms that optimize compression based on sector-specific data characteristics. The batch scheduling functionality implements intelligent scheduling that considers cross-sector dependencies and peak usage patterns across different industrial domains. Error checking and retry mechanisms are enhanced to maintain data integrity across sector-specific validation requirements. This extension is composed of sub-components performing data partitioning, compression, batch scheduling, data transfer, error checking, and retries (if needed).
- **Data stream extension:** This novel component extends traditional Dataspace connector capabilities to handle real-time streaming data across different industrial sectors. It implements robust stream processing capabilities that can handle diverse data formats and protocols specific to different industries. The stream message broker is designed to maintain data consistency and semantic meaning when streaming data crosses sector boundaries, such as when energy consumption data from smart buildings feeds into both energy management and urban planning applications. The stream processing engine implements transformation capabilities that ensure data compatibility across sector-specific standards and formats. This extension is composed of stream processing engine and stream message broker.
- **Data store:** These data stores represent data storage at the participant. Each participant data store may be designed to manage a variety of data types and formats, providing a robust and scalable storage solution. It supports multiple data stores leveraging multi-cloud environments, enabling data to be distributed and replicated across different geographic locations, enhancing accessibility and redundancy.
- **Vast data store:** Like the data store mentioned above, these are specific to storages with vast amounts of data.
- **Data stream source:** This refers to data sources that continuously produce data such as IoT devices (video cameras) and real-time applications (weather apps).

III. WORKFLOWS

The workflows outline key processes enabling efficient, secure, cross-sector, and intelligent data sharing in the Dataspaces. Each workflow is designed to address specific aspects of data sharing, from the discovery and retrieval of data offers to advance mechanisms like real-time data streaming and intelligent caching.

A. Write data offer description

This workflow implements metadata registration within the Dataspace ecosystem. It initiates with the Dataspace provider participant's self-provisioning process, wherein structured metadata is formulated in JSON format,

encompassing comprehensive data specifications and access parameters. It orchestrates secure transmission of this metadata to the Catalogue module, which functions as a centralised registry mechanism. This transmission adheres to strict security protocols, ensuring the integrity of metadata during transfer. The Catalogue module subsequently persists this information in its local database system, facilitating subsequent discovery operations.

B. Read data offer description

With this step, a consumer participant discovers the data offers available in the Dataspace. The sequence is initiated through a structured read request from a consumer participant seeking to discover available data offerings. The workflow executes a query against the Catalogue module, which returns matching data offer descriptions encoded in JSON format. The returned descriptions undergo systematic parsing and interpretation within the consumer participant's architectural components. This represents a critical discovery phase that precedes actual data transfer operations, enabling informed selection of appropriate data sources based on requirements.

C. Data sharing using push/pull data transfer

This workflow implements discrete data exchange mechanisms between Dataspace participants and is activated post discovery, utilising established provider participant's identification. The process initiates when a use case application generates a data retrieval request. The data retrieval service, having obtained provider information during discovery, executes a read request to the designated data store through the push/pull transfer mechanism. This workflow demonstrates efficacy in discrete data transfer scenarios, facilitating direct transmission of complete datasets.

D. Data sharing using data stream transfer

It supports real-time data sharing through the novel data stream extension of the Dataspace connector. It implements continuous data transmission mechanisms (e.g., data pipelining), optimised for real-time data exchange scenarios. This workflow architecture differs fundamentally from push/pull mechanisms, being engineered for persistent data flows. The workflow initialisation occurs through application-specific requests for stream topic access. The data retrieval service establishes connectivity with the appropriate provider, enabling continuous data publication to the stream extension.

E. Intelligent data caching in Dataspace

This operational flow optimises data access and sharing by strategically placing and caching frequently used data closer to consumer participants. These steps are initiated when vast data stores transmit data to the intelligent data placement engine. This engine executes algorithmic decision-making processes regarding optimal data placement, considering multiple parameters including access patterns and network topology. The data placement controller implements these decisions through strategic data forwarding to cache storage locations. The cache invalidation and consistency manager, which executes continuous monitoring operations to maintain data validity and consistency. Upon detection of inconsistencies, the workflow triggers update mechanisms to maintain data accuracy.

IV. VALIDATION METRICS

The paper has developed comprehensive validation metrics that align with the cross-sector trusted data sharing module's key functionalities and technical performance requirements.

A. Data offer description management metrics

Discovery response time - It measures the time taken to discover and retrieve data offer descriptions from the Catalogue module and is critical for efficient service discovery and initialization of data sharing processes. The acceptance criteria for the round trip response time are defined to be <500 ms for single data offer retrieval and at-least 95% of discovery requests must meet the target response time.

Data offer description accuracy - It validates the correctness and completeness of JSON format data offer descriptions and ensures reliable service discovery and appropriate data sharing initialisation. The acceptance criteria have been defined to 100% schema compliance and field accuracy.

B. Streaming data transfer metrics

Stream processing latency - It measures end-to-end latency from data stream source to consumer application and is therefore relevant use cases running real-time applications requiring immediate data processing. The acceptance criteria have been set to <100 ms for standard streaming operations where at-least 95% of streaming data must be processed within the target latency.

Stream throughput stability - It measures the consistency of data streaming rates under various network loads and is drawn up for testing reliable performance for continuous data streaming operations. The acceptance criteria are set to < 5% throughput variation under normal conditions while maintaining stable throughput for 99% of operational time.

C. Intelligent data placement metrics

Cache hit ratio - It calculates the effectiveness of the intelligent data placement engine providing an indication of the efficiency of the caching strategy and data placement decisions. The acceptance criteria have been defined to > 80% cache hit ratio while maintaining target hit ratio over 24-hour operational periods for successful validation.

Cache invalidation accuracy - The precision of cache invalidation decisions is measured in this metric which ensures data consistency and prevents stale data usage. The acceptance criteria are defined to 100% accuracy in invalidation decisions which amounts to zero instances of stale data being served.

D. Vast data transfer metrics

Transfer completion reliability - It is captured in terms of the success rate of completion of large data transfers. The acceptance criteria have been set to 99.99% successful completion rate where complete data transfer with zero corruption or loss is accomplished.

Batch processing efficiency - The effectiveness of data partitioning and compression as a part of the batch processing is crucial and indicates the efficiency of vast data handling mechanisms. The acceptance criteria are set to > 50% reduction in transfer size through compression whilst

achieving target compression while maintaining data integrity is going to be checked.

V. CONCLUSION

In a nutshell, this paper presents an architectural solution that addresses the critical challenges of cross-sector, efficient data sharing in European Dataspaces. The presented module makes several contributions to the European Dataspace paradigm. The introduction of novel extensions to Dataspace connectors and the intelligent data placement and caching mechanisms for efficient data access across sector boundaries significantly advances the state-of-the-art. The architecture presents a foundation for ongoing developments DS2 use case implementations, particularly as cross-sector data sharing becomes more critical for innovation and efficiency.

ACKNOWLEDGMENT

The research leading to the results presented in this paper has received funding from the European Union funded project DS2 under grant agreement no. 101135967.

REFERENCES

[1] Curry, E. (2020). Dataspaces: Fundamentals, Principles, and Techniques. In: Real-time Linked Dataspaces. Springer, Cham. https://doi.org/10.1007/978-3-030-29665-0_3.

[2] Šestak, Martina, and Daniel Copot. 2023. "Towards Trusted Data Sharing and Exchange in Agro-Food Supply Chains: Design Principles for Agricultural Data Spaces" Sustainability 15, no. 18: 13746.

[3] Guo, J., Cheng, Y., Wang, D., Tao, F., & Pickl, S. (2021). Industrial Dataspace for smart manufacturing: connotation, key technologies, and framework. International Journal of Production Research, 61(12), 3868–3883.

[4] S. Meneguzzo, A. Favenza, V. Gatteschi and C. Schifanella, "Integrating a DLT-Based Data Marketplace with IDSA for a Unified Energy Dataspace: Towards Silo-Free Energy Data Exchange within GAIA-X," 2023 5th Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS), Paris, France, 2023, pp. 1-2.

[5] Pampus, J., Jahnke, BF., Quensel, R. (2022). Evolving Data Space Technologies: Lessons Learned from an IDS Connector Reference Implementation. In: Margaria, T., Steffen, B. (eds) Leveraging Applications of Formal Methods, Verification and Validation. Practice. ISoLA 2022. Lecture Notes in Computer Science, vol 13704. Springer, Cham.

[6] S. H. Alsamhi et al., "Empowering Dataspace 4.0: Unveiling Promise of Decentralized Data-Sharing," in IEEE Access, vol. 12, pp. 112637-112658, 2024.

[7] Arne Bröring, Soumya Kanti Datta, and Christian Bonnet. 2016. A Categorization of Discovery Technologies for the Internet of Things. In Proceedings of the 6th International Conference on the Internet of Things (IoT '16). Association for Computing Machinery, New York, NY, USA, 131–139.

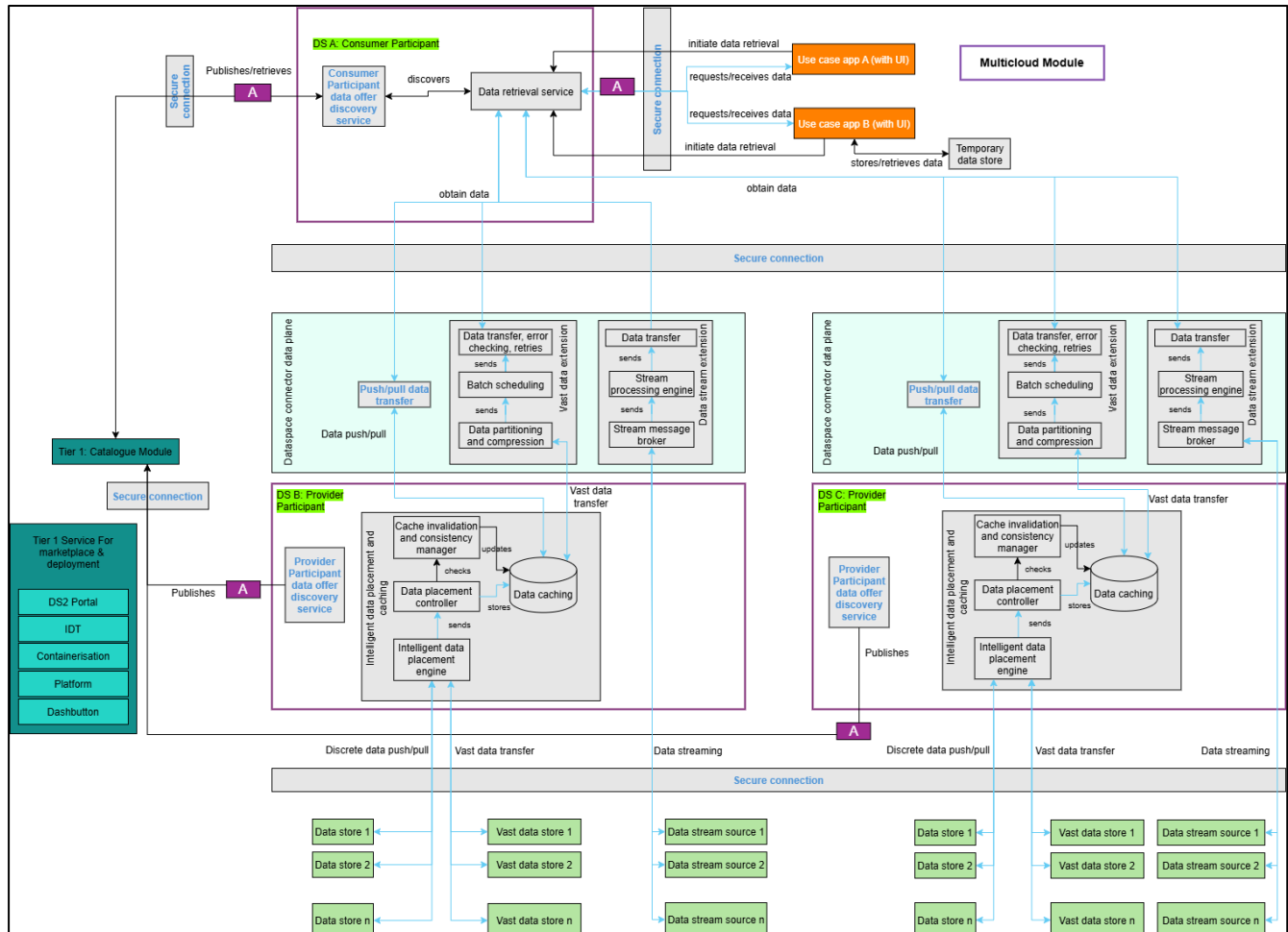


Figure 1 – cross-sector trusted data sharing module architecture and components.