Introduction

The data mesh concept learns from three generations – and failures – of current data management platforms and architectures [Dehghani, 2020]:

- Data warehouses and business intelligence platforms have left companies with large and unmaintainable amounts of ETL jobs, tables and reports;
- the “over promised and under realized” big data ecosystems and data lakes are operated by a central team of specialized data engineers;
- data analytics solutions with no clear domain boundaries fail for large enterprises with a diverse number of data sources and consumers.

Data mesh architectures apply “product thinking” to data management platforms: individual domains are responsible for providing the data. In most current architectures data gets pushed from the domains into a central data lake or platform; in a data mesh the domains have to host and serve their datasets, i.e. own their domain data. “Data products” enclose information and functionalities of the dataset, i.e. the data itself which is served by an operational system (e.g. relational tables, batch files, etc.), the metadata (documentation, semantic descriptions, etc.), and output (e.g. access ports/APIs).

While the individual domain teams own the necessary technology to store, process and serve their data products, a common platform is needed to allow homogeneous interactions with the data products [Priebe et al., 2021].

Challenge & Project Goals

The data mesh paradigm considers a single-organization setup, building on a centralized, single-location infrastructure. However, what about a data product distributed over multiple locations and/or multiple data owners? For instance, in the below figure, a research institution requires data services from a media company and a weather data provider to build machine learning models – which again are offered to the media company.

The goal of the project is to research and develop a solution for a fully distributed data mesh platform which allows federated governance and policy management of the individual data products. The main objectives are:

G1 Collect stakeholder requirements (e.g., how to organize the domains).
G2 Develop a conceptualization and architecture model.
G3 Research technical solutions, e.g., how to reach consensus on a single truth across the distributed platform.
G4 Integrate into the project partner’s Data Hub software solution.

Approach

Datasets are indexed and cataloged in a platform, the Data Hubs. The Data Hubs allow to organize the data products in Domain Data Spaces; the spaces group the datasets and operational capabilities of a domain. A data product in a data space offers connectors to access the data and allows to define policies to control access, visibility, etc. [EDC, 2021].

Sharing data across multiple hubs implies a consensus and governance layer:

- The federated governance layer allows the execution of (access) policies and decisions across the data space members.
- The consensus layer allows to scale the distribution to multiple hubs by resolving synchronization, versioning, and state conflicts.

Project Partner

Project partner Nexyo is a Vienna-based data-tech startup that develops technology for decentralized data ecosystems. Its product combines data management with intuitive data sharing and secures data sovereignty to enable data-driven business. Nexyo is member of IDSA (International Data Space Association) [Otto et al., 2016] and also involved in several research initiatives regarding data sharing in mobility, agriculture, tourism and industry [EDC, 2021].

References