TEADAL

D2.3 Pilot cases' final description and intermediate architecture of the platform

Revision: v.1.0

Work package	WP 2
Task	Task T2.1, 2.2, 2.3
Due date	31/08/2024
Submission date	31/08/2024
Deliverable lead	Cybernetica (CYB)
Version	1.0
Authors	Eduardo Brito, Pille Pullonen-Raudvere (CYB) Victor Casamayor Pujol, Cynthia Marcelino, Boris Sedlak (TUW), Katherine Barabash (IBM, WP4), all pilots and technical partners
Reviewers	POLIMI, ALMAVIVA
Abstract	This deliverable summarises the changes to all TEADAL pilot use-cases and describes the expected final demonstration for each of the pilots. It also introduces one new pilot idea that replaces the financial governance demonstrator. In addition, the deliverable continues describing the TEADAL architecture.

Grant Agreement No.: 101070186
Call: HORIZON-CL4-2021-DATA-01
Topic: HORIZON-CL4-2021-DATA-01-01
Type of action: HORIZON-RIA

Keywords	pilots, architecture, demo

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	01/06/2024	Document structure and template setup	Pille Pullonen-Raudvere (CYB)
V0.2	06/08/2024	Contributions from all partners	See Authors
V0.3	08/08/2024	Cleaned version for final internal review	Eduardo Brito (CYB)
V1.0	06/09/2024	Final version after review	Pille Pullonen-Raudvere (CYB)

DISCLAIMER

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

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. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

COPYRIGHT NOTICE

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	**R**	
Dissemination Level		
PU	*Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)*	✔
SEN	*Sensitive, limited under the conditions of the Grant Agreement*	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

DATA: Data sets, microdata, etc.

DMP: Data management plan

ETHICS: Deliverables related to ethics issues.

SECURITY: Deliverables related to security issues

OTHER: Software, technical diagram, algorithms, models, etc.

EXECUTIVE SUMMARY

The TEADAL project is focused on creating a system for federated data sharing in order to benefit data owners and data users. In the centre of TEADAL there are our pilot use-cases: medical data sharing, access point for public transportation data, smart viticulture monitoring, efficient industry KPIs and regional planning for environmental sustainability. These pilots are documented throughout the series of deliverables D2.1 and D2.2 and the current D2.3 deliverable. In addition to the previous pilots, this deliverable introduces a new pilot idea for financial predictions to replace the previous financial data governance pilot. The goal of the document is to summarise the requirements of the pilots, the expected demo stories and used data, as well as which components of TEADAL technical tools are used by the pilots.

Further, the architecture work continues and some advances are given in this deliverable. First, D2.3 gives an overview of the revised TEADAL architecture and presents the main components that are being developed for the TEADAL's toolset. Then, TEADAL's cluster runtime view is presented as a new architectural view to complement those of deliverable D2.2. Finally, the TEADAL architecture is compared with available Data Spaces initiatives to analyse the possible integration of TEADAL within other Data Spaces architectures.

Funded by the European Union

Table of Contents

Funded by the European Union

LIST OF FIGURES

Funded by
the European Union

LIST OF TABLES

Funded by the European Union

ABBREVIATIONS

FDP Federated Data Product

sFDP Shared Federated Data Product

BPMN Business Process Model and Notation

OHDSI Observational Health Data Sciences and Informatics

OMOP Observational Medical Outcomes Partnership

RAP Regional Access Point

NAP National Access Point

AVM Automatic Vehicle Monitoring

AVL Automatic Vehicle Location

GTFS General Transit Feed Specification

TSP Transport Service Provider

KPI Key Performance Indicator

AOI Area of Interest

CDD Customer Due Diligence

CI/CD Continuous Integration / Continuous Deployment

DAG Directed Acyclic Graph

DLO Data Lake Owner

ETL Extract, Transform, Load

FDP Federated Data Product

GA General Assembly

GDPR General Data Protection Regulation

GPS Global Positioning System

GTFS-RT General Transit Feed Specification Realtime

IaC Infrastructure as Code

IAM Identity and Access Management

JWT JSON Web Token

KPI Key Performance Indicator

Funded by the European Union

KYC Know Your Customer

MDM Master Data Management

ML Machine Learning

OLAP Online Analytical Processing

OLTP Online Transaction Processing

PDP Policy Decision Point

PII Personally Identifiable Information

SaaS Software as a Service

C2B Customer to business

C2C Customer to customer

Funded by the European Union

1 INTRODUCTION

The TEADAL project aims to provide a toolset to allow data providers and consumers more trustworthy, privacy-aware and energy-efficient data usage in a federated setting. The project entails pilot use-cases from different domains - healthcare, mobility, viticulture, industry, and regional environmental development. Each pilot provides unique requirements that the proposed solution should fulfil. Previously, the TEADAL pilot requirements and architecture have been documented in deliverables D2.1 and D2.2. The goal of this deliverable is to document the final updates to the pilot requirements and to continue the documentation of the architecture. For each pilot, the deliverable also lists the TEADAL technologies that are especially relevant for the given pilot.

The Trust Plane anchors the evidence generated by the different services in a TEADAL node. It is offered as an infrastructure for trustworthiness and the enabling of trustworthy mechanisms. The infrastructure and mechanisms are described in detail in D5.2 and TEADAL architecture components are also described in Chapter 8. Part of the Trust Plane components are the Catalogue, for metadata and data event handling, and Advocate, for evidence collection. By leveraging the Trust Plane, pilots can systematically collect verifiable evidence, ensuring the authenticity and integrity of data from the initial stages of software development through to its deployment and operation. Hence, pilots requiring either Advocate or Catalogue make use of the TEADAL Trust Plane.

The role of the TEADAL Control Plane, similarly to the role of the TEADAL Trust Plane explained above, is central to TEADAL federation and is thus required by all the pilots. The purpose of the Control Plane is runtime orchestration of all the workloads deployed in TEADAL federation, across all its locations and all its nodes. The workloads can be both the TEADAL platform components, such as, for example, components implementing metrics collection or policy checking and enforcement, as well as the data pipeline components, such as FDPs, SFDPs, and code producing SFDPs out of FDPs based on signed TEADAL contracts. Runtime orchestration mostly consists of managing workload lifecycles, starting with the deployment of individual components on suitable infrastructure elements, overseeing components operations while running, and decommissioning the resources as soon as the work is completed. As was described in the initial architecture documents (D2.2 and D4.1), the Control Plane is not the sole decision plane of the project. Rather, it works in cooperation with other components of the TEADAL platform by executing actions based on these other components' inputs, to achieve federation's goals in the best possible way. For example, in cooperation with the Trust Plane that might determine that certain components misbehave, the Control Plane is responsible to stop or to quarantine the culprits. Another example is making workload placement decisions based on performance and energy consumption data collected from the nodes and analysed by the data governance components developed in WP3. Details of the TEADAL Control Plane architecture and its kubernetes based implementation were already presented in D4.1 and during the first periodic review. At the current stage of the project, the TEADAL Control Plane needs to be extended to support cooperation with the decision making components as required by the individual pilots. This will be summarised in D4.2 and implemented as part of pilot integrations in the next stage of the project. In addition, WP4 team began automating data pipeline creation processes using generational AI technologies so that, for example, given a signed SFDP contract, the data product corresponding to sFDP can be generated and deployed automatically by the TEADAL Control Plane.

This document focuses on the requirements and goals, the actual progress of the pilot setup and demonstration development is documented in deliverable D6.2. Note that this document updates the lists of requirements for each of the pilots, however the actual demonstrator of the pilot in the TEADAL project may not fulfil all these requirements.

Funded by the European Union

The main chapters of the deliverables can be divided into two distinct parts. First, chapters 2 to 8 focus on the pilots. Then Chapter 9 is focused on the TEADAL architecture.

Chapter 2 USE CASE PILOT #1: EVIDENCE-BASED MEDICINE gives the final set of requirements and description of the pilot by RIBERA SALUD.

Chapter 3 USE CASE PILOT #2: MOBILITY gives an overview and the final demo idea for the pilot by AMTS and UITP.

Chapter 4 USE CASE PILOT #3: SMART VITICULTURE gives a brief overview and update on the current state of the pilot by TERRAVIEW.

Chapter 5 USE CASE PILOT #4: INDUSTRY 4.0 gives an overview of the pilot, led by ERT.

Chapter 6 USE CASE PILOT #6: REGIONAL PLANNING FOR ENVIRONMENTAL SUSTAINABILITY gives an overview of the pilot and demo by Regione Toscana and BOX2M.

Chapter 7 NEW USE CASE PILOT: OPTIMISING FINANCIAL RETURNS FROM RENEWABLE ENERGY SOURCES introduces the replacement of the previous financial data governance pilot. This new pilot is led by BOX2M.

Chapter 8 describes the second iteration of the TEADAL general architecture, presenting the main components of the TEADAL's toolset, the runtime view of the architecture, and the alignment of TEADAL with other Data Spaces.

Funded by the European Union

2 USE CASE PILOT #1: EVIDENCE-BASED MEDICINE

A thorough description of the evidence-based medicine pilot, its stakeholders, goals, and project requirements can be found in deliverable D2.1, under the "USE CASE PILOT #1: EVIDENCE-BASED MEDICINE" chapter. Some updates were defined also in previous deliverable D2.2 regarding the data used for the pilot. This chapter gives further updates and describes the expected final demonstrator for this pilot. The responsibility for this pilot has shifted from MARINA SALUD to RIBERA SALUD.

Pilot Overview

The objective of the pilot in evidence-based medicine is to elevate the current stage of data analytics in the healthcare sector, making it easier to share and analyse medical data. This pilot will primarily address working with data privacy restrictions and the need to ask data subjects' consent for data processing. Privacy restrictions lead most studies to work with anonymised data. As illustrated in Figure 1, RIBERA SALUD will simulate the sharing of data among healthcare organisations in a federated manner, while TEADAL tools will establish mechanisms to navigate data privacy restrictions and manage data access. These actions will occur both organisationally, setting up trust between organisations when data access requirements are satisfied, and individually, supplying tools to manage or limit the use of data solely from those participants who gave consent for inclusion in the medical investigations.

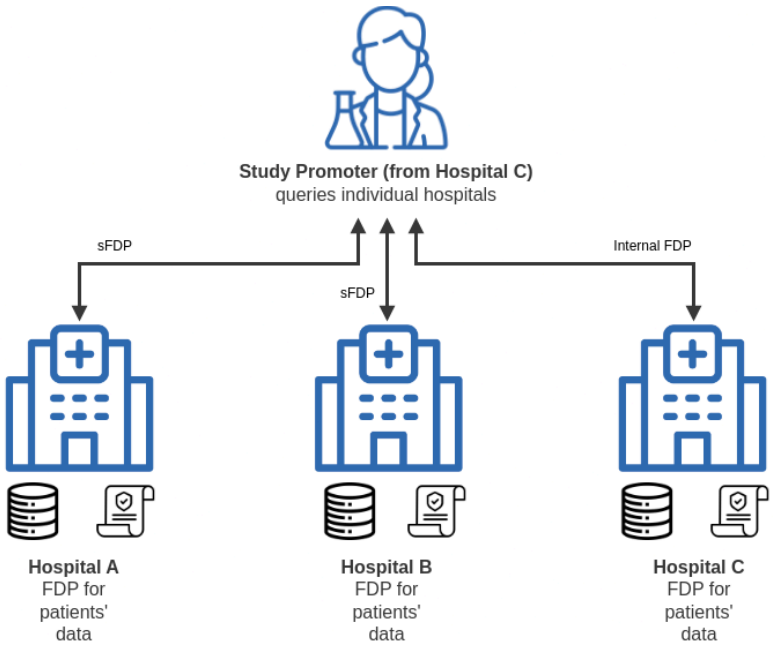


FIGURE 1: OVERVIEW OF THE EVIDENCE-BASED MEDICINE USE CASE PILOT.

Demo Story

The Final Demo for the evidence-based medicine pilot will run a clinical study through the TEADAL node by selecting some medical parameters in a web-based client, showing the processes involved to run this evidence-based medicine study and return only the requested data to the federated user. The steps will be selecting affected patients, asking for their consent to participate in the study and then collecting the desired data.

Funded by the European Union

The final topology has three hospital nodes, two that are simply Hospitals and one that is both Hospital and Researcher who will run the study.

Data Description and Access Policy

RIBERA SALUD Hospital A uses 6 datasets of around 69.2 MB of total size for the pilot.

The data follows the OHDSI (Observational Health Data Sciences and Informatics) standards data model, OMOP (Observational Medical Outcomes Partnership), developing 6 tables:

1. **Drug Exposure:** The drug exposure domain captures records about the utilisation of a Drug when ingested or otherwise introduced into the body. Drugs include prescription and over-the-counter medicines, vaccines, and large-molecule biologic therapies. Radiological devices ingested or applied locally do not count as Drugs.

2. **Observation:** This table captures clinical facts about a person obtained in the context of examination, questioning or a procedure. Any data that cannot be represented by any other domains, such as social and lifestyle facts, medical history, family history, etc. are recorded here.

3. **Person:** This domain contains records that uniquely identify each patient in the data source who is time-at-risk to have clinical observations recorded within the source systems.

4. **Procedure Occurrence:** This table contains records of activities or processes ordered by, or carried out by a healthcare provider on the patient to have a diagnostic or therapeutic purpose. Procedures are present in various data sources in different forms with varying levels of standardisation.

5. **Measurement:** This table contains records of measurement, i.e. structured values (numerical or categorical) obtained through systematic and standardised examination or testing of a person or person's sample. It contains both orders and results of such measurements as laboratory tests, vital signs, quantitative findings from pathology reports, etc.

6. **Condition Occurrence:** Conditions are records of a person suggesting the presence of a disease or medical condition stated as a diagnosis, a sign or a symptom, which is either observed by a provider or reported by the patient. Conditions are recorded in different sources and levels of standardisation, for example:

    - Medical claims data include diagnoses coded in ICD-9-CM that are submitted as part of a reimbursement claim for health services and

    - EHRs may capture personal conditions in the form of diagnosis codes or symptoms.

All the TEADAL nodes were loaded with these datasets, each hospital node containing no less than 100,000 unique patients. In terms of access policies, data access inside the hospitals' federation was defined for federated & non-federated users. Non-federated users won't have access to any data. Federated users can access all the datasets. We also described a "root" user for maintenance tasks (system administration and related tasks, exclusive for each hospital).

Funded by the European Union

Final Requirements

User requirements were collected in D2.1, and then confirmed or updated in D2.2. Table 1 collects the main and final requirements for this pilot.

Req. ID	Description
P1-Gen02	Allow hospitals to easily advertise the data in the data catalogue.
P1-Privacy01	Help to cope with GDPR (represent consent to enable GDPR like access, or policies based on data to enable compliant access).
P1-Privacy02	Allow to share only data from patients that consented (see if data contains consent information and if it is related to personal identifiable information).
P1-Arch01	Support a field-specific ontology in the catalogue.
P1-Arch02	The architecture has to facilitate as much as possible the automation of the process to set up the study.
P1-Arch03	The architecture has to tune the data lake to be able to offer the data as agreed.
P1-Arch04	Offer a variety of computation methods with different datasets (including with privacy preserving mechanisms).
P1-Mgmt01	Provide a data catalogue that tells the study promoter where he will find the patients that it can study.
P1-Mgmt02	Provide a variety of data sanitization methods with callbacks to data owners (some in TEADAL, others are plugged in by use-cases)
P1-Gen03	Offer means to measure sustainability related metrics for each action taken by TEADAL.
P1-Gen04	The request for a new consent to the patient needs to be simple.
P1-Gen05	Facilitate the bureaucracy of a study proposal.
P1-Arch05	Implement orchestration of the computations.
P1-Policy01	Define policies on how data could be copied and stored.
P1-Mgmt03	Allow to extract information about the amount of patients affected (SELECT COUNT) to know the strength of the study.
P1-Mgmt04	Manage the process to store and copy data, enabling accountability about data access, transformation and use in each study.
P1-Gen06	Allow to approve access to personal data to certain "study promoters" by comparing basic rules.
P1-Privacy03	Support for other privacy protection laws and schemes (GDPR has to be one of the options).

Funded by the European Union

P1-Arch06	Provide technical components to enable GDPR like compliant data access, e.g., tools to anonymize, tools to check anonymity or tools to check informed consent.
P1-Arch09	Provide some kind of persistency for multiple-step studies.

TABLE 1: FINAL REQUIREMENTS OF THE EVIDENCE-BASED MEDICINE USE CASE PILOT.

TEADAL Technologies

The Catalogue component is used in two separate flows in the storyboard of this pilot:

- To let doctors discover FDPs made available in the federation.
- To manage consent permissions therefore starting the process leading to the creation of SFDPs.

This orchestration is achieved by defining appropriate BPMN processes which are deployed in the Catalogue, and relies on the Catalogue federation features which are going to be implemented.

Advocate, as introduced in D5.1, is part of the trust plane of TEADAL and responsible for collecting interaction evidence. The purpose of the evidence is later auditing and for building a chain of trust between data providers and data consumers. Such evidence is particularly important for this medical use case, where highly sensitive patient data is shared within the federation. Here, Advocate, will observe the entire data sharing life-cycle. From the selection and access in the Catalogue, the creation of an SFDP, to the access in all TEADAL nodes interacting with the data. The data provider has the option to describe policies that should continuously hold true, e.g., a SFDP should only be accessed by a consumer of a specific organisation and should only be deployed at a specific location. The Advocate will continuously review the collected evidence to attest these policies for a simplified auditing process.

Since this use case may involve several energy-intensive processes applied to the data, AI-based energy optimizations will also be employed, in accordance with the monitoring, scheduling and deployment strategies detailed in D3.2 and in the upcoming deliverable D4.2. Firstly, transformations such as querying and anonymization are performed. Additionally, data is transferred from hospitals to the study promoter. Finally, the data is stored on various servers for a specified period, depending on the likelihood of its future reuse. The energy consumed by these processes can be estimated using the AI-based energy optimization module, which also determines the optimal deployment of the data pipeline.

Data Synthesis

To ensure data consistency across all nodes, we synthetically generated all available pilot data following OMOP standards. This approach guarantees that the data remains coherent and standardised across the TEADAL federation. Once the datasets were created, they were uploaded to their corresponding TEADAL nodes, ensuring each node received the appropriate dataset for its further FDP advertisement in the catalogue.

Funded by the European Union

The OpenAPI definition for the FDP, which includes detailed schemas, endpoints, and response structures, can already be accessed[1]. The study promoter will then query the data from the TEADAL federation, retrieving the SFDP, to fulfil its specific research needs. By querying the "patients" table, they can retrieve targeted information, such as identifying patients over the age of 33 who have been exposed to a specific drug. This capability allows for precise, trustworthy, and eventually privacy-preserving data analysis, enabling the study promoter to later conduct in-depth research and draw meaningful conclusions from the data.

1 https://gitlab.teadal.ubiwhere.com/teadal-pilots/medicine-pilot/fdp-medicine/-/blob/main/api/fdp-medicine.yaml

Funded by the European Union

3 USE CASE PILOT #2: MOBILITY

A thorough description of the pilot, its stakeholders, goals, and project requirements can be found in deliverable D2.1, under the "USE CASE PILOT #2: MOBILITY" chapter. Some updates to the pilot were also described in deliverable D2.2 and final pilot description and expected demo story are described in this chapter.

Pilot Overview

The mobility pilot by AMTS and UITP leverages TEADAL technologies to showcase data sharing among four Italian public transport stakeholders: the regional transportation operator (AMTS), national transport operator (open data from Trenitalia), Regional Access Point (RAP), and National Access Point (NAP). This pilot simulates these components to showcase the capabilities of the TEADAL tools. However, Italy is currently working out a similar system to collect and manage public transport data. Urban area data compilation initiatives at the regional level face restrictions due to inconsistent cross-border cooperation. To tackle this, Italy has allocated the responsibility for transport data gathering to individual regions, fostering a three-tiered structure. This arrangement involves the RAP collecting data from transport operators and infrastructure managers, representing edge nodes, subsequently distributing it to the NAP. NAP can then be used as a single access point for all public transport data in Italy, to be accessed by the country's service providers and other interested users. An illustration of the main participants and data of the pilot is given in Figure 2 and the next section details the interactions.

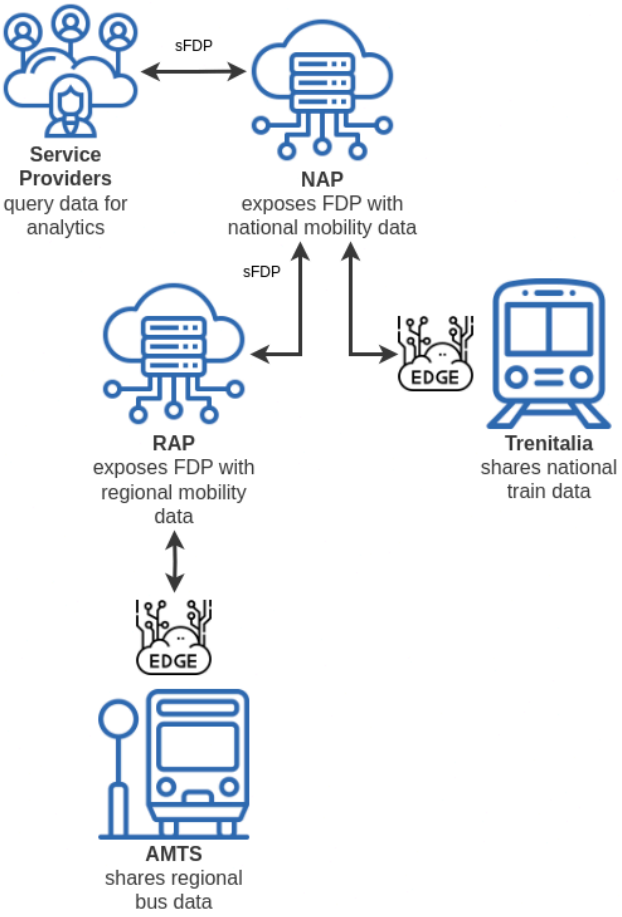


FIGURE 2: OVERVIEW OF THE MOBILITY USE CASE PILOT.

Funded by the European Union

Demo Story

The demo lies around the architecture outlined above, where data providers make available both static and dynamic information to the upper layers, in order to facilitate their accessibility and use upon specific policies. The data flows is designed as follows:

- AMTS data is uploaded to RAP and from there to the NAP.
- The NAP and RAP nodes are implemented with TEADAL technologies.
- Static mobility data, for different modes of transport, may be provided to the multimodal NAP by transport authorities, transport operators, infrastructure managers, or transport service providers, on demand, as is the example of open data from Trenitalia, the national railway company of Italy.
- AMTS will include bus static mobility data and real time data.

The NAP for multimodal mobility serves as a central point that collects and provides data related to various modes of transport, simplifying data consultation through a single portal. In this pilot and current demo, AMTS is the sole transport service provider represented. However, to enhance the platform's capabilities, simulated data sourced from Open Data portals, such as Trenitalia, has been incorporated into the NAP. This additional data enables the combination of diverse datasets, allowing for enriched data analytics and testing of multimodal and multi-operator services.

After uploading data into the NAP several manipulations and combinations of such datasets could be carried out to extract insightful analytics and potentially enable new business such as:

- analysis on **punctuality and reliability**, to check points of failure of the service and to assess LoS contractually agreed;
- analysis on **operational efficiency,** to measure route and service efficiency
- Analysis of **potential transport offer**, to define improved multi-operator schemes such as Mobility-as-a-Service or door-to-door transport

Data Description and Access Policy

Automatic Vehicle Monitoring (AVM) is a system that allows transport operators to monitor, in real time, various variables related to transport services (position detection, route description, speed, stops, and delays). AVM features are not limited to monitoring such variables but, in the context of local public transport, they also monitor the service performed by the vehicle (in the form of individual vehicle shifts). The AVM is based on Automatic Vehicle Location (AVL) technology, which is the subsystem that takes care of vehicle remote location monitoring (e.g. with GPS).

The data produced is provided in General Transit Feed Specification (GTFS) format. A GTFS feed is a collection of CSV files contained in a zip file. Together, the related CSV tables describe operations of transport services. GTFS consists of two main parts: GTFS Schedule and GTFS Real-Time. GTFS Schedule contains information about routes, schedules, fares, and geographic transit details, and it is presented in simple text files. This straightforward format allows for easy creation and maintenance without relying on complex or proprietary

Funded by the European Union

software. GTFS Real-Time contains trip updates, vehicle positions, and service alerts. It is based on Protocol Buffers, which are standard mechanisms for serialising structured data, and is constantly produced and acquired (approximately every minute).

AMTS adheres to the national Open Data policy to stimulate the integration of travel planners' methods of movement within the territory, to increase the penetration and capillarity of information on company services, and to stimulate technological innovation in the sector. For this reason, the data that is included in the pilot is provided without restriction policies. Within the Mobility Pilot, UITP will take care of retrieving static datasets (updated on a semestral basis for an approximate file size of 2MB) and a stream of dynamic ones for a period sufficient to carry out consistent data analysis (updated every minute, with each file update having a size of few KB).

Final Requirements

User requirements were collected in D2.1, and then confirmed or updated in D2.2. Table 2 collects the main and final requirements for this pilot.

Req. ID	Description
P2-Gen01	To access to Open Street Map for the street network of the city/region involved.
P2-Gen02	To be able to take information from the NAP about the schedules of all public transport operators in a specific area.
P2-Gen03	To access data from the operator on the stop arrival times.
P2-Gen04	Data (or running of queries) in the right place at the right time (for example, when performing analysis workflows or training AI models).
P2-Mgmt01	To manage both batch and real time data (e.g. from GPS).
P2-Gen06	Include also information about the average price of the trip, without the need to give the opportunity to book the trip.
P2-Gen07	To offer transparency of data access across NAP, RAP, Local, etc. (e.g. allow the NAP to see what happened to the data downstream).
P2-Privacy01	Each Transport Service Provider (TSP) should be able to prevent a subset of his data to be shared with other TSPs (e.g. competitors).
P2-Arch01	To allow for joining the RAP or NAP TEADAL federation to access available data.

Funded by the European Union

P2-Arch02	To provide only updated data, i.e. datasets can be dynamic so that only last versions should be available.
P2-Arch03	We should avoid copying data to avoid outdated or out-of-sync information.
P2-Policy01	Policies should be established if one of the federated data lakes has stretched components on a public cloud (i.e. Amazon).
P2-Mgmt02	The TSP should have a way to check who accessed or requested data at the national level.
P2-Mgmt03	The TSP should have a way to decide whether a dataset should be shared in the federation.
P2-Mgmt04	Each TSP should be able to access data belonging to other TSPs through the RAP or NAP.
P2-Policy02	To aid in verifying correct application of data policies.
P2-Policy03	Enable a mechanism to prohibit access or sharing of data that should not be available to the NAP or RAP.

TABLE 2: FINAL REQUIREMENTS OF THE MOBILITY USE CASE PILOT.

TEADAL Technologies

Since this pilot mimics the relationship between the NAP and RAP for multimodal transportation data, the Catalogue is used in place of the NAP and RAP user interfaces, for instance, to demonstrate that an FDP published on the AMTS data lake can be visible both at the regional and at the national level.

AI-based energy optimizations are also necessary for this pilot, since FDPs are being transferred continuously, for example, from the RAP to the NAP, and from there to NAP users, consuming energy in each data movement. The number of transfers can be reduced, for instance, if a copy of the most demanded FDP can be stored in the NAP domain. The AI-based energy optimization module can determine the optimal timing and location for storing these copies to reduce overall energy consumption.

Funded by the European Union

4 USE CASE PILOT #3: SMART VITICULTURE

A thorough description of the pilot, its stakeholders, goals, and project requirements can be found in deliverable D2.1, under the "USE CASE PILOT #3: SMART VITICULTURE" chapter. Some updates to the pilot were also described in deliverable D2.2. The final pilot description and expected demo story are described in this chapter.

Pilot Overview

Over the past 5-10 years, vineyard operators and winemakers have been faced with a growing number of operational challenges. These challenges stem from climate change, tighter regulatory compliance, and shifting expectations of contemporary consumers. In response to these challenges, Terraview launched a Software as a Service (SaaS) platform named TerraviewOS. This platform equips vineyard operators with a sophisticated tool for more effective asset management. TerraviewOS enhances decision-making processes for vineyard operators by providing data-driven forecasts and timely alerts for dealing with unforeseen circumstances like pests, diseases, and weather-related issues. This offering has been complemented by the release of Aquaview, a service that empowers any customer with agricultural land to comprehend their water usage via detailed water moisture maps. Each map comprises a series of adjoining water moisture probes with an accuracy of +/- 1.5% in comparison to hardware-based control sensors.

TEADAL tools extend the capabilities of these tools and facilitate the creation of a solution for enabling data sharing across different vineyards, particularly those in close proximity to one another. This aims to enable swift monitoring of changes that could potentially impact nearby vineyards (e.g., disease alerts, weather data, specific spray logs via traceability, etc.) with a specific focus on water moisture profiles (surface level, and depth soil moisture). An illustration of the main participants and data of the pilot is given in Figure 3 and 4.

Funded by the European Union

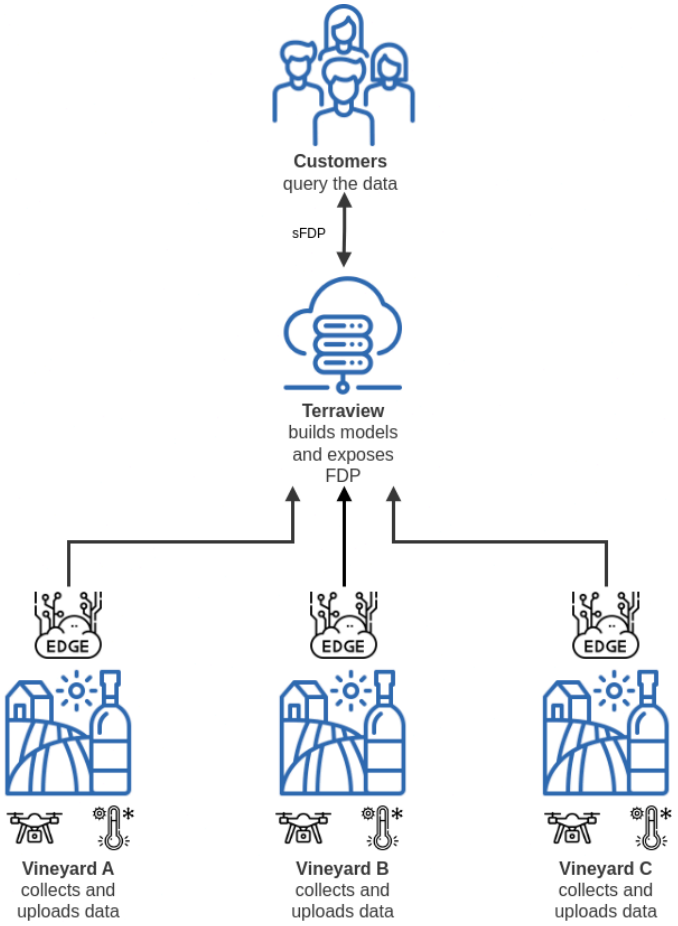


FIGURE 3: OVERVIEW OF THE SMART VITICULTURE USE CASE PILOT.

Demo Story

Using a viticultural plot, the key aim of this pilot is to collect data from the plot to allow the calculation of soil moisture information without the use of hardware sensors. Once this information is created, we will be able to share that data using the TEADAL components and protect access to the data through policies.

In this revision of our prototype, we will provide a tool that allows an end user access to the data available to them. From a provider perspective, a tool to expose the data and set access policies will be provided. From an architectural perspective, the existing work will be extended to provide SFDP capabilities. Other aspects include the use of satellite data and a focus on business policies.The demonstrator uses a centralised data lake with some virtual edge devices. These edge devices simulate the actions of the customer's data mediators. Additionally, a single client can manage multiple vineyards.

Funded by
the European Union

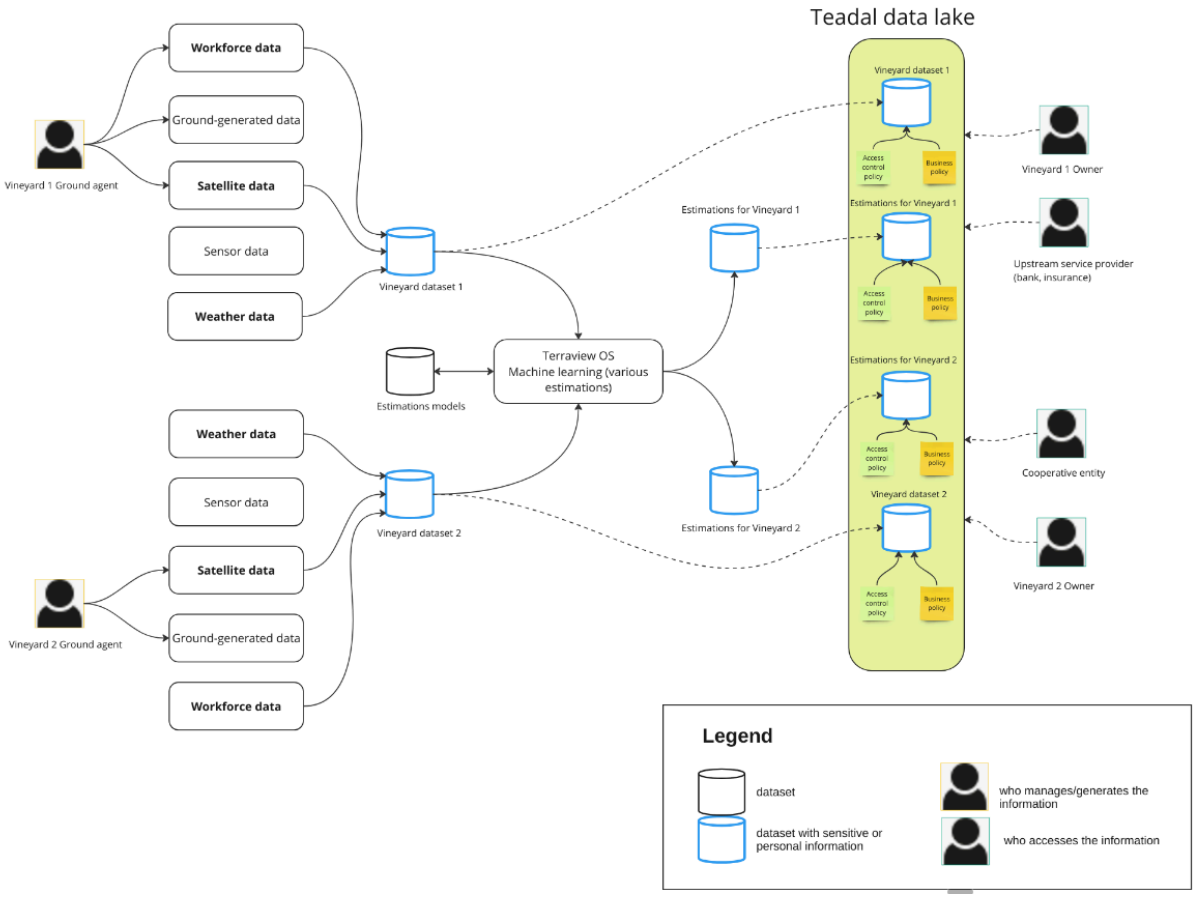


FIGURE 4: DATA FLOW AND INTERACTIONS OF THE SMART VITICULTURE USE CASE PILOT.

Data Description and Access Policy

The description of the data and its generation is contained within the previous deliverable D2.2, and the access policies for this pilot will be developed and showcased within the context of the TEADAL Role-Based Access Control framework, detailed further in this document.

Final Requirements

User requirements were collected in D2.1, and then confirmed or updated in D2.2. Table 3 collects the main and final requirements for this pilot.

Req. ID	Description
P3-Gen01	Support C2B.
P3-Gen02	Support C2C (e.g. Vineyard operator-A to Vineyard operator-B).

Funded by the European Union

P3-Policy01	Define policies for the control of data lakes. From the owner perspective, the solution must implement an interface for managing data sharing consent policy for vineyard owners. From the external user perspective, there must be a way to consume data given to him/her.
P3-Policy02	Ensure that data is accessed according to policies.
P3-Gen03	Support B2B sharing (e.g. Terraview to insurance company).
P3-Arch01	The datasets should be placed in the continuum according to pilot-case specific notions, including geography.
P3-Arch02	Let different data lakes federate at all levels of the continuum (especially at the edge).
P3-Mgmt02	Sharing with 3rd party upstream providers, e.g finance and insurance service providers (no payment).

TABLE 3: FINAL REQUIREMENTS OF THE SMART VITICULTURE USE CASE PILOT.

TEADAL Technologies

Data access policies will be implemented using the TEADAL Role-Based Access Control framework documented in D4.1. Summarised, satellite data and soil moisture measurements are modelled as REST resources and access to these resources is governed by rules that specify what HTTP methods a given role may request on each resource. There are roles both for vineyard owners and external users who would like to consume soil moisture measurements. Actual users are defined in Keycloak (TEADAL's IdM service) and then mapped to RBAC roles.

Funded by the European Union

5 USE CASE PILOT #4: INDUSTRY 4.0

A thorough description of the pilot, its stakeholders, goals, and project requirements can be found in deliverable D2.1, under the "USE CASE PILOT #4: INDUSTRY 4.0" chapter. Some updates to the pilot were also described in deliverable D2.2. The final pilot description and expected demo story are described in this chapter.

Pilot Overview

The industry pilot will be focused on the necessity of computing a batch of key performance indicators (KPIs) that are shared between two ERT Group facilities located in different countries (Portugal and Czech Republic). Given that data is typically collated based on each facility separately, the KPIs for the entire corporation must be calculated and compiled according to the unified standard of the group as a whole. The objective of TEADAL is to enhance and automate the process of calculating KPIs that are pertinent to the company's management-related aspects (operational, commercial, quality, etc…). TEADAL will offer tools to automate and fine-tune the tech-impact related to data sharing within the ERT Group. An illustration of the main participants and data of the pilot is given in Figure 5.

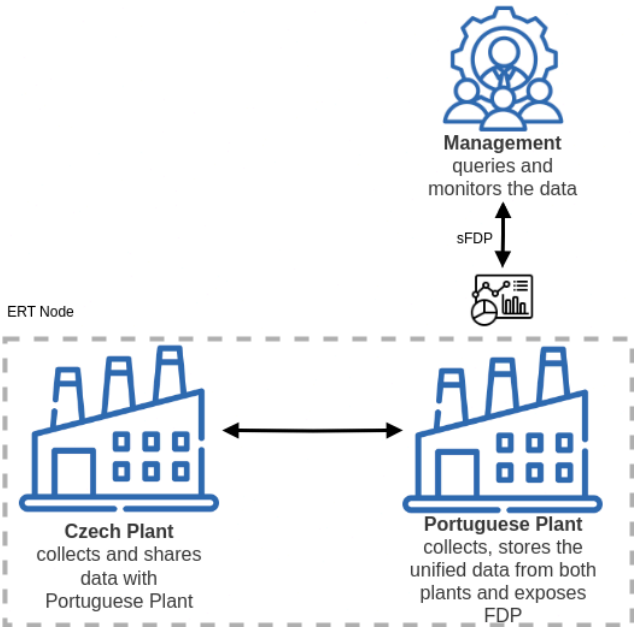


FIGURE 5: OVERVIEW OF THE INDUSTRY 4.0 USE CASE PILOT.

Demo Story

The final demo of this pilot is an interface accessible via web to graphically present the KPIs. The access and information viewed on the site should be controlled based on access policies. After entering, the user can access the KPIs based on consolidated data from both plants. The focus of the demo will be on security and access control, with the Portuguese plant functioning both as a plant and as the headquarters. The demonstration will centre around a dashboard, with at least two users accessing it to view different sets of data. The headquarters manager will have full access, while plant managers will be restricted to viewing

Funded by
the European Union

data only from their respective plants. Overall, the demo requirements have remained consistent with those documented in the previous deliverables.

Data Description and Access Policy

ERT plans to use datasets of 200 MBs of total size for the demo. In the real production database, there are 100GBs of data available. The data is based on the four largest departments in ERT (Sales, Production, Logistics, Quality). There are five key access rules:

1. Top Management has access to the four departments.
2. Sales Management has access only to the Sales department.
3. Production Management has access only to the Production department.
4. Logistics Management has access only to the Logistics department.
5. Quality Management has access only to the Quality department.

Final Requirements

User requirements were collected in D2.1, and then confirmed or updated in D2.2. Table 4 collects the main and final requirements for this pilot.

Req. ID	Description
P4-Arch01	Support the normalisation of ingested data across the distributed databases.
P4-Arch02	Providing an easy-access interface for querying the data to generate the reports (e.g., API for accessing data).
P4-Arch03	Create an harmonised/standard protocol for data collecting, processing and information sharing with different plants, departments and teams.
P4-Arch04	Support data located in more than one storage.
P4-Arch05	Support data loading from an SQL database.
P4-Gen01	Standardising information coming from different plants.
P4-Gen02	Reduce costs and improve the operation efficiency for generating reports.
P4-Policy01	Define the level of "confidentiality" of data, KPIs and reports.

Funded by the European Union

P4-Policy02	Define policies (read, use, edit, forward) for accessing specific data, KPI values and/or KPI categories, and reports.
P4-Gen03	Allow sharing of reports with external companies involved in the supply chain.
P4-Gen04	Ensuring the security of accessing data whenever they are exchanged.
P4-Gen05	Ensure that data is processed and stored using ERTs infrastructure.
P4-Mgmt01	Compliance with regulations especially for security. Compliance with Portuguese legislation, ISO 27001 and IATF16949.
P4-Policy03	Define rules (aggregation, obfuscation, etc.) for visualising data, KPI values and/or KPI categories, reports that are accessed by users with restricted rights.

TABLE 4: FINAL REQUIREMENTS OF THE INDUSTRY 4.0 USE CASE PILOT.

TEADAL Technologies

The data access policy detailed earlier will be implemented using the TEADAL Role-Based Access Control framework documented in D4.1. Department data are modelled as REST resources and access to these resources is governed by rules that specify what HTTP methods a given role may request on each resource. There is a role in correspondence of each kind of manager specified earlier: Sales manager, Production manager, and so on. Actual users are defined in Keycloak (TEADAL's IdM service) and then mapped to RBAC roles.

Martel will also play a role in the policy implementation for the Industry 4.0 trial. The motivation behind selecting this trial lies in the similarities between ERT's data access patterns and the multi-tenant setup of Martel's own IoT platform, Orchestra Cities. Martel aims to implement a security layer in Orchestra Cities that mirrors the one developed for TEADAL. By gaining practical experience with ERT's security policies, Martel expects to gather valuable insights that will inform the security policy implementation for their platform.

Funded by the European Union

6 USE CASE PILOT #6: REGIONAL PLANNING FOR ENVIRONMENTAL SUSTAINABILITY

A thorough description of the pilot, its stakeholders, goals, and project requirements can be found in deliverable D2.1, under the "USE CASE PILOT #6: REGIONAL PLANNING FOR ENVIRONMENTAL SUSTAINABILITY" chapter. Some updates to the pilot were also described in deliverable D2.2. The final pilot description and expected demo story are described in this chapter.

Pilot Overview

The aim of the pilot is to connect sensor data relating to environment and energy consumption monitoring of buildings deployed by a private enterprise with building energy profiles managed by public authorities. This pilot involves two partners: BOX2M, a private firm, and RT, a public authority of the Tuscany Region, Italy. The main goal is to facilitate the reconstruction of both static and dynamic energy profiles for public and private structures, along with the delineation of local energy efficiency and air quality trends. Open data about weather conditions and air quality is also incorporated into the analysis. The goal is for RT to get a better overview if the certification documents and real energy consumption are in accordance with each other. An illustration of the main participants and data of the pilot is given in Figure 6.

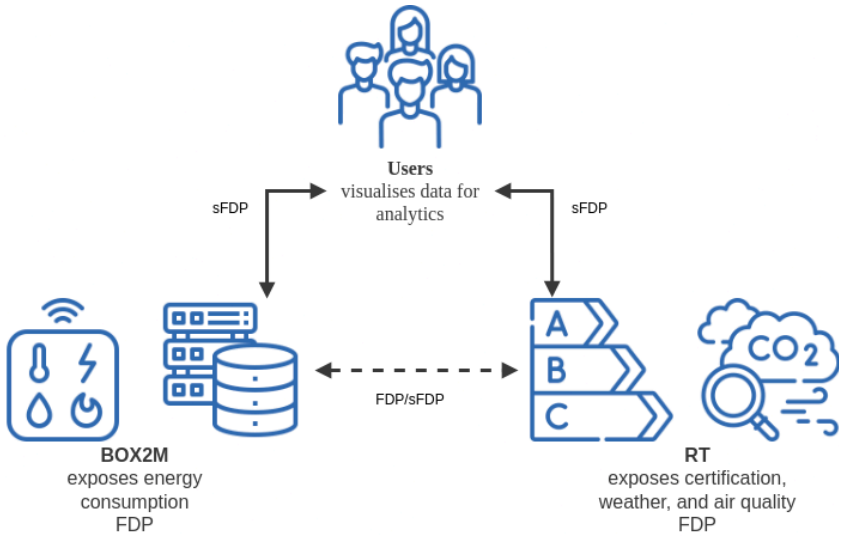


FIGURE 6: OVERVIEW OF THE REGIONAL PLANNING USE CASE PILOT.

Demo Story

The final demo will feature a data visualisation tool that aggregates information from static sources on the RT node and API calls from the BOX2M node. The dashboard will be accessible to multiple users with the same access levels, displaying data in aggregated formats. Additionally, the RT node could share the integration results with BOX2M through an RT FDP/SFDP channel, as illustrated in Figure 6. In the demo, there are no external users accessing the FDP from BOX2M but they are supported by the approach.

Funded by the European Union

Data Description and Access Policy

The data belonging to RT was described in full detail in D2.1 and D2.2, including its sources, formats, volumes, and synthetic data generation strategies. The BOX2M dataset is produced by sensors deployed in the field. For the demonstration purpose, BOX2M is going to deploy 11 sensors. Every 15 minutes, their energy consumption index would be read in real time by the custom BOX2M devices. This information is parsed and sent via MQTT to the BOX2M broker in Microsoft Azure cloud. The data is further processed and is accessible in real time from the BOX2M FDP that is built to integrate this data set with the needs of RT. The size of each message should be 1Kb, having approx 1000 messages per day, and a daily datasets volume of 1 MB. Hence, the overall data volume depends on the observation period.

Final Requirements

User requirements were collected in D2.1, and then confirmed/updated in D2.2. Table 5 collects the main requirements for this pilot.

Req. ID	Description
P6-Gen01	Data about plants and buildings coming from the Tuscany Region must be enriched with BOX2M dynamic data coming from sensors in an aggregated territorial perspective.
P6-Gen02	The aggregated data must be used for decision support system development.
P6-Gen03	Users must benefit from the combined use of RT and BOX2M data and from the analytics produced on top of these datasets.
P6-Privacy01	The system must implement policies to ensure data confidentiality according to access rules defined by the pilot, e.g., normal users must not be able to access raw data.
P6-Privacy02	The aggregation of data has a minimum threshold of 3 units. No analysis can be performed if less than 3 records are affected.
P6-Privacy03	The system must forbid the users to see confidential data about buildings and plants stored in the RT data lake.
P6-Privacy04	The solution must allow sensor owners to accept or decline the use of their data for specific analytics purposes. The way for gathering consent must be an "Opt-in" solution: when an installation or certification of a plant happens, the owner can

Funded by the European Union

	give his/her consent for sharing data for that specific purpose. The gathered consent must be stored in a database.
P6-Privacy05	The must be 3 level of consent referring to P6-Privacy04: BOX2M is allowed to collect building's data and move them on cloud. (mandatory) Plant owners give to BOX2M the consent to analyse data, but not to share them (optional) Plant owners give to BOX2M the consent to analyse data and share them (optional)
P6-Arch01	The solution must define a sort of ecosystem where RT is at the centre and BOX2M is one of the actors who is providing data. RT is one of the main consumers of data.
P6-Arch02	The solution must define a logical component which federates SIERT dataset, open data and BOX2M sensor data without replication.
P6-Arch03	The solution must define one node for RT and one node for BOX2M.
P6-Policy01	Data transfer from edge to cloud must be case dependent. It depends on the type of analytics to be performed.
P6-Policy02	The solution must define some policies for data movement according to the use case. Therefore, the type of analysis to be done (data involved and expected results/analysis) must be clearly defined from the beginning.
P6-Policy03	The solution must manage data retention, data management and data sharing policies.
P6-Policy04	Metadata must be assigned to datasets. Metadata consists in one or more tags describing the data contained (sensitive data, sensor data, environmental data…) and they may be specific for each use case. Data policies could be based on the tags assigned to the datasets.
P6-Mgmt01	The solution must provide the ability to mark certain data elements as invalid in a particular interval, since some field sensors could be altered by malice, neglect or vandalism
P6-Mgmt03	Dataset should be accessible in their original format (xml, json, csv and relational data).
P6-Mgmt04	Data sanitization (data can be contaminated and need to be cleared) should be a conceptual feature of the TEADAL toolset. Since the cleaning service is strongly

Funded by the European Union

	dependent on the contents and data schemas of the data sources, the specific implementation of the data sanitization must be implemented by each pilot.
P6-Arch04	The open data node should be implemented as a real node, to simulate the fact that it can be continuously accessed by other actors (simulation of workloads).

TABLE 5: FINAL REQUIREMENTS OF THE REGIONAL PLANNING USE CASE PILOT.

TEADAL Technologies

Advocate, as introduced in D5.1, is part of the trust plane of TEADAL and responsible for collecting interaction evidence. For this pilot, where data is accessed from two data providers, such evidence is crucial to build mutual trust between BOX2M and RT. Advocate will ensure that both FDP's will provide audit logs in a verifiable and immutable way accessible by both parties. This way, both can verify the adherence to prior commitments and evaluate specific policies, even if the consumed data is only aggregated.

Funded by the European Union

7 NEW USE CASE PILOT: OPTIMISING FINANCIAL RETURNS FROM RENEWABLE ENERGY SOURCES

BOX2M proposes substituting the ING's financial data governance pilot proposal with a similar financial-focused integration, driven by costs in the energy commodities market. The capabilities to be developed under TEADAL include a dedicated predictive AI and national grid integration system within the scope of FDP/SFDP, as well as using data aggregation policies to reduce the amount of data in transit and limit potential exposure. Figure 7 depicts an overview of the expected pilot capabilities, which are summarised as follows:

- **FDP:** Measuring energy financial value for photovoltaic parks through edge computing, running IoT devices with custom firmware.
- **Aggregated SFDP:** Calculating optimal production capacity to maximise revenue and assessing in real time the cost of decarbonization
- **SFDP:** Acting based on financial criteria and production forecast.
- **Financial Modelling and Analytics:** Assessing the financial viability and return on investment (ROI) during daily operations for renewable energy projects.
- **AI-Based Prediction Mechanisms:** Developing AI Based prediction mechanisms for dynamic adjustment based on financial criteria and production forecasts. The plan is to leverage the national energy market pricing data for decision-making.
- **FDP-SFDP Virtual Coupling Mechanisms:** Creating virtual coupling mechanisms for park as a production/supply source for multiple consumers. Minimise the costs associated with rebalancing by avoiding overpayment for excess energy export into the network.

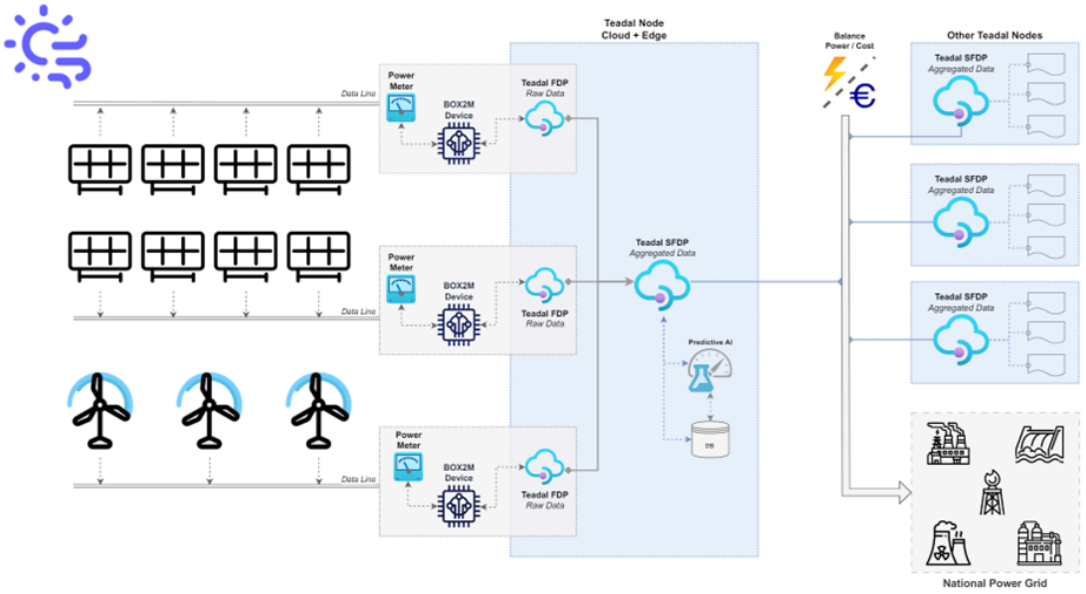


FIGURE 7: OVERVIEW OF THE NEW USE CASE PILOT.

Funded by the European Union

8 ARCHITECTURE

In this chapter, first the architectural concepts and ideas from recent progress are summarised so that new concepts introduced in this new set of deliverables have a solid foundation. This overview of the existing architecture focuses in particular on the requirements that the architecture must fulfil, as well as the phases that the data products undergo to be shared between data providers and consumers.

Afterward, this chapter gives a 360-degree overview of the tools and components that are developed within the different WPs of TEADAL, thus creating a common ground where all the tools are collected for the reader's convenience.

Then, the deliverable presents the runtime view of the TEADAL's architecture, to complement the other architectural views provided in deliverable D2.2. This view provides an understanding on how the different services of the cluster are layered according to their functionalities.

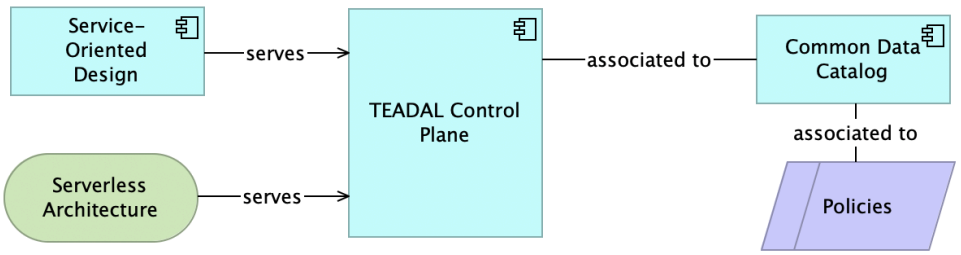
Finally, the last part of this chapter is to position TEADAL within the landscape of existing large-scale data sharing platforms, such as Gaia-X and IDSA, and discuss how TEADAL's functionalities match these approaches. Thus, we can assure that we stay well aligned with existing standards and put the emphasis of our contributions and novelty on features that have not yet been explored sufficiently by other platforms.

ARCHITECTURAL OVERVIEW OF TEADAL

Motivation and Requirements

In this section, the requirements that drove the development of the TEADAL platform are presented; to this extent, a summary of the requirements from D2.2 is provided, which includes explanations on how requirements are reflected within the architecture. Accordingly, the main requirements that the architecture, its tools, and the embedded processes need to fulfil, can be summarised under three points: (1) automation of data sharing, (2) optimization of data sharing, and (3) trust between partners.

1. **Automation of data sharing**: To simplify the infrastructure management and transformation of data products, the TEADAL platform should facilitate data sharing through a service-oriented, serverless design, managed by TEADAL's control plane (Figure 8). This should ease the policy definitions and translation from human-readable language to machine-understandable policies, where a common data catalogue allows for data discovery without accessing actual data, preserving privacy and confidentiality.



Funded by the European Union

FIGURE 8: SERVICE-ORIENTED SERVERLESS ARCHITECTURE, MANAGED BY TEADAL CONTROL PLANE.

2. **Optimization of data sharing**: Inter-organizational data sharing requires transforming the shared data according to policies attached to it; running this transformation produces what is called data friction. TEADAL's architecture must include performance monitoring tools to control and minimise this friction, but also optimise data and computation placement along a stretched data lake using its control plane to facilitate processing near data sources (Figure 9).

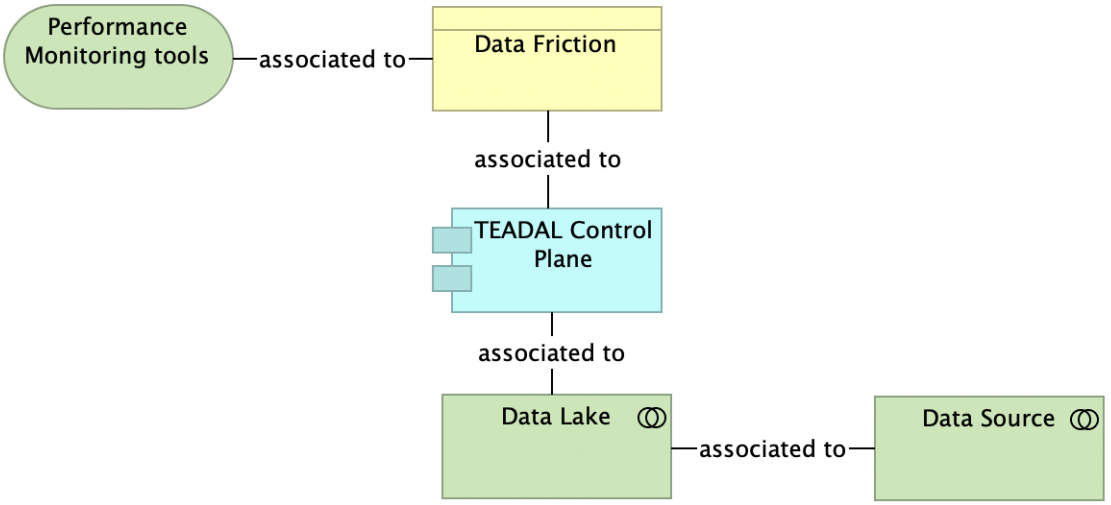


FIGURE 9: DATA FRICTION MANAGEMENT FOR OPTIMAL DATA SHARING.

3. **Trust between partners**: As a prerequisite for data sharing, trust must be guaranteed through policy enforcement and verification. For this, confidentiality and privacy of the data are maintained by controlling and fine-tuning data visibility through privacy-preserving computations. The trust plane should continuously monitor and verify these processes; in that regard, the architecture ensures data integrity and provenance with its control plane and blockchain technology, providing traceability and enabling audits throughout the data lifecycle (Figure 10).

Funded by the European Union

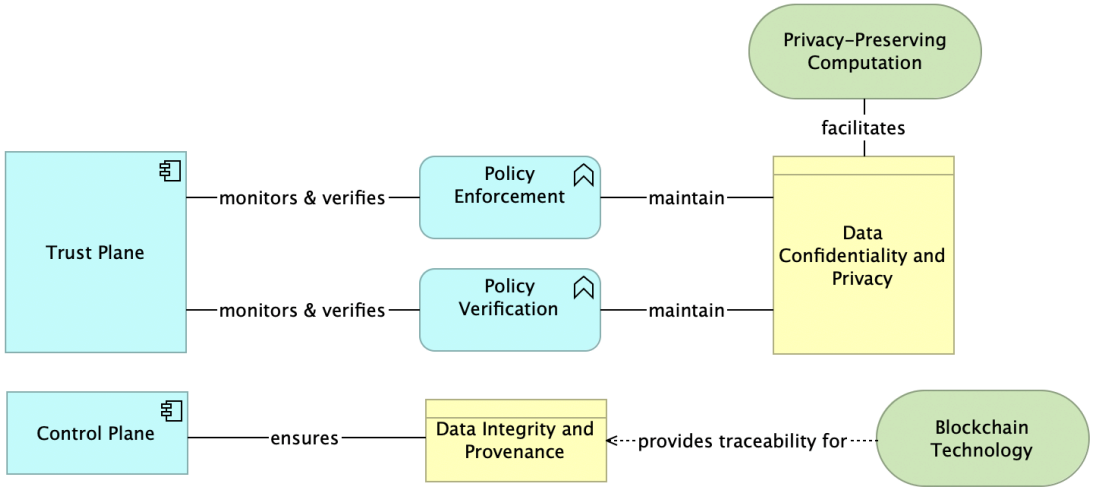


FIGURE 10: ESTABLISHING TRUST THROUGH DATA CONFIDENTIALITY AND PRIVACY MECHANISMS.

Underlying Concepts

This section focuses on two fundamental concepts in data sharing and service provisioning that are applied within TEADAL, namely data mesh and service mesh. For each of them, it is highlighted why they were included and how they are instantiated within TEADAL.

Data Mesh is a design principle that emphasises decentralised data governance and lifecycle management in large-scale organisations. Key concepts include defining the minimal unit of shareable data as a data product by domain experts, ensuring domain ownership of data, managing the data lifecycle through a self-service platform, and implementing federated computational governance with automated policy enactment. TEADAL extends these principles to interactions between organisations, introducing the federated data product (FDP) as the minimal unit of shareable data, addressing associated challenges in cross-organization data sharing.

Service Mesh uses layers of proxies connected to TEADAL to intercept communications, enabling features like security, policy enforcement, and traceability. In a service mesh, proxies handle both inbound and outbound service communications, isolating services from each other and the network. This allows proxies to inspect, route, and possibly alter service requests and responses, enriching functionality without altering service code. TEADAL's service mesh enhances data security, tracks FDP and SFDP life cycles to produce verifiable evidence, and improves observability of complex metrics such as gravity and friction.

Data Sharing Processes

This section summarises the data sharing process implemented by TEADAL and the various types of artefacts that are created throughout this process to facilitate the data sharing. In particular, this focuses on what a data product is – the smallest unit in TEADAL's data sharing process – and what are the phases through which a data product goes.

The high-level conceptual view of TEADAL's architecture includes multiple components and entities, which are detailed below. Further, it contains their responsibilities, communication interfaces, and interactions to fulfil requirements. Although most of them were introduced in

Funded by the European Union

previous deliverables, we summarise the data exchange process at this point to verify the extent to which the actual implementation matches it.

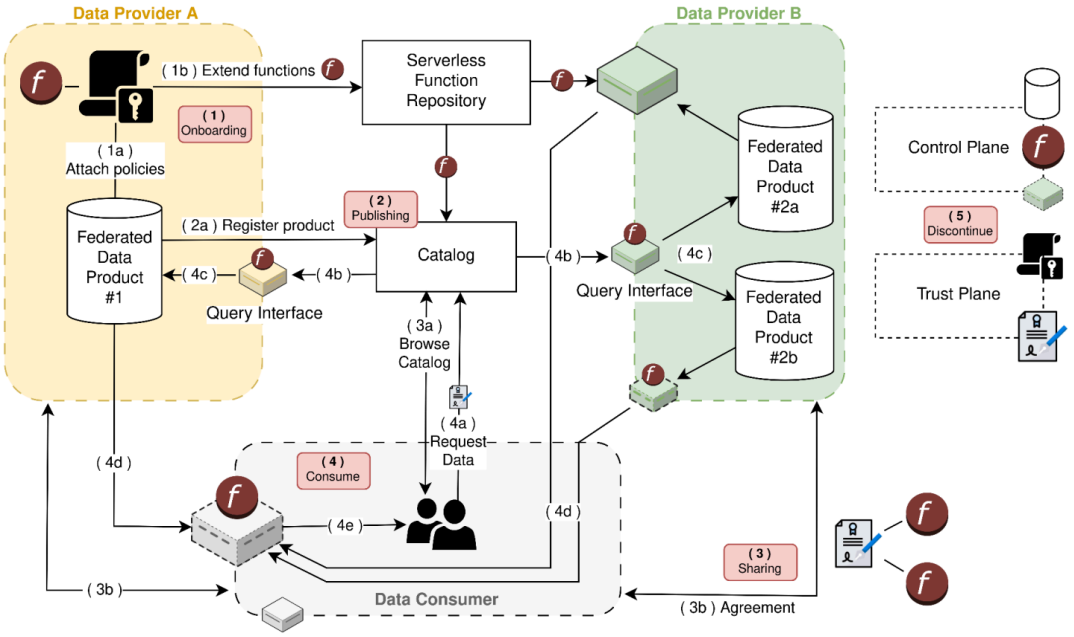


FIGURE 11: CONCEPTUAL (COMPONENTS AND PROCESSES) VIEW OF TEADAL'S ARCHITECTURE.

TEADAL outlines a five-phase lifecycle for a Federated Data Product (FDP) which are also included in Figure 11:

1. **Data Onboarding**: Providers prepare their data products, i..e, extending them with policies and finding a respective storage location in the federation. This transforms the data products into FDPs with policies attached.

2. **Publishing**: The data providers register the FDPs in a federation-wide and decentralised catalogue, which allows potential data consumers within the federation to browse the catalogue and discover FDPs based on their custom requirements.

3. **Sharing**: Data provider and consumer agree on the terms for data sharing through a contract, which defines the data processing pipelines, creating a Shared Federated Data Product (SFDP), a derived instance of the FDP that is provided for the consumer.

4. **Consumption**: Data consumers request access to the SFDP by providing their sharing agreement, which is validated and executed by the trust and the control plane, transforming data as per policies before consumption.

5. **Discontinue**: Agreements end due to various conditions, ceasing access and releasing associated resources using the control plane's data lineage capabilities.

These phases ensure controlled and efficient data sharing between organisations, supported by TEADAL's core components like FDPs, the catalogue, the control plane, and the trust

Funded by the European Union

plane. Below, we now summarise two of the core components in this architecture, the FDP and the SFDP in more detail.

A Federated Data Product (FDP) is a data product according to the data mesh principle that is shared between members of a federation. The FDP allows data to be stored on provider's premises or federation resources, managed by the control plane. After policies are registered in the onboarding phase, a data product is registered in the Data Catalog and federated within TEADAL. Each FDP undergoes a custom transformation process, resulting in a Shared Federated Data Product (SFDP) tailored to a usage agreement. This process is further detailed in Figure 12, where the FDP-SFDP pipeline provides an SFDP for each consumer. On the left, it is visible how the raw data product is combined with data sharing policies to create the FDP. For each particular consumer, the pipeline then creates a derived view into the FDP, which ensures that specific policies and requirements from their data-sharing contract are fulfilled.

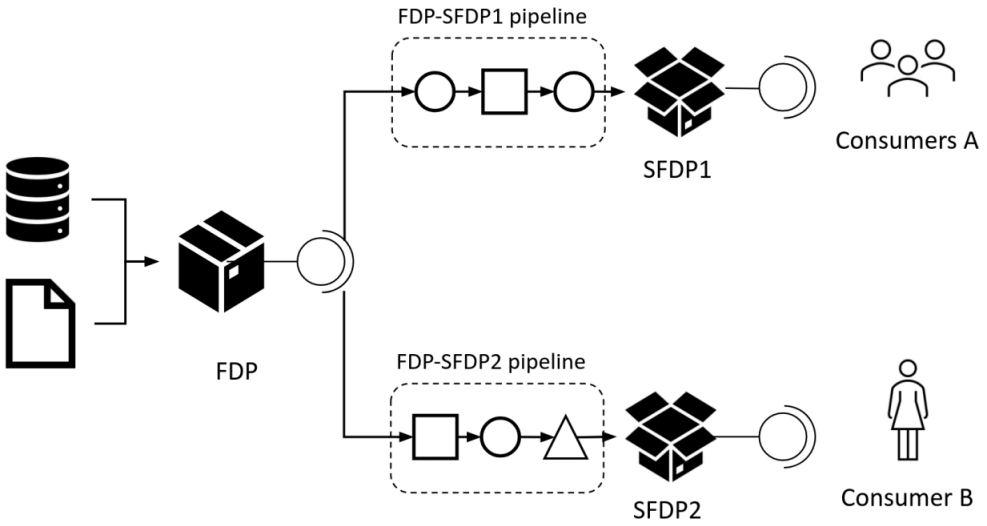


FIGURE 12: FDP TO SFDP PIPELINES. EACH SFDP IS BOUND TO A SPECIFIC AGREEMENT BETWEEN THE DATA OWNER AND THE CONSUMER.

Main components

This section presents the main components of the TEADAL architecture. The reader can delve further in each component by following the references at the end of each description.

Catalogue

The Catalog is a central piece for TEADAL's architecture. It provides a systematic and detailed view of every data asset within the federation. Further, its integration with an identity management system limits the visibility of the data assets to those with the required credentials. Hence, the catalogue brings discoverability to the federation, enabling members to browse the FDP's available. Further, TEADAL's Catalog can automate several steps during the generation of both FDPs and SFDPs, thanks to the integration of BPMN processes.

From a data provider perspective the Catalog allows publishing federated data products (FDP) metadata. Then, once a data provider and a data consumer have an agreement

Funded by the European Union

(contract) over the use of an FDP, TEADAL's Catalog enables the discoverability of the related SFDP to the consumer.

New information and details of this component can be found in deliverable D4.2.

Control plane

TEADAL Control Plane is created to provide seamless integration, centralization, and management of all the TEADAL nodes across all the locations within TEADAL's federation. Note that to enable deployment over different infrastructures, such as cloud IaaS, on-prem data centres and edge facilities, TEADAL Node architecture is based on cloud-native principles and is realised using Kubernetes (k8) technology, so that TEADAL Node is basically a k8s cluster extended with several layers of TEADAL services. Thus, all the services running on each specific TEADAL Node are controlled natively by the local k8s control plane and the related infrastructure level services such as istio as well as by a cluster-local argocd service, following GitOps procedures developed in WP6. As stated in D2.1, early in the project it was decided that TEADAL Control Plane will not be built from scratch but will leverage k8s control plane as much as possible, extending it as needed. Several candidate technologies were then considered to become the base of the TEADAL Control Plane, one of them is Kubestellar[2], an open source project created to manage workloads over multiple k8s clusters while itself being based on k8s. After the initial experimentation, and taking into account, among other considerations, the maturing pilot requirements and the invent of the argocd based TEADAL GitOps automation described in D5.2, additional technology candidates, such as multi-cluster argocd or terraform, need to be evaluated and might be found more beneficial for the project. After the evaluation planned as part of pilot integration and validation phases, the best fit will be chosen for the final realisation of the TEADAL platform and will be reported in a final set of deliverables.

In addition, the initial control plane architecture and implementation is being extended in several ways. Here are some examples:

1. Integrate components developed in WP3 that analyse system performance in terms of friction and gravity and provide information on optimal placement of TEADAL data pipeline related services, to improve energy footprint and reduce performance impacts related to data sharing.

2. Integrate components developed in WP5 that continuously observe and audit data accesses throughout the system and provide information on which services can be deployed and which must be stopped or quarantined.

3. Develop and integrate software for automating the process of generating SFDP based on existing FDP and the contract between the data provider owning the FDP and the data consumer requesting the SFDP. This new software is based on generative AI technologies.

4. Further develop and integrate the machine learning-based tools for enhanced monitoring capability, timely detection of anomalies and prediction of resource utilisation in the clusters, to detect and resolve any issues popping up in the system.

More details about the current state of the TEADAL Control Plane capabilities and about the planned extensions such as listed above are presented in deliverable D4.2.

[2] https://docs.kubestellar.io/release-0.23.1/

Funded by the European Union

Data pipelines

TEADAL's data pipelines are the core data-based computational operations within a TEADAL federation responsible for transforming the data as it moves through its lifecycle stages, based on the data governance as well on the security and trust rules. As described in the Section [Architectural overview of TEADAL], in TEADAL data goes through the stages of onboarding, publishing, sharing, consumption, and discontinue. Some of these stages involve contract negotiations and are inherently manual while some can be realised with code. Striving for maximum automation, TEADAL includes data pipelines as code that can be invoked to realise the required data transformations when needed..

There are two major types of data pipelines in TEADAL. First, there are pipelines that ingest a dataset (or a data product) from a specific data provider organisation into the TEADAL federation and transform this data into a TEADAL FDP. Second, there are pipelines that convert the FDP already existing in the TEADAL federation and available in the TEADAL Catalogue into an SFPD based on the contract agreement between the data provider and the data consumer.

As computational components, data pipelines follow the usual software lifecycle when code needs to be developed, tested, deployed, run, etc. In TEADAL, data pipelines are realised as modular k8s deployments and services deployed on demand by the TEADAL Control Plane as Kubeflow[3] workflows. The deployment is informed by the metadata and the annotations available in the TEADAL Catalogue ,for example, hardware requirements, task characteristics, transformations needed, or policies to be applied, to optimise the data pipeline workflows based on a set objectives and constraints.

In addition to managing FDP-to-SFDP data pipelines that are created from manually developed software (e.g., when a developer implements an SFDP from an FDP upon contract approval), TEADAL includes capabilities to create this type of pipelines automatically using advanced generative AI technologies.

More information on data pipelines related automation is presented in deliverable D4.2

Security policies and service mesh

Security policies use interception proxies to manage communication between consumers and FDP services. These proxies enforce access control through policy decisions and enforcement points in the policy store. Data providers create access control policies, which are stored and evaluated against consumer requests. The system uses Istio proxies, which intercept HTTP traffic and seamlessly work with the TEADAL data mesh. They route requests through the proxy for access control checks before reaching the data product service.

The security policies component is built on open-source components such as Istio, Envoy, and Open Policy Agent (OPA). The Istio proxy is paired with each data product service to intercept and route requests based on configured policies in our component. The Envoy proxy uses an External Authorization Filter to connect with OPA, which evaluates policies written in Rego, fetched from a dedicated policy store. This process ensures that only authorised

[3] https://www.kubeflow.org/

Funded by the European Union

requests reach the data product services. Additionally, TEADAL provides a Role-Based Access Control (RBAC) framework, simplifying access control implementation for RESTful services, and supports alternative policy decision points like the TEADAL Datalog interpreter and Anubis for flexible and secure access management.

New information and details of this component can be found in deliverable D3.2 and upcoming deliverable D4.2.

Trust plane

TEADAL's trust plane is in charge of gathering evidence of all processes, interactions, and data exchanges of the TEADAL's data lake. Further, evidence is stored in an immutable manner, so that a posteriori verification is viable and trustworthy. Finally, evidence data is accessible to all members of the federation, allowing independent verification.

TEADAL's trust plane is composed of several components, the next subsection introduces the Advocate, which is the main component of the plane. Other components are the claims registry, the immutable evidence storage, the DLT TEADAL, and the TNS contract. More detailed information about the trust plane can be found in D5.2

Advocate

The advocate component in the TEADAL framework plays a crucial role in ensuring the integrity and trustworthiness of federated data exchanges. It is responsible for orchestrating the collection, summarization, and publication of verifiable evidence regarding data transactions and interactions within the TEADAL system and storing its results in the Shared Evidence Plane. By leveraging cryptographic methods, such as verifiable credentials and public/private key pairs, the advocate component generates tamper-proof records of every significant event in the data lifecycle. These records are then securely stored in an immutable storage system and anchored in a blockchain-based claims registry to ensure they remain unaltered and accessible for future audits.

In practice, the advocate component ensures that each data lake within the federation adheres to the established protocols for data integrity and trust. It provides a comprehensive audit trail that links each action back to a specific user or process, thereby enabling transparency and accountability. By integrating closely with the observability and control planes, the advocate verifies and attests to the authenticity of all interactions, from data ingestion to consumption, ensuring compliance with regulatory requirements and organisational policies. This component is essential for building and maintaining trust among all parties involved in the federated data ecosystem, facilitating secure and reliable data sharing across organisational boundaries.

New information and details of this component can be found in deliverables D5.2.

Energy optimization

TEADAL's toolset will incorporate means to monitor and optimise the energy consumption of the data lake. The energy optimization, more than a component, will be a feature given by the monitoring and estimation of energy consumption, which will be incorporated as metadata to the pipeline deployment strategies, as well as to the SFDP candidate storage locations. This way, the control plane will leverage the metadata to make the best decisions in terms of energy efficiency. New information and details of this feature can be found in deliverables D3.2 and soon upcoming D4.2.

Funded by the European Union

RUNTIME VIEW

This section presents a conceptual view of the TEADAL cluster runtime. It is an abstract description of the services that typically run in a TEADAL cluster instance in terms of the functionality they provide without reference to actual software implementing that functionality. D6.2 complements this abstract description with the software products that have been selected to provide the functionality detailed next.

The TEADAL cluster runtime comprises four layers of processes and hardware. This is a classic layered architecture where each layer depends on and builds on the functionality provided by the layer directly beneath whereas it has no knowledge of the layers above. From bottom to top, the four layers are: hardware, mesh infrastructure, core services and products. Figure 13 depicts these layers along with the services in each layer and should be kept at hand as we describe each layer in detail.

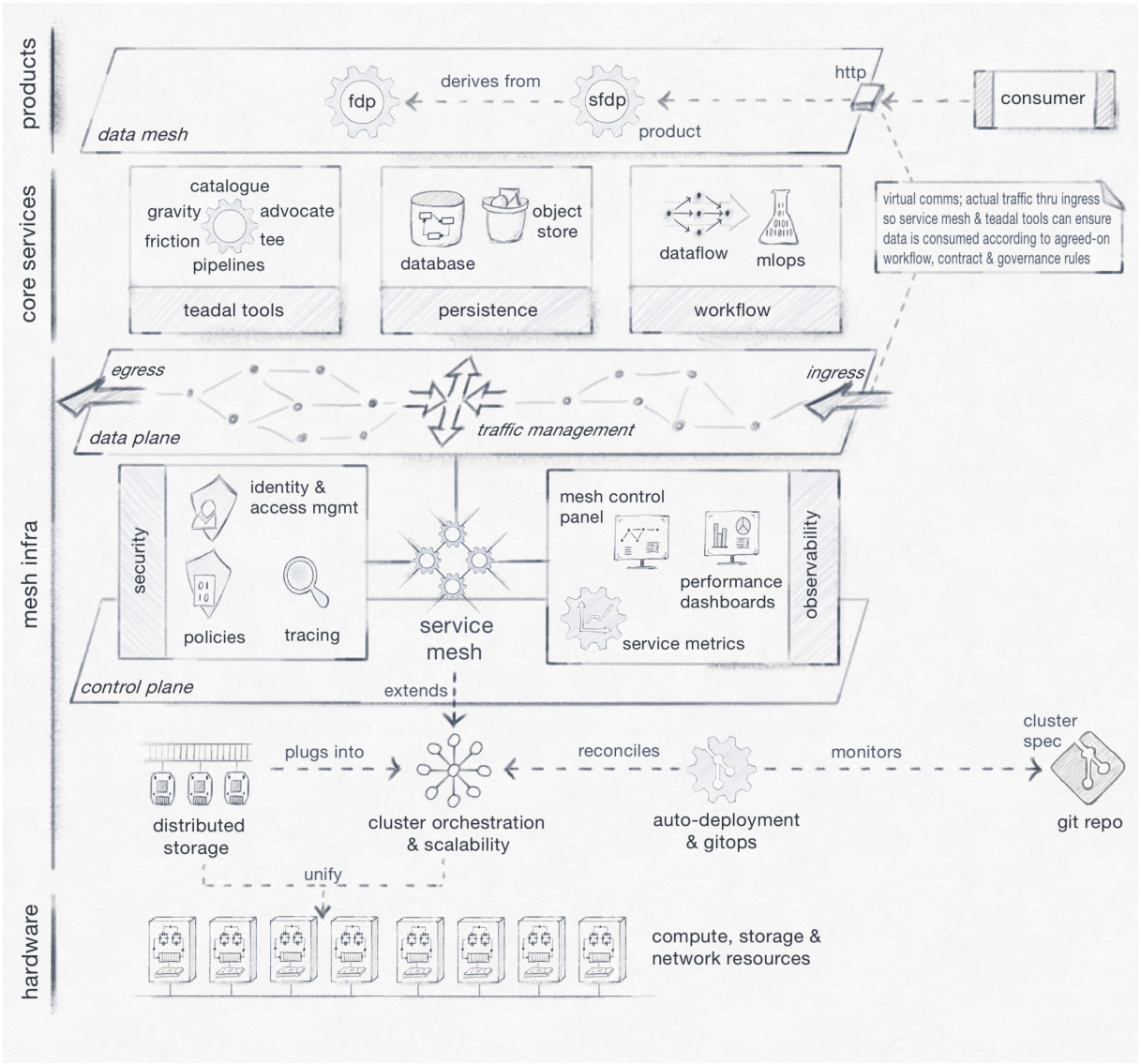


FIGURE 13: TEADAL CLUSTER RUNTIME LAYERED ARCHITECTURE.

Hardware layer

The hardware layer comprises the physical or virtual computing, storage and network resources on which all the TEADAL cluster software runs. In the case of a public cloud

Funded by the European Union

deployment, typically a hypervisor would provide the hardware layer as a set of virtual resources. On the other hand, an on-premises deployment may entail provisioning physical machines. The number of hardware resources in the cluster depends on the processing power required for a given deployment but notice that it is also possible to host the whole TEADAL cluster runtime on a single physical or virtual machine in cases where a simpler deployment is desirable—e.g. for testing or evaluating TEADAL.

Mesh infrastructure layer

The mesh infrastructure layer interfaces with the hardware layer to provide the service mesh functionality. A core service in this layer provides cluster orchestration and scalability by managing computational resources (CPU, memory, storage), allocating them to processes in the upper layers and orchestrating the deployment and operation of services by means of containers. A distributed storage facility provides the orchestration service with uniform, high-level access to the underlying disks attached to cluster nodes. Together, orchestration and distributed storage unify the underlying hardware resource in a high-level, aggregated computing facility.

A TEADAL cluster is instantiated and then subsequently updated through a GitOps approach. An automated deployment and GitOps service in the mesh infrastructure layer monitors the desired cluster runtime specification as declared in an online Git repository associated with the cluster. On detecting a new revision in the Git repository, the service automatically reconciles the desired new runtime specification with the actual live state of the cluster.

Service mesh software extends cluster orchestration with control and data planes. The control plane manages a network of proxies, the data plane, which captures and processes cluster inbound and outbound traffic as well as internal service traffic. This allows to augment service functionality at runtime without requiring any modifications to the services themselves. The TEADAL cluster exploits this to transparently route and balance service traffic, secure communication and access to service resources, and monitor service operation. In fact, a set of security services plug into the control plane to provide identity and access management, security policies, and tracing. Likewise, a set of observability services complement the control plane's core functionality with service metrics, performance dashboards and a mesh control panel.

Core services layer

The core services layer runs on the mesh infrastructure to provide the TEADAL core functionality which enables federated data products. A set of TEADAL tools and services allow multiple TEADAL clusters to be joined in a federation where producers and consumers can share data in a trustworthy and secure way, according to agreed-upon governance, privacy and energy-efficiency policies. Catalogue, Advocate, FDP Pipelines, Trusted Execution Environment as well gravity and friction policies are part of the TEADAL tools.

General-purpose persistence and workflow services are also part of the core services layer. Persistence services include a relational database and an object store, whereas workflow services support dataflow programming for engineering data pipelines and MLOps for managing machine learning operations.

Products layer

The product layer hosts data products and services—i.e., federated data products (FDPs), shared federated data products (SFDP), etc. As detailed in D3.1, a federated data product (FDP) extends the notion of data mesh product to cater for sharing data in a data lake federation according to the governance rules of that federation. A shared federated data

Funded by the European Union

product (SFDP) encapsulates a consumer-producer agreement (contract) about sharing a part of an FDP and provides the means for the consumer to process the shared data only within the bounds of the agreed-upon contract.

Notice a consumer accesses an SFDP through a REST API. Whereas it is convenient, from a conceptual standpoint, to think of the consumer directly connecting to the SFDP through HTTP, actual network traffic goes through the data plane ingress so that TEADAL can process HTTP requests before they reach the SFDP and responses before they reach the consumer. This way, the service mesh and the TEADAL tools can ensure data is consumed according to the consumer-producer agreed workflow and contract as well as federation governance rules.

Alignment of the TEADAL data lake architecture with available Dataspaces architectures

Starting from the glossary provided by the DataSpace Support Center, we will identify the similarities and differences between the conceptual model of dataspaces and the architecture of TEADAL. This analysis will help in understanding how TEADAL aligns with and diverges from the dataspace model.

According to the DataSpace Support Center (DSSC), a dataspace is a distributed system defined by a governance framework that enables secure and trustworthy data transactions between participants while supporting trust and data sovereignty. A dataspace is implemented by one or more infrastructures and enables one or more use cases. It facilitates value creation for its participants (and for the economy, society, and environment) through easy, secure, and trustworthy data transactions.

The general scenario as described by DSSC is: "The parties engaged in data sharing need reassurance that their data is used in ways they can accept and that the data they need from others will have the promised qualities. To practically enable such trustworthy data transactions, groups of parties may organise themselves as a data space initiative and start deploying and maintaining a dataspace. A dataspace is defined by a governance framework and implemented by one or more interlinked dataspace infrastructures. The dataspace governance authority creates and maintains the governance framework that all dataspace participants must commit to."

One of the two main objectives of TEADAL is to reduce the so-called "friction" to facilitate data sharing between different organisations. This goal closely aligns with the purpose of dataspaces, which is to enable secure and efficient data sharing across organisational boundaries. TEADAL's aims in the context of data sharing are comparable to those of dataspaces, particularly in promoting interoperability, trust, and data sovereignty.

Phases of Data Sharing in a Dataspace

A high-level sequence of operations performed by two organisations when engaging in a data sharing activity in a dataspace can be summarised as follows:

1. Data product offering: Describe the data product offering using a shared metadata vocabulary. The data product offering can include policies related to data usage and visibility of the data product itself.

2. Data product advertisement: metadata describing the data product offering are advertised to the members of the dataspace.

3. Exploration: Discover existing data products.

Funded by the European Union

4. Business Negotiation: Negotiate contracts between data providers and consumers, including pricing and intermediary services.
5. Operational Negotiation: Agree on technical policies and configurations.
6. Compliance: Ensure all parties comply with relevant policies and regulations.
7. Data Access and Transfer: Transfer data between provider and user, and to other designated participants.
8. Data Usage: Access and use data by the consumer.

The concepts of a dataspace as defined by DSSC and this high-level sequence are currently being implemented by various initiatives (like IDSA, Gaia-X, FIWARE). Even though the concepts are the same, the way each organisation implements them makes each implementation not interoperable with the others, neither at API nor at data and metadata schema level.

IDSA / Eclipse Dataspace Connector

Illustrated in Figure 14, the approach taken by IDSA revolves around the concept of a software component called a "dataspace connector". Such component implements basically all the interactions required to put in place a successful data exchange:

- metadata publishing
- federated metadata discovery
- metadata access control policy enforcement
- contract negotiation
- data provisioning
- data access
- data usage control policy enforcement

As a result, the connector can be seen as the overall gateway of a company which provides everything that is required to participate in a dataspace, as visible in the following diagram taken from IDS Reference Architecture Model (RAM).

Funded by the European Union

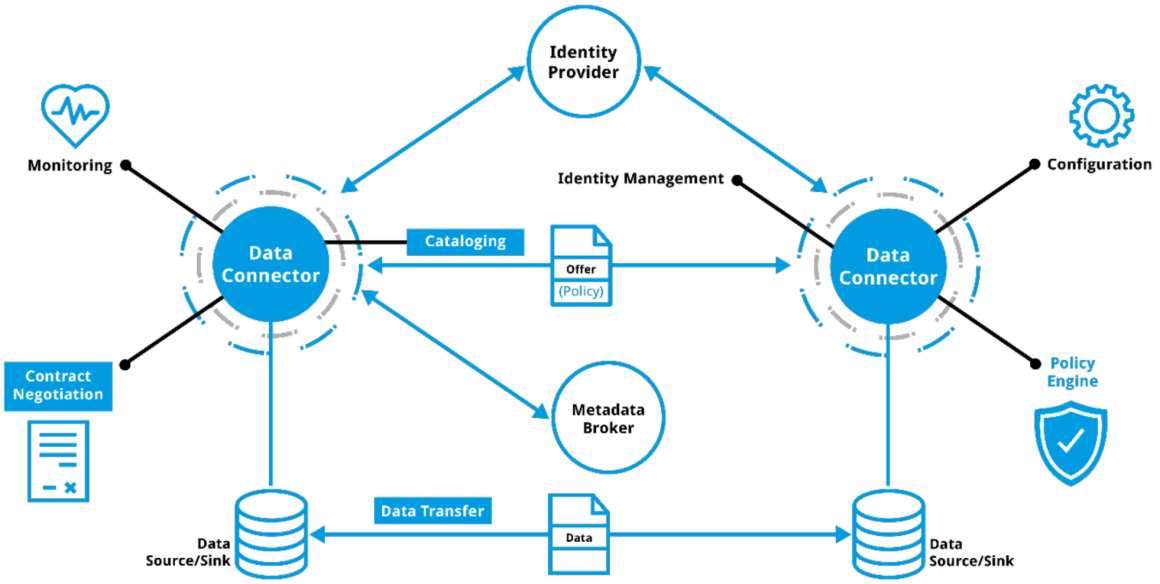


FIGURE 14: IDSA / ECLIPSE DATASPACE CONNECTOR ARCHITECTURE OVERVIEW.

Participating in an IDSA-based dataspace therefore means adopting the identity specifications (based on X.509 certificates) and being able to interact with an IDSA connector. Figure 15 shows an Archimate diagram that describes which would be the TEADAL components involved and their role in a dataspace-enabled data exchange.

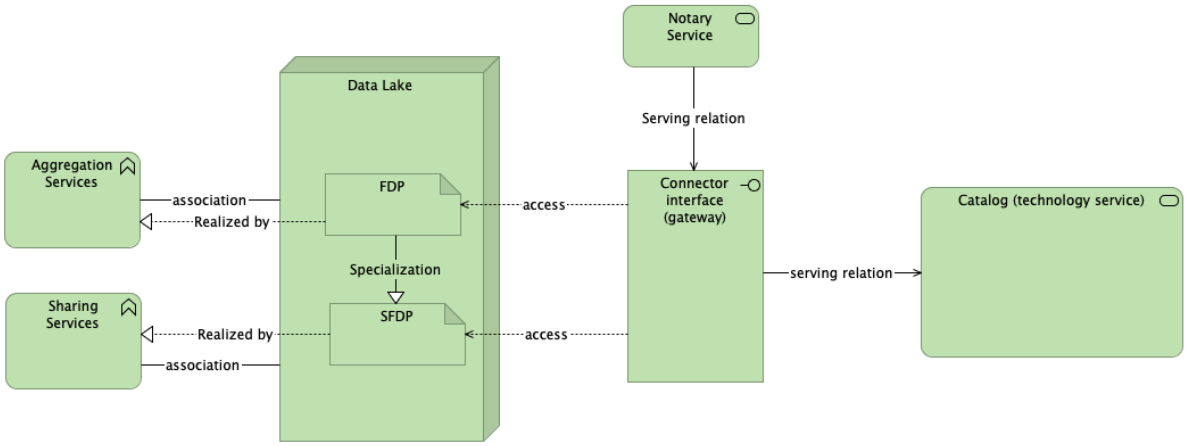


FIGURE 15: TEADAL COMPONENTS FOR DATASPACE-ENABLED DATA EXCHANGE, IN ARCHIMATE NOTATION.

Data Lake

We start from a TEADAL data lake, which is a centralised repository that allows to store structured and unstructured data at any scale. Internally, the data lake contains a variety of data sources, which can include raw and processed data.

- **ArchiMate Element: Node**

Federated Data Product (FDP)

Funded by the European Union

We aggregate these data sources through services that expose a combined view of the data. This aggregation involves processes such as data integration, filtering, and processing to create a unified dataset that can be used for analytics and reporting. We refer to this comprehensive dataset as a Federated Data Product (FDP).

- **ArchiMate Element:** `Artifact`

Shared Federated Data Product (SFDP)

When there is a need to share a Federated Data Product with external entities, additional data processing steps are necessary. This includes further data cleansing, anonymization, and other data transformations to ensure compliance with data privacy and security standards. The resultant dataset, which is ready for external sharing, is called the Shared Federated Data Product (SFDP).

- **ArchiMate Element:** `Artifact`

Connector

The core component of the data sharing landscape is the connector. The connector acts as an intermediary that manages the publication, visibility, and access of the FDP and SFDP. It facilitates secure data exchange between internal and external stakeholders, ensuring that only authorised users can access the data.

- **ArchiMate Element:** `Technology Interface`

Catalogue

Within the connector, there is a catalogue service. This catalogue lists all available data products, allowing users to browse, search, and request access to the data. The catalogue provides metadata about each data product, such as its description, source, and usage policies.

- **ArchiMate Element:** `Technology Service`

Clearing House

A critical component in the data sharing ecosystem is the clearing house. This notary service keeps track of authorization levels, records contracts between data providers and consumers, and ensures that all data sharing activities comply with agreed-upon policies and regulations. The clearing house also helps in dispute resolution and auditing.

- **ArchiMate Element:** `Technology Service`

Aggregation Services

Aggregation services are responsible for integrating data from various sources within the data lake. These services perform operations such as data extraction, transformation, and loading (ETL) to create a consolidated view of the data. Aggregation services are crucial for generating the FDP.

- **ArchiMate Element:** `Technology Function`

Sharing Services

Funded by the European Union

To prepare the FDP for external sharing, sharing services are employed. These services handle additional data processing tasks, including data cleansing, anonymization, and compliance checks. The output of these services is the SFDP, which is suitable for external consumption.

- **ArchiMate Element:** `Technology Function`

Data Lake Hosts FDP and SFDP

The data lake serves as the repository that hosts both the FDP and SFDP. The data lake's infrastructure supports the storage, processing, and management of these data artifacts.

- **ArchiMate Relationship:** `Assignment` between `Node` (Data Lake) and `Artifact` (FDP), and between `Node` (Data Lake) and `Artifact` (SFDP).

FDP Transforms into SFDP

The FDP undergoes further processing to become the SFDP. This transformation includes additional data cleansing, anonymization, and other necessary adjustments to ensure the data can be shared externally.

- **ArchiMate Relationship:** `Specialization` between `Artifact` (FDP) and `Artifact` (SFDP).

Connector Manages the Publication, Visibility, and Access of FDP and SFDP

The connector is the central component that facilitates access to both the FDP and SFDP. It ensures that data products are published correctly and that only authorised users can view and access the data.

- **ArchiMate Relationship:** `Accessed by` between `Artifact` (FDP) and `Technology Interface` (Connector), and between `Artifact` (SFDP) and `Technology Interface` (Connector).

Catalog Connected to the Connector

The catalog service, embedded within the connector, provides an interface for users to discover available data products. It serves as the front-end service that allows users to interact with the data offerings.

- **ArchiMate Relationship:** `Serving` between `Technology Service` (Catalog) and `Technology Interface` (Connector).

Services Aggregate Data from Sources within the Data Lake

Aggregation services work in conjunction with the data lake to compile data from various sources. These services ensure that the data is integrated and prepared for analysis and reporting.

- **ArchiMate Relationship:** `Association` between `Node` (Data Lake) and `Technology Function` (Aggregation Services).

Sharing Services Transform FDP into SFDP

Funded by the European Union

Sharing services are responsible for the additional processing required to transform the FDP into the SFDP. They perform necessary data operations to ensure the data meets external sharing standards.

- **ArchiMate Relationship:** `Association` between `Technology Function` (Sharing Services) and `Artifact` (SFDP).

Clearing House Records Contracts

The clearing house service is essential for recording and managing the contracts and policies associated with data sharing. It interacts with the connector to maintain a comprehensive record of all data sharing agreements.

- **ArchiMate Relationship:** `Serving` between `Technology Service` (Clearing House) and `Technology Interface` (Connector).

To summarise, TEADAL can be made compatible with the IDSA architecture of a dataspace by carefully distinguishing between FDPs and SFDPs during the different interactions. A possible sequence of interactions for ensuring compatibility would be:

- the data product owner describes the FDP via the TEADAL Catalog;
- the TEADAL Catalog interacts with the IDSA Connector, ordering the advertisement of the FDP with a "default policy" forcing all possible customers to start a negotiation;
- once a negotiation request is received by the IDSA Connector, it is forwarded to the TEADAL Catalog, which activates a corresponding "Negotiation BPMN";
- the two parties engage in the negotiation process via the TEADAL Catalog and out of band communication;
- when the agreement is reached, the SFDP is created for the customer;
- the SFDP owner describes it on the TEADAL Catalog. As a result, the TEADAL Catalog pushes all the required metadata and policy to the IDSA Connector (stating that only the specific customer can access the specific SFDP), and notifies the customer that the SFDP is available.

FIWARE Dataspace Connector

While IDSA conceives the "Connector" as a single component which acts as an agent inside the dataspace on behalf of an organisation, FIWARE calls "Connector" an interconnected set of software components, each of them dedicated to a specific feature, as depicted in Figure 16.

Funded by the European Union

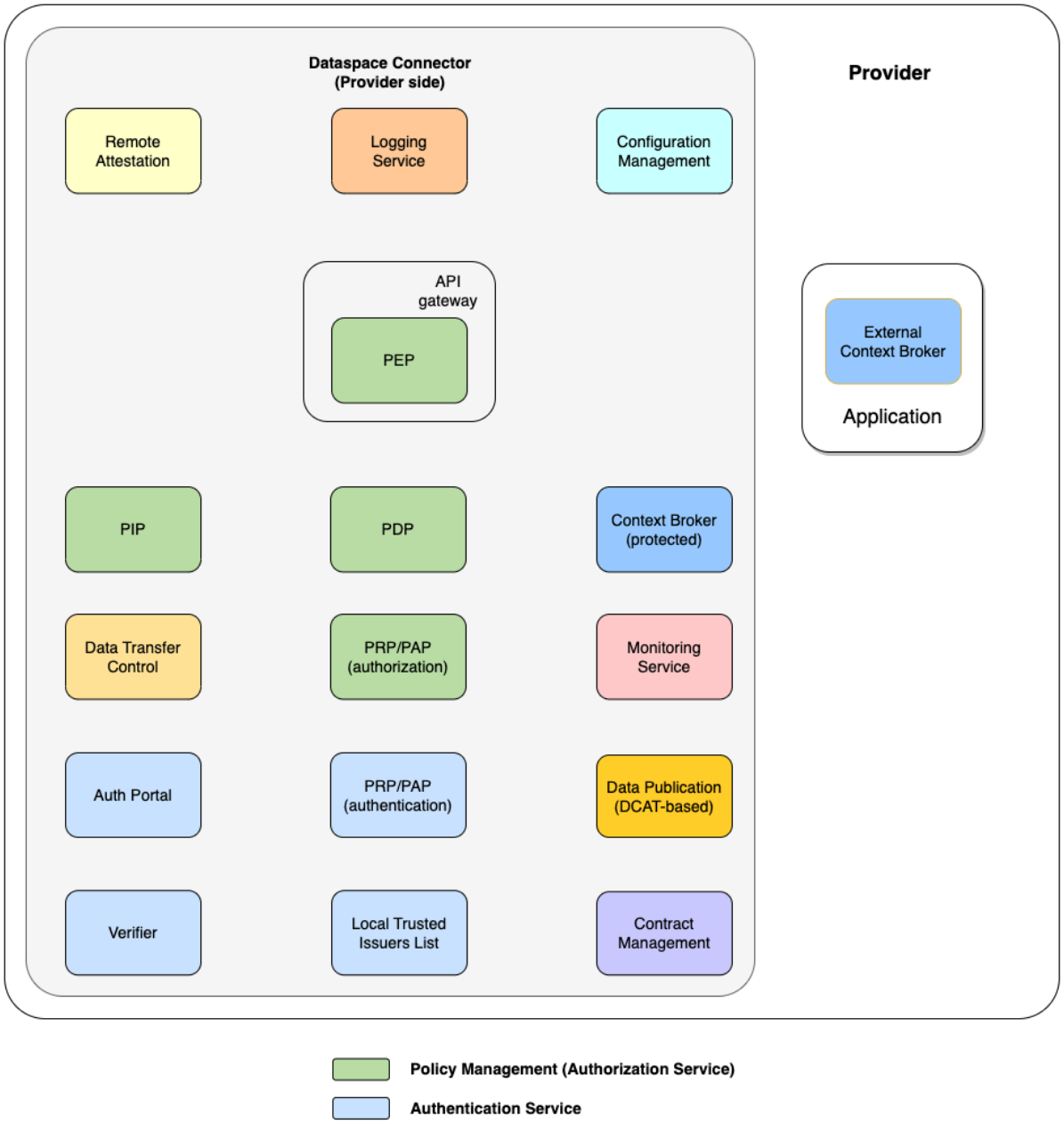


FIGURE 16: FIWARE DATASPACE CONNECTOR OVERVIEW.

Inside the FIWARE Connector we can group the components into three families:

- identity management
- authorisation, access control and data provisioning
- catalogue

Identity management is based on Verifiable Credentials and implemented via an extension of Keycloak. Authorisation, access control and data provisioning are implemented using a combination of an API Gateway (Apache APISix) and an OPA agent which enforces Rego policies. The catalogue, differently from the IDSA implementation, is not a feature embedded inside the Connector component, but a dedicated and independent component which takes care of metadata publishing and advertising, and of contract management. While in IDSA the

Funded by the European Union

Connector takes care of every aspect related to data sharing, the FIWARE environment basically splits the sharing into two separate moments:

- First, metadata is published to the Catalog, which operates as a marketplace. Interested customers can discover datasets, request access and start negotiation processes;
- Second, when an agreement is found between the dataset owner and the customer, the API gateway is configured (using a Rego policy) to allow access to the customer according to the terms and conditions created during the negotiation phase.

The FIWARE Dataspace Connector architecture, albeit way more complex, features a lot of similarities with the TEADAL architecture:

- the usage of the API gateway is similar to the usage of the data lake ingress in TEADAL;
- both FIWARE and TEADAL use OPA and Rego for policy definition and enforcement;
- both FIWARE and TEADAL use Keylcloak for users management;
- both FIWARE and TEADAL have a Catalogue component which is used for data products discovery, access requests and for starting negotiation processes;
- FIWARE exposes a standard API (TM Forum API[4]) for interacting with the Catalog/Marketplace and for dealing with access requests and negotiation requests. TEADAL Catalog exposes proprietary API.
- The splitting of the metadata advertisement phase and the data provisioning phase is compatible with the TEADAL distinction between FDPs and SFDPs. In an hypothetical FIWARE-compliant TEADAL environment, FDPs would be advertised on the FIWARE Catalog, while SFDPs would be made available through the API Gateway after signing data usage deals.

[4] TM Forum API https://www.tmforum.org/oda/open-apis/

Funded by the European Union

9 CONCLUSIONS

This deliverable has provided a comprehensive overview and final updates on the pilot use cases, detailing the specific requirements and demonstration plans of each pilot. From the evidence-based medicine pilot to the newly introduced financial pilot, each chapter has outlined the final requirements and advancements in these different use cases. The document presents the corresponding final demo ideas, pilot statuses, and how these pilots contribute to shaping the overall set of TEADAL technologies.

In terms of the TEADAL's architecture, this deliverable introduces the main technical components that are being developed as the TEADAL's tool set. Further, it provides the cluster runtime view to complete the views presented in the previous deliverable, providing a complete view of the TEADAL's general architecture. Finally, this deliverable discusses the alignment of the TEADAL data lake with dataspaces architectures, showing that TEADAL would be capable of aligning with them. The following, and last deliverable of the architecture (deliverable D2.4) will present the integration of all TEADAL's tools with the general architecture to present a final and complete view. Further, it will specify for each use case how this final general architecture will fit their use case and how they will leverage it.

Funded by the European Union