TEADAL: Trustworthy, Energy-Aware federated DAta Lakes along the computing continuum*

Pierluigi Plebani[1,*], Ronen Kat[2], Frank Pallas[3], Sebastian Werner[3], Giacomo Inches[4], Peeter Laud[5] and Rita Santiago[6]

[1] *Politecnico di Milano, Italy*
[2] *IBM Research, Israel*
[3] *Technical University of Berlin, Germany*
[4] *Martel Innovate, Switzerland*
[5] *Cybernetica, Estonia*
[6] *UbiWhere, Portugal*

Abstract

While the value that data analytics has in any organization is undoubted and widely recognized, most approaches and solutions are mainly based on processing data that are internally produced or, if available from the outside, are publicly available with almost no restrictions. The situation becomes, however, more challenging when data analytics can take advantage of data owned by other organizations, which are only available to share under some conditions that must not affect the value that such data has for them. In this context, the need for data sharing has to deal with the need for data sovereignty, i.e., the possibility of an organization to exert full control over their data. This introduces barriers at both organizational and technological level: data must be shared under an agreement established among the parties which could affect the locations – on premise or on cloud – where the storage and processing of the data are performed.

The goal of the TEADAL project is to propose and provide a set of tools to enable the creation of a federation of data lakes, particularly reducing the barriers in defining the terms under which data can be shared. The actual sharing shall optimize the exploitation of resources owned by the members to foster efficient, energy-aware, and trusted data management.

Keywords

Data sharing, Data sovereignty, Federated data lakes

RPE@CAiSE'23: Research Projects Exhibition at the International Conference on Advanced Information Systems Engineering, June 12–16, 2023, Zaragoza, Spain

* Duration: September 2022 - August 2025. Consortium: UbiWhere (PT) (Project Coordinator), Politecnico di Milano (IT) (Scientific coordinator), Almaviva (IT), AMTS (IT), Box2M (RO), Cefriel (IT), Cybernetica (EE), ERT (ES), I2CAT (ES), IBM (IL), ING (NL), Marina Salud (ES), Martel Innovate (CH), Regione Toscana (IT), Technical University of Berlin (DE), Technical University of Wien (AT), Terraview (CH), UITP (BE). Funding Agency: European Union's Horizon Europe research and innovation programme. Website: https://www.teadal.eu

*Corresponding author.

pierluigi.plebani@polimi.it (P. Plebani); ronenkat@il.ibm.com (R. Kat); fp@ise.tu-berlin.de (F. Pallas); sw@ise.tu-berlin.de (S. Werner); giacomo.inches@martel-innovate.com (G. Inches); peeter.laud@cyber.ee (P. Laud); rsantiago@ubiwhere.com (R. Santiago)

1. Project overview

For almost two decades now, researchers and practitioners have devoted a lot of effort to addressing the so-called Big Data phenomenon. As a result, tools and technologies able to deal with the unprecedented generation of data have reached, using the terms of the famous Gartner hype cycle, the “plateau of productivity”. Focusing on Big Data Analytics, now the market is full of solutions for efficiently storing and processing data created by organizations, for increasing their value and, in particular, for extracting meaningful insights.

From a technology perspective, there has been a paradigm shift towards a “one size does not fit all” approach [1], which has produced a plethora of tools and platforms extremely specialized for a specific aspect in data management. For instance, while relational databases are still perfect in transactional settings, for data analytics different models (e.g., column-based, key-value) have been proposed, each of them with a specific realm of application. A tangible effect, in organizations, can be seen by the more and more frequent replacement of data warehouses – which were seen as the preferable way to perform data analytics in organizations – with data lakes [2]. With the term data lake, we refer to a platform composed of a set of software tools supporting the acquisition, the governance, and the provisioning of heterogeneous datasets to improve the effectiveness and the efficiency of data analytics, especially when considering the Big Data domain. This architectural change was dictated by the paradigmatic shift from a “schema-on-write” (i.e., schema is defined before data are written, thus at design-time) approach that governs data warehouses, towards a less rigid “schema-on-read” (i.e., schema is defined after data is read, thus at run-time) approach typical in a data lake setting [3].

Although data lakes are gaining momentum, current solutions do still not properly unleash the full potential of data analysis for four reasons:

- Cross-organizational data sharing support: data lakes are usually deployed and managed by a single party, fed by data generated internally or already publicly available. In a federated setting, companies should be able to share data across organizational boundaries to improve the effectiveness of data analytics.
- Efficient data management: the computing continuum (i.e., the resources located at the edge, fog, and cloud) is not fully exploited. To minimize the impact of avoidable data transfers, data should be processed where they are generated. At the same time, consciously exploiting the computing continuum can also help addressing security/privacy and governance concerns [4] albeit not without the possibility of giving rise to new ones.
- Complete data control: data sovereignty [5] must be preserved, thus personal data or business sensitive data cannot leave the boundaries of the organization unless a proper data transformation is performed in compliance with the organization’s policies and general norms, which often limits the data sharing.
- Sustainability: because of the illusion created by low-cost storage devices and the assumption that data have a huge value for companies [6], operators do not discriminate if data is already stored or if it is useful, resulting in a data mess — data duplications and storage of unused data.

On this basis, the project TEADAL will enable the creation of trusted, verifiable, and energy-efficient data flows, both inside a data lake and across federated data lakes, based on a shared

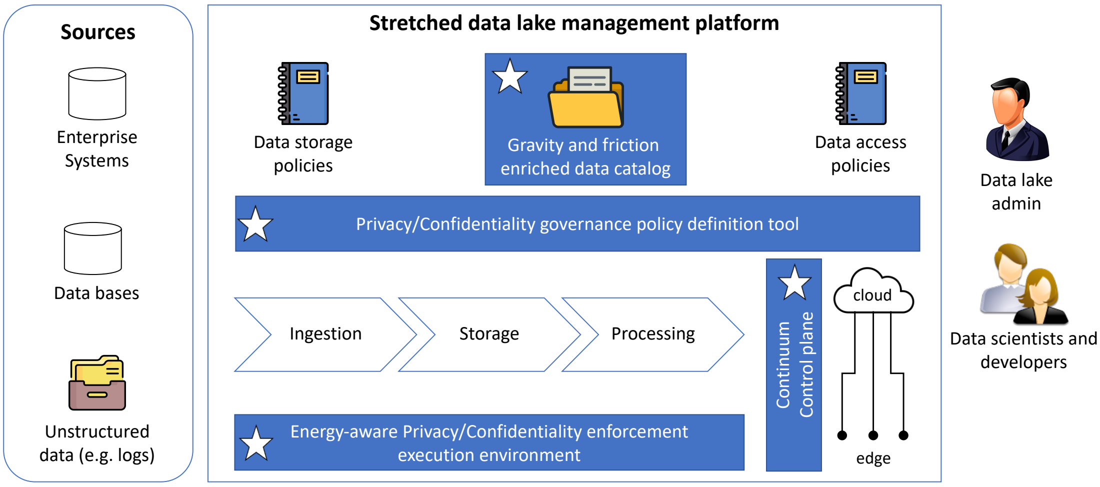
Sources
Enterprise Systems
Data bases
Unstructured data (e.g. logs)
Stretched data lake management platform
Data storage policies
Gravity and friction enriched data catalog
Data access policies
Privacy/Confidentiality governance policy definition tool
Ingestion
Storage
Processing
Continuum Control plane
cloud
edge
Energy-aware Privacy/Confidentiality enforcement execution environment
Data lake admin
Data scientists and developers

Figure 1: TEADAL stretched data lake

approach for defining, enforcing, and tracking data governance requirements with specific emphasis on privacy/confidentiality.

2. Project objectives

Generally speaking, the TEADAL project aims to achieve four main objectives:

- To propose a *stretched data lake* to efficiently handle data across the continuum.
- To enable the construction of trustworthy data lakes and mediatorless *data lake federations*.
- To reduce the environmental impact of data analytics by carefully managing how data are stored, reused, moved, and processed in a federation of stretched data lakes.
- To simplify the specification and enforcement of privacy/confidentiality requirements, constraints and policies for federated stretched data lakes to be compliant with regulations, norms, and organizations' policies.

The envisioned stretched data lake (see Figure 1) is a data lake platform powered by the innovative tools (reported as boxes with a star) proposed in TEADAL.

Notably, the *privacy/confidentiality governance policy definition tool* will be available to the data lake administrator to define data storage, movement and processing policies with respect to the owned/managed resources along the continuum, thus considering resources at the edge (where usually data are produced), compute and storage systems on-premises, as well as cloud resources. Similarly, it will be possible for the data lake governance officer to define data access policies, thus, to specify which are the internal and external actors that have the right to access the data and in which format. These policies will be mainly influenced by trustworthiness requirements which can be modeled through a conceptual framework which, and this is particularly useful in a federated environment, provides a common ground used to avoid misinterpretations among different stakeholders involved in the project itself as well as in the application of its outcomes.

These policies will contribute to enrich the usual *data catalog* with meta-data concerning data gravity [7] and data friction that takes inspiration from physics, modelling the forces that regulate, respectively, the way in which data can be distributed along the continuum, i.e., vertically from the edge to the cloud, and the possibility to share data with other organizations, i.e., horizontally, where involved resources are owned/managed by the organization that has the right to use the data. In this way, the data catalog will be not only an inventory but also a source of information about the privacy/confidentiality requirements attached to the data, which specifies the locations, the possible or necessary transformations, and the processing that can be done on such data.

The *data lake continuum control plane* will create a data management environment able to provide an efficient data solution focused on performance, privacy/confidentiality, and energy efficiency. Knowing which are the available resources and influenced by its gravity, datasets can be placed in one or more locations in the continuum and managed with technologies, e.g., persistent memory media or in-memory, which ensure the satisfaction of performance, privacy, and security requirements. If dictated by privacy requirements, some transformations, e.g., (pseudo-)anonymization, filtering, or encryption, could be applied on this data before moving/copying them to other locations as well as to keep the different copies aligned to the original data source – considering both data-at-rest and data streams.

Finally, the *energy-aware privacy/confidentiality enforcement execution environment* will provide an orchestrated environment deployed along the continuum which allows to automatically satisfy privacy compliance while delivering efficient data access and an adequate performance level and also considering energy efficiency, especially in terms of data movement. In this way, data consumed and produced by applications that are used or developed by data scientists and developers will be monitored to avoid, for instance, data leakage. Policies may require monitoring all data outputted by the application to recognize some data patterns (such as SSNs or credit card numbers) and prevent their output or alert the system. Other orchestration aspects will guard the credentials to data sources from the application by injecting them directly instead of getting them from the application, preventing credentials leakage and human mistakes.

The proposed data lake platform will enable the creation of a federation of data lakes (see Figure 2). The members of the envisioned federation can be both normal data lakes and stretched data lakes where, in the latter case, the data distribution approach will also consider that all the resources belonging to the computing continuum under the responsibility of a given organization can be exploited. Analytical computation can be distributed across such federated data lakes in a privacy-/confidentiality-preserving and trustworthy manner in different configurations which balance the possibility of moving the data among the organizations and potential risks for privacy/confidentiality when data from different sources are integrated (defined by the data friction) and the possibility to employ the computational resources offered by the members of the federation which could require data transformation.

To this aim, a *trustless mediation and computation* environment will be created to provide a more agile approach as organizations are not required to accept – in order to become a member – a particular coordinator which could be a competitor. In the proposed approach, trust among the members of the federation will be established through an appropriate, to be selected (e.g., permissioned/permissionless, proof-of-work/proof-of-stake) blockchain which is easily accessible and used to specify – via smart contracts – the agreements among members about data

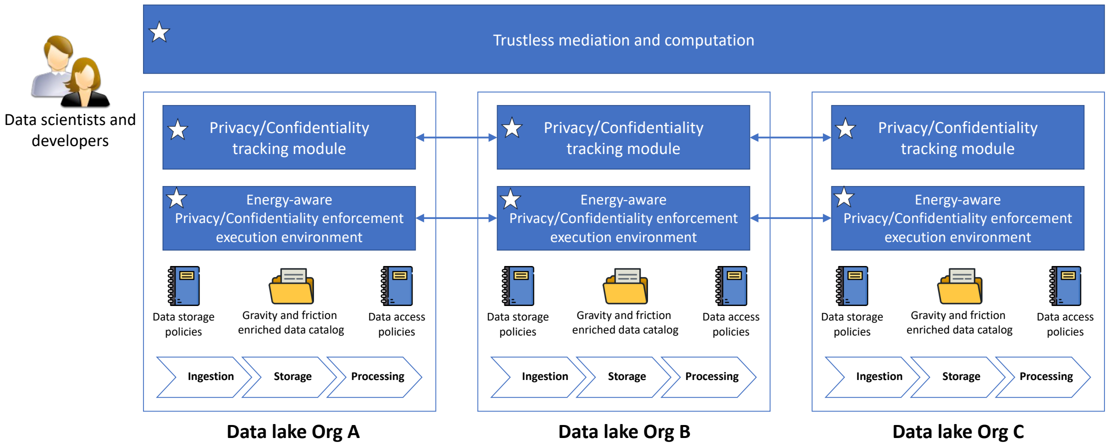
Trustless mediation and computation
Data scientists and developers
Privacy/Confidentiality tracking module
Energy-aware Privacy/Confidentiality enforcement execution environment
Data storage policies
Gravity and friction enriched data catalog
Data access policies
Ingestion
Storage
Processing
Data lake Org A
Data lake Org B
Data lake Org C

Figure 2: Trustworthy data lake federation

sharing: e.g., which data, which format, for which purpose. The consensus algorithms governing the blockchain -– combined with privacy-preserving and scalability-enhancing mechanisms for off-chaining (e.g., based on zero-knowledge proofs) that allow harnessing the benefits of blockchains without publicly revealing underlying data — will ensure a decentralized, secure, and tamper-proof approach for data lake federation. Together with the advanced privacy-related tracking techniques and tools built on top of and consciously tailored to these mediatorless federation mechanisms, this allows verifying if, in particular, shared data have been managed properly, thus verifying the fulfillment of privacy requirements. The approach will also mitigate the drawbacks of blockchain in terms of cost and performance by providing a good balance between on-chain and off-chain data without reducing the provided level of trust. Environmental aspects will be duly taken into account in the selection of the blockchain technology to be used, particularly avoiding energy-consumptive ones like PoW-Ethereum.

3. Current results

The project is currently closing its eighth month and so far the main activities have been focused on (i) identifying the requirements of the pilot cases and, based on them, (ii) defining the general architecture of the proposed solution.

Considering the *requirements elicitation*, one of the project's main challenges is to consider six pilot cases that cover a broad spectrum of the Common European Data Spaces [1] related to the following domains: healthcare, industry, agriculture, finance, mobility, green deal, energy, and public administration. Starting with managing highly sensitive data in the healthcare pilot case, TEADAL must enable medical study promotors to explore, review and select patient records in a federation of hospitals while removing the friction of enforcing policies and navigating different legal frameworks across diverse medical data spaces. Moreover, pilots in the mobility, energy and public administration space require support for managing highly confidential data

[1] http://dataspaces.info/common-european-data-spaces/

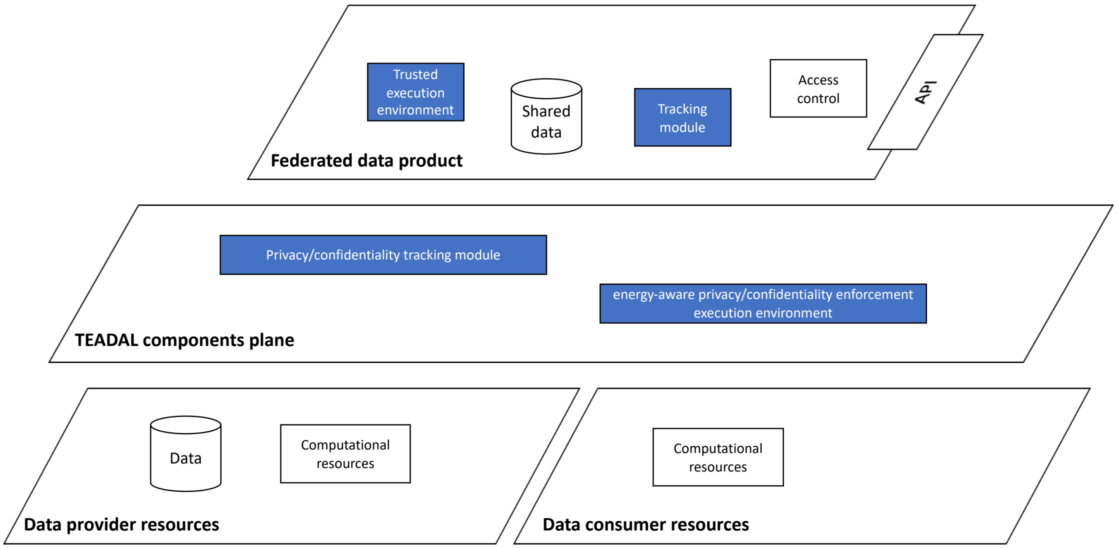

Figure 3: Federated data product

while still enabling large-scale data aggeration. Analogously, pilots in the agriculture and industry 4.0 space explore the combination of data sources distributed across the cloud-edge-continuum, requiring TEADAL to adapt the deployment, transformation and consumption of data to the capabilities of each environment while maintaining transparency about the execution and access to all data. Lastly, the pilots in the financial case introduce the challenge of managing data that is stretched across many units of a company located everywhere in the world while maintaining performant, compliant and cost-efficient access to data for mandatory data analysis on historical and real-time transaction data. Overall, the pilots in TEADAL reveal that the required architecture must balance the data and computational movement both along the cloud-edge-continuum and across administrative domains while maintaining compliant, energy-efficient, transparent and performant data sharing in a broad set of environments.

Based on the requirements that have emerged, it becomes clear that trustworthy data sharing must combine organizational and technical aspects. In this sense, TEADAL is investigating the adoption of a mixed approach based on *data mesh* and *serverless computing* to define an agile solution to enrich a data lake with tools for efficiently and effectively sharing data. Notably, data meshes are emerging as socio-technical systems that can support organizations in better managing data for analytical purposes. Initially proposed by Dehghani [8], this paradigm envisions data management based on an architecture able to manage *data products*, each of them associated with a specific domain. A data product is under the responsibilities of a team that has to take care of the complete life cycle: from the ingestion to the publication. Although data meshes have been proposed to make data sharing inside the organization more efficient, we claim that the same principles can be adopted and extended to regulate data sharing across organizations. In particular, the concept of a data product is extended to a federated data product (see Figure 3): a logical component that embeds data to be shared and the modules that regulate the access to the data and the tracking of data usage.

As one of the main objectives of the project is to make data sharing easier for the organization, the serverless computing paradigm is adopted to design and implement the envisioned federated data product. In this way, the federated data product will be accessible by the data consumer via an API which hides the complexity required to regulate the access, the computation, and the monitoring of the data sharing. Notably, by exploiting the capabilities offered by TEADAL, resources made available by both the data providers and also data consumers could be used to make data management more efficient through, for instance, reducing the data movement.

4. Outlook

In conclusion, to fully realize the potential of currently locked-away data, we must overcome entry barriers: (i) Organisational barriers such as defining privacy and confidentiality governance policies, and (ii) technical barriers such as policy enforcement, resource management across the computing continuum, and trustless mediation of organizations in sharing networks. The TEADAL project aims to tackle these challenges by adopting a stretched data lake approach, incorporating the data mesh and serverless computing concepts, and utilizing new data qualities such as data gravity and data friction to facilitate automatic and mediatorless data sharing federations. TEADAL addresses data-sharing barriers across various European data spaces, including the medical, financial, and agricultural sectors, by carefully selecting and adapting relevant technologies and organizational processes. The adaptable tools developed by TEADAL are expected to reduce entry barriers and unlock the value of hidden data across numerous European organizations.

Acknowledgments

Funded by the European Union (TEADAL, 101070186). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] M. Stonebraker, U. Cetintemel, "one size fits all": an idea whose time has come and gone, in: 21st International Conference on Data Engineering (ICDE'05), 2005, pp. 2–11. doi:10.1109/ICDE.2005.1.

[2] A. Gorelik, The Enteprise Big Data Lake, O' Reilly, 2019.

[3] R. S. Xin, J. Rosen, M. Zaharia, M. J. Franklin, S. Shenker, I. Stoica, Shark: Sql and rich analytics at scale, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 13–24. doi:10.1145/2463676.2465288.

[4] F. Pallas, P. Raschke, D. Bermbach, Fog computing as privacy enabler, IEEE Internet Computing 24 (2020) 15–21. doi:10.1109/MIC.2020.2979161.

[5] C. Cappiello, A. Gal, M. Jarke, J. Rehof, Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391), Dagstuhl Reports 9 (2020) 66–134. doi:10.4230/DagRep.9.9.66.

[6] F. Lucivero, Big data, big waste? A reflection on the environmental sustainability of big data initiatives, Sci. Eng. Ethics 26 (2020) 1009–1030. doi:10.1007/s11948-019-00171-7.

[7] C. Madera, A. Laurent, The next information architecture evolution: The data lake wave, in: Proceedings of the 8th International Conference on Management of Digital EcoSystems, MEDES, Association for Computing Machinery, New York, NY, USA, 2016, p. 174–180. doi:10.1145/3012071.3012077.

[8] Z. Dehghani, Data mesh principles and logical architecture, 2020. URL: https://martinfowler.com/articles/data-mesh-principles.html.